Score test variable screening

Sihai Dave Zhao
Department of Biostatistics and Epidemiology, University of Pennsylvania
Perelman School of Medicine, Philadelphia, Pennsylvania, U.S.A.

Yi Li
Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, U.S.A.

Abstract

Variable screening has emerged as a crucial first step in the analysis of high-throughput data, but existing procedures can be computationally cumbersome, difficult to justify theoretically, or inapplicable to certain types of analyses. Motivated by a high-dimensional censored quantile regression problem in multiple myeloma genomics, this paper makes three contributions. First, we establish a score test-based screening framework, which is widely applicable, extremely computationally efficient, and relatively simple to justify. Secondly, we propose a resampling-based procedure for selecting the number of variables to retain after screening according to the principle of reproducibility. Finally, we propose a new iterative score test screening method which is closely related to sparse regression. In simulations we apply our methods to four different regression models and show that they can outperform existing procedures. We also apply score test screening to an analysis of gene expression data from multiple myeloma patients using a censored quantile regression model to identify high-risk genes.

Keywords: High-dimensional data; Feature selection; Projected subgradient method; Score test; Variable screening

1 Introduction

High-dimensional datasets are now common in clinical genomics research. Though regularized estimation can consistently estimate sparse regression parameters even when $p > n$ (Bühlmann et al., 2011), in practice these methods still perform poorly if $p \gg n$ (Fan and Lv, 2008). Variable screening is crucial for quickly reducing tens of thousands of covariates to a more manageable size. Our interest in screening is motivated by our work with censored quantile regression in the study of the genomics of multiple myeloma, a blood cancer characterized by the hyperproliferation of plasma cells in the bone marrow. We are interested in identifying genes highly associated with the 10% quantile of the conditional survival distribution in order to better understand the biological basis of high-risk myeloma, in view of personalized medicine.

Perhaps the most popular screening framework is marginal screening, where each covariate is individually evaluated for association with the outcome and those with associations above some threshold are retained. Currently three major classes of marginal screening methods have been proposed. Wald screening retains covariates with the most significant marginal parameter estimates, and has been theoretically justified for generalized linear models and the Cox model (Fan and Lv, 2008; Fan and Song, 2010; Zhao and Li, 2012). Semiparametric screening assumes a functional form for the regression model but not for the probability distribution, and uses model-free statistics to quantify the associations between covariates and the outcome. Such methods have been proposed for single-index hazard models, linear transformation models, and general single-index models (Fan

and Song, 2010; Zhu et al., 2011; Li et al., 2012). Finally, nonparametric screening does not assume a functional form for the regression model and instead approximates it, using for example a B-spline basis. It retains covariates whose estimated functional relationships have the largest L_2-norms. Such methods have been studied for linear additive models and censored quantile regression (Fan et al., 2011; He et al., 2013). The distance correlation-based screening method of Li et al. (2012) requires very few assumptions about both the regression model and the probability model. It is well-known that marginal screening can miss covariates that are only associated with the outcome conditional on other covariates. To address this difficulty, iterative versions of several of these procedures have been proposed, though without theoretical support.

However, there are several issues that make existing screening methods unsuitable for application to our multiple myeloma analysis. Wald screening using censored quantile regression estimators, such as those of Honore et al. (2002), Portnoy (2003), Peng and Huang (2008), or Wang and Wang (2009), has not been theoretically justified. Semiparametric screening is not appropriate because the probability model is actually critical in our case: we are interested only in genes that affect the 0.1 quantile, whereas semiparametric screening would identify genes that affect any quantile of the survival distribution. There were no nonparametric screening methods for censored quantile regression until very recently, with the work of He et al. (2013), but in practice their approach is still computationally cumbersome, especially for resampling or cross-validation procedures where screening must be repeated multiple times. There is also no efficient iterative screening procedure for this model.

To address these issues, we propose in this paper a marginal *score testing* framework, where we use score tests rather than Wald tests to effect variable screening. This has several advantages. First, score screening is a general approach which can be applied to any model that can be fit using an estimating equation, including censored quantile regression, as well as to semi- and nonparametric regression models. Second, theoretical justification for score test screening is much simpler than for other screening methods and generally requires only concentration inequalities. Third, because they only require fitting the null model, score tests are exceedingly computationally efficient. Finally, the score test perspective suggests a new method for iterative screening that is easy to implement and turns out to be closely related to sparse regression, suggesting a possible approach to a theoretical justification.

In this paper we make three contributions. First, in Section 2 we propose marginal score test screening procedure and illustrate its application to several popular models. We give theoretical justifications for these procedures in Appendix A. Second, in Section 3 we propose a resampling-based method for choosing the number of covariates to retain after screening, based on the principle of reproducible screening. This procedure is only practical because score screening can be quickly computed. Third, in Section 4 we propose an iterative score test screening procedure that turns out to be related to projected subgradient methods from the numerical optimization literature. We illustrate our procedures on simulated data in Section 5, use them in our MM analysis in Section 6, conclude with a discussion in Section 7.

2 Score test screening

2.1 Method

Let $\mathbf{X}_{ik} = (X_{ik1}, \ldots, X_{ikp_n})^T$ be the vector of covariates measured at the k^{th} observation on the i^{th} subject, where $k = 1, \ldots, K_i$ and $i = 1, \ldots, n$, and let $\boldsymbol{\beta}_0$ be a set of possibly infinite-dimensional parameters quantifying the association of the $\mathbf{X}_{ik}$ with the outcome. For example, in linear models the outcome is a function of $\mathbf{X}_{ik}^T\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_0$ is a vector of scalar coefficients, and in additive models

the outcome is a function of $\sum_{j=1}^{p_n} f_j(X_{ikj})$ and $\boldsymbol{\beta}_0$ is the set of functions f_j. We will say that $\beta_{0j} = 0$ implies that the j^{th} covariate is not functionally associated with the outcome and is thus unimportant. This is a slight abuse of notation, as β_{0j} for irrelevant covariates would equal the scalar zero in linear models but the zero function in additive models, but the appropriate meaning will be clear from the context. Finally, let $\mathbf{U}(\boldsymbol{\beta})$ be an estimating equation for $\boldsymbol{\beta}_0$, such that $\mathbf{U}(\boldsymbol{\beta}_0) \to \mathbf{0}$ in probability as $n \to \infty$.

Denote the set of important covariates by $\mathcal{M} = \{j : \beta_{0j} \neq 0\}$. We assume that its size $|\mathcal{M}| = s_n$ is small and fixed or growing slowly. Our proposed marginal score test screening proceeds as follows:

1. Center and standardize each covariate to have mean 0 and variance 1.

2. For each covariate j, construct an estimating equation for β_{0j} assuming the marginal model that all other covariates are unrelated to the outcome. Denote this marginal estimating equation by $U_j^M(\beta_j)$.

3. Retain the parameters $\hat{\mathcal{M}} = \{j : |U_j^M(0)| \geq \gamma_n\}$ for some threshold γ_n.

Each $|U_j^M(0)|$ is the numerator of the score test statistic for $H_0 : \beta_{0j} = 0$ under the j^{th} marginal model and thus is a sensible screening statistic. We could also screen after dividing each $U_j^M(0)$ by an estimate of its standard deviation. However, this would add computational complexity to our procedure, and even without doing so we will be able to achieve good results and give finite-sample performance guarantees. In the presence of nuisance parameters, such as intercept terms, we propose using profiled score tests, where we first estimate the nuisance parameters under the null model and then evaluate the $U_j^M(0)$ fixing the value of nuisance terms. To avoid theoretical difficulties we will assume that nuisance parameters are either known, or can be well-estimated in independent datasets, so that in the screening step they can be treated as constants.

In order for score screening to have desirable theoretical properties, we need the sample $U_j^M(0)$ to quickly approach its population limit. Let $u_j^M(\beta_j)$ be the limiting marginal estimating equation, such that $U_j^M(\beta_j) \to u_j^M(\beta_j)$.

Condition 1 *For $\kappa \in (0, 1/2)$ and $c_2 > 0$, $p_n \mathrm{P}(|U_j^M(0) - u_j^M(0)| \geq c_2 n^{-\kappa}) \to 0$.*

In Appendix A we discuss the verification of Condition 1, which is often a simple consequence of a concentration inequality, and explicitly verify it for censored quantile regression. We also show that under this condition and a few other mild assumptions:

Theorem 1 *If $\gamma_n = c_1 n^{-\kappa}/2$, then $\mathrm{P}(\mathcal{M} \subseteq \hat{\mathcal{M}}) \to 1$.*

Theorem 2 *If $\gamma_n = c_1 n^{-\kappa}/2$, then $\mathrm{P}\{|\hat{\mathcal{M}}| \leq O(\sigma_{\max}^* n^{2\kappa})\} \to 1$, where $\sigma_{\max}^*$ is related to the largest singular value of the negative Jacobian of the limiting estimating equation.*

Theorem 1 shows that marginal score testing can capture all of the important covariates with high probability. This holds even if p_n grows exponentially in n. Theorem 2 shows that the number of selected covariates is not too large, with high probability. For example, if $\sigma_{\max}^*$ increased only polynomially in n, $|\hat{\mathcal{M}}|$ would increase polynomially, and the false positive rate decreases quickly to zero.

2.2 Examples

When applied to the models studied thus far in the screening literature, score test screening gives procedures that are very similar to previously proposed procedures. Throughout this section we let $K_i = 1$, with each covariate vector $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip_n})^T$. We also assume that the X_{ij} have mean 0 and variance 1.

First consider the usual ordinary least squares model studied by Fan and Lv (2008), where Y_i is a continuous outcome. The full model is $\mathrm{E}(Y_i \mid \mathbf{X}_i) = \mathbf{X}_i^T \boldsymbol{\beta}_0$, so the j^{th} marginal score equation is $U_j^M(\beta_j) = n^{-1} \sum_i X_{ij}(Y_i - X_{ij}\beta_j)$. Score test screening then retains $\hat{\mathcal{M}} = \{j : n^{-1} | \sum_i X_{ij} Y_i | \geq \gamma_n\}$, which is exactly the correlation screening procedure originally proposed by Fan and Lv (2008).

Next consider the Cox model. Let T_i be the survival time, C_i the censoring time, $Y_i = \min(T_i, C_i)$, $\delta_i = I(T_i \leq C_i)$, $\tilde{N}_i(s) = I(T_i \leq s, \delta_i = 1)$, and $\tilde{Y}_i(s) = I(Y_i \geq s)$. The marginal score equations are $U_j^M(\beta_j) =$

$$\frac{1}{n} \sum_{i=1}^{n} \int \left\{ X_{ij} - \frac{\sum_{i=1}^n X_{ij} \exp(X_{ij}\beta_j) \tilde{Y}_i(s)}{\sum_{i=1}^n \exp(X_{ij}\beta_j) \tilde{Y}_i(s)} \right\} d\tilde{N}_i(s),$$

and $\hat{\mathcal{M}} = \{j : |U_j^M(0)| \geq \gamma_n\}$. This is exactly the screening procedure used by Gorst-Rasmussen and Scheike (2013).

Finally consider a nonparametric model, where we assume only that $\mathrm{P}(Y_i < y \mid \mathbf{X}_i)$ has a continuous distribution function $F_0(y; \mathbf{X}_i, \boldsymbol{\beta}_0)$ whose dependence on $\mathbf{X}_i$ is parametrized by $\boldsymbol{\beta}_0$. Conditional on $\mathbf{X}_l$ and $\mathbf{X}_m$, $F_0(Y_l; \mathbf{X}_l, \boldsymbol{\beta}_0)$ and $F_0(Y_m; \mathbf{X}_m, \boldsymbol{\beta}_0)$ are independent and identically distributed uniform random variables. This motivates defining $\mathbf{U}(\boldsymbol{\beta}) =$

$$\frac{1}{n^2} \sum_{m=1}^{n} \sum_{l=1}^{n} \mathbf{X}_l \left[I\{F_0(Y_l; \mathbf{X}_l, \boldsymbol{\beta}) < F_0(Y_m; \mathbf{X}_m, \boldsymbol{\beta})\} - \frac{1}{2} \right].$$

Since $\mathrm{E}\{\mathbf{U}(\boldsymbol{\beta}_0)\} = \mathbf{0}$, this is an unbiased estimating equation for $\boldsymbol{\beta}_0$. Though it cannot be used to estimate $\boldsymbol{\beta}_0$ because the functional form of F_0 is unknown, it is still useful for constructing a screening procedure. The marginal score equations are $U_j^M(\beta_j) =$

$$\frac{1}{n^2} \sum_{m=1}^{n} \sum_{l=1}^{n} X_{lj} \left[I\{F_0(Y_l; X_{lj}, \beta_j) < F_0(Y_m; X_{mj}, \beta_j)\} - \frac{1}{2} \right].$$

When $\beta_j = 0$, $F_0(y; X_{lj}, 0)$ is a monotone function that does not depend on X_{lj}, which implies that $U_j^M(0) = n^{-2} \sum_{lm} X_{lj}\{I(Y_l < Y_m) - 1/2\}$ and therefore $\hat{\mathcal{M}} = \{j : |n^{-2} \sum_{lm} X_{lj} I(Y_l < Y_m)| \geq \gamma_n\}$. This is very similar to proposal of Zhu et al. (2011), who suggested $\hat{\mathcal{M}} = [j : n^{-1} \sum_m \{n^{-1} \sum_l X_{lj} I(Y_l < Y_m)\}^2 \geq \gamma_n]$.

Each of these screening procedures can be implemented as or more quickly than the corresponding Wald screening. In addition, the nonparametric screening procedure is impossible in the Wald framework. Each of these screening procedures can be theoretically justified by verifying Condition 1 and applying Theorems 1 and 2.

3 Threshold for reproducible screening

In practice, it is unclear how best to choose the screening threshold γ_n. Fan and Lv (2008) suggested retaining the top $n/\log n$ covariates. Zhao and Li (2012) proposed a method to choose γ_n based on the desired false positive rate of the set of retained covariates. Similarly, Zhu et al. (2011)

suggested simulating auxiliary variables and setting the threshold so that no auxiliary variables are retained, and proved that this procedure controls the false positive rate of screening. Finally, He and Lin (2011) used the stability selection approach of Meinshausen and Bühlmann (2010) to retain covariates that are frequently selected when screening is performed on multiple subsamples of the data.

Though controlling the false positive rate is important, we believe that in practice the more relevant issue is the reproducibility of the screening procedure. Let $\hat{\mathcal{M}}^{(j)}$ be the top j variables retained after screening our observed data, and let $\mathcal{M}^{(j)}$ be the top j variables we would retain after screening a new sample of the same size. We propose choosing j such that

$$P(j^{-1}|\mathcal{M}^{(j)} \cap \hat{\mathcal{M}}^{(j)}| \geq p \mid \mathcal{D}) \geq r,$$

where p is the proportion of reproducibly retained covariates, r is a probability quantifying the degree of reproducibility. This probability is conditional on the observed dataset $\mathcal{D}$. Here $\mathcal{M}^{(j)}$ is a random set, and we approximate its distribution using the bootstrap. When both p and r are high, we will say that retaining the top j covariates after screening gives a highly reproducible screening procedure.

Thus our threshold for reproducible screening is calculated as follows:

1. Screen observed data to obtain the sets $\hat{\mathcal{M}}^{(j)}$ for $j = 1, \ldots, p_n$.
2. Generate B bootstrap samples and screen the b^{th} sample to get $\mathcal{M}_b^{(j)}$.
3. Let $o_b^{(j)} = |\mathcal{M}_b^{(j)} \cap \hat{\mathcal{M}}^{(j)}|$, so that $r^{(j)} = B^{-1} \sum_b I(j^{-1} o_b^{(j)} \geq p)$ is the reproducibility of keeping the top j covariates.
4. Choose $j^* = \min\{j = 1, \ldots, p_n : r^{(j)} \geq r\}$ and retain $\hat{\mathcal{M}}^{(j^*)}$.

Reproducible screening is similar to stability selection except that the latter is concerned with individual variables, while here we identify *sets* of variables that are reproducible. In Appendix B we describe a dynamic programming algorithm to find j^* without calculating $r^{(j)}$ for all $j = 1, \ldots, j^*$.

In some cases we may find that j^* is very small, indicating that there are a few covariates which are consistently highly ranked after score test screening across the bootstrap samples. For example, the same covariate may have the highest $|U_j^M(0)|$ in each bootstrap sample, in which case $j^* = 1$, and it is likely that we will miss some important covariates. In these cases we can let $j^* = \min\{j = s, \ldots, p_n : r^{(j)} \geq r\}, s > 1$, so that we are guaranteed to retain at least s variables. We can choose s according to how many truly important covariates we think there are. We can be fairly imprecise in choosing this number because for moderate s, such as $s \approx 10$, it is unlikely that the top s covariates will all be highly reproducible, meaning that j^* will frequently be much larger than s.

4 Iterative score test screening

When the covariates are highly correlated, marginal screening may incur a large number of false positives, and may miss covariates that are only important conditional on other covariates. Fan and Lv (2008) and Fan et al. (2009) therefore proposed iterative screening: an initial set of covariates is first identified using marginal screening. Next a multivariate regularized selection procedure is used to further select a subset of these covariates. Finally the remaining covariates are again screened individually, but this time controlling for the covariates in the subset. All selected covariates are

subjected to multivariate selection again, and the procedure iterates until some stopping rule is achieved.

However, this algorithm requires fitting regularized regression estimates at each step, which for complicated models can be difficult to implement and computationally intensive. Furthermore, its theoretical properties are very difficult to analyze. Zhu et al. (2011) proposed an alternative method which at each step performs marginal screening on the projections of each remaining covariate onto the orthogonal complement of the columns space of the already selected covariates. This method is akin to forward selection, so a covariate cannot be dropped from the selected set once it has been added.

Our score-test screening perspective suggests a new approach to iterative screening:

1. Set $\boldsymbol{\beta}^{(0)} = \mathbf{0}$.
2. For $k = 1, \ldots, K$:
   (a) Let $\mathbf{b}^{(k)} = \boldsymbol{\beta}^{(k-1)} - \alpha_k \mathbf{U}(\boldsymbol{\beta}^{(k-1)})$ for some step size α_k.
   (b) Let $\boldsymbol{\beta}^{(k)} = \Pi_R(\mathbf{b}^{(k)})$, where $\Pi_R : \mathbb{R}^{p_n} \to \mathbb{R}^{p_n}$ is the Euclidean projection onto the ℓ_1-ball of radius R.
3. Retain covariates $\hat{\mathcal{M}} = \{j : \beta_j^{(K)} \neq 0\}$, where $\beta_j^{(k)}$ is the j^{th} component of $\boldsymbol{\beta}^{(k)}$.

The intuition is that when $k = 1$, step 2(a) is equivalent to calculating the marginal score statistics $U_j^M(0)$ and step 2(b) sets all but the largest of them to zero. Thus after a single iteration, this procedure is identical to score test screening. When $k > 1$, step 2(a) controls for the covariates selected in $\boldsymbol{\beta}^{(k-1)}$ by using $-\alpha_k \mathbf{U}(\boldsymbol{\beta}^{(k-1)})$ to update the importance of the covariate. Step 2(b) then again selects only the top covariates. In the ideal case where the sample size is infinite and $\boldsymbol{\beta}^{(k-1)} = \boldsymbol{\beta}_0$, step 2(a) gives $\mathbf{b}^{(k)} = \boldsymbol{\beta}_0$ and step 2(b) selects the largest components of $\boldsymbol{\beta}_0$.

Our algorithm has several advantages. First, it does not require fitting any regularized regression estimates and is relatively computationally convenient. The evaluations of the $\mathbf{U}(\boldsymbol{\beta}^{(k-1)}$ are quick to compute, and a simple algorithm for implementing the projection Π_R can be found in Daubechies et al. (2008), with a more efficient procedure proposed by Duchi et al. (2008). Second, covariates can be dropped from the retained set as the iteration progresses, which is an improvement over forward selection. Third, our algorithm exactly corresponds to projected subgradient methods for minimizing nonsmooth functions. In fact, if $\mathbf{U}(\boldsymbol{\beta})$ is the subdifferential of some loss function $f(\boldsymbol{\beta})$, it has been shown that

$$\lim_{k \to \infty} f(\boldsymbol{\beta}^{(k)}) = \inf_{\|\boldsymbol{\beta}\|_1 \leq R} f(\boldsymbol{\beta})$$

for certain choices of α_k (Shor et al., 1985). The minimization problem on the right-hand side is exactly equivalent to the lasso (Tibshirani, 1996) with loss function f, and this links our iterative screening algorithm to sparse regression methods. Finally, when f is smooth, Agarwal et al. (2012) proved that $\boldsymbol{\beta}^{(k)}$ converges to $\boldsymbol{\beta}_0$ under certain conditions, and if a similar result holds for nonsmooth f, this connection may allow for a theoretical analysis of iterative score test screening.

There are three tuning parameters we must set when implementing iterative screening: the radius R, the step sizes α_k, and the maximum number of iterations. We can choose R by either guessing the ℓ_1-norm of the true $\boldsymbol{\beta}_0$ or using the reproducible screening criterion described in Section 3. Since our algorithm can be viewed as a regression estimator, we can also minimize information criteria or cross-validated prediction errors. Since iterative screening tends to be time-consuming in high-dimensions, it is easiest avoid resampling or cross-validation and to use a liberal

guess for $\|\boldsymbol{\beta}_0\|_1$. To set the step sizes, one popular rule is to let the α_k be square summable but not summable, with $\alpha_k = \gamma/k$. To choose γ, we first note that it can be shown that

$$\min_{k=1,\ldots,K} f(\boldsymbol{\beta}^{(k)}) - \inf_{\|\boldsymbol{\beta}\|_1 \leq R} f(\boldsymbol{\beta}) \leq \frac{D^2 + G^2 \sum_{k=1}^K \alpha_k^2}{2 \sum_{k=1}^K \alpha_k},$$

where D is the Euclidean distance from $\boldsymbol{\beta}^{(0)}$ to a point that minimizes f and G is an upper bound on $\mathbf{U}(\boldsymbol{\beta}^{(k)})$ for all k (Shor et al., 1985). When $\alpha_k = \gamma/k$, this converges to zero as $K \to \infty$, but fixing K we can derive that the right-hand side is minimized at $\gamma^2 = D^2(G^2 \sum_{k=1}^K \alpha_k^2)^{-1} \to D^2(G^2\pi^2/6)^{-1}$. We propose approximating D by R and G by $\|\mathbf{U}(\boldsymbol{\beta}^{(0)})\|_2$ to get step sizes $\alpha_k = R\sqrt{6}/\{k\pi\|\mathbf{U}(\mathbf{0})\|_2\}$. Finally, the maximum number of iterations should ideally be as large as possible, with the speed of convergence depending on the restricted convexity and smoothness of f (Agarwal et al., 2012). In practice we stop after either $\mathbf{U}(\boldsymbol{\beta}^{(k)}) \approx \mathbf{0}$, $\boldsymbol{\beta}^{(k-1)} \approx \boldsymbol{\beta}^{(k)}$, or $K = 250$ iterations. Early stopping can be viewed as another way of regularizing the regression estimate $\boldsymbol{\beta}$.

5 Simulations

5.1 Settings

We illustrate our marginal and iterative score test screening on data simulated from four models, described below along with their corresponding estimating equations. We ran 100 simulations, each with $n = 200$ observations and $p_n = 10,000$ covariates. We compared our methods to the semiparametric screening of Zhu et al. (2011), and when possible we also compared to Wald and nonparametric screening.

Example 1 (accelerated failure time model). The accelerated failure time model is a useful alternative to the Cox model for survival outcomes (Wei, 1992) and posits that $\log(T_i) = \mathbf{X}_i^T\boldsymbol{\beta}_0 + \epsilon_i$, where T_i are the survival times, $\mathbf{X}_i$ are $p_n \times 1$ covariate vectors, and ϵ_i are independent of $\mathbf{X}_i$. We only observe follow-up times $Y_i = \min(T_i, C_i)$ and censoring indicators $\delta_i = I(T_i \leq C_i)$, but the $\boldsymbol{\beta}_0$ can be estimated using the estimating equation $\mathbf{U}(\boldsymbol{\beta}) =$

$$n^{-1} \sum_{l=1}^n \sum_{m=1}^n (\mathbf{X}_m - \mathbf{X}_l) I\{e_l(\boldsymbol{\beta}) \leq e_m(\boldsymbol{\beta})\}\delta_i,$$

where $e_i(\boldsymbol{\beta}) = \log(Y_i) - \mathbf{X}_i^T\boldsymbol{\beta}$ (Tsiatis, 1996; Jin et al., 2003; Cai et al., 2009).

Score test screening retains

$$\hat{\mathcal{M}} = \{j : |\sum_{lm} (X_{mj} - X_{lj}) I(Y_l \leq Y_m)\delta_l| \geq \gamma_n\},$$

and it is simple to verify Condition 1 for this procedure using Berstein's inequality for U-statistics (Hoeffding, 1963). We implemented Wald test screening using the estimator of Jin et al. (2003), available in the R package `lss`. Nonparametric screening has not been developed for this model.

We generated the covariates from a p-dimensional multivariate normal with a covariance matrix whose jk^{th} entry equaled $0.5^{|j-k|}$. We then let the first 10 elements of $\boldsymbol{\beta}_0$ increase from 1 to 3.25 in increments in 0.25 and set the rest equal to zero. We generated ϵ_i from a standard normal distribution, T_i according to the model, and C_i from an exponential distribution with rate parameter 0.25 to give 40% censoring.

Example 2 (linear censored quantile regression). For a quantile $\tau \in (0, 1)$, censored quantile regression models posit $h(T_i) = \beta_{int}(\tau) + \mathbf{X}_i^T\boldsymbol{\beta}_0(\tau) + e_i(\tau)$, where the intercept $\beta_{int}(\tau)$ and the

coefficients $\boldsymbol{\beta}_0(\tau)$ depend on τ and $e_i(\tau)$ has τ^{th} quantile equal to 0 conditional on $\mathbf{X}_i$. The h function is a known monotone transformation, and here we let it be the log function. In contrast to global models such as the Cox or accelerated failure time model, this censored quantile regression directly models the τ conditional quantile and makes no assumptions about the other quantiles. Honore et al. (2002) proposed the estimating equation $\mathbf{U}(\boldsymbol{\beta}) =$

$$\begin{aligned}&\tfrac{1}{n}\textstyle\sum_i \mathbf{X}_i \tau I\{h(Y_i) > \beta_{int} + \mathbf{X}_i^T\boldsymbol{\beta}\} - \\ &\tfrac{1}{n}\textstyle\sum_i \mathbf{X}_i (1-\tau)\hat{S}_{h(C)}\{h(Y_i)\}^{-1} I\{h(Y_i) \le \beta_{int} + \mathbf{X}_i^T\boldsymbol{\beta}\}\delta_i \hat{S}_{h(C)}(\beta_{int} + \mathbf{X}_i^T\boldsymbol{\beta}),\end{aligned}$$

where $\hat{S}_{h(C)}$ is an estimate of $S_{h(C)}(t) = \mathrm{P}\{h(C_i) \ge t \mid \mathbf{X}_i\}$. This estimate could be obtained by positing a regression model for $h(C_i)$ conditional on the $\mathbf{X}_i$, but for theoretical and practical simplicity we will make the common assumption that C_i is completely independent of T_i and $\mathbf{X}_i$ and use the Kaplan-Meier estimator (see for example Cheng et al. (1995), Uno et al. (2011), and He et al. (2013)).

Score test screening retains the parameters $\{j : |U_j^M(0)| \ge \gamma_n\}$, where $U_j^M(0) =$

$$\frac{1}{n}\sum_i X_{ij}\tau I\{h(Y_i) > \beta_{int}\} - \frac{1}{n}\sum_i X_{ij}(1-\tau)\hat{S}_{h(C)}\{h(Y_i)\}^{-1} I\{h(Y_i) \le \beta_{int}\}\delta_i \hat{S}_{h(C)}(\beta_{int}).$$

In Appendix A we verify Condition 1 for this screening procedure. To use score test screening, we first estimated the nuisance parameter β_{int} under the null model in an independently simulated dataset. We implemented Wald screening using the estimator of of Peng and Huang (2008), available in the package `quantreg`. He et al. (2013) developed a nonparametric screening method for quantile regression, which we also applied.

We generated the covariates from a p-dimensional multivariate normal with a covariance matrix whose jk^{th} entry equaled $0.8^{|j-k|}$. We let $(\beta_{0,5}, \beta_{0,10}, \beta_{0,15}, \ldots, \beta_{0,50})$ equal

$$(0.356, 0.480, -1.507, 0.937, -1.660, -0.021, -0.491, -1.071, 0.693, 0.478)$$

and simulated $\log(T_i) = \mathbf{X}_i^T\boldsymbol{\beta}_0 + \epsilon_i\{5 + (0.5X_{i55} - 0.356X_{i5})/\Phi^{-1}(0.25)\}$, where ϵ_i followed a standard normal distribution. Under this construction the covariates $j \in \{5, \ldots, 50\}$ are associated with the $\tau = 0.5$ conditional quantile, and the covariates $j \in \{10, \ldots, 55\}$ are relevant to the $\tau = 0.25$ conditional quantile. We separated the nonzero entries of $\boldsymbol{\beta}_0$ so that important covariates would be fairly correlated with a few unimportant covariates. The nonzero magnitudes of $\boldsymbol{\beta}_0$ are such that when $\tau = 0.5$, the 20^{th} covariate is marginally unimportant, and when $\tau = 0.25$ the 10^{th} and 20^{th} covariates are marginally unimportant. We simulated C_i from an exponential distribution with rate 0.1 to give 40% censoring.

Example 3 (nonlinear censored quantile regression). We generated survival times from a nonlinear censored quantile regression model similar to the one used by He et al. (2013). If $g_1(x) = x$, $g_2(x) = (2x-1)^2$, $g_3(x) = \sin(2\pi x)/\{2 - \sin(2\pi x)\}$, $g_4(x) = 0.1\sin(2\pi x) + 0.2\cos(2\pi x) + 0.3\sin(2\pi x)^2 + 0.4\cos(2\pi x)^3 + 0.5\sin(2\pi x)^3$, we simulated

$$\log(T_i) = 5g_1(X_{i5}) + 3g_2(X_{i10}) + 4g_3(X_{i15}) + 6g_4(X_{i20}) + \epsilon_i\left[40 + \frac{5\{g_1(X_{i25}) - g_1(X_{i5})\}}{\Phi^{-1}(0.25)}\right],$$

where ϵ_i followed a standard normal distribution.

Under the null hypothesis the functions $g_j = 0$ for all j, so the marginal estimating equations $U_j^M(0)$ evaluated at zero are identical to those from Example 2. The theoretical justifications thus also follow from Example 2. We applied the nonparametric screening of He et al. (2013) as well, which was designed for this nonlinear setting.

Table 1: Average runtime (seconds) of different screening methods.

Example	Wald	Score	Zhu et al. (2011)	He et al. (2013)
1	3249.68	1.53	5.19	
2	344.60	2.10	5.10	125.04
3		2.11	5.19	125.74
4		2.22	5.98	

Under our construction, the covariates $j \in \{5, \ldots, 20\}$ are relevant to the $\tau = 0.5$ conditional quantile and the covariates $j \in \{10, \ldots, 25\}$ are relevant to the $\tau = 0.25$ conditional quantile. We generated the $\mathbf{X}_i$ as in Example 3, and $\log(C_i)$ from the mixture distribution $0.4N(0, 4) + 0.1N(10, 1) + 0.5N(100, 1)$ to give 40% censoring.

Example 4 (Cox model with measurement error). The Cox model is the most popular method for modeling the effect of covariates on survival, but in many cases the covariates may be measured with errors, where instead of observing $\mathbf{X}_i$ we observe only $\mathbf{W}_i = \mathbf{X}_i + \boldsymbol{\epsilon}_i$. Not accounting for measurement error can result in bias, and to address this issue Song and Huang (2005) proposed the corrected score equation $\mathbf{U}(\boldsymbol{\beta}) =$

$$\frac{1}{n}\sum_{i=1}^{n}\int\left[\mathbf{W}_i + D(\boldsymbol{\beta}) - \frac{\sum_{i=1}^{n}\tilde{\mathbf{W}}_i(\boldsymbol{\beta}, s)\exp\{\tilde{\mathbf{W}}_i(\boldsymbol{\beta}, s)^T\boldsymbol{\beta}\}\tilde{Y}_i(s)}{\sum_{i=1}^{n}\exp\{\tilde{\mathbf{W}}_i(\boldsymbol{\beta}, s)^T\boldsymbol{\beta}\}\tilde{Y}_i(s)}\right]d\tilde{N}_i(s),$$

where $\tilde{\mathbf{W}}_i(\boldsymbol{\beta}, s) = \mathbf{W}_i + D(\boldsymbol{\beta})d\tilde{N}_i(s)$, $D(\boldsymbol{\beta}) = \mathrm{E}\{\boldsymbol{\varepsilon}_i\exp(\boldsymbol{\varepsilon}_i^T\boldsymbol{\beta})\}/\mathrm{E}\{\exp(\boldsymbol{\varepsilon}_i^T\boldsymbol{\beta})\} - \mathrm{E}(\boldsymbol{\varepsilon}_i)$, $\tilde{N}_i(s) = I(T_i \leq s, \delta_i = 1)$ is the observed failure process, and $\tilde{Y}_i(s) = I(Y_i \geq s)$ is the at-risk process. The $D(\boldsymbol{\beta})$ term is unknown in general unless the distribution of $\boldsymbol{\varepsilon}_i$ is known.

Under the null hypothesis of $\boldsymbol{\beta}_0 = \mathbf{0}$, $D(\mathbf{0}) = \mathbf{0}$, so score test screening retains

$$\hat{\mathcal{M}} = \left[j : \left|\frac{1}{n}\sum_{i=1}^{n}\int\left\{W_{ij} - \frac{\sum_{i=1}^{n}W_{ij}\tilde{Y}_i(s)}{\sum_{i=1}^{n}\tilde{Y}_i(s)}\right\}d\tilde{N}_i(s)\right| \geq \gamma_n\right]$$

regardless of the distribution of $\boldsymbol{\epsilon}_i$. Condition 1 can be verified using Lemmas 2 and 3 of Gorst-Rasmussen and Scheike (2013). Wald screening is not possible without knowing the distribution of $\boldsymbol{\epsilon}_i$, and nonparametric screening has not been developed for this model.

We generated the covariates and set $\boldsymbol{\beta}_0$ as in Example 2, so that again the 20^{th} covariate appears marginally unimportant. We then generated the T_i from the usual Cox model with baseline hazard function equal to 1. Next we let $\mathbf{W}_i = \mathbf{X}_i + \boldsymbol{\varepsilon}_i$, where the $\boldsymbol{\varepsilon}_i$ were independent of the $\mathbf{X}_i$ and normally distributed with a compound symmetry covariance matrix with correlation parameter 0.5. We generated C_i from an exponential distribution with rate parameter 0.2 to give 40% censoring.

5.2 Results

These simulations were run on machines with 2 GHz Intel Xeon cores with 4GB of memory per core. Table 1 reports the average runtimes of these various screening methods and shows that our marginal score test procedure is by far the most computationally efficient. In Example 1 it is nearly 2000 times faster than Wald screening, and in Examples 2 and 3 it is 60 times faster than the nonparametric method of He et al. (2013). In each example it is also twice as fast as the semiparametric estimator of Zhu et al. (2011).

Table 2 compares score test screening to existing methods in terms of the minimum number of variables that need to be retained in order to capture all of the important covariates. In Example 1,

Table 2: Medians (interquartile ranges) of minimum model sizes required to retain the covariates in the second column. In Example 2, $\beta_{0,5}$ is relevant only when $\tau = 0.5$ and $\beta_{0,55}$ is relevant only when $\tau = 0.25$. Similarly, in Example 3 $\beta_{0,5}$ is relevant only when $\tau = 0.5$ and $\beta_{0,25}$ is relevant only when $\tau = 0.25$.

Covariates	Wald	Score	Zhu et al. (2011)	He et al. (2013)
Example 1				
All	60.5 (228.75)	44 (234.25)	77 (348.25)	
Example 2, $\tau = 0.5$				
All	8575.5 (2037.5)	8526.5 (2543.5)	7500 (2584.75)	9539 (642)
$\beta_{0,5}$	3808 (4587)	3683.5 (4869.75)	2668.5 (4298.75)	5088 (4908.5)
$\beta_{0,55}$	4012 (5558)	3878 (4297.25)	2626 (4609)	4959 (6458.5)
Example 2, $\tau = 0.25$				
All	8012 (3131.75)	8049 (2608)	7664 (2518.5)	9710.5 (613)
$\beta_{0,5}$	4986.5 (4955.25)	5010 (4592.25)	2668.5 (4298.75)	3947.5 (4443)
$\beta_{0,55}$	1044 (2603.25)	1496 (3203)	2626 (4609)	4761.5 (6826.25)
Example 3, $\tau = 0.5$				
All		7792.5 (3253)	7400.5 (3274.25)	7827 (3038.25)
$\beta_{0,5}$		3452.5 (4538.75)	4195.5 (4909.25)	4349.5 (5049)
$\beta_{0,25}$		4674 (4982)	2979.5 (3455.5)	5194.5 (5211.25)
Example 3, $\tau = 0.25$				
All		7361 (3134.75)	7002.5 (3694.5)	7421 (3437)
$\beta_{0,5}$		4945.5 (5553.5)	4195.5 (4909.25)	4221 (4905)
$\beta_{0,25}$		1919 (4644)	2979.5 (3455.5)	2770 (4482.75)
Example 4				
All		7879 (2708.75)	7552.5 (2736.5)	

score test screening performed better than both Wald and the semiparametric screening of Zhu et al. (2011). In Example 2, score screening was comparable to Wald screening and outperformed the nonparametric screening of He et al. (2013). Semiparametric screening performed the best but was unable to identify the fact that $\beta_{0,5}$ was important only to the 0.5 quantile and $\beta_{0,55}$ was important only to the 0.25 quantile. In contrast, the other methods correctly assigned their relative rankings for the different quantiles. The same trends held in Example 3, though Wald screening was not possible. In Example 4 the only two screening methods that could accommodate the unknown measurement error distribution were score and semiparametric screening, which performed similarly.

Table 3 compares the performance of our threshold for reproducible screening, described in Section 3, to the $n/\log n$ rule of Fan and Lv (2008) and the auxiliary variables method of Zhu et al. (2011). To implement reproducible screening we generated 100 bootstrap samples and found the threshold at which 70% of the covariates retained after screening the observed data were also retained after screening in 70% of the bootstrap samples. Because the signals were relatively strong in Example 1, we set $s = 10$ in our reproducible screening algorithm to guarantee that we would retain at least 10 covariates. We see that except for in Example 1, where all methods performed well, our reproducible screening procedure had the highest average true positive rates, at close to 80%. At the same time it also reduced the number of covariates by about half. Since screening procedures are typically followed by a second variable selection step, such as regularized regression, they are designed to capture as many important covariates as possible, with a high true discovery rate as a secondary concern. While the auxiliary variables method had the highest true discovery rates, it had very low true positive rates, which is undesirable in a screening procedure.

Table 4 reports the performance of our iterative screening procedure from Section 4, which we applied to the parametric models in Examples 1 and 2 with $R = 20$. In Example 1 all of the important variables were also marginally important, and iterative screening was able to capture nearly all of them. In Example 2, one of the important variables appeared marginally irrelevant to the 0.5 quantile and two were marginally irrelevant to the 0.25 quantile. However, iterative screening was indeed able to capture at least one of the marginally unimportant variables after retaining only around 800 variables, as opposed to marginal score screening, which had to retain thousands of variables.

6 Data analysis

6.1 Analysis methods

We were interested in identifying genes highly associated with the 10% conditional quantile of the survival distribution of MM patients, because these genes are likely to important in high-risk MM. Previous studies have searched for genes associated with patient survival (Shaughnessy et al., 2007; Decaux et al., 2008), but their analyses did not recognize that some genes may only affect certain quantiles of the conditional survival distribution.

We used gene expression and survival outcome data from newly diagnosed multiple myeloma patients who were recruited into clinical trials UARK 98-026 and UARK 2003-33, which studied the total therapy II (TT2) and total therapy III (TT3) treatment regimes, respectively. These data are described in Zhan et al. (2006) and Shaughnessy et al. (2007), and can be obtained through the MicroArray Quality Control Consortium II study (Shi et al., 2010), available on GEO (GSE24080). Gene expression profiling was performed using Affymetrix U133Plus2.0 microarrays, and we averaged the expression levels of probesets corresponding to the same gene, resulting in 33,326 covariates. We used the TT2 arm as a training set, giving us 340 subjects and 126 observed deaths, we validated the results on the TT3 arm.

Table 3: Performance of different methods for choosing the screening threshold. Methods: RS = reproducible screening, described in Section 3; Auxiliary = auxiliary variables method of Zhu et al. (2011). Average performance metrics (standard deviation): TP = true positive rate, TD = true discovery rate. Median size is reported (interquartile range).

Screening	Threshold	TP	TD	Size
Example 1				
Wald	$n/\log n$	93.2 (6.8)	25.19 (1.84)	37 (0)
Score	RS	89.5 (7.16)	78.52 (28.36)	10 (0)
Zhu et al. (2011)	Auxiliary	84.6 (9.26)	88.86 (12.04)	9 (2.25)
Example 2, $\tau = 0.5$				
Wald	$n/\log n$	26.9 (10.22)	7.27 (2.76)	37 (0)
Score	RS	80.4 (20.3)	5.38 (21.96)	6344 (41)
Zhu et al. (2011)	Auxiliary	16.3 (9.6)	41.87 (28.26)	4 (4.25)
He et al. (2013)	$n/\log n$	0.6 (2.39)	0.16 (0.65)	37 (0)
Example 2, $\tau = 0.25$				
Wald	$n/\log n$	25.4 (9.89)	6.86 (2.67)	37 (0)
Score	RS	87.3 (12.54)	0.39 (2.49)	6352.5 (46.75)
Zhu et al. (2011)	Auxiliary	16.3 (9.6)	41.87 (28.26)	4 (4.25)
He et al. (2013)	$n/\log n$	0.5 (2.19)	0.14 (0.59)	37 (0)
Example 3, $\tau = 0.5$				
Score	RS	77.5 (19.3)	0.05 (0.01)	6367 (35.5)
Zhu et al. (2011)	Auxiliary	0.5 (3.52)	2 (14.07)	0 (1)
He et al. (2013)	$n/\log n$	7.25 (11.94)	0.78 (1.29)	37 (0)
Example 3, $\tau = 0.25$				
Score	RS	79.25 (19.48)	0.05 (0.01)	6377.5 (45.25)
Zhu et al. (2011)	Auxiliary	0.5 (3.52)	2 (14.07)	0 (1)
He et al. (2013)	$n/\log n$	9.25 (12.64)	1 (1.37)	37 (0)
Example 4				
Score	RS	84.3 (20.85)	3.14 (9.85)	6393 (299.75)
Zhu et al. (2011)	Auxiliary	28.1 (11.25)	33.83 (13.76)	8 (7)

Table 4: Performance of iterative screening. The second column reports the average percentage of times (SD) the marginally unimportant variables (see Section 5.1) were capture by iterative screening. Average performance metrics (standard deviation): TP = true positive rate, TN = true negative rate, TD = true discovery rate, TND = true nondiscovery rate. Median size is reported (interquartile range).

Hidden	TP	TD	Size
Example 1			
	96 (4.92)	16.85 (5.59)	61.5 (24)
Example 2, $\tau = 0.5$			
9 (28.76)	47.6 (13.72)	0.73 (0.41)	860.5 (297.25)
Example 2, $\tau = 0.25$			
15 (35.89)	50.1 (14.6)	0.93 (0.58)	830.5 (858)

To identify these high-risk genes we used the censored quantile regression of Honore et al. (2002), described earlier in Example 2 in Section 5.1, with the transformation function $h = \log$. First, in the screening step we compared Wald screening with the estimator of Peng and Huang (2008), marginal score screening, the semiparametric method of Zhu et al. (2011), the nonparametric method of He et al. (2013), and iterative score screening. In the score screening procedures we estimated the nuisance intercept parameter from another MM dataset collected by Avet-Loiseau et al. (2009). For iterative score screening we set $R = 20$.

Second, to set a screening threshold we retained the top $n/\log n$ covariates from Wald and nonparametric screening, used our reproducible screening threshold for score screening, and used the auxiliary variables procedure of Zhu et al. (2011) for semiparametric screening. For reproducible screening we searched for the threshold for which at least 90% of the retained covariates were found in at least 90% of the 100 bootstrap samples, while specifying that the result should contain at least 10 covariates.

Finally, we used the screened covariates to estimate regression models. To our knowledge there do not exist any computationally convenient procedures for censored quantile regression for arbitrary quantiles that can be computed in high-dimensions, so we used our projected subgradient method from Section 4 to serve as a regression estimator. We tuned the procedure by selecting the value of R that minimized an approximate Bayesian Information Criterion, which we calculated as $\|n\mathbf{U}(\hat{\boldsymbol{\beta}}_R)\|_2^2 + \|\hat{\boldsymbol{\beta}}_R\|_0 \log n$ with $\mathbf{U}$ the estimating equation of Honore et al. (2002) and $\hat{\boldsymbol{\beta}}_R$ the regression estimate for a given value of R.

6.2 Results

Wald screening required 930 seconds, the nonparametric screening of He et al. (2013) required 240 seconds, iterative score screening required 84 seconds, the semiparametric screening of Zhu et al. (2011) required 44 seconds, and marginal screening took only 5 seconds. Because of the computational efficiency of score screening, calculating the reproducible screening threshold required only 685 seconds, which was still faster than Wald screening.

Table 5 reports the genes selected in the final censored quantile regression models, which share no common genes. One possible reason is that the correlations between the selected genes were not low. For example, among the top 100 genes selected by Wald screening, 20% of the pairwise correlations were above 0.25 and the largest reached 0.73, and for score screening 20% of the correlations were

Table 5: Final regression models for the 0.1 conditional quantile of MM patient survival. The regression model after marginal score screening contained 49 genes and is not shown. Validation metrics: PE = prediction error (1), Coefficient = coefficient of the regression of the true 0.1 quantile on the predicted quantile, using Wang and Wang (2009), P-value = bootstrap p-value of the coefficient.

Validation	Wald	Zhu et al. (2011)	He et al. (2013)	Iterative	Score
	CDK13	ADAR	ATP6	hnRNPK	
	MAPKAP1	ATP6V0E1	CARD8	hnRNPKP4	
	PEX11B	DPY30	CTCF	MATR3	⋮
	VCP	HNRNPU	DDX3X	OAZ1	
		NOLC1		RAB10	
				RPS3A	⋮
				SPCS1	
				SPCS2	
				SUMO2	⋮
				TMBIM4	
				UBC	
PE	0.66	1.16	0.43	**0.08**	0.59
Coefficient	-1.98	-1.04	3.86	**5.89**	2.97
P-value	0.2431	0.5396	0.0231	**0.0005**	0.0811

at least 0.58 and reached 0.99. In other words, the different screening methods most likely selected blocks of correlated covariates together, and the same covariates could be ranked very differently by different methods if they were in different blocks. This highlights the importance of quantifying the reproducibility of a screening procedure. With our reproducible screening threshold approach, we can be confident that at least score screening will retain similar sets of variables when used with different samples observed from the same data generating mechanism.

To choose between the four models, we used the fitted regression models to predict the 0.1 conditional quantiles in the TT3 arm and calculated validation metrics in two ways. First, to estimate the quantile prediction error we used the loss function

$$n^{-1}\sum_i \frac{\delta_i}{\hat{S}_{h(C)}\{h(Y_i)\}}\{\tau - I(Y_i - \hat{Y}_i < 0)\}Y_i, \tag{1}$$

where δ_i is the censoring indicator, Y_i is the observed follow-up time, $\tau = 0.1$ is the target quantile, and $\hat{Y}_i$ is the predicted τ conditional quantile. A similar loss function was described by Honore et al. (2002). Second, we used the locally weighted censored quantile regression approach of Wang and Wang (2009) to estimate the associations between the predicted quantiles and the true 0.1 quantile. Table 5 shows that the model selected after iterative score screening performs the best under both validation metrics, followed by semiparametric screening and marginal score screening. In contrast, the quantiles predicted after Wald and semiparametric screening were actually negatively associated with the true quantile. This suggests that the true relationship between the genes and the quantile may be significantly nonlinear. This nonlinearity can still be detected by the score screening methods.

7 Discussion

Motivated by our analysis of genomic factors influencing the high risk multiple myeloma patients, we introduced a new framework for variable screening based on score tests. Score screening is widely applicable to parametric, semiparametric, and nonparametric models, relatively easy to theoretically justify, and computationally efficient. Using score test screening in our MM analysis resulted in a predictive model for the conditional 10% quantile (high risk group) which was more accurate the models obtained using other screening methods.

We introduced a method for selecting the number of covariates to retain based on the principle of reproducible screening. We have so far used the bootstrap to estimate the reproducibility of different selection thresholds, and it would be of interest to establish theoretical guarantees that such a procedure can indeed give reproducible screening. Our score testing framework also suggested a new approach to iterative screening based on projected subgradient methods, which can be applied even to nonsmooth estimating equations. It is related to sparse regression techniques and it is possible that this connection can lead to better a theoretical understanding of iterative screening, which is still elusive.

Acknowledgements

We are grateful to the editor, the associate editor, and the anonymous referee for their helpful comments. We also thank Professors Lee Dicker and Julian Wolfson for reading an earlier version of this manuscript. This research is partially supported by an NIH grant.

A Theoretical properties of score test screening

Theoretical justifications can be easier to derive for score test-based screening compared to Wald test-based screening. The main task is to find a finite-sample bound for $\mathrm{P}\{|U_j^M(0)| \geq \gamma_n\}$, which can often be done by applying Bernstein-type inequalities. In contrast, Wald test-based screening requires deriving non-asymptotic tail bounds for the marginal estimators, which can be considerably more involved. We will give a sufficient condition that, under certain assumptions, will guarantee sure screening and false positive control.

We assume throughout that the covariates have mean 0 and variance 1. Let $u_j^M(\beta_j)$ be the limiting marginal estimating equation, such that $U_j^M(\beta_j) \to u_j^M(\beta_j)$.

Assumption 1 *There exists some constant* $c_1 > 0$ *such that* $\min_{j \in \mathcal{M}} |u_j^M(0)| \geq c_1 n^{-\kappa}$ *with* $0 < \kappa < 1/2$.

Assumption 2 $\|\mathbf{u}(\mathbf{0})\|_2^2 = \sqrt{\sum_j u_j^M(0)^2}$.

Assumption 3 *The negative Jacobian* $\mathbf{i}(\boldsymbol{\beta}) = -\partial \mathbf{u}/\partial \boldsymbol{\beta}$ *exists.*

Assumption 4 *There exists some constant* $c_3 > 0$ *such that* $\|\boldsymbol{\beta}_0\|_2 \leq c_3$.

Assumption 1 ensures that the limiting numerator of the marginal score test for $H_0 : \beta_{0j} = 0$, is still large enough to detect. For example, for generalized linear models when $K_i = 1$, $u_j^M(\beta_j) = n^{-1} \sum_i \mathrm{E}\{X_{ij}(Y_i - g^{-1}(X_{ij}\beta_j)\}$ with g equal to the canonical link function, so Assumption 1 is equivalent to assuming that $|\mathrm{cov}\{g^{-1}(\mathbf{X}_i^T \boldsymbol{\beta}_0), X_{ij}\}| \geq c_1 n^{-\kappa}$. Fan and Song (2010) make exactly

this assumption to prove the sure screening property in their Theorem 4(ii). Assumption 2 relates the marginal expected estimating equations to the full expected estimating equation. This is a mild assumption, because frequently $u_j\{(\mathbf{0})\} = u_j^M(0)$, where u_j is the j^{th} component of $\mathbf{u}$. This holds, for example, for generalized estimating equations when $\mathrm{E}(\mathbf{Y}_i \mid \mathbf{X}_i) = \mu(\mathbf{X}_i\boldsymbol{\beta}_0)$ for some mean function μ (Zeger et al., 1988). Assumption 3 can hold even if the sample estimating equation $\mathbf{U}$ is nondifferentiable. Assumption 4 merely requires that there exist an upper bound on the size of the true $\boldsymbol{\beta}_0$ that does not grow with n, which is a reasonable condition

Next we give a sufficient condition which will ensure good screening properties.

Condition 1 For κ from Assumption 1 and any constant $c_2 > 0$, $p_n\mathrm{P}(|U_j^M(0) - u_j^M(0)| \geq c_2 n^{-\kappa}) \to 0$.

Condition 1 requires that that the probability that $U_j^M(0)$ is not close to $u_j^M(0)$ approaches 0 faster than p_n approaches ∞, so that we can use $U_j^M(0)$ for screening in high dimensions. This condition must be separately verified for different regression models. Condition 1 will often hold even when p_n grows exponentially in n.

For many estimating equations, to verify Condition 1 we need additional assumptions on the tails of $\mathbf{X}_i$ and $\mathbf{Y}_i$ and on the rate of p_n, such as the following:

Assumption 5 *There exist constants $l_0, l_1, \eta > 0$ such that for all j, $\mathrm{P}(|X_{ij}| > s) \leq l_0 \exp(-l_1 s^\eta)$ for sufficiently large s.*

Tail conditions of this type were also assumed in Fan and Song (2010) and Gorst-Rasmussen and Scheike (2013). When $U_j^M(0)$ is a sum of independent random variables, we can appeal to the usual Bernstein's inequality. A similar approach applies when $U_j^M(0)$ is a U-statistic (Hoeffding, 1963). Establishing this condition for more complicated $U_j^M(0)$, such as the marginal score equations for the Cox model, is more involved (Gorst-Rasmussen and Scheike, 2013).

Under these assumptions, and given the sufficient condition, score test screening has the following theoretical guarantees:

Theorem 1 Let $\gamma_n = c_1 n^{-\kappa}/2$. Under Assumption 1, if Condition 1 is satisfied, then $\mathrm{P}(\mathcal{M} \subseteq \hat{\mathcal{M}}) \to 1$.

Theorem 2 Let $\gamma_n = c_1 n^{-\kappa}/2$. Under Assumptions 1–4, if Condition 1 is satisfied, then $\mathrm{P}(|\hat{\mathcal{M}}| \leq 16 c_3^2 \sigma_{\max}^{*2} / c_1^2 n^{-2\kappa}) \to 1$, where $\sigma_{\max}(\mathbf{A})$ is the largest singular value of the matrix $\mathbf{A}$ and $\sigma_{\max}^* = \sup_{0<t<1} \sigma_{\max}\{\mathbf{i}(t\boldsymbol{\beta}_0)\}$.

Theorem 1 shows that marginal score testing maintains the sure screening property, and is thus an attractive alternative to marginal Wald testing. Theorem 2 shows that the number of selected covariates is not too large, with high probability. For example, if $\sigma_{\max}^*$ increased only polynomially in n, $|\hat{\mathcal{M}}|$ would increase polynomially. At the same time, p_n can frequently be allowed to increase exponentially in n. Thus the false positive rate would decrease quickly to zero.

The presence of $\sigma_{\max}^*$ in Theorem 2 reflects the dependence of $|\hat{\mathcal{M}}|$ on the degree of collinearity of our data. The collinearity of estimating equations not only depends on the design matrix, but also varies across the parameter space. For example, Mackinnon and Puterman (1989) and Lesaffre and Marx (1993) showed that generalized linear models can be collinear even if their design matrices are not, and vice versa. In our situation, we are concerned with collinearity along the line segment between $\boldsymbol{\beta}_0$ and $\mathbf{0}$.

A.1 Proof of Theorem 1

The event $\{\mathcal{M} \subseteq \hat{\mathcal{M}}\}$ equals $\{\min_{j\in\mathcal{M}} |U_j^M(0)| \geq \gamma_n\}$, so it is easy to see that

$$\mathrm{P}(\mathcal{M} \subseteq \hat{\mathcal{M}}) \geq 1 - \sum_{j\in\mathcal{M}} \mathrm{P}(|U_j^M(0)| < \gamma_n).$$

By the triangle inequality, we know that for all j, $|u_j^M(0)| \leq |U_j^M(0) - u_j^M(0)| + |U_j^M(0)|$, and by Assumption 1 we see that $c_1 n^{-\kappa} - |U_j^M(0)| \leq |U_j^M(0) - u_j^M(0)|$ for all $j \in \mathcal{M}$. Therefore, $|U_j^M(0)| < \gamma_n$ for $j \in \mathcal{M}$ implies $|U_j^M(0) - u_j^M(0)| \geq c_1 n^{-\kappa}/2$, so that

$$\mathrm{P}(\mathcal{M} \subseteq \hat{\mathcal{M}}) \geq 1 - s_n \mathrm{P}(|U_j^M(0) - u_j^M(0)| \geq c_1 n^{-\kappa}/2).$$

The right-hand side goes to 1 if Condition 1 is satisfied. □

A.2 Proof of Theorem 2

If Condition 1 is satisfied, then

$$\mathrm{P}\{\max_j |U_j^M(0) - u_j^M(0)| \geq c_1 n^{-\kappa}/4\} \geq 1 - p_n \mathrm{P}\{|U_j^M(0) - u_j^M(0)| \geq c_1 n^{-\kappa}/4\} \to 1.$$

On the event $\max_j |U_j(0) - u_j(0)| \leq c_1[n/m]^{-\kappa}/4$, $|U_j(0)| \geq \gamma_n$ implies that $|u_j(0)| \geq c_1[n/m]^{-\kappa}/4$. This means that

$$\begin{aligned} |\hat{\mathcal{M}}| &= |\{j : |U_j(0)| \geq \gamma_n\}| \leq |\{j : |u_j(0)| \geq c_1[n/m]^{-\kappa}/4\}| \\ &\leq \sum_j u_j^M(0)^2/(c_1 n^{-\kappa}/4)^2 = \|\mathbf{u}(\mathbf{0})\|_2^2 16/c_1^2 n^{-2\kappa}, \end{aligned}$$

where the last equality follows from Assumption 2. Using the generalization of the mean value theorem to vector-valued functions (Hall and Newell, 1979) and Assumptions 3 and 4,

$$\|\mathbf{u}(\mathbf{0})\|_2 = \|\mathbf{u}(\boldsymbol{\beta}_0) - \mathbf{u}(\mathbf{0})\|_2 \leq \sup_{0<t<1} \|\mathbf{i}(t\boldsymbol{\beta}_0)\|_2 \|\boldsymbol{\beta}_0\|_2 \leq c_3 \sup_{0<t<1} \sigma_{\max}\{\mathbf{i}(t\boldsymbol{\beta}_0)\} = c_3\sigma_{\max}^*,$$

which implies that $|\hat{\mathcal{M}}| \leq 16c_3^2\sigma_{\max}^{*2}/c_1^2 n^{-2\kappa}$. □

A.3 Verifying Condition 1 for censored quantile regression

Define

$$Z_i^{(1)} = X_{ij}\left[\tau I\{h(Y_i) > \beta_{int}\} - (1-\tau)\frac{I\{h(Y_i) \leq \beta_{int}\}\delta_i}{S_{h(C)}\{h(Y_i)\}} S_{h(C)}(\beta_{int})\right] - u_j^M(0)$$

and

$$Z_i^{(2)} = X_{ij}(1-\tau) I\{h(Y_i) \leq \beta_{int}\}\delta_i \left[\frac{S_{h(C)}(\beta_{int})}{S_{h(C)}\{h(Y_i)\}} - \frac{\hat{S}_{h(C)}(\beta_{int})}{\hat{S}_{h(C)}\{h(Y_i)\}}\right].$$

We assume that β_{int} is either known or has been estimated from an independent dataset, so that in the remainder of the proof we can treat it as a constant. Then

$$\mathrm{P}(|U_j^M(0) - u_j^M(0)| \geq 2t) \leq \mathrm{P}(n^{-1}|\sum_i Z_i^{(1)}| \geq t) + \mathrm{P}(n^{-1}|\sum_i Z_i^{(2)}| \geq t).$$

To bound the term containing $Z_i^{(1)}$ we first note that $\mathrm{E}(Z_i^{(1)}) = 0$ by the definition of $u_j^M(0)$. Also, by assumption $S_{h(C)}(\beta_{int}) > 0$, so the term $I\{h(Y_i) \leq \beta_{int}\}\delta_i / S_{h(C)}\{h(Y_i)\}$, which is nonzero only when $h(Y_i) \leq \beta_{int}$, can be at most $S_{h(C)}(\beta_{int})^{-1}$. Therefore when $|X_{ij}| \leq M$ for all i, j, where $M > 0$, $|Z_i^{(1)}| \leq 2M$. Using Bernstein's inequality van der Vaart and Wellner (1996) and Assumption 5,

$$\mathrm{P}(n^{-1}|Z_1^{(1)} + \ldots + Z_n^{(1)}| \geq t) \leq 2\exp\left(-\frac{1}{2}\frac{t^2 n}{4M^2 + 2Mt/3}\right) + nl_0\exp(-l_1 M^\eta).$$

To bound the term containing $Z_i^{(2)}$, we first note that

$$\mathrm{P}(n^{-1}|Z_1^{(2)} + \ldots + Z_n^{(2)}| \geq t) \leq \mathrm{P}(\max_i |Z_i^{(2)}| \geq t) \leq n\mathrm{P}(|Z_i^{(2)}| \geq t).$$

Since $Z_i^{(2)} = 0$ when $h(Y_i) > \beta_{int}$, $\mathrm{P}(|Z_i^{(2)}| \geq t) = \mathrm{P}\{|Z_i^{(2)}| \geq t \cap h(Y_i) \leq \beta_{int}\}$. For notational convenience let $S_{int} = S_{h(C)}(\beta_{int})$, $\hat{S}_{int} = \hat{S}_{h(C)}(\beta_{int})$, $S_Y = S_{h(C)}\{h(Y_i)\}$, and $\hat{S}_Y = \hat{S}_{h(C)}\{h(Y_i)\}$. Then

$$\begin{aligned}
\mathrm{P}(|Z_i^{(2)}| \geq t) &\leq \mathrm{P}\left\{\left|\frac{S_{int}}{S_Y} - \frac{\hat{S}_{int}}{\hat{S}_Y}\right| \geq \frac{t}{M(1-\tau)} \cap h(Y_i) \leq \beta_{int}\right\} \\
&\leq \mathrm{P}\left\{|S_{int}\hat{S}_Y - \hat{S}_{int}S_Y| \geq \frac{tS_Y\hat{S}_Y}{M(1-\tau)} \cap h(Y_i) \leq \beta_{int}\right\}.
\end{aligned}$$

Now let $\varepsilon_{int} = |\hat{S}_{int} - S_{int}|$ and $\varepsilon_Y = |\hat{S}_Y - S_Y|$. Then

$$\begin{aligned}
\mathrm{P}(|Z_i^{(2)}| \geq t) \leq& \mathrm{P}\left\{S_{int}\varepsilon_Y + S_Y\varepsilon_{int} \geq \frac{tS_Y(S_Y - \varepsilon_Y)}{M(1-\tau)} \cap h(Y_i) \leq \beta_{int}\right\} \\
\leq& \mathrm{P}\left[\left\{S_{int} + \frac{tS_Y}{M(1-\tau)}\right\}\varepsilon_Y + S_Y\varepsilon_{int} \geq \frac{tS_Y^2}{M(1-\tau)} \cap h(Y_i) \leq \beta_{int}\right] \\
\leq& \mathrm{P}\left[\varepsilon_Y \geq \frac{tS_Y^2}{2M(1-\tau)}\left\{S_{int} + \frac{tS_Y}{M(1-\tau)}\right\}^{-1} \cap h(Y_i) \leq \beta_{int}\right] + \\
& \mathrm{P}\left[\varepsilon_{int} \geq \frac{tS_Y}{2M(1-\tau)} \cap h(Y_i) \leq \beta_{int}\right] \\
\leq& \mathrm{P}\left[\varepsilon_Y \geq \frac{tS_{int}^2}{2M(1-\tau)}\left\{S_{int} + \frac{t}{M(1-\tau)}\right\}^{-1}\right] + \\
& \mathrm{P}\left\{\varepsilon_{int} \geq \frac{tS_{int}}{2M(1-\tau)}\right\},
\end{aligned}$$

where the last inequality follows because $h(Y_i) \leq \beta_{int}$ implies that $S_Y \geq S_{int}$. Now by the theorem of Bitouzé et al. (1999), there exists some constant C such that

$$\mathrm{P}(n^{1/2}\|S_{h(T)}(\hat{S}_{h(C)} - S_{h(C)})\|_\infty \geq \lambda) \leq 2.5\exp(-2\lambda^2 + C\lambda),$$

where $S_{h(T)}$ is the survival function of the $h(T_i)$. When $h(Y_i) \leq \beta_{int}$, $S_{h(T)}\{h(Y_i)\} \geq S_{h(T)}(\beta_{int})$, so we can apply this theorem to

$$\mathrm{P}\left[n^{1/2}S_{h(T)}\{h(Y_i)\}\varepsilon_Y \geq \frac{tn^{1/2}S_{int}^2}{2M(1-\tau)}\left\{S_{int} + \frac{tS_{h(T)}\{\beta_{int}\}}{M(1-\tau)}\right\}^{-1}\right]$$

and

$$\mathrm{P}\left\{ n^{1/2} S_{h(T)}(\beta_{int})\varepsilon_{int} \geq \frac{t n^{1/2} S_{int} S_{h(T)}(\beta_{int})}{M(1-\tau)} \right\}$$

to bound $\mathrm{P}(n^{-1}|Z_1^{(2)} + \ldots + Z_n^{(2)}| \geq t)$.

By setting $t = c_2 n^{-\kappa}/2$ and $M = n^{(1-2\kappa)/(\eta+2)}$ and combining the previous tail bounds, we can conclude that $\mathrm{P}(|U_j^M(0) - u_j^M(0)| \geq 2t) \leq O\{\exp(-n^{(1-2\kappa)\eta/(\eta+2)})\}$. $\square$

B Algorithm for finding the reproducible screening threshold

Our reproducible screening procedure is as follows:

1. Screen observed data to obtain the sets $\hat{\mathcal{M}}^{(j)}$ for $j = 1, \ldots, p_n$.
2. Generate B bootstrap samples and screen the b^{th} sample to get $\mathcal{M}_b^{(j)}$.
3. Let $o_b^{(j)} = |\mathcal{M}_b^{(j)} \cap \hat{\mathcal{M}}^{(j)}|$ be the size of the b^{th} overlap so that $r^{(j)} = B^{-1} \sum_b I(j^{-1} o_b^{(j)} \geq p)$ is the reproducibility of keeping the top j covariates.
4. Choose $j^* = \min\{j = 1, \ldots, p_n : r^{(j)} \geq r\}$ and retain $\hat{\mathcal{M}}^{(j^*)}$.

It is important to note that $r^{(j)}$ is not monotonic in j. It's possible for $r^{(j^*)} \geq r$ but $r^{(j^*+1)} < r$.

This procedure can be implemented by calculating the $\mathcal{M}_b^{(j)}$ for each j, but when p is large this is time-consuming. Instead, we can use the following algorithm:

1. Given the $o_b^{(j)}$, derive an upper bound on $r^{(j+s)}$ for any step size s. Denote this upper bound by $U^{(j+s)}$.
2. Find the smallest s such that $U^{(j+s)} \geq r$.
3. Check whether indeed $r^{(j+s)} \geq r$. If so, $j^* = j+s$. If not, return to step 1, this time beginning with $o_b^{(j+s)}$.

This can save a great deal of time if the step sizes s are large.

To derive a $U^{(j+s)}$, note that $r^{(j+s)}$ will have its largest possible value if the next s variables retained in each bootstrap sample all overlap with previously retained covariates in the original data, and if the next s variables retained in the original data all overlap with previously retained covariates from the bootstrap sample. In other words, $r^{(j+s)}$ cannot be larger than its value if

$$\mathcal{M}_b^{(j+s)} \backslash \mathcal{M}_b^{(j)} \subseteq \hat{\mathcal{M}}^{(j)} \text{ and } \hat{\mathcal{M}}^{(j+s)} \backslash \hat{\mathcal{M}}^{(j)} \subseteq \mathcal{M}_b^{(j)} \text{ for all } b = 1, \ldots, B. \tag{2}$$

This increases the size of $o_b^{(j)}$ by $2s$. Under this condition, if $o_b^{(j)} \geq pj$, then $o_b^{(j+s)} = o_b^{(j)} + 2s \geq pj + s \geq p(j+s)$. Therefore

$$U^{(j+s)} = B^{-1} \sum_{\{b: o_b^{(j)} < pj\}} I\{o_b^{(j+s)} \geq p(j+s)\} + B^{-1} |\{j : o_b^{(j)} \geq pj\}|,$$

so $U^{(j+s)} \geq r$ if

$$\sum_{\{b: o_b^{(j)} < pj\}} I\{o_b^{(j+s)} \geq p(j+s)\} \geq Br - |\{j : o_b^{(j)} \geq pj\}|.$$

Next, for the b such that $o_b^{(j)} < pj$, under condition (2), $I\{o_b^{(j+s)} \geq p(j+s)\} \leq I\{o_b^{(j)}+2s \geq p(j+s)\}$, which equals 1 if $s \geq (jp - o_b^{(j)})/(2-p) = s_b$. Sort these s_b from smallest to largest so that if s is larger than or equal to the x^{th} smallest one, $\sum_{\{b:o_b^{(j)}<pj\}} I\{o_b^{(j)} + 2s \geq p(j+s)\} = x$. Therefore the smallest s such that $U^{(j+s)} \geq r$ occurs when s is at least the $[Br - |\{j : o_b^{(j)} \geq pj\}|]^{th}$ largest of the s_b.

References

Agarwal, A., Negahban, S., and Wainwright, M. J. (2012). Fast global convergence of gradient methods for high-dimensional statistical recovery. *The Annals of Statistics* **40,** 2452–2482.

Avet-Loiseau, H., Li, C., Magrangeas, F., Gouraud, W., Charbonnel, C., Harousseau, J.-L., Attal, M., Marit, G., Mathiot, C., Facon, T., et al. (2009). Prognostic significance of copy-number alterations in multiple myeloma. *Journal of Clinical Oncology* **27,** 4585–4590.

Bitouzé, D., Laurent, B., and Massart, P. (1999). A Dvoretzky–Kiefer–Wolfowitz type inequality for the Kaplan–Meier estimator. In *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, volume 35, pages 735–763. Elsevier.

Bühlmann, P. L., van de Geer, S. A., and Van de Geer, S. (2011). *Statistics for high-dimensional data*. Springer.

Cai, T., Huang, J., and Tian, L. (2009). Regularized estimation for the accelerated failure time model. *Biometrics* **65,** 394–404.

Cheng, S., Wei, L., and Ying, Z. (1995). Analysis of transformation models with censored data. *Biometrika* **82,** 835–845.

Daubechies, I., Fornasier, M., and Loris, I. (2008). Accelerated projected gradient method for linear inverse problems with sparsity constraints. *Journal of Fourier Analysis and Applications* **14,** 764–792.

Decaux, O., Lodé, L., Magrangeas, F., Charbonnel, C., Gouraud, W., Jézéquel, P., Attal, M., Harousseau, J. L., Moreau, P., Bataille, R., Campion, L., Avet-Loiseau, H., and Minvielle, S. (2008). Prediction of survival in multiple myeloma based on gene expression profiles reveals cell cycle and chromosome instability signatures in high-risk patients and hyperdiploid signatures in low-risk patients: a study of the Intergroupe Francophone du Myélome. *Journal of Clinical Oncology* **26,** 4798–4805.

Duchi, J., Shalev-Shwartz, S., Singer, Y., and Chandra, T. (2008). Efficient projections onto the ℓ_1-ball for learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning*, pages 272–279. ACM.

Fan, J., Feng, Y., and Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association* **106,** 544–557.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society, Ser. B* **70,** 849–911.

Fan, J., Samworth, R., and Wu, Y. (2009). Ultrahigh dimensional feature selection: beyond the linear model. *The Journal of Machine Learning Research* **10,** 2013–2038.

Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models and NP-dimensionality. *The Annals of Statistics* **38,** 3567–3604.

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

Hall, W. S. and Newell, M. L. (1979). The mean value theorem for vector valued functions: a simple proof. *Mathematics Magazine* **52,** 157–158.

He, Q. and Lin, D.-Y. (2011). A variable selection method for genome-wide association studies. *Bioinformatics* **27,** 1–8.

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

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association* **58,** 13–30.

Honore, B., Khan, S., and Powell, J. L. (2002). Quantile regression under random censoring. *Journal of Econometrics* **109,** 67–105.

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

Lesaffre, E. and Marx, B. D. (1993). Collinearity in generalized linear regression. *Communications in Statistics – Theory and Methods* **22,** 1933–1952.

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

Li, R., Zhong, W., and Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association* **107,** 1129–1139.

Mackinnon, M. J. and Puterman, M. L. (1989). Collinearity in generalized linear models. *Communications in Statistics – Theory and Methods* **18,** 3463–3472.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72,** 417–473.

Peng, L. and Huang, Y. (2008). Survival analysis with quantile regression models. *Journal of the American Statistical Association* **103,** 637–649.

Portnoy, S. (2003). Censored regression quantiles. *Journal of the American Statistical Association* **98,** 1001–1012.

Shaughnessy, J., Zhan, F., Burington, B. E., Huang, Y., Colla, S., Hanamura, I., Stewart, J. P., Kordsmeier, B., Randolph, C., Williams, D. R., 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.

Shi, L., Campbell, G., Jones, W. D., et al. (2010). The MAQC-II Project: A comprehensive study of common practices for the development and validation of microarray-based predictive models. *Nature Biotechnology* **28,** 827–838.

Shor, N. Z., Kiwiel, K. C., and Ruszcayski, A. (1985). *Minimization methods for non-differentiable functions*. Springer-Verlag New York, Inc.

Song, X. and Huang, Y. (2005). On corrected score approach for proportional hazards model with covariate measurement error. *Biometrics* pages 702–714.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Ser. B* **58,** 267–288.

Tsiatis, A. A. (1996). Estimating regression parameters using linear rank tests for censored data. *The Annals of Statistics* **18,** 354–372.

Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B., and Wei, L. J. (2011). On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine* **30,** 1105–1117.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York.

Wang, H. J. and Wang, L. (2009). Locally weighted censored quantile regression. *Journal of the American Statistical Association* **104,** 1117–1128.

Wei, L. J. (1992). The accelerated failure time model: a useful alternative to the cox regression model in survival analysis. *Statistics in Medicine* **11,** 1871–1879.

Zeger, S. L., Liang, K.-Y., and Albert, P. S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics* pages 1049–1060.

Zhan, F., Huang, Y., Colla, S., Stewart, J., Hanamura, I., Gupta, S., Epstein, J., Yaccoby, S., Sawyer, J., Burington, B., et al. (2006). The molecular classification of multiple myeloma. *Blood* **108,** 2020.

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