Statistica Sinica **27** (2017), 1967-1985
doi:https://doi.org/10.5705/ss.202016.0024

CONCORDANCE MEASURE-BASED FEATURE SCREENING AND VARIABLE SELECTION

Yunbei Ma[1], Yi Li[1], Huazhen Lin[1] and Yi Li[2]

[1]*Southwestern University of Finance and Economics*
and [2]*University of Michigan*

Abstract: The C-statistic, measuring the rank concordance between predictors and outcomes, has become a standard metric of predictive accuracy and is therefore a natural criterion for variable screening and selection. However, as the C-statistic is a step function, its optimization requires brute-force search, prohibiting its direct usage in the presence of high-dimensional predictors. We propose a smoothed C-statistic sure screening (C-SS) method for screening ultrahigh-dimensional data, and a penalized C-statistic (PSC) variable selection method for regularized modeling based on the screening results. We show that these procedures form an integrated framework for screening and variable selection: the C-SS possesses the sure screening property, and the PSC possesses the oracle property. Our simulations reveal that, compared to existing procedures, our proposal is more robust and efficient. Our procedure has been applied to analyze a multiple myeloma study, and has identified several novel genes that can predict patients response to treatment.

Key words and phrases: C-statistic, false positive rates, sparsity, ultra-high dimensional predictors, variable selection, variable screening.

1. Introduction

Modern technologies yield abundant data with ultrahigh-dimensional risk predictors from diverse scientific fields. Developing sound risk score systems that can function as accurate diagnostic tools in this environment has become a requirement. For example, in microarray-based risk prediction studies, arrays usually number in the tens, while the potential predictors can be tens of thousands of gene expressions.

Traditional variable selection methods include Bridge regression in Frank and Friedman (1993), Lasso in Tibshirani (1996), SCAD in Fan and Li (2001), the Elastic net in Zou and Hastie (2005), and the Dantzig selector in Candes and Tao (2007). When the number of covariates far exceeds the sample size, these traditional methods incur difficulties in speed, stability, and accuracy (Fan and Lv (2008)). Sure independence screening methods, e.g. those proposed by Fan and Lv (2008) and Fan, Samworth and Wu (2009), have emerged as a powerful

means to effectively eliminate unimportant covariates.

As successful as they are, the validity of these proposals hinges upon the proximity of the working models to the truth. To relax such restrictions, Hall and Miller (2009) extended the Pearson correlation learning by considering polynomial transformations of predictors. Fan, Feng and Song (2011) considered nonparametric independence screening in sparse ultrahigh-dimensional additive models, and Li et al. (2012) proposed a robust screening method by Kendall τ rank correlation (RRCS) and its iterative version (IRRCS) for transformation models. Within a fully nonparametric model framework, Li, Zhong and Zhu (2012) developed a sure independence screening procedure based on the distance correlation (DC-SIS). Other screening methods for ultrahigh dimensional discriminant analysis can be found in Mai and Zou (2013), among many others.

In summary, parametric methods are stable but rely heavily on assumptions, and have potentially high bias, while nonparametric methods can adapt to various situations, but estimators depend heavily on a handful of input observations, and are unstable. As a compromise between fully parametric and fully nonparametric modeling, we consider feature screening and variable selection in a semiparametric framework. A typical semiparametric model is an index model in which the response is associated with predictors through an unknown function of linear combinations. Zhu et al. (2011) and Zhong et al. (2012), among many others, have developed methods to simultaneously perform dimension reduction and variable selection for index models. These shrinkage-based variable-selection methods often perform poorly on the index model when p is large.

We propose to conduct variable screening and selection based on the C-statistic. The C-statistic (Harrell and Davis (1982)), measuring the rank concordance between predictors and outcomes, has become a standard measure of predictive accuracy. However, as a step function, its optimization requires a brute-force search. We employ a smoothed C-statistic screening (C-SS) for ultrahigh-dimensional data, followed by a penalized smoothing C-statistic (PSC) based on the screening results to further select and estimate the regression coefficients. We show that these procedures form an integrated framework for screening and variable selection: the C-SS possesses the sure screening property of Fan and Lv (2008), and the PSC possesses the oracle property of Fan and Li (2001) under a sparse assumption. We prove that the PSC achieves oracle properties if $m_n = o(n^{1/4})$, where m_n is the cardinality of the set of predictors captured by the C-SS. Compared with existing procedures, our procedure has practical appeal. Being semiparametic, while the link function relating the outcomes to the covari-

ates can be unspecified, while the existing SIS for the linear regression model (Fan and Lv (2008)) and the ISIS for the generalized linear model (Fan, Samworth and Wu (2009)) typically assume a linear link for continuous outcomes, and a logit or log link for ordinal outcomes. Such parametric assumptions can result in improper feature screening and estimation (Hettmansperger and McKean (2010)). Nonparametric screening methods, such as SIRS (Zhu et al. (2011)), RRCS (Li et al. (2012)) and DC-SIS (Li, Zhong and Zhu (2012)), can be unstable, especially with ultrahigh-dimensional data, while our C-SS method is stable and applicable to various types of data (continuous, count, ordinal and categorical). Further, our procedure leads to a selection of significant risk predictors without calling for additional modelling as required by nonparametric approaches. Finally, our method does not require a linearity condition on the predictors and does not require calculation of the $p \times p$ covariance matrix and its inverse.

This article is organized as follows. In Section 2, we develop the C-SS for feature screening by ranking a semi-robust measure of marginal utility. The sure screening property and model selection consistency under certain technical conditions are established. In Section 3, the PSC for selection and estimation of the regression coefficients is proposed; it allows the dimension of variables after screening to diverge to infinity. Development of iterative procedures, PC-SS and GC-SS, is discussed in Section 4. In Section 5, report on numerical studies conducted to evaluate the performance of our methods. We describs in Section 6 an analysis of a multiple myeloma study using the proposed methods. Concluding remarks are in Section 7. Proofs are in the online supplementary materials.

2. Screening Method Based on Smoothed C-Statistic

Consider a study with n independent subjects, where Y_i denotes the response variable (continuous, binary, ordinal or count) and $\mathbf{X}_i = (X_{i1}, \cdots, X_{ip})^T$ is a length p covariate vector containing, for example, all gene expressions for individual i. We assume that each component of $\mathbf{X}_i$ has been standardized such that $E(X_{ij}) = 0, Var(X_{ij}) = 1$ for $j = 1, \ldots, p$. We seek a feature $\mathbf{X}_i^T\boldsymbol{\beta}$ that predicts the response Y_i, as accurately as possible, through the use of the C-statistic, $C(\boldsymbol{\beta}) = \Pr(\mathbf{X}_i^T\boldsymbol{\beta} > \mathbf{X}_j^T\boldsymbol{\beta} | Y_i > Y_j)$, that can be estimated by $\widehat{\mathrm{C}}(\boldsymbol{\beta}) = (\sum_{i,j} I(Y_i > Y_j) I(\mathbf{X}_i^T\boldsymbol{\beta} > \mathbf{X}_j^T\boldsymbol{\beta}))/(\sum_{i,j} I(Y_i > Y_j))$, where $I(\cdot)$ is the indicator function. An estimator of $\boldsymbol{\beta}$ can be obtained by maximizing $\widehat{\mathrm{C}}(\boldsymbol{\beta})$ or

$$\mathrm{C}_n(\boldsymbol{\beta}) = \frac{1}{n(n-1)} \sum_{i \neq j} I(Y_i > Y_j) I(\mathbf{X}_i^T \boldsymbol{\beta} > \mathbf{X}_j^T \boldsymbol{\beta}). \tag{2.1}$$

Here $C_n(\boldsymbol{\beta})$ is the maximum rank correlation (MRC) defined in Han (1987). For a binary response, $C_n(\boldsymbol{\beta})$ is the Wilcoxon-Mann-Whitney statistic, and is identical to the area under a receiver operating characteristic curve for comparing predictions in the two groups. As $\boldsymbol{\beta}$ is only identifiable up to a constant multiplier, we take $\|\boldsymbol{\beta}\| = 1$.

2.1. Smoothed C-statistic

The indictor $I(\mathbf{X}_i^T \boldsymbol{\beta} > \mathbf{X}_j^T \boldsymbol{\beta})$ at (2.1) is discrete, presenting computational as well as theoretical challenges; see Han (1987) and Sherman (1993). The optimization requires a search that grows at the order of n^p and becomes impossible for ultra-high p. If $\Phi(\cdot)$ denotes the distribution function of the standard normal, we use $\Phi\left\{(\mathbf{X}_i^T \boldsymbol{\beta} - \mathbf{X}_j^T \boldsymbol{\beta})/h\right\}$ as a smooth approximation to the indicator function $I(\mathbf{X}_i^T \boldsymbol{\beta} > \mathbf{X}_j^T \boldsymbol{\beta})$, where the bandwidth h converges to zero as the sample size increases. A smoothed $C_n(\boldsymbol{\beta})$ is thus

$$C_s(\boldsymbol{\beta}) \;=\; \frac{1}{n(n-1)} \sum_{i \neq j} I(Y_i > Y_j) \Phi\left\{ \frac{(\mathbf{X}_i^T \boldsymbol{\beta} - \mathbf{X}_j^T \boldsymbol{\beta})}{h} \right\}. \tag{2.2}$$

When p is finite, it can be shown that when h is small enough the difference between $C_s(\boldsymbol{\beta})$ and $C_n(\boldsymbol{\beta})$ can be ignored. Hence the maximizer of $C_s(\boldsymbol{\beta})$ agrees well with those of $C_n(\boldsymbol{\beta})$. Because $C_s(\boldsymbol{\beta})$ is a smoothing function of $\boldsymbol{\beta}$, the computation of the maximizer of $C_s(\boldsymbol{\beta})$ is straightforward and can be accomplished through Newton-Raphson iteration. Other approximation methods, including the sigmoid approximation proposed by Ma and Huang (2005), can also be used to approximate the indicator function $I(\mathbf{X}_i^T \boldsymbol{\beta} > \mathbf{X}_j^T \boldsymbol{\beta})$.

Under some regular conditions for the binary response (Lin et al. (2011)), the estimator based on maximizing (2.2) is consistent when p is finite. Li et al. (2012) considered a penalized version of $C_s(\boldsymbol{\beta})$ when p goes to infinity, and proposed RRCS (Robust Rank Correlation Screening) to deal with ultra-high dimensional problems. The former lacks speed and stability while the later may not work well for discrete data. This has been confirmed by our simulation studies.

2.2. Screening method based on the smoothed C-statistic

Assume that the parameter $\boldsymbol{\beta}$ is sparse, and let $\mathcal{M}_0 = \{k : \beta_k \neq 0\}$ be the true sparse model with size $s_0 = |\mathcal{M}_0|$, where s_0 is small or grows slowly with n. We allow p to grow with n and denote it by p_n whenever necessary.

We estimate $\boldsymbol{\beta}$ by maximizing (2.2), or solving

$$\frac{\partial C_s(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{1}{n(n-1)} \sum_{i \neq j} I(Y_i > Y_j) \phi \left\{ \frac{(\mathbf{X}_i^T \boldsymbol{\beta} - \mathbf{X}_j^T \boldsymbol{\beta})}{h} \right\} \frac{(\mathbf{X}_i - \mathbf{X}_j)}{h} = 0,$$

where $\phi(\cdot)$ is the standard normal density function. If $\widehat{g}_k(\boldsymbol{\beta})$ is the kth component of $h\sqrt{2\pi}\partial C_s(\boldsymbol{\beta})/\partial \boldsymbol{\beta}$, $\widehat{g}_k(\boldsymbol{\beta}) = h\sqrt{2\pi}\partial C_s(\boldsymbol{\beta})/\partial \beta_k$ for $k = 1, \cdots, p_n$, then $(\widehat{g}_1(\boldsymbol{\beta}), \cdots, \widehat{g}_{p_n}(\boldsymbol{\beta})) = 0$ are estimating equations for $\boldsymbol{\beta}$.

For the k-th covariate, we construct an estimating equation for β_k, assuming a marginal model that has all other covariates unrelated to the outcome: $U_k(\beta_k) = \widehat{g}_k(0, \ldots, \beta_k, \ldots, 0) = 0$. Then each $|U_k(0)| \equiv |\widehat{g}_k(0_p)|$, where 0_p is a p-dimensional zero vector, is the numerator of the score statistic for a hypothesis: $\beta_k = 0$ under the k-th marginal model and can be a sensible screening statistic. The general theory for such score-test based screening statistics has been given by Zhao and Li (2014).

For a given thresholding value γ_n, we screen the covariates as

$$\widehat{\mathcal{M}}_{\gamma_n} = \{1 \leq k \leq p : |\widehat{g}_k(0_p)| \geq \gamma_n\}.$$

If $g_k(\boldsymbol{\beta}) = E[\widehat{g}_k(\boldsymbol{\beta})]$, then $\widehat{g}_k(0_p) = 1/n(n-1) \sum_{i \neq j}^n I(Y_i > Y_j)(X_{ik} - X_{jk})$ and $g_k(0_p) = E[I(Y_1 > Y_2)(X_{1k} - X_{2k})]$. Hence, $g_k(0_p)$, then $\widehat{g}_k(0_p)$, can be regarded as a surrogate measure of the nonparametric rank correlation between the response Y and the kth covariate X_k. For example, independence between Y and X_k implies $g_k(0_p) = 0$. Under some regularity conditions, the smoothed C-statistic sure screening (C-SS) procedure reduces the full model of size p to a submodel $\widehat{\mathcal{M}}_{\gamma_n}$ with size less than n. The procedure only requires a single evaluation of the smoothed C-statistic at $\boldsymbol{\beta} = 0$ instead of p separate models as is commonly used by the existing screening methods. Compared to the existing model-free sure screening methods such as SIRS (Zhu et al. (2011)), RRCS (Li et al. (2012)) and DC-SIS (Li, Zhong and Zhu (2012)), the method utilizes the linear structure of the predictors, $\mathbf{X}'\boldsymbol{\beta}$, and hence is more efficient.

2.3. Sure screening properties

For the sure screening properties, we need some conditions.

(C.1) For all $1 \leq k \leq p$, there exists a positive constant K_1 and $r_1 \geq 1$, such that

$$\Pr(|X_k| > t) \leq \exp(1 - (\frac{t}{K_1})^{r_1}), \text{ for any } t \geq 0. \tag{2.3}$$

(C.2) For all $k \in \mathcal{M}_0$, there exist positive constants δ and $\kappa < 1$, such that $|E[I(Y_1 > Y_2)(X_{1k} - X_{2k})]| > \delta n^{-\kappa}$.

(C.3) If $\boldsymbol{\beta}_0$ is the true value of $\boldsymbol{\beta}$, there exists a monotonic increasing function $m(\cdot)$ such that

$$E(Y|\mathbf{X}) = m(\mathbf{X}^T\boldsymbol{\beta}_0). \tag{2.4}$$

Condition (C.2) guarantees that the marginal signal of the active components $\{g_k(0_p)\}_{k\in\mathcal{M}_0}$ does not vanish as the sample size grows. Condition (C.3) supports the idea of using $\widehat{g}_k(0_p)$ to screen covariates based on the C-statistic $C(\boldsymbol{\beta}) = \Pr(\mathbf{X}_i^T\boldsymbol{\beta} > \mathbf{X}_j^T\boldsymbol{\beta}|Y_i > Y_j)$.

Theorem 1. *Suppose* (C.1) *and* (C.3) *hold.*

(1) *If* $0 < \kappa < 1/2$, *then for any* $c_1 > 0$, *there exist positive* c_2 *and* c_3 *such that*

$$P\left(\max_{1\le k\le p_n} |\widehat{g}_k(0) - E\widehat{g}_k(0)| > c_1 n^{-\kappa}\right) \le 4p_n \exp\left\{-\frac{c_1^2 n^{1-2\kappa}}{2(2c_2 + c_1c_3n^{-\kappa})}\right\}. \tag{2.5}$$

(2) *If* (C.2) *also holds, then with* $\gamma_n = \delta n^{-\kappa}/2$,

$$P\left(\mathcal{M}_0 \subset \widehat{\mathcal{M}}_{\gamma_n}\right) \ge 1 - 4s_0 \exp\left\{-\frac{\delta^2 n^{1-2\kappa}}{4(4c_2 + \delta c_3 n^{-\kappa})}\right\}.$$

The sure screening property then holds for the non-polynomial (NP) dimensionality of covariates with $\log p_n = o(n^{1-2\kappa})$; this is the rate for the linear regression model in Fan and Lv (2008).

Theorem 1(1) reveals that the signal level of the important effectors is of the same rate as that of their approximations, i.e., $O(n^{-\kappa})$. The ideal case for a vanishing false-positive rate is when $E[I(Y_1 > Y_2)(X_{1k} - X_{2k})] = o(n^{-\kappa})$ for $k \notin \mathcal{M}_0$, so that there is a natural separation between important and unimportant variables. When $p_n \exp\{-c_1^2 n^{1-2\kappa}/4(4c_2 + c_1c_3n^{-\kappa})\}$ tends to zero, we have, with probability going to 1, that $\max_{k\notin\mathcal{M}_0} |\widehat{g}_k(0_p)| \le cn^{-\kappa}$, for any $c > 0$. Thus, by choosing γ_n as in Theorem 1(2), the proposed screening method can achieve model selection consistency, $P(\mathcal{M}_0 = \widehat{\mathcal{M}}_{\gamma_n}) = 1 - o(1)$.

Theorem 2. *Under the conditions of Theorem* 1, *for* $\gamma_n = c_4 n^{-\kappa}$, *there exist positive constants* c_2 *and* c_3 *such that*

$$\Pr\left\{\|\widehat{\mathcal{M}}_{\gamma_n}\|_0 \le O(n^{\kappa}\sum_{k=1}^{p} |E[I(Y_1 > Y_2)(X_{1k} - X_{2k})]|)\right\}$$
$$\ge 1 - 4p\exp\left\{-\frac{c_4^2 n^{1-2\kappa}}{4(4c_2 + c_4c_3n^{-\kappa})}\right\}.$$

Where $\|\cdot\|_0$ *denotes the cardinality of a set.*

Thus, as long as $\sum_{k=1}^{p} |E[I(Y_1 > Y_2)(X_{1k} - X_{2k})|$ is of a polynomial order of sample size, the number of selected variables is also of polynomial order of sample size, and a variable selection procedure is conducted for parameters of a polynomial order.

3. Variable Selection and Parameter Estimation Based on the Penalized Smoothed C-Statistic

For variable selection with finite covariates, penalization methods such as LASSO, SCAD, and adaptive LASSO, among others, have routinely been used. Fan and Peng (2004) extended the SCAD penalized likelihood estimation to the situation where the number of parameters is of the order $o(n^{1/5})$.

Without loss of generality, we suppose that the first m_n variables are kept after screening: $\tilde{\mathbf{X}} = (X_1, \cdots, X_{m_n})^T$ with coefficients $\tilde{\boldsymbol{\beta}} = (\beta_1, \cdots, \beta_{m_n})^T$. We then further assume that the first s_0 variables of $\tilde{\mathbf{X}}$ are the important selectors: $\tilde{\mathbf{X}}^{(1)} = (X_1, \cdots, X_{s_0})^T$ with coefficients $\boldsymbol{\beta}^{(1)} = (\beta_1, \cdots, \beta_{s_0})^T$.

3.1. Penalized smoothed C-statistic

We rewrite the smoothed C-statistic after screening as

$$\widetilde{C}_s(\tilde{\boldsymbol{\beta}}) = \frac{1}{n(n-1)} \sum_{i \neq j} \left[I(Y_i > Y_j) \Phi \left\{ \frac{(\tilde{\mathbf{X}}_i^T \tilde{\boldsymbol{\beta}} - \tilde{\mathbf{X}}_j^T \tilde{\boldsymbol{\beta}})}{h} \right\} \right], \tag{3.1}$$

and estimate $\tilde{\boldsymbol{\beta}}$ by

$$\widehat{\boldsymbol{\beta}} = \arg \max_{\tilde{\boldsymbol{\beta}} \in \Omega, \|\tilde{\boldsymbol{\beta}}\| = 1} \{\widetilde{C}_s(\tilde{\boldsymbol{\beta}}) - \sum_{j=1}^{m_n} p_{\lambda_n}(|\beta_j|)\},$$

where $p_{\lambda_n}(\cdot)$ is a prespecified penalty function with a regularization parameter λ_n.

As the SCAD penalty satisfies all three properties of unbiasedness, sparsity and continuity (Fan and Li (2001)), we choose SCAD as the penalty function: for some $a > 0$ and $\beta > 0$, it satisfies $p'_{\lambda_n}(\beta) = \lambda_n \{I\{\beta \leq \lambda_n\} + (a\lambda_n - \beta)_+/(a-1)\lambda_n I\{\beta > \lambda_n\}\}$, with $p'_{\lambda_n}(0) = 0$.

3.2. Oracle property

We establish the asymptotic theory for the penalized smoothed estimation of $\tilde{\boldsymbol{\beta}}$ when m_n diverges. Let $\tilde{\boldsymbol{\beta}}_0 = (\boldsymbol{\beta}_0^{(1)T}, \boldsymbol{\beta}_0^{(2)T})^T$ be the true values of coefficients. Then $\boldsymbol{\beta}_0^{(2)} = 0_{m_n - s_0}$. We consider a generalized nonconcave penalty function, and let $a_n = \max\{p'_{\lambda_n}(|\beta_{j0}|) : \beta_{j0} \neq 0\}$ and $b_n = \max\{p''_{\lambda_n}(|\beta_{j0}|) : \beta_{j0} \neq 0\}$.

Theorem 3. *(Consistency) Under Conditions* (C.1*)-(C.4*) *in the supplementary materials, if the penalty function* $p_{\lambda_n}(\cdot)$ *satisfies conditions* (P.1) *and* (P.2) *there, and if* $nh \to \infty$, $nh^4 \to 0$, $m_n^4/n \to 0$ *as* $n \to \infty$, *then there exists a maximizer* $\widehat{\boldsymbol{\beta}}$ *of* $PC_s(\tilde{\boldsymbol{\beta}})$ *satisfying* $\|\widehat{\boldsymbol{\beta}}\| = 1$ *and*

$$\|\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}_0\| = O_p\{\sqrt{m_n}(n^{-1/2} + a_n)\}.$$

Thus if $a_n = O(n^{-1/2})$, the penalized smoothed estimator is root-(n/m_n) consistent. This rate is the same as for the M-estimator with diverging parameters in Huber (1973). For the SCAD penalty, by $(C.5^*)$ in the supplementary materials, $a_n = 0$ when n is large enough. Hence, there exists a root-(n/m_n)-consistent penalized smoothed estimator with probability tending to 1, and no requirements for the convergence rate of λ_n.

Let $G(Z_1, Z_2) = \int_{-\infty}^{\infty}\int_{y^*}^{\infty} dF(y|Z_1)dF(y^*|Z_2)$, $I^*(\boldsymbol{\beta}_0^{(1)}) = E[G^2(Z,Z)\, Cov(\tilde{\mathbf{X}}^{(1)}|Z)]$, $\Sigma_{\lambda_n}(\boldsymbol{\beta}_0^{(1)}) = \mathrm{diag}\{p''_{\lambda_n}(|\beta_{10}|), \cdots, p''_{\lambda_n}(|\beta_{s_0 0}|)\}$, and $\mathbf{b} = (p'_{\lambda_n}(|\beta_{j0}|)\, sgn(\beta_{j0}), j = 1, \cdots, s_0)^T$.

Theorem 4. *(Oracle property). If* (C.1*)-(C.5*) *and* (P.1)-(P.4) *in the supplementary materials hold, and if* $\lambda_n \to 0$, $\sqrt{n/m_n}\lambda_n \to \infty$, $nh \to \infty$, $nh^4 \to 0$ *and* $m_n^4/n \to 0$ *as* $n \to \infty$, *then with probability tending to 1, the* $\sqrt{n/m_n}$*-consistent local maximizer* $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}^{(1)T}, \widehat{\boldsymbol{\beta}}^{(2)T})^T$ *in Theorem* 3 *satisfies:*

(i) *(Sparsity)* $\widehat{\boldsymbol{\beta}}^{(2)} = 0$, *and*

(ii) *(Asymptotic normality)*

$$\sqrt{n}[I(\boldsymbol{\beta}_0^{(1)}) + \Sigma_{\lambda_n}(\boldsymbol{\beta}_0^{(1)})]\left(\widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}_0^{(1)} + [I(\boldsymbol{\beta}_0^{(1)}) + \Sigma_{\lambda_n}(\boldsymbol{\beta}_0^{(1)})]^{-1}\mathbf{b}\right)$$
$$\xrightarrow{\mathcal{L}} N(0_{s_0}, I^*(\boldsymbol{\beta}_0^{(1)})).$$

Sparsity and asymptotic normality then are still valid when the number of parameters after screening diverges. For the SCAD penalty, Condition (C.5*) implies that $\Sigma_{\lambda_n} = 0$ and $\mathbf{b} = 0$ for large enough n. Then Theorem 4(ii) becomes $\sqrt{n}I(\boldsymbol{\beta}_0^{(1)})(\widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}_0^{(1)}) \xrightarrow{\mathcal{L}} N(0_{s_0}, I^*(\boldsymbol{\beta}_0^{(1)}))$, which implies that the penalized smoothed estimator of $\boldsymbol{\beta}^{(1)}$ performs as well as a maximized rank correlation estimator when $\boldsymbol{\beta}_0^{(2)} = 0$ is known. This demonstrates that the penalized smoothed estimator with diverging m_n parameters possesses the oracle property.

4. Iterative Algorithm and Relative Issues

To reduce false negatives and false positives, we adopt an iterative framework to enhance model performance by repeatedly applying our variable screening

and variable selection. These result in a conditional random permutation C-SS (PC-SS) method that performs conditional random permutation in the screening step to determine the threshold, and a Greedy C-SS (GC-SS) method that is a greedy version of the iterative screening-SCAD procedure. These algorithms are similar to the INIS procedure of Fan, Ma and Dai (2014), the details are in the supplementary materials.

We need to select the tuning parameters, (λ_n, a) for the SCAD penalty function, and h for smoothing function $\widetilde{C}_s(\tilde{\boldsymbol{\beta}})$. To reduce the computational burden in our simulation studies and examples, we took $a = 2\sqrt{3}$, as recommend by Fan and Li (2001). The selection of λ_n is governed by the BIC-criterion - we chose λ_n as the maximizer of $\mathrm{BIC}_{\lambda_n} = \log\{\widetilde{C}_s(\widehat{\boldsymbol{\beta}})\} - 1/2\mathrm{df}_{\lambda_n} \log n/n$, where df_{λ_n} is the number of nonzero coefficient estimates. We chose $h = n^{-1/3}$ in the same way as in Lin et al. (2011).

5. Simulation Studies

We examined the finite sample performance of the proposed method. We investigated the screening capacity by comparing it with the parametric methods of SIS (Fan and Lv (2008)) for linear regression models and GLM-SIS (Fan, Samworth and Wu (2009)) for generalized linear models, nonparametric methods including RRCS (Li et al. (2012)), SIRS (Zhu et al. (2011)) and DC-SIS (Li, Zhong and Zhu (2012)), and the robust screening methods QaSIS (He, Wang and Hong (2013))and DC-RoSIS (Zhong et al. (2016)). We compared the estimation accuracy of the proposed selection method with that of SIS-SCAD (Fan and Lv (2008)) for linear models, and that of vanilla-SIS-SCAD and permutation-SIS-SCAD (Fan, Samworth and Wu (2009)) for generalized linear models.

5.1. Comparison of screening methods

Screening was assessed according to the minimum model size (MMS) needed to include all the true variables. The response variable Y was generated from either a linear regression model $Y = X'\beta + \varepsilon$ with a normal error (Model 1), or a nonlinear regression model $Y = \exp(X'\beta) + \varepsilon$ with a normal error (Model 2). The variance of the error in Models 1 and 2 was taken to make the SNR between 8 and 10. To check the effect of misspecification of link functions in generalized linear models, we generated ordinal responses via a 3-class ordinal model $P(Y < j) = g^{-1}(c_j + X'\beta)$ $(j = 1, 2)$ with $g(x) = -\log(-\log(x))$ and cut-off points $c_1 = -3$ and $c_2 = 2$ (Model 3). In addition, we considered the case with discrete covariates; configuration was based on Model 1 but with binary covariates $I(X >$

0), where each component of X was an original continuous covariate (Model 4). Finally, we conducted a simulation study with a weak variable or signal, similar to the setting of Example III of Fan and Lv (2008). This simulation setup was based on Model 1 except that the last two nonzero coefficients were the same value as the standard deviation of the error to investigate the performance of the proposed method in the case of a weak signal (Model 5). The predictors were set as follows for all of the models: $X_{ij} = (tU_i + \epsilon_{ij})/\sqrt{1+t^2}, i = 1, \ldots, n, j = 1, \ldots, p$, where the U_i and ϵ_{ij} were independent standard normal variables, and t was chosen to control the correlation among predictors with 0 as the independent case. We chose $n = 100, 200$, $p = 1,000, 4,000$, and the size s of the true models to be 4 and 8. The non-zero components of the p-vectors $\boldsymbol{\beta}$ were randomly chosen as follows. We set $a = 4\log(n)/n^{1/2}$ and picked non-zero coefficients of the form $(-1)^u(a + |z|)$ for each model, where u was a Bernoulli with parameter 0.4 and z was standard normal. For each model, we simulated 200 data sets. The boxplots of the minimum number of selected variables that required to include the true model are reported in Fig. 1-Fig. 5, with Fig. 4 and Fig. 5 in the supplementary materials. We estimated the generalized linear model with the correct link as well as the mis-specified probit link.

Our findings are as follows.

(1) The proposed C-SS performed slightly worse than the SIS when the estimators were implemented under the linear regression model; see Fig. 1, Fig. 4 and Fig. 5. This is not surprising as the SIS was carried out under the true model. However, if the true model was not a linear regression model, the SIS performed the worst among all the competing methods as shown by Fig. 2. The comparison of Fig. 1 and Fig. 5 suggests that the relative performance of various methods is similar in the cases of weak and strong signal.

(2) The RRCS failed for Models 3 and 4 (see Fig. 3 and Fig. 4) because the response or the covariates were discrete. Our method performed better than the QaSIS for all the simulations.

(3) The DC-SS and DC-RoSIS performed worse than the C-SS for all the simulations, worse than the SIS for the linear regression model and the GLM-SIS for the generalized linear regression model (see Fig. 1, Fig. 3 and Fig. 5). This is not surprising as the DC-SS and DC-RoSIS were designed to accommodate fully nonparametric settings, while the other methods were designed under the semiparametric or parametric settings. Fig. 1-Fig. 5 also illustrate that our method was superior to SIRS under the linear regression, and was slightly better than SIRS under the generalized linear model and the nonlinear regression model.

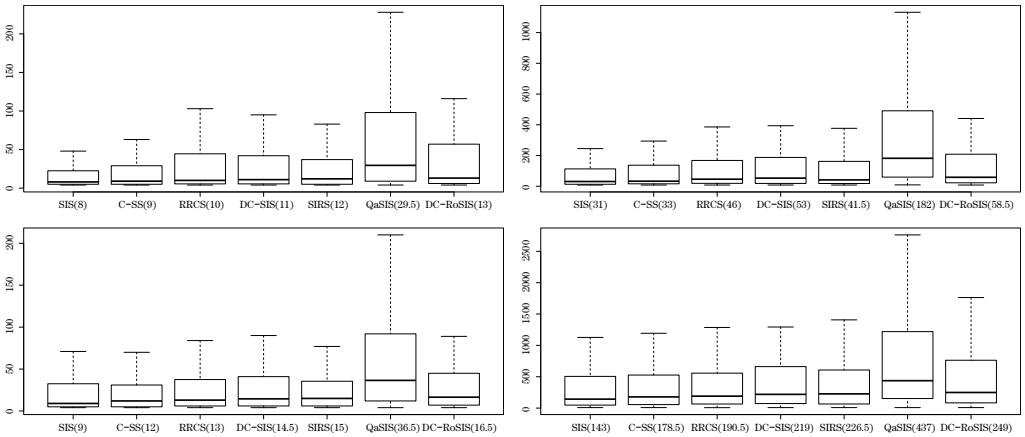

Figure 1. Boxplot of the minimum number of selected variables required to include the true linear regression model (Model 1) by using SIS, proposed C-SS method, RRCS, DC-SIS, SIRS, QaSIS and DC-RoSIS when (a) $n = 100$, $p = 1,000$, $s = 4$, $Cor(X_j, X_k) = 0$, $(j \neq k)$; (b) $n = 200$, $p = 4,000$, $s = 8$, $Cor(X_j, X_k) = 0$, $(j \neq k)$; (c) $n = 100$, $p = 1,000$, $s = 4$ $Cor(X_j, X_k) = 0.2$, $(j \neq k)$ and (d) $n = 200$, $p = 4,000$, $s = 8$, $Cor(X_j, X_k) = 0.2$, $(j \neq k)$. The number in brackets is the median of distribution for the minimum number of selected variables.

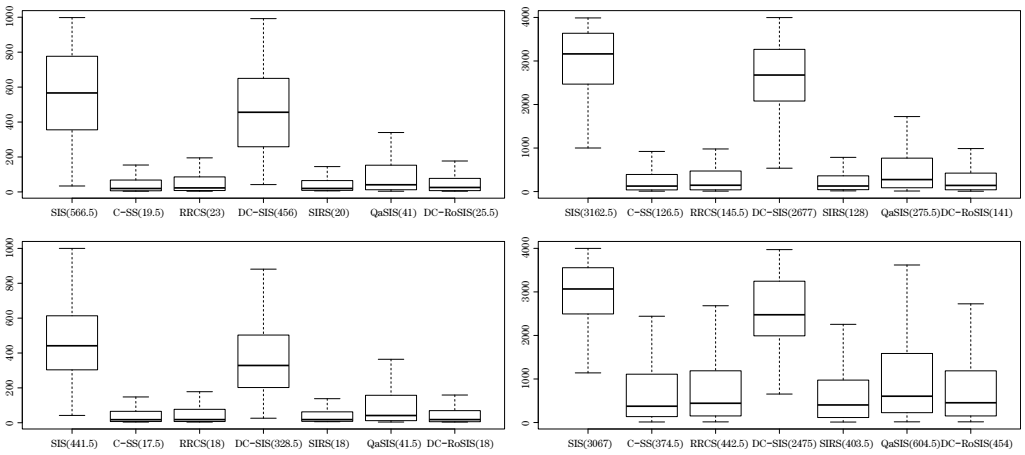

Figure 2. Boxplot of the minimum number of selected variables required to include the true nonlinear regression model (Model 2) by using SIS, proposed C-SS method, RRCS, DC-SIS, SIRS, QaSIS and DC-RoSIS with outliers excluded when (a) $n = 100$, $p = 1,000$, $s = 4$, $Cor(X_j, X_k) = 0$, $(j \neq k)$; (b) $n = 200$, $p = 4,000$, $s = 8$, $Cor(X_j, X_k) = 0$, $(j \neq k)$; (c) $n = 100$, $p = 1,000$, $s = 4$ $Cor(X_j, X_k) = 0.15$, $(j \neq k)$ and (d) $n = 200$, $p = 4,000$, $s = 8$, $Cor(X_j, X_k) = 0.5$, $(j \neq k)$. The number in brackets is the median of distribution for the minimum number of selected variables.

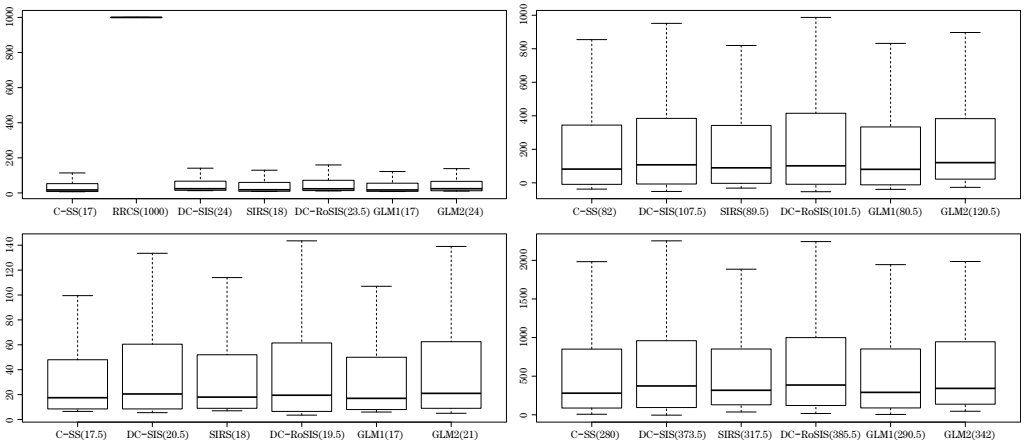

Figure 3. Boxplot of the minimum number of selected variables required to include the true ordinal model (Model 3) by using proposed C-SS method, RRCS, DC-SIS, SIRS, DC-RoSIS, GLM-SIS with correct link function (named GLM1 in the graph) and GLM-SIS with link function misspecified to probit link (named GLM2 in the graph) when (a) $n = 100$, $p = 1,000$, $s = 4$, $Cor(X_j, X_k) = 0$, $(j \neq k)$, (b) $n = 200$, $p = 4,000$, $s = 8$, $Cor(X_j, X_k) = 0$, $(j \neq k)$, (c) $n = 100$, $p = 1,000$, $s = 4$, $Cor(X_j, X_k) = 0.2$, $(j \neq k)$ and (d) $n = 200$, $p = 4,000$, $s = 8$, $Cor(X_j, X_k) = 0.2$, $(j \neq k)$. The number in brackets is the median of distribution for the minimum number of selected variables.

(4) Fig. 3 shows that the C-SS performed similarly to the GLM-SIS when the link function was correctly specified, and outperformed the GLM-SIS when the link function was misspecified.

We compared the computing time of the various methods. For example, for Model 1 with $n = 100$, $p = 1,000$ and $s = 4$, the average computing time of SIS, C-SS, SIRS, DC-SIS, RRCS, QaSIS, and DC-RoSIS per simulation is 1.38s, 1.95s, 1.75s, 2.75s, 1.73s, 2.22s and 2.78s, respectively. It appears that our method is on a par with these methods in terms of computing time.

5.2. Comparison of estimation accuracy for the variable selection

We used $s = 5$ as the size of the true models that, without loss of generality, are $\beta_1, \cdots, \beta_5$. The non-zero components of the p-vectors $\boldsymbol{\beta}$ were randomly chosen as in Section 5.1. To let $\boldsymbol{\beta}$ have a unit norm, we took the final non-zero parameters as $\boldsymbol{\beta}/\|\boldsymbol{\beta}\|$. To generate covariates, we randomly generated an $s \times s$ symmetric positive definite matrix A with a condition number $n^{1/2}/\log(n)$, and took s predictors $X_1, \cdots, X_s \sim N(0, A)$. Then, by letting $r = 1 - 4\log(n)p$, we generated $Z_{s+1}, \cdots, Z_p$ from $N(0, I_{p-s})$ and defined the predictors $X_{s+1}, \cdots, X_p$ as $X_i = Z_i + rtX_{i-s}, i = s+1, \cdots, 2s$, and $X_i = Z_i + (1-r)tX_1, i = 2s+1, \cdots, p$,

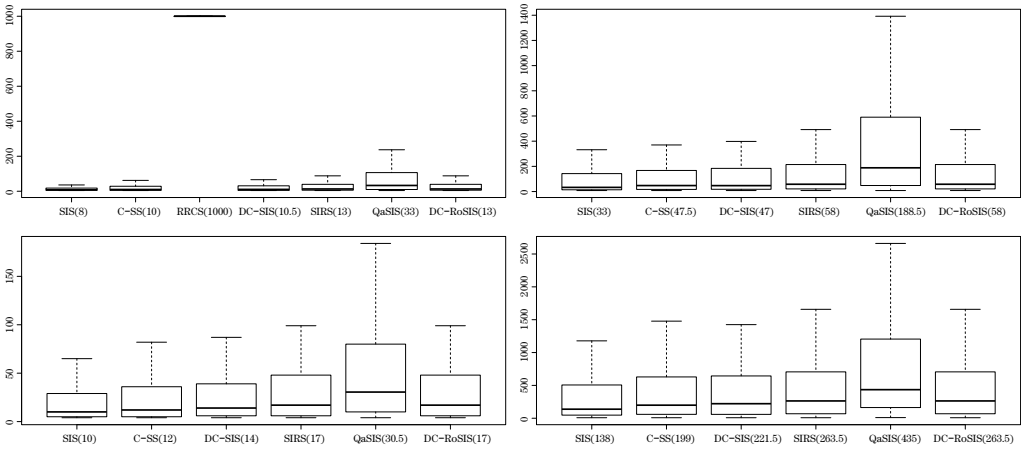

Figure 4. Boxplot of the minimum number of selected variables required to include the true model with discrete covariates (Model 4) by using SIS, the proposed C-SS method, RRCS, DC-SIS, SIRS, QaSIS and DC-RoSIS when (a) $n = 100$, $p = 1,000$, $s = 4$, $Cor(X_j, X_k) = 0$, $(j \neq k)$, (b) $n = 200$, $p = 4,000$, $s = 8$, $Cor(X_j, X_k) = 0$, $(j \neq k)$, (c) $n = 100$, $p = 1,000$, $s = 4$ $Cor(X_j, X_k) = 0.2$, $(j \neq k)$ and (d) $n = 200$, $p = 4,000$, $s = 8$, $Cor(X_j, X_k) = 0.2$, $(j \neq k)$. The number in brackets is the median of distribution for the minimum number of selected variables.

with $t = 0$ for independent predictors, and $t = 1$ for correlated predictors. If not otherwise stated, the common parameters for the following simulations were sample size $n = 200$, number of covariates $p = 1,000$, and Monte Carlo repetitions $N = 100$.

We considered the regression model, $Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + e$, where the noise e was generated as $N(0, \sigma^2)$, $X_1^2 \cdot N(0, \sigma^2)$ with $\sigma = 0.5$, or $0.1 \cdot t(1)$. Here $t(1)$ is the t distribution with degree of freedom 1. Four methods were compared, including conditional permutation screening-SCAD methods based on a smoothed C-statistic (PC-SS) as in Section 1.1 of the supplementary materials with $K = 0$, $\mathcal{M}^0 = \emptyset$, Greedy screening-SCAD methods based on a smoothed C-statistic (GC-SS) as in Section 1.2 of the supplementary materials with $p_0 = 1$, Permutation-SIS-SCAD (PSIS) as in Fan and Lv (2008), and Vanilla-SIS-SCAD (VSIS) as in Fan, Samworth and Wu (2009). Several performance measures are reported in Table 1. The med.$\|\widehat{\boldsymbol{\beta}}_{oracle} - \boldsymbol{\beta}\|$ is also presented. In Table 1, $\boldsymbol{\beta}$ and $\widehat{\boldsymbol{\beta}}$ had been normalized to have unit 1 norm.

Table 1 reveals that the proposed PC-SS and GC-SS methods yielded results similar to PSIS and VSIS under the normal noise of the same distribution. However, when the noise was heteroskedastic, PSIS had a low true-positive rate and missed important predictors, while VSIS had a high false-positive rate and

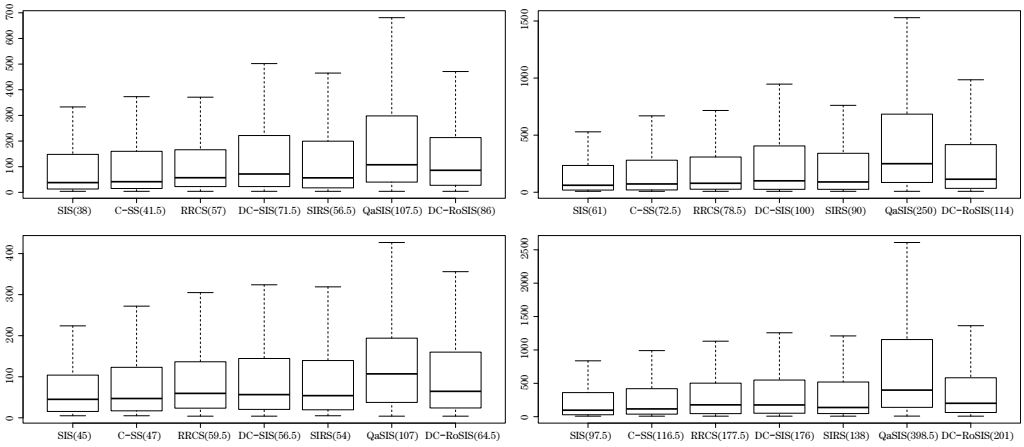

Figure 5. Boxplot of the minimum number of selected variables required to include the true model with weak signal (Model 5) by using SIS, the proposed C-SS method, RRCS, DC-SIS, SIRS, QaSIS and DC-RoSIS when (a) $n = 100$, $p = 1,000$, $s = 4$, $Cor(X_j, X_k) = 0$, $(j \neq k)$, (b) $n = 200$, $p = 4,000$, $s = 8$, $Cor(X_j, X_k) = 0$, $(j \neq k)$, (c) $n = 100$, $p = 1,000$, $s = 4$ $Cor(X_j, X_k) = 0.2$, $(j \neq k)$ and (d) $n = 200$, $p = 4,000$, $s = 8$, $Cor(X_j, X_k) = 0.2$, $(j \neq k)$. The number in brackets is the median of distribution for the minimum number of selected variables.

identified a large number of unimportant predictors, and failed when the noise was fat-tailed. Our method had a high true-positive rate, a low false-positive rate, and a small prediction error for either heteroskedasticity or fat-tailed noise. Note the closeness of med.$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|$ and med.$\|\widehat{\boldsymbol{\beta}}_{oracle} - \boldsymbol{\beta}\|$.

We have conducted more simulation studies on several generalized linear models, including a non-linear regression model, a poisson regression model and an ordinal regression model, with comparable results. For more details, see the supplementary materials.

6. A Study of the Intergroupe Francophone du Myelome

Multiple myeloma is a progressive blood cancer often diagnosed through the presence of an excessive numbers of abnormal plasma cells in the bone marrow, and overproduction of intact monoclonal immunoglobulin. Myeloma patients are typically characterized with wide clinical and pathophysiologic heterogeneities, and exhibit various levels of response to the same treatments. Extensive studies have revealed that the achievement of complete or partial response to treatment will substantially prolong progression-free and overall survival. Gene expressions of patients have been offered as effective prognostic tool for treatment response, and have informed the design of appropriate gene therapies. Further identifying

Table 1. Simulation results for linear regression model.

	$e \sim N(0,\sigma^2)$, med.$\|\widehat{\boldsymbol{\beta}}_{oracle} - \boldsymbol{\beta}\| = 0.074$							
	$t=0$				$t=1$			
	PC-SS	GC-SS	PSIS	VSIS	PC-SS	GC-SS	PSIS	VSIS
perc.incl.true	0.96	0.93	1.00	1.00	0.96	0.90	1.00	1.00
med. model size	5	5	5	5	5	5	5	5
aver. model size	5.02	4.81	5.30	5.17	4.99	4.99	5.18	5.14
med.$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|$	0.082	0.078	0.087	0.145	0.083	0.084	0.081	0.135
	$e \sim 0.1 \cdot t(1)$, med.$\|\widehat{\boldsymbol{\beta}}_{oracle} - \boldsymbol{\beta}\| = 0.031$							
	$t=0$				$t=1$			
	PC-SS	GC-SS	PSIS	VSIS	PC-SS	GC-SS	PSIS	VSIS
perc.incl.true	1.00	0.95	0.23	0.48	0.99	0.97	0.27	0.40
med. model size	5	5	3	37	5	5	4	37
aver. model size	5.01	4.81	3.69	29.23	4.99	4.90	4.20	28.92
med.$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|$	0.031	0.033	0.556	2.905	0.032	0.030	0.666	2.840
	$e \sim X_1^2 N(0,\sigma^2)$, med.$\|\widehat{\boldsymbol{\beta}}_{oracle} - \boldsymbol{\beta}\| = 0.060$							
	$t=0$				$t=1$			
	PC-SS	GC-SS	PSIS	VSIS	PC-SS	GC-SS	PSIS	VSIS
perc.incl.true	0.97	0.93	0.84	0.94	0.96	0.93	0.76	0.87
med. model size	5	5	5	7	5	5	5	7
aver. model size	4.95	4.80	5.44	10.76	4.89	4.80	5.43	11.77
med.$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|$	0.060	0.058	0.289	0.788	0.056	0.063	0.280	0.788

genes that predict treatment efficacy can boost our capabilities for personalized medicine.

For previously untreated multiple myeloma patients, high-dose therapy with autologous stem cell transplantation (HDT-ASCT) is the standard of care. Bortezomib-based therapy has recently emerged as a useful induction treatment prior to HD-ASCT. A recent trial by the Intergroupe Francophone du Myelome investigated the efficacy of receiving bortezomib therapy before HDT-ASCT. A total of 136 newly-diagnosed patients were enrolled and, for each patient, gene expression files with 44,280 probes were obtained. The goal of the study was to identify genes that were predictive of the response to treatment (coded values of 0 = no response, 1 = partial response, 2 = complete response were assigned). We applied the proposed PC-SS and GC-SS methods to analyze the data and obtained similar results; we only report the results for GC-SS. For comparison, we also applied the Vanilla-SIS-SCAD (VSIS-G) method proposed by Fan, Samworth and Wu (2009) for generalized linear models. The selected genes and their descriptions are presented in Table 2. It appears that GC-SS selected several novel genes that were predictive of response to treatment, such as CD74, ma-

Table 2. Gene selection for the Intergroupe Francophone du Myelome Study.

GC-SS	
Probeset	Gene name
228093_at	Zinc finger protein 599
243695_at	Transcribed locus
1567628_at	CD74 molecule, major histocompatibility complex, class II invariant chain
206094_x_at	UDP glucuronosyltransferase 1 family, polypeptide A1 A3-10
208306_x_at	Major histocompatibility complex, class II, DR beta 1, 3, 4
205004_at	NFKB repressing factor
217389_s_at	Activating transcription factor 5
1554161_at	Solute carrier family 25, member 27
230499_at	Baculoviral IAP repeat containing 3
206408_at	Leucine rich repeat transmembrane neuronal 2

VSIS-G	
Probeset	Gene name
205549_at	Purkinje cell protein 4
222285_at	Immunoglobulin heavy constant delta
241226_at	Transcribed locus
229941_at	Family with sequence similarity 166, member B
206094_x_at	UDP glucuronosyltransferase 1 family, polypeptide A1 A3-10
220622_at	Leucine rich repeat containing 31
206679_at	Amyloid beta (A4) precursor protein-binding, family A, member 1
228093_at	Zinc finger protein 599
217389_s_at	Activating transcription factor 5
214608_s_at	Eyes absent homolog 1 (Drosophila)

jor histocompatibility complex/class II, and NFkB, which have all been known to regulate the proliferation of multiple myeloma cells; see Burton et al. (2004) and Demchenko and Kuehl (2010). These important genes were missed by the VSIS-G method.

To study the predictive performance of selected genes, we applied a K-fold cross-validation method to compare the estimated predictive accuracy, the estimated C-statistic. Our approach was similar to that of Tian et al. (2007) that assessed model performance based on absolute prediction error. We randomly split the data into K disjoint subsets of equal sizes and labeled them $I_k, k = 1, \ldots, K$. For each k, we used all the observations, excluding I_k, to obtain an estimate $\widehat{\boldsymbol{\beta}}_{(-k)}$ for the final set of genes shown in Table 2, by maximizing the smoothed C-statistic (3.1). We then computed the estimated C-statistic $\widehat{C}_{(k)}(\widehat{\boldsymbol{\beta}}_{(-k)})$ via (2.1) based on observations in I_k. An average C-statistic could be computed as $\widehat{C} = K^{-1}\sum_{k=1}^{K}\widehat{C}_{(k)}(\widehat{\boldsymbol{\beta}}_{(-k)})$. Taking $K = 32$, we obtained the

averaged C-statistics $\widehat{C}_{GC-SS} = 0.84$ and $\widehat{C}_{VSIS-G} = 0.81$, based on the GC-SS and VSIS-G methods, respectively. Both sets of genes gave very high predictive power, though our proposed method showed even higher predictive accuracy.

7. Discussion

We have proposed an integrated framework that combines screening and variable selection based on the smoothed C-statistic, a rank concordance measure between predictors and outcomes, and have established the sure screening properties and model consistency property of the proposed method. Future research lies in extending the results to encompass censored outcome data, with applications in identifying novel biomarkers that can predict disease progression or risk of death. We will report the results elsewhere.

Supplementary Materials

The supplementary materials consist of: (i) some details of the iterative screening-SCAD procedure; (ii) further simulation studies; (iii) some technical lemmas used in the proofs of Theorem 1 and 2; (iv) the proofs of Theorems 1 and 2; (v) the conditions and the proof for the oracle property.

Acknowledgment

Lin, Ma and Li's research were partially supported by National Natural Science Foundation of China (No. 11528102, 11571282 and 11301424) and Fundamental Research Funds for the Central Universities (No. JBK141111, JBK141121, JBK120509 and 14TD0046) of China. Li was also partially supported by the U.S. National Institutes of Health (No. RO1CA050597). We thank our editorial assistant, Ms. Martina Fu from Stanford University, for proofreading the manuscript and for many useful suggestions. We are grateful for the helpful comments of the Co-Editor, an associate editor, and referees that substantially improved the presentation of the paper. Corresponding to: linhz@swufe.edu.cn.

References

Burton, J. D., Ely, S., Reddy, P. K., Stein, R., Gold, D. V., Cardillo, T. M. and Goldenberg, D. M. (2004). CD74 is expressed by multiple myeloma and is a promising target for therapy. *Clinical Cancer Research* **10**, 6,606-6,611.

Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n (with discussion). *Ann. Statist.* **35**, 2,313-2,404.

Demchenko, Y. N., Kuehl, W. M. (2010). A critical role for the NFkB pathway in multiple myeloma. *OncoTarget* **5**, 59-68.

Fan, J., Feng, Y. and Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *J. Am. Statist. Assoc.* **106**, 544-557.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Assoc.* **96**, 1,348-1,360.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). *J. Roy. Statisti. Soc. B.* **70**, 849-911.

Fan, J., Ma, Y. and Dai, W. (2014). Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models. *J. Am. Statist. Assoc.* **109**, 1,270-1,284.

Fan J. and Peng H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32**, 928-961.

Fan, J., Samworth, R. and Wu, Y. (2009). Ultrahigh dimensional feature selection: beyond the linear model. *Journal of Machine Learning Research* **10**, 2,013-2,038.

Frank, I.E. and Friedman, J.H. (1993). A statistical view of some chemometrics regression tools (with discussion). *Technometrics* **35**, 109-148.

Hall, P. and Miller, H. (2009). Using generalised correlation to effect variable selection in very high dimensional problems. *Journal of Computational and Graphical Statistics* **18**, 533-550.

Han, A. (1987). Non-parametric analysis of a generalized regression model: The maximum rank correlation estimato. *Journal of Econometrics* **35**, 303-316.

Harrell, F. E. and Davis, C. E. (1982). A new distribution-free quantile estimator. *Biometrika* **69**, 635-640.

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

Hettmansperger, T. P. and McKean, J. W. (2010). *Robust Nonparametric Statistical Methods.* CRC Press.

Huber, P. J. (1973). Robust regression: asymptotics, conjectures and Monte Carlo. *Ann. Statist.* **1**, 799-821.

Li, G, Peng, H., Zhang, J. and Zhu, L. (2012). Robust rank correlation based screening. *Ann. Statist.* **40**, 1,846-1,877.

Li, R., Zhong, W. and Zhu, L. (2012). Feature screening via distance correlation learning. *J. Am. Statist. Assoc.* **107**, 1,129-1,140.

Lin, H., Zhou, L., Peng, H. and Zhou, X. (2011). Selection and combination of biomarkers using ROC method for disease classification and prediction. *Canadian Journal of Statistics* **39**, 324-343.

Ma, S. and Huang, J. (2005). Regularized ROC method for disease classification and biomarker selection with microarray data. *Bioinformatics* **21**, 4,356-4,362.

Mai, Q. and Zou, H. (2013). A note on the connection and equivalence of three sparse linear discriminant analysis methods. *Technometrics* **55**, 243-246.

Sherman, R. P. (1993). The limiting distribution of the maximum rank correlation estimator. *Econometrica*, **61**, 123-137.

Tian, L., Cai, T., Goetghebeur, E. and Wei, L. J. (2007). Model evaluation based on the sampling distribution of estimated absolute prediction error. *Biometrika* **94**, 297-311.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statisti. Soc. B.*

58, 267-288.

Zhao, D. S. and Li, Y. (2014). Score test variable screening. *Biometrics* **70**, 862-871.

Zhong, W., Zhang, T., Zhu, Y. and Liu, J. S. (2012). Correlation pursuit: Forward stepwise variable selection for index models. *J. Roy. Statisti. Soc. B.* **74**, 849-870.

Zhong, W., Zhu, L., Li, R. and Cui, H. (2016). Regularized quantile regression and robust feature screening for single index models. SS. **26**, 69-95.

Zhu, L. P., Li, L., Li, R. and Zhu, L. X. (2011). Model-free feature screening for ultrahigh-dimensional data. *J. Am. Statist. Assoc.* **106**(496).

Zou, H. and Hastie, T. (2005). Addendum: Regularization and variable selection via the elastic net. *J. Roy. Statisti. Soc. B.* **67**, 301-320.

Center of Statistical Research, School of Statistics Southwestern University of Finance and Economics, Chengdu, China.

E-mail: myb@swufe.edu.cn

Center of Statistical Research, School of Statistics Southwestern University of Finance and Economics, Chengdu, China.

E-mail: liy@swufe.edu.cn

Center of Statistical Research, School of Statistics Southwestern University of Finance and Economics, Chengdu, China.

E-mail: linhz@swufe.edu.cn

Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109-20, USA.

E-mail: yili@med.umich.edu

(Received January 2016; accepted October 2016)