Stat
The ISI's Journal for the Rapid
Dissemination of Statistics Research
(wileyonlinelibrary.com) DOI: 10.100X/sta.0000

A data-driven approach to conditional screening of high dimensional variables

Hyokyoung G. Hong[a], Lan Wang[b], Xuming He[c]*

Received 00 Month 2016; Accepted 00 Month 2016

Marginal screening is a widely applied technique to handily reduce the dimensionality of the data when the number of potential features overwhelms the sample size. Due to the nature of the marginal screening procedures, they are also known for their difficulty in identifying the so-called hidden variables that are jointly important but have weak marginal associations with the response variable. Failing to include a hidden variable in the screening stage has two undesirable consequences: (1) important features are missed out in model selection; and (2) biased inference is likely to occur in the subsequent analysis. Motivated by some recent work in conditional screening, we propose a data-driven conditional screening algorithm, which is computationally efficient, enjoys the sure screening property under weaker assumptions on the model, and works robustly in a variety of settings to reduce false negatives of hidden variables. Numerical comparison with alternatives screening procedures are also made to shed light on the relative merit of the proposed method. We illustrate the proposed methodology using a leukemia microarray data example.

Keywords: Conditional screening; False negative; Feature screening; High dimension; Sparse principal component analysis; Sure screening property

1. Introduction

1.1. Background

A popular approach for analyzing big data is to first apply a computationally expedient screening procedure to reduce the dimensionality to a moderate size. More sophisticated but often more computationally intensive statistical methods, such as penalized regression, can then be applied in the second stage. This practice has become routine in many fields such as genomics and finance, where the number of available features in the data is often huge comparing to the sample size.

[a] **Department of Statistics and Probability, Michigan State University, East Lansing, Michigan 48824, U.S.A.**
[b] **School of Statistics, University of Minnesota, Minneapolis, Minnesota 55455, U.S.A.**
[b] **Department of Statistics, University of Michigan, Ann Arbor, Michigan 48109, U.S.A.**
***Email: xmhe@umich.edu**

Prepared using staauth.cls [Version: 2012/05/12 v1.00]

Fan and Lv (2008) proposed the sure independence screening (SIS) methodology for linear regression, which screens variables by ranking their marginal correlations with the response variable. Under some regularity conditions, these authors proved that sure independence screening has the ability to keep all the important variables with probability tending to one. This desirable property is often referred to as the sure screening property. The marginal screening procedure has been further developed in a series of recent papers for a variety of settings; see Hall and Miller (2009), Fan and Song (2010), Bühlmann *et al.* (2010), Fan *et al.* (2011), Zhu *et al.* (2011), Li *et al.* (2012), Li *et al.* (2012), Mai and Zou (2013), He *et al.* (2013), Liu *et al.* (2014), Shao and Zhang (2014), among others. A different screening procedure is recently proposed by Wang and Leng (2015), which aims to target joint associations in a covariate space in the dimension of the sample size.

1.2. False negatives in marginal screening

Naturally, the success of any marginal screening procedure depends on how well the marginal utility, correlation coefficient between the response and each individual predictor, captures the importance of the predictors in a joint model. A variable may be retained by the screening procedure when it is marginally important but not jointly important (false positive); or a variable that is jointly important but not marginally important can be screened out, resulting in a false negative.

False negatives have two potentially serious consequences. Firstly, important features may be screened out and will not be reinstated by the second-stage analysis. Secondly, the false negatives can lead to bias in subsequent inference. The risk of false negatives is widely recognized. An active variable can be hidden when the correlation is estimated marginally. One illustrative example was given by Guyon and Elisseeff (2003, Section 3.3) for a two-class classification problem, where individual variables have no separation power, but jointly the variables provide good class separation. As another example, we consider the following model from Barut *et al.* (2016), $Y = b_1X_1 + \cdots + b_qX_q - \alpha(b_1 + \cdots + b_q)X_{q+1} + e$, where $q \ll p$, $b_1, \ldots, b_q$ are nonzero constants, and e has a standard normal distribution. The vector of covariates $(X_1, \ldots, X_p)^{\mathsf{T}}$ has a multivariate normal distribution, and the covariance matrix of which has an equally-correlated structure with the correlation coefficient α. In this case, $\mathrm{cov}(Y, X_{q+1}) = 0$ even if X_{q+1} has a large coefficient. Hence, we expect that marginal screening will give little priority to X_{q+1}.

Existing marginal screening procedures share the simplistic assumption that jointly important variables are also marginally important. This assumption is critical to ensuring the sure screening property since if it is violated, false negatives are likely to arise. To alleviate this problem, Fan and Lv (2008) suggested an iterative procedure (ISIS) by repeatedly using the residuals from the previous iteration, which was subsequently adopted by many other marginal screening procedures. For a generalized linear model version of ISIS is studied in Fan and Song (2010) and Fan *et al.* (2009). Due to the iterative nature, the computational costs are higher and the statistical properties of the resulting model are more difficult to analyze. Perhaps more importantly, the performance of the iterative sure independence screening depends a lot on the underlying model. In some settings, particularly when the signal to noise ratio is not very high, iterative sure independence screening can underperform sure independence screening, as shown in our empirical comparisons.

Recently, Barut *et al.* (2016) proposed a conditional screening technique which uses the prior knowledge that a certain set of predictors (denoted by C) are relevant to the response variable. Each remaining variable X_j, where $X_j \notin C$, is evaluated by fitting a regression model using X_j and the predictors in C. The variable X_j is considered important if the magnitude of its coefficient in the above regression model exceeds a given threshold. They derived the sure screening property and demonstrated the prior knowledge of a "good" set C can be very helpful for identifying hidden variables.

Prepared using staauth.cls

However, the question remains how to select a good set C. Our simulation studies suggest that the performance of such a procedure is sensitive to the choice of C.

Motivated by the work of Barut *et al.* (2016) and recent developments on sparse principal component analysis, we propose a data-driven algorithm for conditional screening with generalized linear models. Our goal is to reduce false negatives due to marginal screening without relying on a cherry-picked set of C on which the conditional screening is based. To illustrate our proposed methodology, we analyzed the popular leukemia data (Golub *et al.*, 1999). In this data set, expression levels are measured on 7129 genes for each of the 72 patients. The classification task is to discriminate acute lymphoblastic leukemia from acute myeloid leukemia. We compare the proposed method with several competing procedures with respect to their abilities to select a small subset of genes to build an interpretable and effective predictive model. This real data example demonstrates that subjective or random choice of the conditioning set often leads to unstable performance. The proposed data-adaptive method produced meaningful and more reproducible results in the analysis.

2. The Proposed 3-Step Method

2.1. Preliminaries

Throughout the paper, we assume that the conditional density of the response variable Y given the vector of predictors $X = x$ belongs to the following exponential family

$$f(y; x, \theta) \quad = \quad \exp\left[y\theta(x) - b\{\theta(x)\} + c(x, y)\right], \tag{1}$$

where $b(\cdot)$ and $c(\cdot, \cdot)$ are known functions and $\theta(x)$ is the canonical parameter. We write $x = (x_1, \ldots, x_p)^{\mathsf{T}}$. The popular generalized linear model assumes that there exists a $(p+1)$-vector of parameters $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^{\mathsf{T}}$ such that $E(Y \mid X = x) = b'\{\theta(x)\} = g^{-1}(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)$, where $g(\cdot)$ is the canonical link function, i.e., $g = (b')^{-1}$ and $b'(\theta)$ is the first derivative of $b(\theta)$ with respect to θ. In (1), no dispersion parameter is included since we focus on the mean regression. Furthermore, we standardize the covariates so that they have mean zero and standard deviation one. To emphasize the dependence of p on the sample size, we use $p = p_n$ in the remaining of the paper, and let $\boldsymbol{\beta}_* = (\beta_{*0}, \beta_{*1}, \ldots, \beta_{*p_n})^T$ be the vector of true parameter values. Furthermore, assume that $\boldsymbol{\beta}_*$ is sparse in the sense that the size of the set $M_* = \{j : \beta_{*j} \neq 0, 1 \leq j \leq p_n\}$ is small relative to n even when p_n is large.

In a nutshell, our new algorithm consists of three steps. First, we perform pre-cleaning by standard marginal regression to reduce dimensionality of the problem. Second, we perform sparse principal component analysis on the variables surviving the first step and obtain a set of variables corresponding to those having large loadings on the leading eigenvectors. Finally, using the set of predictors obtained in Step 2, we perform conditional marginal screening. Each step of the new algorithm is computationally fast. We will describe the details of the three steps later in this section.

Step 2 of the proposed method relies on the recent work on sparse principal component analysis. Recently, it has been revealed by several authors, including Li (2007), Artemiou and Li (2009, 2013) and Ni (2011), that the response is often highly correlated with the leading principal components of the covariate vector, and hence principal component analysis is valuable for the purpose of finding a low dimensional summary of relevant predictors in the regression setting. In Section 3, we demonstrate that with the assistance of the sparse principal component analysis, the proposed method yields a robust variant of the conditional marginal screening and works well in a variety of settings to reduce false negatives without relying on a pre-selected set of variables with a priori information to perform the conditional analysis.

Prepared using staauth.cls

2.2. Step 1: Pre-cleaning

In some applications, the number of available predictors can be in the order of tens of thousands or more but the sample size is limited. To expedite the computation, we first perform a pre-screening step. In this step, the number of predictors we retain is allowed to be larger than the sample size, but usually significantly less than the candidate number of predictors. Although the pre-cleaning is based on the marginal utility as the traditional marginal screening methods, we do not require the assumption that jointly important variables are also marginally important to hold. It is important to note that a variable that is screened out at this stage may still be identified in Step 3.

Let $\hat{\beta}_j$ be the maximum likelihood estimator of the coefficient of X_j from fitting a marginal generalized linear model using only the intercept and X_j. Denote $(\hat{\beta}_{0j}, \hat{\beta}_j) = \underset{(\beta_0,\beta_j)}{\operatorname{argmin}} \mathbb{P}_n l(\beta_0 + X_j\beta_j, Y)$, where $l(\theta(x), Y) = -\big[\theta(x)Y - b\{\theta(x)\}\big]$, and for a measurable function g, $\mathbb{P}_n g(X,Y) = n^{-1}\sum_{i=1}^n g(X_i, Y_i)$. We retain the variables whose estimated marginal magnitude is sufficiently large. For a given threshold γ_n, let

$$M_{n1} = \{j : |\hat{\beta}_j| > \gamma_n\}$$

be the index set of the predictors that survive the pre-cleaning.

Define $(\beta_{0j}^*, \beta_j^*) = \arg\min_{(\beta_0,\beta_j)} E\{l(\beta_0 + X_j\beta_j, Y)\}$ as the population version of the marginal regression coefficients. Assume that there exists $A_n \subset \{1, \ldots, p_n\}$ such that $\min_{j\in A_n} |\beta_j^*| \geq c_1 n^{-\kappa}$ and $\sup_{j\in A_n^c} |\beta_j^*| \leq c_2 n^{-\kappa-\delta}$, for some $0 < \kappa < 1/2$, $\delta > 0$ and some positive constants c_1 and c_2. Let $q_n = |A_n|$, where $|A_n|$ denotes the cardinality of A_n. We assume that $q_n \to \infty$ but $q_n = o(p_n)$.

Proposition 2.1
Assume conditions 1-5 in Appendix are satisfied. For $\gamma_n = c_3 n^{-\kappa}$ with $c_3 \leq c_1/2$, there exist positive constants $\xi_i > 0$, $i = 1, 2, 3$, such that $\text{pr}(M_{n1} = A_n) \geq 1 - p_n\{\exp(-\xi_1 n^{1-2\kappa} k_n^{-2} K_n^{-2}) + \exp(-\xi_2 K_n^{\xi_3})\}$, *for all n, where K_n and k_n are defined in conditions 2 and 5, respectively.*

Remark 1
This result follows from a direct application of the exponential bound for the marginal maximum likelihood estimator in the generalized linear model shown in Fan and Song (2010). Under relatively weak conditions the above probability bound converges to 1 as $n \to \infty$.

2.3. Step 2: Sparse principal component analysis

Next, we apply sparse principal component analysis to the variables surviving the pre-cleaning to construct a subset of covariates to condition on. Let X_{A_n} denote the subset of the components of X whose indice are in A_n. Let Σ be the population covariance matrix of X_{A_n}, and consider its spectral decomposition $\Sigma = \sum_{j=1}^{q_n} \lambda_j u_j u_j^{\mathsf{T}}$, where $\lambda_1 \geq \cdots \geq \lambda_{q_n}$ are eigenvalues, and $u_1, \ldots, u_{q_n} \in \mathbb{R}^{q_n}$ constitute an orthonormal basis of the eigenvectors. For a given positive integer $k < q_n$, we consider the principal subspace spanned by the k leading eigenvectors of Σ, that is, the space spanned by $u_1, \ldots, u_k$. Our working assumption is that the leading eigenvectors are sparse in the sense that most of their components are zero.

For a vector v, let $\text{supp}(v)$ be the index set corresponding to non-zero entries of v. Then $B_n = \cup_{j=1}^k \text{supp}(u_j)$ represents the collection of the indices of the predictors corresponding to the nonzero components of the first k eigenvectors. In our numerical examples, a small positive integer k, say $k = 1, 2$, is found to work well. The set of predictors X_{B_n} is what we will use in the conditional screening step.

Prepared using staauth.cls

There are a number of different algorithms for sparse principal component analysis; we refer to Jolliffe *et al.* (2003), Zou *et al.* (2006), Shen and Huang (2008), Witten *et al.* (2009), Johnstone and Lu (2009), Ma (2013), Vu and Lei (2013), She (2014), and the references therein. We use the recently developed fantope projection and selection algorithm of Vu *et al.* (2013) in our work. Unlike some of the alternatives, fantope projection and selection algorithm can be applied to correlation matrices, and has proved theoretical guarantee for consistently identifying B_n. Let $U = [u_1, \ldots, u_k]$. Then the projection matrix associated with the principal subspace is $\Pi = UU^{\mathsf{T}}$. Let S_n be the sample covariance matrix of $X_{M_{n1}}$. Vu *et al.* (2013) proposed to estimate Π by $\hat{H} = \arg\max\{\langle S_n, H\rangle - \rho\|H\|_{1,1}\}$ subject to the $|M_{n1}| \times |M_{n1}|$ matrix $H \in F^k$, where $F^k = \{H : 0 \preceq H \preceq I \text{ and } \mathrm{tr}(H) = k\}$, which is called the trace-k Fantope, ρ is a tuning parameter, and $\langle S_n, H\rangle = \mathrm{tr}(S_n{}^{\mathsf{T}} H)$ is the trace inner product. The matrix (1,1)-pseudonorm $\|H\|_{1,1} = (\|H_{1*}\|_1, \ldots, \|H_{k*}\|_1)_1$ with H_{j*} denoting the jth row of H, $j = 1, \ldots, k$, that is $\|H\|_{1,1}$ is the L_1 norm of the vector that consists of row-wise L_1 norms of H. We then estimate B_n by

$$M_{n2} = \mathrm{supp}\{\mathrm{diag}(\hat{H})\},$$

which is the index set corresponding to the nonzero diagonal elements of $\hat{H}$.

Proposition 2.2
Assume conditions 1-8 in Appendix are satisfied. Then $\mathrm{pr}(M_{n2} = B_n) \geq (1 - 2q_n{}^{-2})\big[1 - p_n\{\exp(-\xi_1 n^{1-2\kappa} k_n^{-2} K_n^{-2}) + \exp(-\xi_2 K_n^{\xi_3})\}\big]$, *for all n.*

Remark. This result follows directly from Theorem 2 on sparsistency of Lei and Vu (2015). The validity of this result does not depend on any assumption on the joint model. Lei and Vu (2015) also showed that even without assuming sparsity, fantope projection and selection method provides a sparse, linear dimension-reducing transformation that is close to the best possible in terms of maximizing the predictive covariance. Vu *et al.* (2013) recommended to dynamically update ρ after each iteration of their algorithm to keep the primal and dual residual norms within a constant factor of each other.

2.4. Step 3: Conditional screening

In the last step, we perform the conditional screening of Barut *et al.* (2016) by conditioning on $X_{M_{n2}}$. This step allows us to include variables with large additional contributions, which may help recruit predictors that are missed in Step 1. More specifically, for each $j \notin M_{n2}$, let $(\hat{\eta}_{0j}, \hat{\eta}_{M_{n2}}, \hat{\eta}_j) = \underset{(\eta_{0j}, \eta_{M_{n2}}, \eta_j)}{\mathrm{argmin}} \ \mathbb{P}_n l(\eta_0 + X_{M_{n2}}^{\mathsf{T}} \eta_{M_{n2}} + X_j \eta_j, Y)$.

Conditioning on M_{n2}, we keep the variables in the following set

$$M_{n3} = \{j : |\hat{\eta}_j| > \gamma_2, j \in M_{n2}^c\}$$

for a given threshold γ_2. Under weak conditions, it can be shown that M_{n3} enjoys the sure screening property. In practice, the threshold γ_2 may need no prior specification if we choose to retain a fixed number of variables with the largest coefficients $|\hat{\eta}_j|$. At the end of the algorithm, we keep the predictors in the set $\hat{M}_n = M_{n2} \cup M_{n3}$.

Theorem 2.1
(Sure screening property) Assume conditions 1-11 in Appendix are satisfied. For the κ' *defined in condition 9, let* $\gamma_2 = a_1 n^{-\kappa'}$ *with* $a_1 \leq a_2$, *where* $a_1 > 0$, $a_2 > 0$ *are constants. Then there exist positive constants* ξ_i, $i = 4, 5, 6$ *such that* $\mathrm{pr}(M^* \subset \hat{M}_n) \geq \big[1 - d_n\{\exp(-\xi_4 n^{1-2\kappa} k_n'^{-2} K_n^{-2}) + n\exp(-\xi_5 K_n^{\xi_6})\}\big](1 - 2q_n{}^{-2})\big[1 - p_n\big\{\exp(-\xi_1 n^{1-2\kappa} k_n^{-2} K_n^{-2}) + \exp(-\xi_2 K_n^{\xi_3})\big\}\big]$ *for all n, where* $d_n = |M^*|$ *and* k_n' *are defined in condition 11.*

Remark. This result follows from Proposition 2 and Theorem 3 of Barut *et al.* (2016). For linear and logistic models, the optimal order of K_n is a positive power of $n^{1-2\kappa}$. It is easy to see the probability in Theorem 1 goes to one even when p_n is allowed to grow exponentially fast.

3. Monte Carlo studies

In the simulation studies, we compare the proposed 3-step method (denoted by 3S in Table 1) with the following alternatives: (i) sure independence screening (SIS) of Fan and Lv (2008), (ii) iterative sure independence screening (ISIS) of Fan *et al.* (2009), (iii) conditional sure independence screening (CSIS) of Barut *et al.* (2016) with different choices of the conditioning set C, and (iv) high-dimensional ordinary least-squares projector (HOLP) of Wang and Leng (2015). Sure independence screening and iterative sure independence screening are both implemented using the R-package SIS on CRAN (version of 0.7-5). We fix $\rho = 0.5$, the tuning parameter in the fantope projection. In the real analysis, we can determine a value of ρ by cross-validation.

We evaluate different screening methods on B simulations runs, where $B = 200$ for $(p, n) = (1000, 100)$ and $B = 100$ for $(p, n) = (10000, 200)$. Across B simulation runs, we compare the different methods according to two criteria: i) true model inclusion ratio (TMR), the proportion of times when the first n variables retained from screening include all the active variables; ii) average active variable ratio (AAR), the proportion of active variables in the set of n retained variables after screening. Higher values of TMR and AAR indicate better screening procedures.

In this study, Examples 1–4 are adapted or modified from Barut *et al.* (2016) and Fan and Lv (2008) and Example 6 is used in Wang (2009) and Wang and Leng (2015). The random error ϵ follows $N(0, \sigma^2)$ with σ^2 being adjusted to achieve a pre-specified R^2 defined as $\text{var}(X^{\mathsf{T}}\beta)/\text{var}(Y)$.

Example 1. $Y = X^{\mathsf{T}}\beta + \epsilon$, where all covariates follow the standard normal distribution with equal correlation 0.5 and $\beta = (3, 3, 3, 3, 3, -7.5, 0, \ldots, 0)^{\mathsf{T}}$. In this setting, X_6 is the hidden active variable since $\text{cov}(X_6, Y) = 0$ although β_6 has a large contribution to the response variable.

Example 2. The conditional distribution of Y given $X = x$ follows the binomial distribution with $\text{pr}(Y = 1 \mid X = x) = \exp(x^{\mathsf{T}}\beta)/\{1 + \exp(x^{\mathsf{T}}\beta)\}$, where X and β are the same as in Example 1.

Example 3. $Y = X^{\mathsf{T}}\beta + \epsilon$, where all the covariates except X_1–X_7 follow the independent standard normal distribution and $\beta = (10, 0, \ldots, 0, 1)^{\mathsf{T}}$. The first 7 covariates are normal with equi-correlation 0.9.

Example 4. $Y = X^{\mathsf{T}}\beta + \epsilon$, where $\beta = (1, 1.3, 1, 1.3, 1, 1.3, 1, 1.3, 1, 1.3, 1, 1.3, 0, \ldots, 0)^{\mathsf{T}}$, $X_j = \epsilon_j$, $\{\epsilon_j\}_{j=1,\ldots,[p/3]}$ are independent and identically distributed standard normal variables, $\{\epsilon_j\}_{j=[p/3]+1,\ldots,[(2p)/3]}$ are independent and identically distributed double exponential variables with location parameter zero and scale parameter one, and $\{\epsilon_j\}_{j=[(2p)/3]+1,\ldots,p}$ are independent and identically distributed with a mixture normal distribution with two components $N(-1, 1)$ and $N(1, 0.5)$ of equal mixture proportion, where $[x]$ denotes the integer part of x.

Example 5. $Y = X^{\mathsf{T}}\beta + \epsilon$, where all the covariates except X_2–X_{50} are independent standard normal random variables and $\beta = (0.7, 0.2, -0.2, -0.2, 0, \ldots, 0)^{\mathsf{T}}$. The covariates X_2–X_{50} are normal with equi-correlation 0.9.

Example 6. $Y = X^{\mathsf{T}}\beta + \epsilon$, where each X_i follows a multivariate normal distribution with mean 0 and

Prepared using staauth.cls

$\text{cov}(X_{ij_1}, X_{ij_2}) = 0.3^{|j_1-j_2|}$ and $\beta = (3, 0, 0, 1.5, 0, 0, 2, 0, \ldots, 0)^{\mathsf{T}}$.

Put Table 1 about here.

We summarize the simulation results from the above examples in Table 1, but note that Wang and Leng (2015) is not developed for binary responses so is excluded in Example 2. The performance of conditional sure independence screening depends on the prior knowledge of a conditioning set C. We consider the case when such an informative set is available as well as the case where we do not have such knowledge. In Example 1, we consider two choices of the conditioning sets, $C_1 = \{1, 2\}$ and $C_2 =$\{a random choice of 2 inactive variables\}; in Examples 2–6 we consider the choices $C_1 = \{1\}$ and $C_2 =$\{a random choice of 1 inactive variable\}. Here, C_1 is a favorable choice, but C_2 is not. For the proposed 3S method, we retain n covariates in the pre-cleaning step. Although the number of predictors kept in this stage can be larger than the sample size, we did not observe significant gains from holding larger sets of covariates such as sizes of $2n$ and $3n$ in the above examples. To control the effect due to the size of the conditioning set, we let the size of M_{n2} of our proposed procedure be the same as that of the conditioning set used for conditional sure independence screening. We observe from Table 1 that

- The iterative sure independence screening tends to perform well when the signal-to-noise ratio is high (with $R^2 = 0.9$) but it does not always improve over sure independence screening. In more realistic setting with $R^2 = 0.5$ or lower, the iterative screening often performs worse, and sometimes substantially so (Examples 2, 5 and 6).
- The benefit of conditional sure independence screening is clear when a favorable choice of the conditioning set is given, such as the case with C_1. However, when the conditioning set is not well chosen, it becomes less robust and its performance could degenerate easily.
- In all the examples, the proposed method 3S remains highly competitive. Often, it performs significantly better than sure independence screening when the latter has difficulty selecting the hidden active variables (Examples 1–3). When compared with conditional sure independence screening, it is much closer to the behavior of conditional sure independence screening under a favorable conditioning set than that without a well-informed choice of the conditioning set.
- High-dimensional ordinary least-squares projector is often a competitive method, however, it has very poor performance in Example 5 and it has not been developed for binary regression (Example 2).

In summary, for a wide range of settings, the proposed method remains competitive. It can be viewed as a data-adaptive version of the conditional sure independence screening. Our empirical work demonstrates the potential of the proposed method to reduce the false negatives when hidden variables are present.

4. Real Data Example

We illustrate the proposed methodology using the widely analyzed leukemia microarray data set from Golub *et al.* (1999). The data set contains measurements on specimens from bone marrow or peripheral blood samples taken from 72 patients, who had either acute lymphoblastic leukemia (ALL) or acute myeloid leukemia (AML). Gene expression levels were measured using Affymetrix high-density oligonucleotide arrays containing p =7129 human genes.

It is important to identify signature genes for distinguishing ALL and AML. We first split the data into 38 training samples (27 ALL and 11 AML) and 34 testing samples (20 ALL and 14 AML) as done in the original investigation (Golub *et al.*, 1999), and then apply the proposed method to select genes and evaluate the performance of classification

power. We standardize the gene expression data so that the arrays have mean 0 and variance 1 across genes, which is a common pre-prcessing method used in expression data analysis. The predictive performance of the selected genes on the training and testing data is assessed for several competing methods under consideration. They are

- SIS: using top three genes by marginal screening.
- ISIS: using top three genes by iterative screening.
- CSIS-i: the CSIS method conditional on $C_1 = \{X95735, D26156\}$, two genes used in Barut *et al.* (2016) as good choices.
- CSIS-ii: the CSIS method conditional on $C_2 = \{X95735, M27783\}$, two top genes from marginal screening SIS.
- CSIS-iii: the CSIS method conditional on C_3 which consists of two randomly chosen genes.
- 3S: the proposed 3-step screening method.

Put Table 2 about here.

The results on the mis-classification performance (evaluated on the testing data) are reported in Table 2, where the misclassification rates are computed using two different approaches. The first one *Test error*^{1} uses the training data to both select the genes and fit the logistic regression model which is then used for classification, while the second one *Test error*^{2} uses the training data to select the genes and then use the full data to estimate the logistic model coefficients for classification.

We retain three genes for SIS to compare directly with the results in Barut *et al.* (2016). For the CSIS-iii, the mis-classification rates are averaged over 200 randomly chosen sets of C_3. For our proposed method 3S, we choose the tuning parameter ρ needed in Step 2 (the fantope projection) by cross-validation. The ρ value of 0.45 for the training data set is used.

The proposed data-driven conditional screening method 3S identified genes D88422 (CYSTATIN A), X95735 (Zyxin) in Step 2 and selected an additional gene M21624 (TCRD T-cell receptor, delta) in Step 3. We observe that the three genes identified by the 3S procedure have satisfactory predictive power. In fact, the misclassification rates (*Test error*^{1}=2/34, *Test error*^{2}=1/34) are the same as those for CSIS-i, where a knowledgeable choice of the conditioning set based on information from the medical literature is used (Barut *et al.*, 2016). The original paper of Golub *et al.* (1999) used 50 genes to build a predictive model but only achieved correct classification on 29 of the 34 cases in the testing data set. In comparison, the proposed 3S method achieve better classification performance with far fewer genes.

It is also informative to see what happens when we apply different methods to the full data, that is, the full data are used to select the genes, fit the model and evaluate the classification performance. When CSIS-i with the same C_1 as in Table 2 is applied to the full data, it selects an additional gene HG651-HT4201 and results in a misclassification rate of 4/72. This is clearly not as favorable as the results in Table 2 and demonstrates that even though C_1 (subjective choice) gives good performance for the training data, it would lead to less favorable results on the full data. When the proposed method S3 is used, we obtain C ={M27891 (Cystatin C)} for $k = 1$, and C ={M27891 (Cystatin C), D88422 (Cystatin A)} for $k = 2$ in Step 2 using the fantope projection tuning parameter $\rho = 0.49$. In Step 3, an additional gene U62136 (putative enterocyte differentiation promoting factor mRNA) is selected conditioning on either $\{M27891, D88422\}$ or $\{M27891\}$. In either case, the method identified a model with no misclassification on the same data. Note that U62136 is removed in Step 1 of the 3S method due to its small marginal utility, but it is retrieved in Step 3 by conditioning on the gene(s) selected in Step 2. Those three genes by 3S were all identified as important in Golub *et al.* (1999). Previous studies (Dramiński, and Vannucci, 2008; Guan and Zhao, 2005) also confirmed U62136 as one of the top-ranked reference genes for leukemia.

Prepared using staauth.cls

With SIS, U62136 is ranked as 63 out of the total number of 7129 genes. Therefore, it can be easily missed by marginal screening when only a small subset of genes are chosen in the final model. It is also missed by CSIS with the subjective choice of C_1 used by Barut *et al.* (2016). Overall, we conclude that the proposed 3S method with adaptive choice of conditioning set provides reliable results in this example, whether the procedure is applied to the training data alone or to the full data.

Appendix

A. Technical conditions

We first introduce some additional notation. Let $\boldsymbol{\beta}_j = (\beta_{0j}, \beta_j)^{\mathsf{T}}$, $\mathbf{X}_j = (1, X_j)^{\mathsf{T}}$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{p_n})^{\mathsf{T}}$ and $\mathbf{X} = (1, X_1, \ldots, X_{p_n})^{\mathsf{T}}$. Let $\mathbf{X}_{B_n}$ be the subvector of $\mathbf{X}$ consisting of those components in B_n, where B_n is defined in Section 2.3. For $j \notin B_n$, let $\mathbf{X}_{B_nj} = (\mathbf{X}_{B_n}^{\mathsf{T}}, X_j)^{\mathsf{T}}$. Let $\boldsymbol{\eta} = (\eta_0, \eta_1, \ldots, \eta_{p_n})^{\mathsf{T}}$. For $j \notin B_n$, let $\boldsymbol{\eta}_{B_nj} = (\boldsymbol{\eta}_{B_n}^{\mathsf{T}}, \eta_j)^{\mathsf{T}}$ and $\boldsymbol{\eta}^*_{B_nj} = \arg\min_{\boldsymbol{\eta}_{B_nj}} EI(\mathbf{X}_{B_nj}^{\mathsf{T}}\boldsymbol{\eta}_{B_nj}, Y)$. For a vector $v \in \mathbb{R}^k$, $||v||$ denotes the Euclidean norm and $||v||_\infty = \max_{1\le i\le k} |v_i|$. For a matrix $A \in \mathbb{R}^{n\times m}$, and index sets $J_1 \subset \{1, \ldots, n\}$ and $J_2 \subset \{1, \ldots, m\}$, $A_{J_1J_2}$ denotes the submatrix consists of rows in J_1 and columns in J_2. For $1 \le i \le n$ and $1 \le j \le m$, A_{ij} denotes the (i, j)th entry of A. The matrix $(2, \infty)$-norm $||A||_{2,\infty}$ is defined as $||(||A_{1*}||, ||A_{2*}||), \ldots, ||A_{n*}||)||_\infty$, where A_{i*} denotes the ith row of A, $i = 1, \ldots, n$. Let $M_{*B_n} = M_* \cap B_n$. Let $E_L(X_j|\mathbf{X}_{B_n})$ denote the best linear regression fit of X_j regressed on $\mathbf{X}_{B_n}$. Let $\mathrm{Cov}_L(Y, X_j|\mathbf{X}_{B_n}) = E(X_j - E_L(X_j|\mathbf{X}_{B_n}))(Y - E_L(Y|\mathbf{X}_{B_n}))$.

Condition 1. The marginal Fisher information $I_j(\boldsymbol{\beta}_j) = E\{b''(\mathbf{X}_j^{\mathsf{T}}\boldsymbol{\beta}_j)\mathbf{X}_j\mathbf{X}_j^{\mathsf{T}}\}$ is finite and positive definite at $\boldsymbol{\beta}_j^* = (\beta_{0j}^*, \beta_j^*)^{\mathsf{T}}$, where $(\beta_{0j}^*, \beta_j^*) = \arg\min_{(\beta_0,\beta_j)} E\{I(\beta_0 + X_j\beta_j, Y)\}$ for $j = 1, \ldots, p_n$. Moreover, $\sup_{\boldsymbol{\beta}_j\in\mathbb{B}, ||\mathbf{X}||=1} ||I_j(\boldsymbol{\beta}_j)^{1/2}\mathbf{X}||$ is bounded from above, where $||\cdot||$ is the Euclidean norm, $\mathbb{B} = \{\boldsymbol{\beta}_j : |\beta_{0j}| \le B, |\beta_j| \le B\}$ is a sufficiently large set for which $\boldsymbol{\beta}_j^*$ is an interior point.

Condition 2. The second derivative of $b(\theta)$ is continuous and positive. There exists an $\epsilon_1 > 0$ such that for all $j = 1, \ldots, p_n$, for some sufficiently large positive constant K_n, $\sup_{\boldsymbol{\beta}_j\in\mathbb{B}, ||\boldsymbol{\beta}_j - \boldsymbol{\beta}_j^*||\le\epsilon_1} \left|Eb(\mathbf{X}_j^{\mathsf{T}}\boldsymbol{\beta}_j)I(|X_j| > K_n)\right| = o(n^{-1})$.

Condition 3. For all $\boldsymbol{\beta}_j \in \mathbb{B}$, we have $E(I(\mathbf{X}_j^{\mathsf{T}}\boldsymbol{\beta}_j, Y) - I(\mathbf{X}_j^{\mathsf{T}}\boldsymbol{\beta}_j^*, Y)) \ge V||\boldsymbol{\beta}_j - \boldsymbol{\beta}_j^*||^2$, for some positive constant V, bounded from below uniformly over $j = 1, \ldots, p_n$.

Condition 4. There exist some positive constants m_0, m_1, s_0, s_1 and α, such that for sufficiently large t, $P(|X_j| > t) \le (m_1 - s_1)\exp(-m_0 t^\alpha)$, for $j = 1, \ldots, p_n$, and that $E\left[\exp\{b(\mathbf{X}^{\mathsf{T}}\boldsymbol{\beta}_* + s_0) - b(\mathbf{X}^{\mathsf{T}}\boldsymbol{\beta}_*)\}\right] + E\left[\exp\{b(\mathbf{X}^{\mathsf{T}}\boldsymbol{\beta}_* - s_0) - b(\mathbf{X}^{\mathsf{T}}\boldsymbol{\beta}_*)\}\right] \le s_1$.

Condition 5. Let $k_n = b'(K_nB + B) + m_0K_n^\alpha/s_0$. Assume that $n^{1-2\kappa}/(k_n^2K_n^2) \to \infty$.

Condition 6. The matrix Σ defined in Section 2.3 satisfies the sparse principal subspace condition (SPS) in Lei and Vu (2015) with some positive integer k and a support set J of size s; Σ also satisfies the limited correlation condition (LCC) in Lei and Vu (2015) with a constant $\delta \in (0, 1]$. The matrix S_n defined in Section 2.3 satisfies the maximum error bound condition in Lei and Vu (2015) with a scaling constant σ.

Condition 7. Assume that $s\sqrt{\log q_n/n} \le \{\delta(\lambda_k - \lambda_{k+1})^2\}/\{4\sigma(8\lambda_1 + \lambda_k - \lambda_{k+1})\}$. The tuning parameter ρ in estimating $\hat{H}$ satisfies: $\rho = \sigma\delta^{-1}\sqrt{\log q_n/n}$.

Condition 8. Either of the following two conditions holds:
(1) $4s\sigma\sqrt{\log q_n/n} < \delta(\lambda_k - \lambda_{k+1})\min_{j\in J}\sqrt{\Pi_{jj}}$, where Π is defined in Section 2.3;
(2) $\mathrm{rank}(\mathrm{sign}(\Sigma_{JJ})) = 1$ and $2\sigma\sqrt{\log q_n/n} < \delta\min_{i\in J, j\in J}\Sigma_{ij}$.

Condition 9. There exist $c_1' > 0$ and $0 < k' < 1/2$ such that $\mathrm{cov}_L(Y, X_j|\mathbf{X}_{B_n}) \geq c_1' n^{-\kappa'}$, $\forall\ j \in M_{*B_n}$. Let $\phi_j = \{b'(\mathbf{X}_{B_nj}^\top \boldsymbol{\eta}_{B_nj}^*) - b'(\mathbf{X}_{B_n}^\top \boldsymbol{\eta}_{B_n}^*)\}/(\mathbf{X}_{B_nj}^\top \boldsymbol{\eta}_{B_nj}^* - \mathbf{X}_{B_n}^\top \boldsymbol{\eta}_{B_n}^*)$. Assume that $E(\phi_j X_j^2) \leq c_2'$ for some positive constant c_2, uniformly in $j \in B_n^c$.

Condition 10. (i) The marginal Fisher information $I_j(\boldsymbol{\eta}_{B_nj}) = E\{b''(\mathbf{X}_{B_nj}^\top \boldsymbol{\eta}_{B_nj}^*)\mathbf{X}_{B_nj}\mathbf{X}_{B_nj}^\top\}$ is bounded. Moreover, $\sup_{\boldsymbol{\beta}_{B_nj} \in \mathbb{B}', ||\mathbf{X}_{B_nj}||=1} ||I_j(\boldsymbol{\eta}_{B_nj})^{1/2}\mathbf{X}_{B_nj}||$ is bounded, where $||\cdot||$ is the Euclidean norm, $\mathbb{B}' = \{\boldsymbol{\eta}_{B_nj} : |(\boldsymbol{\eta}_{B_nj})_l| \leq B',\ \forall\ l \in B_n \cup \{j\}\}$ is a sufficiently large set for which $\boldsymbol{\eta}_{B_nj}^*$ is an interior point, B' is a sufficiently large positive constant, and $(\boldsymbol{\eta}_{B_nj})_l$ denotes the lth component of $\boldsymbol{\eta}_{B_nj}$.

(ii) There exists an $\epsilon_1' > 0$ such that for all $j \in B_n^c$, for the K_n specified in condition 2, $\sup_{\boldsymbol{\eta}_{B_nj} \in \mathbb{B}', ||\boldsymbol{\eta}_{B_nj} - \boldsymbol{\eta}_{B_nj}^*|| \leq \epsilon_1'} \left|Eb(\mathbf{X}_{B_nj}^\top \boldsymbol{\eta}_{B_nj})I(|X_j| > K_n)\right| = o(n^{-1})$. The function $l(x^\top\beta, y)$ satisfies the Lipschitz condition with positive constant k_n, $|l(x^\top\beta, y) - l(x^\top\beta', y)|I_n(x,y) \leq k_n|x^\top\beta - x^\top\beta'|I_n(x,y)$, for $\beta, \beta' \in B$, a compact, convex parameter set, where $I_n(x,y) = I\{(x,y) \in \Omega_n\}$ and $\Omega_n = \{(x,y) : ||x||_\infty \leq K_n, |y| \leq K_n^*\}$, for some positive constants K_n and $K_n^* = m_0 K_n^\alpha / s_0$.

(iii) For all $\boldsymbol{\eta}_{B_nj} \in \mathbb{B}'$, we have $E(l(\mathbf{X}_{B_nj}^\top \boldsymbol{\eta}_{B_nj}, Y) - l(\mathbf{X}_{B_nj}^\top \boldsymbol{\eta}_{B_nj}^*, Y)) \geq V'||\boldsymbol{\eta}_{B_nj}^* - \boldsymbol{\eta}_{B_nj}^*||^2$, for some positive constant V'.

Condition 11. Let $k_n' = b'(K_n B(|B_n| + 1)) + m_0 K_n^\alpha / s_0$. Assume that $n^{1-2\kappa'}/(k_n'^2 K_n^2) \to \infty$.

Remark. Conditions 1-5 are those that appear in Theorem 4 of Fan and Song (2010) for the exponential bound for the marginal maximum likelihood estimator in the generalized linear model. As they have discussed, conditions 1-3 are satisfied by many examples of generalized linear models; condition 4 ensures the tail of Y to be exponentially light; condition 5 is necessary for the exponential inequality of the marginal MLE. Conditions 6-8 are those that appear in Theorem 2 of Lei and Vu (2015). Conditions 9-11 are those that appear in Theorem 3 of Barut et al. (2016), which are parallel to conditions 1-5 for the unconditional case.

B. Technical derivations

Proof of Proposition 1. It is embedded in the proof for Theorem 4(i) of Fan & Song (2010) that for $\gamma_n = c_3 n^{-\kappa}$ with $c_3 \leq c_1/2$ there exist positive constants $\xi_i > 0$, $i = 1, 2, 3$, such that the marginal likelihood estimator satisfies

$$\mathrm{pr}(|\hat{\beta}_j - \beta_j^*| > \gamma_n) \leq \exp(-\xi_1 n^{1-2\kappa} k_n^{-2} K_n^{-2}) + \exp(-\xi_2 K_n^{\xi_3}), \tag{2}$$

$\forall\ 1 \leq i \leq p_n$. Consider the event $E_n = \{\max_{1 \leq i \leq p_n} |\hat{\beta}_j - \beta_j^*| \leq \gamma_n\}$. Note that on this event, we have $M_{n1} = A_n$. Hence, $\mathrm{pr}(M_{n1} = A_n) \geq \mathrm{pr}(E_n) = 1 - \mathrm{pr}(E_n^c)$ and the result follows by an application of (2) and the union bound. □

Proof of Proposition 2. Note that $\mathrm{pr}(M_{n2} = B_n) \geq \mathrm{pr}(M_{n2} = B_n|M_{n1} = A_n)\mathrm{pr}(M_{n1} = A_n)$. By Theorem 2 of Lei and Vu (2015), $\mathrm{pr}(M_{n2} = B_n|M_{n1} = A_n) \geq 1 - 2q_n^{-2}$. The result then follows by an application of Proposition 1. □

Proof of Theorem 1. Note that $\mathrm{pr}(M^* \subset \hat{M}_n) \geq \mathrm{pr}(M^* \subset \hat{M}_n|M_{n2} = B_n)\mathrm{pr}(M_{n2} = B_n)$. By Theorem 3 of Barut et al. (2016), $\mathrm{pr}(M^* \subset \hat{M}_n|M_{n2} = B_n) \geq \left[1 - d_n\{\exp(-\xi_4 n^{1-2\kappa} k_n'^{-2} K_n^{-2}) + n\exp(-\xi_5 K_n^{\xi_6})\}\right]$. The result then follows by an application of Proposition 2. □

Acknowledgement

We would like to thank Dr. Emre Barut and Dr. Vincent Vu for helpful discussions, Dr. Zongming Ma for sharing his codes for sparse principal component analysis; and Dr. Chenlei Leng and Dr. Yiyuan She for sharing their unpublished papers. Hyokyoung G. Hong is supported by NSA grant H98230-15-1-0260. Lan Wang is supported by NSF Grant DMS-1512267. Xuming He is supported by NSF grant DMS-1307566.

Prepared using staauth.cls

References

Artemiou, A. and Li, B. (2009). On principal components and regression: a statistical explanation of a natural phenomenon. *Statist. Sinica* **19**, 1557–1566.

Artemiou, A. and Li, B. (2012). Predictive power of principal components for single-index model and sufficient dimension reduction. *J. Multivariate Anal.* **119**, 176–184.

Barut, E., Fan, J. and Verhasselt, A. (2016). Conditional sure independence screening. *Journal of the American Statistical Association, to appear*

Bühlmann, P., Kalisch, M. and Maathuis, M. H. (2010). Variable selection in high-dimensional linear models: partially faithful distributions and the PC-simple algorithm. *Biometrika* **97**, 261–278.

Do, K-A., Müller, P. and Vannucci, M. (2006). Bayesian Inference for Gene Expression and Proteomics. *Cambridge University Press, NY*

Dramiński, M., Rada-Iglesias, A., Enroth, S., Wadelius, C., Koronacki, J. and Komorowski, J. (2008). Monte Carlo feature selection for supervised classification *Bioinformatics* **24**, 110–117.

Dudoit, S., Fridlyand, J. X. and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association* **97**, 77–87.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Assoc.* **96** 1348–1360.

Fan, J. and Lv, J. (2008). Sure independence screening for ultra-high dimensional feature space (with discussion). *J. R. Statist. Soc.* B **70**, 849–911.

Fan, J., Feng, Y. and Song, R. (2011). Nonparametric independence screening in sparse ultra-high dimensional additive models. *J. Am. Statist. Assoc.* **116**, 544–557.

Fan, J., Samworth, R. and Wu, Y. (2009). Ultrahigh Dimensional Feature Selection: Beyond The Linear Model. *Journal of Machine Learning Research* **10**, 2013–2038.

Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *Ann. Statist.* **38**, 3567–3604.

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C. and Lander, E. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286 (5439)** 531–537.

Guan, Z. and Zhao, H. (2005). A semiparametric approach for marker geneselection based on gene expression data *Bioinformatics* **21**, 529–536.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research* **3**, 1157–1182.

Hall, P. and Miller, H. (2009). Using the bootstrap to quantify the authority of an empirical ranking. *Ann. Statist.* **37**, 3929–3959.

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.

Johnstone, I. M. and Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Am. Statist. Assoc.*, **104**, 682–693.

Jolliffe, I. T., Trendafilov, N. T. and Uddin, M. (2003). A modified principal component technique based on the lasso. *J. Comput. Graph. Statist.*, **12**, 531–547.

Lei, J. and Vu, V. Q. (2015). Sparsistency and agnostic inference in sparse PCA. *Ann. Statist.* **43**, 299–322.

Li, B. (2007). Comment on Cook (2007). *Statist. Sinica* **22**, 32–35.

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

Li, R., Zhong, W. and Zhu, L. (2012). Feature screening via distance correlation learning. *J. Am. Statist. Assoc.* **107**, 1129–1139.

Liu, J., Li, R. and Wu, R. (2014). Feature selection for varying coefficient models with ultrahigh dimensional covariates. *J. Am. Statist. Assoc.* **109**, 266–274.

Mai, Q. and Zou, H. (2013). The Kolmogorov filter for variable screening in high-dimensional binary classification. *Biometrika* **100**, 229–234.

Ni, L. (2011). Principal component regression revisited. *Statist. Sinica* **21**, 741–747.

Ma, Z. (2013). Sparse principal component analysis and iterative thresholding. *Ann. Statist.* **41**, 772–801.

Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statist. Sinica*, **17**, 1617–1642.

Saldana, D. F. and Feng, Y. (2015). Sure Independence Screening in Ultrahigh Dimensional Statistical Models. *manuscript*

Shao, X. and Zhang, J. (2014). Martingale difference correlation and its use in high dimensional variable screening. *J. Am. Statist. Assoc.*, **109**, 1302–1318

She, Y. Y. (2014). Selectable factor extraction in high dimensions. *Technical report*.

Shen, H. and Huang, J. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *J. Multivariate Anal.*, **99**, 1015–1034.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. R. Statist. Soc.* B **58**, 267–288.

Vu, V. Q., Cho, J., Lei, J. and Rohe, K. (2013). Fantope projection and selection: a near-optimal convex relaxation of sparse PCA. Advances in Neural Information Processing Systems 26, Curran Associates, Inc., 2670–2678

Vu, V.Q. and Lei, J. (2013). Minimax sparse principal subspace estimation in high dimensions. *Ann. Statist.*, 41, 2905–2947.

Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. *J. Am. Statist. Assoc.*, 104. 1512–1524.

Wang, X. and Leng, C. (2016). High-dimensional ordinary least-squares projector for screening variables. *J. R. Statist. Soc.* B., *to appear*

Witten, D., Tibshirani, R. and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *J. Multivariate Anal.*, **10**, 515–534.

Prepared using staauth.cls

Zhu, L., Li, L., Li, R. and Zhu, L. X. (2011). Model-free feature screening for ultrahigh dimensional data. *J. Am. Statist. Assoc.* **106**, 1464–1475.

Zou, H., Hastie, T. and Tibshirani, R. (2006). Sparse principal component analysis. *J. Comput. Graph. Statist.* **15**, 265–286.

Table 1. Comparison of true model inclusion and average active viable ratios by different screening methods.

		$(p, n) = (1000, 100)$				$(p, n) = (10000, 200)$			
		$R^2 = 50\%$		$R^2 = 90\%$		$R^2 = 50\%$		$R^2 = 90\%$	
	Method	TMR	AAR	TMR	AAR	TMR	AAR	TMR	AAR
Example 1	SIS	0.00	0.60	0.00	0.77	0.00	0.59	0.00	0.80
($d = 6$)	ISIS	0.00	0.41	0.98	1.00	0.01	0.49	1.00	1.00
	CSIS with C_1	0.28	0.81	0.80	0.96	0.29	0.82	0.93	0.99
	CSIS with C_2	0.06	0.68	0.46	0.89	0.07	0.68	0.62	0.94
	HOLP	0.06	0.67	0.58	0.91	0.14	0.69	0.77	0.96
	3S	0.08	0.67	0.65	0.93	0.15	0.74	0.82	0.96
Example 2	SIS	0.00	0.48	0.00	0.63	0.00	0.40	0.00	0.64
($d = 6$)	ISIS	0.00	0.28	0.01	0.52	0.00	0.29	0.04	0.59
	CSIS with C_1	0.12	0.71	0.32	0.83	0.08	0.70	0.48	0.88
	CSIS with C_2	0.02	0.54	0.14	0.74	0.00	0.46	0.10	0.73
	HOLP	-	-	-	-	-	-	-	-
	3S	0.00	0.55	0.23	0.78	0.03	0.52	0.15	0.76
Example 3	SIS	0.13	0.56	0.22	0.61	0.13	0.56	0.19	0.60
($d = 2$)	ISIS	0.06	0.50	0.56	0.78	0.05	0.51	0.79	0.90
	CSIS with C_1	0.24	0.62	0.89	0.94	0.15	0.57	0.95	0.98
	CSIS with C_2	0.14	0.57	0.22	0.61	0.14	0.57	0.21	0.60
	HOLP	0.15	0.57	0.34	0.67	0.13	0.56	0.22	0.61
	3S	0.23	0.62	0.57	0.78	0.13	0.56	0.63	0.82
Example 4	SIS	0.00	0.66	0.07	0.84	0.00	0.69	0.24	0.91
($d = 12$)	ISIS	0.00	0.32	0.85	0.96	0.00	0.37	0.98	0.99
	CSIS with C_1	0.01	0.69	0.17	0.87	0.00	0.75	0.40	0.93
	CSIS with C_2	0.00	0.65	0.06	0.83	0.00	0.69	0.24	0.90
	HOLP	0.00	0.64	0.13	0.86	0.02	0.69	0.30	0.92
	3S	0.00	0.65	0.09	0.84	0.02	0.68	0.28	0.91
Example 5	SIS	0.33	0.63	0.51	0.76	0.50	0.73	0.74	0.91
($d = 4$)	ISIS	0.00	0.28	0.00	0.44	0.00	0.30	0.00	0.48
	CSIS with C_1	0.56	0.79	1.00	1.00	0.78	0.92	1.00	1.00
	CSIS with C_2	0.15	0.62	0.14	0.66	0.14	0.62	0.20	0.69
	HOLP	0.00	0.25	0.00	0.26	0.02	0.30	0.06	0.48
	3S	0.27	0.65	0.36	0.74	0.30	0.67	0.28	0.65
		$R^2 = 30\%$		$R^2 = 50\%$		$R^2 = 30\%$		$R^2 = 50\%$	
Example 6	SIS	0.71	0.90	0.90	0.97	0.83	0.94	0.97	0.99
($d = 3$)	ISIS	0.33	0.70	0.79	0.92	0.37	0.73	0.93	0.94
	CSIS with C_1	0.83	0.94	0.99	0.99	0.83	0.94	1.00	1.00
	CSIS with C_2	0.72	0.90	0.89	0.96	0.70	0.89	0.97	0.99
	HOLP	0.71	0.90	0.89	0.96	0.73	0.91	0.97	0.99
	3S	0.72	0.90	0.91	0.97	0.87	0.96	1.00	1.00

Table 2. Misclassification rates for the Leukemia data with training and testing samples.

Method	*Train error*	*Test error*[1]	*Test error*[2]
1) SIS	0/38	5/34	3/34
2) ISIS	0/38	3/34	3/34
3) CSIS			
i) informative C_1={X95735, D26156}	0/38	2/34	1/34
ii) informative C_2 ={X95735, M27783}	1/38	5/34	3/34
iii) non-informative C_3={Two random choices}	$(0.1 \pm 1.14)/38$	$(6.9 \pm 3.42)/34$	$(6.1 \pm 3.93)/34$
4) Proposed			
data-driven C ={D88422, X95735}	0/38	2/34	1/34