Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *http://www.jstatsoft.org/*

SIS: An R Package for Sure Independence Screening in Ultrahigh Dimensional Statistical Models

Diego Franco Saldana
Columbia University

Yang Feng
Columbia University

Abstract

We revisit sure independence screening procedures for variable selection in generalized linear models and the Cox proportional hazards model. Through the publicly available R package **SIS**, we provide a unified environment to carry out variable selection using iterative sure independence screening (ISIS) and all of its variants. For the regularization steps in the ISIS recruiting process, available penalties include the LASSO, SCAD, and MCP while the implemented variants for the screening steps are sample splitting, data-driven thresholding, and combinations thereof. Performance of these feature selection techniques is investigated by means of real and simulated data sets, where we find considerable improvements in terms of model selection and computational time between our algorithms and traditional penalized pseudo-likelihood methods applied directly to the full set of covariates.

Keywords: Cox model, generalized linear models, penalized likelihood estimation, sparsity, sure independence screening, variable selection.

1. Introduction

With the remarkable development of modern technology, including computing power and storage, more and more high-dimensional and high-throughput data of unprecedented size and complexity are being generated for contemporary statistical studies. For instance, bioimaging technology has made it possible to collect a huge amount of predictor information such as microarray, proteomic, and SNP data while observing survival information and tumor classification on patients in clinical studies. A common feature of all these examples is that the number of variables p can be potentially much larger than the number of observations n, i.e., the number of gene expression profiles is in the order of tens of thousands while the number of patient samples is in the order of tens or hundreds. Following Fan and Lv (2008), we call this

setting ultrahigh dimensional, where the authors gave a precise mathematical formulation of the growth rate of p relative to n. In order to provide more representative and reasonable statistical models, it is typically assumed that only a small fraction of predictors are associated with the outcome. This is the notion of sparsity which emphasizes the prominent role feature selection techniques play in ultrahigh dimensional statistical modeling.

One popular family of variable selection methods for parametric models is based on the penalized (pseudo-)likelihood approach. Examples include the LASSO (Tibshirani 1996, 1997), SCAD (Fan and Li 2001), the elastic net penalty (Zou and Hastie 2005), the MCP (Zhang 2010), and related methods. Nevertheless, in ultrahigh dimensional statistical learning problems, these methods may not perform well due to the simultaneous challenges of computational expediency, statistical accuracy, and algorithmic stability (Fan *et al.* 2009).

Motivated by these concerns, Fan and Lv (2008) introduced a new framework for variable screening via independent correlation learning that tackles the aforementioned challenges in the context of ultrahigh dimensional linear models. Their proposed sure independence screening (SIS) is a two-stage procedure; first filtering out the features that have weak marginal correlation with the response, effectively reducing the dimensionality p to a moderate scale below the sample size n, and then performing variable selection and parameter estimation simultaneously through a lower dimensional penalized least squares method such as SCAD or LASSO. Under certain regularity conditions, Fan and Lv (2008) showed surprisingly that this fast feature selection method has a "sure screening property"; that is, with probability tending to 1, the independence screening technique retains all of the important features in the model. However, the SIS procedure in Fan and Lv (2008) only covers ordinary linear regression models and their technical arguments do not extend easily to more general models such as generalized linear models and hazard regression with right-censored times.

In order to enhance finite sample performance, an important methodological extension, iterative sure independence screening (ISIS), was also proposed by Fan and Lv (2008) to handle cases where the regularity conditions may fail, such as when some important predictors are marginally uncorrelated with the response, or the reverse situation where an unimportant predictor has higher marginal correlation than some important features. Roughly speaking, the original ISIS procedure works by iteratively performing variable selection to recruit a small number of predictors, computing residuals based on the model fitted using these recruited predictors, and then using the residuals as the working response variable to continue recruiting new predictors. With the purpose of handling more complex real data, Fan and Song (2010) extended SIS to generalized linear models; and Fan *et al.* (2009) improved some important steps of the original ISIS procedure, allowing variable deletion in the recruiting process through penalized pseudo-likelihood, while dealing with more general loss based models. In particular, they introduced the concept of conditional marginal regressions and, with the aim of reducing the false discovery rate, proposed two new ISIS variants based on the idea of splitting samples. Other extensions of ISIS include Fan *et al.* (2010) to the Cox proportional hazards model, and Fan *et al.* (2011) to nonparametric additive models.

In this paper we build on the work of Fan *et al.* (2009) and Fan *et al.* (2010) to provide a publicly available package **SIS** (Fan *et al.* 2015), implemented in the R statistical software (R Core Team 2016), extending sure independence screening and all of its variants to generalized linear models and the Cox proportional hazards model. In particular, our codes are able to perform variable selection through the proposed ISIS variants of Fan *et al.* (2009) and through the data-driven thresholding approach of Fan *et al.* (2011). Furthermore, we combine these

sample splitting and data-driven thresholding ideas to provide two novel feature selection techniques.

Taking advantage of the fast cyclical coordinate descent algorithms developed in the packages **glmnet** (Friedman *et al.* 2013) and **ncvreg** (Breheny 2013), for convex and nonconvex penalty functions, respectively, we are able to efficiently perform the moderate scale penalized pseudo-likelihood steps from the ISIS procedure, thus yielding variable selection techniques outperforming direct use of **glmnet** and **ncvreg** in terms of both computational time and estimation error. Our procedures scale favorably in both n and p, allowing us to expeditiously and accurately solve much larger problems than with previous packages, particularly in the case of nonconvex penalties. We would like to point out that the recent package **apple** (Yu and Feng 2015), using a hybrid of the predictor-corrector method and coordinate descent procedures, provides an alternative for the penalized pseudo-likelihood estimation with nonconvex penalties (Yu and Feng 2014a). In the present work, we limit all numerical results to the use of **ncvreg**, noting there are other available options to implement the nonconvex variable selection procedures performed by **SIS**. Similarly, although the package **survHD** (Bernau *et al.* 2014) provides an efficient alternative for implementing Cox proportional hazards regression, in the current presentation, we only make use of the **survival** package (Therneau and Lumley 2015) to compute conditional marginal regressions and of the **glmnet** package (Friedman *et al.* 2013) to fit high-dimensional Cox models.

The remainder of the paper is organized as follows. In Section 2, we describe the vanilla SIS and ISIS variable selection procedures in the context of generalized linear models and the Cox proportional hazards model. Section 3 discusses several ISIS variants, as well as important implementation details. Simulation results comparing model selection performance and run time trials are given in Section 4, where we also analyze four gene expression data sets and work through an example using our package with one of them. The paper is concluded with a short discussion in Section 5.

2. General SIS and ISIS methodological framework

Consider the usual generalized linear model (GLM) framework, where we have independent and identically distributed observations $\{(\boldsymbol{x}_i, y_i) : i = 1, \ldots, n\}$ from the population $(\boldsymbol{x}, y)$, where the predictor $\boldsymbol{x} = (x_0, x_1, \ldots, x_p)^\top$ is a $(p+1)$-dimensional random vector with $x_0 = 1$ and y is the response. We further assume the conditional distribution of y given $\boldsymbol{x}$ is from an exponential family taking the canonical form

$$f(y; \boldsymbol{x}, \boldsymbol{\beta}) = \exp\{y\theta - b(\theta) + c(y)\}, \tag{1}$$

where $\theta = \boldsymbol{x}^\top \boldsymbol{\beta}$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$ is a vector of unknown regression parameters and $b(\cdot)$, $c(\cdot)$ are known functions. As we are only interested in modeling the mean regression, the dispersion parameter is assumed known. In virtue of (1), inference about the parameter $\boldsymbol{\beta}$ in the GLM context is made via maximization of the log-likelihood function

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \{y_i(\boldsymbol{x}_i^\top \boldsymbol{\beta}) - b(\boldsymbol{x}_i^\top \boldsymbol{\beta})\}. \tag{2}$$

For the survival analysis framework, the observed data $\{(\boldsymbol{x}_i, y_i, \delta_i) : \boldsymbol{x}_i \in \mathbb{R}^p, y_i \in \mathbb{R}^+, \delta_i \in \{0, 1\}, i = 1, \ldots n\}$ is an independent and identically distributed random sample from a certain

population $(\boldsymbol{x}, y, \delta)$. Here, as in the context of linear models, $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)^\top$ is a p-dimensional random vector of predictors and y, the observed time, is a time of failure if $\delta = 1$, or a right-censored time if $\delta = 0$. Suppose that the sample comprises m distinct uncensored failure times $t_1 < t_2 < \cdots < t_m$. Let (j) denote the individual failing at time t_j and $\mathcal{R}(t_j)$ be the risk set just prior to time t_j, that is, $\mathcal{R}(t_j) = \{i : y_i \geq t_j\}$. The main problem of interest is to study the relationship between the predictor variables and the failure time, and a common approach is through the Cox proportional hazards model (Cox 1975). For a vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^\top$ of unknown regression parameters, the Cox model assumes a semiparametric form of the hazard function

$$h(t|\boldsymbol{x}_i) = h_0(t)e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}},$$

where $h_0(t)$ is an unknown arbitrary baseline hazard function giving the hazard when $\boldsymbol{x}_i = 0$. Following the argument in Cox (1975), inference about $\boldsymbol{\beta}$ is made via maximization of the partial likelihood function

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{m} \frac{e^{\boldsymbol{x}_{(j)}^\top \boldsymbol{\beta}}}{\sum_{k \in \mathcal{R}(t_j)} e^{\boldsymbol{x}_k^\top \boldsymbol{\beta}}},$$

which is equivalent to maximizing the log-partial likelihood

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \boldsymbol{x}_i^\top \boldsymbol{\beta} - \sum_{i=1}^{n} \delta_i \log \left\{ \sum_{k \in \mathcal{R}(y_i)} \exp(\boldsymbol{x}_k^\top \boldsymbol{\beta}) \right\}. \tag{3}$$

We refer interested readers to Kalbfleisch and Prentice (2002) and references therein for a comprehensive literature review on the Cox proportional hazards model.

For both statistical models, we assume all predictors $x_1, \ldots, x_p$ are standardized to have mean zero and standard deviation one. Additionally, although our variable selection procedures within the **SIS** package also handle the classical $p < n$ setting, we allow the number of covariates p to be much larger than the number of observations n. What makes statistical inference possible in this "large p, small n" scenario is the sparsity assumption; only a small subset of variables among predictors $x_1, \ldots, x_p$ contribute to the response, which implies the parameter vector $\boldsymbol{\beta}$ is sparse. Therefore, variable selection techniques play a pivotal role in these ultrahigh dimensional statistical models.

2.1. SIS and feature ranking by maximum marginal likelihood estimators

Let $\mathcal{M}_\star = \{1 \leq j \leq p : \beta_j^\star \neq 0\}$ be the true sparse model, where $\boldsymbol{\beta}^\star = (\beta_0^\star, \beta_1^\star, \ldots, \beta_p^\star)^\top$ denotes the true value of the parameter vector and $\beta_0^\star = 0$ for the Cox model. In order to carry out the vanilla sure independence screening variable selection procedure, we initially fit marginal versions of models (2) and (3) with componentwise covariates. The maximum marginal likelihood estimator (MMLE) $\hat{\boldsymbol{\beta}}_j^M$, for $j = 1, \ldots, p$, is defined in the GLM context as the maximizer of the componentwise regression

$$\hat{\boldsymbol{\beta}}_j^M = (\hat{\beta}_{j,0}^M, \hat{\beta}_j^M) = \arg\max_{\beta_0, \beta_j} \sum_{i=1}^{n} \{y_i(\beta_0 + x_{ij}\beta_j) - b(\beta_0 + x_{ij}\beta_j)\}, \tag{4}$$

where $\boldsymbol{x}_i = (x_{i0}, x_{i1}, \ldots, x_{ip})^\top$ and $x_{i0} = 1$. Similarly, for each covariate x_j $(1 \leq j \leq p)$, one can define the MMLE for the Cox model as the maximizer of the log-partial likelihood with

a single covariate

$$\hat{\beta}_j^M = \arg\max_{\beta_j} \left(\sum_{i=1}^n \delta_i x_{ij} \beta_j - \sum_{i=1}^n \delta_i \log \left\{ \sum_{k \in \mathcal{R}(y_i)} \exp(x_{kj}\beta_j) \right\} \right), \tag{5}$$

with $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top$. Both componentwise estimators can be computed very rapidly and implemented modularly, avoiding the numerical instability associated with ultrahigh dimensional estimation problems.

The vanilla SIS procedure then ranks the importance of features according to the magnitude of their marginal regression coefficients, excluding the intercept in the case of GLM. Therefore, we select a set of variables

$$\widehat{\mathcal{M}}_{\delta_n} = \{1 \leq j \leq p : |\hat{\beta}_j^M| \geq \delta_n\}, \tag{6}$$

where δ_n is a threshold value chosen so that we pick the d top ranked covariates. Typically, one may take $d = \lfloor n/\log n \rfloor$, so that dimensionality is reduced from ultrahigh to below the sample size. As further discussed in Sections 3.3 and 3.4, the choice of d may also be either data-driven or model-based. Under a mild set of technical conditions, Fan and Song (2010) show the magnitude of these marginal estimators can preserve the nonsparsity information about the joint model with full covariates. In other words, for a given sequence $\{\delta_n\}$, the sure screening property

$$\mathsf{P}(\mathcal{M}_\star \subset \widehat{\mathcal{M}}_{\delta_n}) \to 1 \quad \text{as } n \to \infty \tag{7}$$

holds for SIS, effectively reducing the dimensionality of the model from ultrahigh to below the sample size, and solving the aforementioned challenges of computational expediency, statistical accuracy, and algorithmic stability.

With features being crudely selected by the intensity of their marginal signals, the index set $\widehat{\mathcal{M}}_{\delta_n}$ may also include a great deal of unimportant predictors. To further improve finite sample performance of vanilla SIS, variable selection and parameter estimation can be simultaneously achieved via penalized likelihood estimation, using the joint information of the covariates in $\widehat{\mathcal{M}}_{\delta_n}$. Without loss of generality, by reordering the features if necessary, we may assume that $x_1, \ldots, x_d$ are the predictors recruited by SIS. We define $\boldsymbol{\beta}_d = (\beta_0, \beta_1, \ldots, \beta_d)^\top$ and let $\boldsymbol{x}_{i,d} = (x_{i0}, x_{i1}, \ldots, x_{id})^\top$ with $x_{i0} = 1$. Thus, our original problem of estimating the sparse $(p+1)$-vector $\boldsymbol{\beta}$ in the GLM framework (2) reduces to estimating a sparse $(d+1)$-vector $\boldsymbol{\beta}_d = (\beta_0, \beta_1, \ldots, \beta_d)^\top$ based on maximizing the moderate scale penalized likelihood

$$\widehat{\boldsymbol{\beta}}_d = \arg\max_{\boldsymbol{\beta}_d} \sum_{i=1}^n \{y_i(\boldsymbol{x}_{i,d}^\top \boldsymbol{\beta}_d) - b(\boldsymbol{x}_{i,d}^\top \boldsymbol{\beta}_d)\} - \sum_{j=1}^d p_\lambda(|\beta_j|). \tag{8}$$

Likewise, after defining $\boldsymbol{\beta}_d = (\beta_1, \beta_2, \ldots, \beta_d)^\top$ and setting $\boldsymbol{x}_{i,d} = (x_{i1}, x_{i2}, \ldots, x_{id})^\top$ for survival data within the Cox model, the moderate scale version of the penalized log-partial likelihood problem consists in maximizing

$$\widehat{\boldsymbol{\beta}}_d = \arg\max_{\boldsymbol{\beta}_d} \sum_{i=1}^n \delta_i \boldsymbol{x}_{i,d}^\top \boldsymbol{\beta}_d - \sum_{i=1}^n \delta_i \log \left\{ \sum_{k \in \mathcal{R}(y_i)} \exp(\boldsymbol{x}_{k,d}^\top \boldsymbol{\beta}_d) \right\} - \sum_{j=1}^d p_\lambda(|\beta_j|). \tag{9}$$

Here, $p_\lambda(\cdot)$ is a penalty function and $\lambda > 0$ is a regularization parameter which may be selected using the AIC (Akaike 1973), BIC (Schwarz 1978) or EBIC (Chen and Chen 2008) criteria,

Algorithm 1 VANILLA SIS (VAN-SIS)

1. Inputs: Screening model size d. Penalty $p_\lambda(\cdot)$.
2. For every $j \in \{1, \ldots, p\}$, compute the MMLE $\hat{\beta}_j^M$ from (4) or (5).
3. Choose a threshold value δ_n in (6) such that $\widehat{\mathcal{M}}_{\delta_n}$ consists of the d top ranked covariates.
4. Obtain the parameter estimate $\widehat{\boldsymbol{\beta}}_d$ from the penalized likelihood estimation problems (8) or (9).
5. Outputs: Parameter estimate $\widehat{\boldsymbol{\beta}}_d$ and the corresponding estimate of the true sparse model $\widehat{\mathcal{M}}_1 = \text{supp}\{\widehat{\boldsymbol{\beta}}_d\}$.

or through ten-fold cross-validation and the modified cross-validation framework (Feng and Yu 2013; Yu and Feng 2014b), for example. Penalty functions whose resulting estimators yield sparse solutions to the maximization problems (8) and (9) include the LASSO penalty $p_\lambda(|\beta|) = \lambda|\beta|$ (Tibshirani 1996), the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li 2001), which is a folded-concave quadratic spline with $p_\lambda(0) = 0$ and

$$p'_\lambda(|\beta|) = \lambda\left\{\mathbb{1}_{\{|\beta|\le\lambda\}} + \frac{(\gamma\lambda - |\beta|)_+}{(\gamma-1)\lambda}\mathbb{1}_{\{|\beta|>\lambda\}}\right\},$$

for some $\gamma > 2$ and $|\beta| > 0$, and the minimax concave penalty (MCP), another folded-concave quadratic spline with $p_\lambda(0) = 0$ such that

$$p'_\lambda(|\beta|) = (\lambda - |\beta|/\gamma)_+,$$

for some $\gamma > 0$ and $|\beta| > 0$ (Zhang 2010). For the SCAD and MCP penalties, the tuning parameter γ is used to adjust the concavity of the penalty. The smaller γ is, the more concave the penalty becomes, which means finding a global minimizer is more difficult; but on the other hand, the resulting estimators overcome the bias introduced by the LASSO penalty.

Once penalized likelihood estimation is carried out in (8) and (9), a refined estimate of $\mathcal{M}_\star$ can be obtained from $\widehat{\mathcal{M}}_1$, the index set of the nonzero components of the sparse regression parameter estimator. We summarize this initial screening procedure based on componentwise regressions through Algorithm 1 above.

2.2. Iterative sure independence screening

The technical conditions in Fan and Lv (2008) and Fan and Song (2010) guaranteeing the sure screening property for SIS fail to hold if there is a predictor marginally unrelated, but jointly related with the response, or if a predictor is jointly uncorrelated with the response but has higher marginal correlation with the response than some important predictors in $\mathcal{M}_\star$. In the former case, the important predictor cannot be picked up by vanilla SIS, whereas in the latter case, unimportant predictors in $\mathcal{M}_\star^c$ are ranked too high, leaving out features from the true sparse model $\mathcal{M}_\star$.

To deal with these difficult scenarios in which the SIS methodology breaks down, Fan and Lv (2008) and Fan *et al.* (2009) proposed iterative sure independence screening based on conditional marginal regressions and feature ranking. The approach seeks to overcome the limitations of SIS, which is based on marginal models only, by making more use of the joint covariate information while retaining computational expediency and stability as in the original SIS.

In its first iteration, the vanilla ISIS variable selection procedure begins with using regular SIS to pick an index set $\widehat{\mathcal{A}}_1$ of size k_1, and then similarly applies a penalized likelihood estimation approach to select a subset $\widehat{\mathcal{M}}_1$ of these indices. Note that the screening step only fits componentwise regressions, while the penalized likelihood step solves optimization problems of moderate scale k_1, typically below the sample size n. This is an attractive feature for any variable selection technique applied to ultrahigh dimensional statistical models. After the first iteration, we denote the resulting estimator with nonzero components and indices in $\widehat{\mathcal{M}}_1$ by $\hat{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_1}$.

In an effort to use more fully the joint covariate information, in the second iteration of vanilla ISIS we compute the conditional marginal regression of each predictor j that is not in $\widehat{\mathcal{M}}_1$. That is, under the generalized linear model framework, we solve

$$\underset{\beta_0,\boldsymbol{\beta}_{\widehat{\mathcal{M}}_1},\beta_j}{\arg\max} \sum_{i=1}^{n}\{y_i(\beta_0 + \boldsymbol{x}_{i,\widehat{\mathcal{M}}_1}^\top \boldsymbol{\beta}_{\widehat{\mathcal{M}}_1} + x_{ij}\beta_j) - b(\beta_0 + \boldsymbol{x}_{i,\widehat{\mathcal{M}}_1}^\top \boldsymbol{\beta}_{\widehat{\mathcal{M}}_1} + x_{ij}\beta_j)\}, \tag{10}$$

whereas under the Cox model, we obtain

$$\underset{\boldsymbol{\beta}_{\widehat{\mathcal{M}}_1},\beta_j}{\arg\max} \left(\sum_{i=1}^{n} \delta_i (\boldsymbol{x}_{i,\widehat{\mathcal{M}}_1}^\top \boldsymbol{\beta}_{\widehat{\mathcal{M}}_1} + x_{ij}\beta_j) - \sum_{i=1}^{n} \delta_i \log \left\{ \sum_{k \in \mathcal{R}(y_i)} \exp(\boldsymbol{x}_{k,\widehat{\mathcal{M}}_1}^\top \boldsymbol{\beta}_{\widehat{\mathcal{M}}_1} + x_{kj}\beta_j) \right\} \right) \tag{11}$$

for each $j \in \{1, \ldots, p\} \setminus \widehat{\mathcal{M}}_1$, where $\boldsymbol{x}_{i,\widehat{\mathcal{M}}_1}$ denotes the sub-vector of $\boldsymbol{x}_i$ with indices in $\widehat{\mathcal{M}}_1$ and similarly for $\boldsymbol{\beta}_{\widehat{\mathcal{M}}_1}$. These are again low-dimensional problems which can be solved quite efficiently. Similar to the componentwise regressions (4) and (5), let $\hat{\beta}_j^M$ denote the last coordinate of the maximizer in (10) and (11). The magnitude $|\hat{\beta}_j^M|$ reflects the additional contribution of the jth predictor given that all covariates with indices in $\widehat{\mathcal{M}}_1$ have been included in the model.

Once the conditional marginal regressions have been computed for each predictor not in $\widehat{\mathcal{M}}_1$, we perform conditional feature ranking by ordering $\{|\hat{\beta}_j^M| : j \in \widehat{\mathcal{M}}_1^c\}$ and forming an index set $\widehat{\mathcal{A}}_2$ of size k_2, say, consisting of the indices with the top ranked elements. After this screening step, under the GLM framework, we maximize the moderate scale penalized likelihood

$$\sum_{i=1}^{n}\{y_i(\beta_0 + \boldsymbol{x}_{i,\widehat{\mathcal{M}}_1\cup\widehat{\mathcal{A}}_2}^\top \boldsymbol{\beta}_{\widehat{\mathcal{M}}_1\cup\widehat{\mathcal{A}}_2}) - b(\beta_0 + \boldsymbol{x}_{i,\widehat{\mathcal{M}}_1\cup\widehat{\mathcal{A}}_2}^\top \boldsymbol{\beta}_{\widehat{\mathcal{M}}_1\cup\widehat{\mathcal{A}}_2})\} - \sum_{j\in\widehat{\mathcal{M}}_1\cup\widehat{\mathcal{A}}_2} p_\lambda(|\beta_j|) \tag{12}$$

with respect to $\boldsymbol{\beta}_{\widehat{\mathcal{M}}_1\cup\widehat{\mathcal{A}}_2}$ to get a sparse estimator $\hat{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_1\cup\widehat{\mathcal{A}}_2}$. Similarly, in the Cox model, we obtain a sparse estimator by maximizing the moderate scale penalized log-partial likelihood

$$\begin{aligned} &\sum_{i=1}^{n} \delta_i (\boldsymbol{x}_{i,\widehat{\mathcal{M}}_1\cup\widehat{\mathcal{A}}_2}^\top \boldsymbol{\beta}_{\widehat{\mathcal{M}}_1\cup\widehat{\mathcal{A}}_2}) - \sum_{i=1}^{n} \delta_i \log \left\{ \sum_{k \in \mathcal{R}(y_i)} \exp(\boldsymbol{x}_{k,\widehat{\mathcal{M}}_1\cup\widehat{\mathcal{A}}_2}^\top \boldsymbol{\beta}_{\widehat{\mathcal{M}}_1\cup\widehat{\mathcal{A}}_2}) \right\} \\ &\quad - \sum_{j\in\widehat{\mathcal{M}}_1\cup\widehat{\mathcal{A}}_2} p_\lambda(|\beta_j|). \end{aligned} \tag{13}$$

The indices of the nonzero coefficients of $\hat{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_1\cup\widehat{\mathcal{A}}_2}$ provide an updated estimate $\widehat{\mathcal{M}}_2$ of the true sparse model $\mathcal{M}_\star$.

Algorithm 2 VANILLA ISIS (VAN-ISIS)

1. Inputs: Screening model size d. Penalty $p_\lambda(\cdot)$. Maximum iteration number $l_{\max}$.
2. For every $j \in \{1, \ldots, p\}$, compute the MMLE $\hat{\beta}_j^M$ from problems (4) or (5). Select the k_1 top ranked covariates to form the index set $\widehat{\mathcal{A}}_1$.
3. Apply penalized likelihood estimation on the set $\widehat{\mathcal{A}}_1$ to obtain a subset of indices $\widehat{\mathcal{M}}_1$.
4. For every $j \in \widehat{\mathcal{M}}_1^c$, solve the conditional marginal regression problem (10) or (11) and sort $\{|\hat{\beta}_j^M| : j \in \widehat{\mathcal{M}}_1^c\}$. Form the index set $\widehat{\mathcal{A}}_2$ with the k_2 top ranked covariates, and apply penalized likelihood estimation on $\widehat{\mathcal{M}}_1 \cup \widehat{\mathcal{A}}_2$ as in (12) or (13) to obtain a new index set $\widehat{\mathcal{M}}_2$.
5. Iterate the process in step 4 until we have an index set $\widehat{\mathcal{M}}_l$ such that $|\widehat{\mathcal{M}}_l| = d$ or $\widehat{\mathcal{M}}_l = \widehat{\mathcal{M}}_j$ for some $j < l$ or $l = l_{\max}$.
6. Outputs: $\widehat{\mathcal{M}}_l$ from step 5 along with the parameter estimate from (12) or (13).

In the screening step above, an alternative approach is to substitute the fitted regression parameter $\hat{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_1}$ from the first step of vanilla ISIS into problems (10) and (11), so that the optimization problems become again componentwise regressions. This approach is exactly an extension of the original ISIS proposal of Fan and Lv (2008) to generalized linear models and the Cox proportional hazards model. Even for the ordinary linear model, the conditional contributions of predictors are more relevant for variable selection in the second ISIS iteration than the original proposal of using the residuals $\hat{r}_i = y_i - \boldsymbol{x}_{i,\widehat{\mathcal{M}}_1}^\top \hat{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_1}$ as the new response (see Fan *et al.* 2009). Furthermore, the penalized likelihood steps (12) and (13) allow the procedure to delete some features $\{x_j : j \in \widehat{\mathcal{M}}_1\}$ that were previously selected. This is also an important deviation from Fan and Lv (2008), as their original ISIS procedure does not contemplate intermediate deletion steps.

Lastly, the vanilla ISIS procedure, which iteratively recruits and deletes predictors, can be repeated until some convergence criterion is reached. We adopt the criterion of having an index set $\widehat{\mathcal{M}}_l$ which either has the prescribed size d, or satisfies $\widehat{\mathcal{M}}_l = \widehat{\mathcal{M}}_j$ for some $j < l$. In order to ensure that iterated SIS takes at least two iterations to terminate, in our implementation we fix $k_1 = \lfloor 2d/3 \rfloor$, and thereafter at the lth iteration we set $k_l = d - |\widehat{\mathcal{M}}_{l-1}|$. A step-by-step description of the vanilla ISIS procedure is provided in Algorithm 2.

We conclude this section providing a simple overview of the main features of the vanilla SIS and ISIS procedures for applied practitioners. In the ultrahigh dimensional statistical model setting where $p \gg n$, and even in the classical $p < n$ setting with $p > 30$, variable screening is an essential tool in helping eliminate irrelevant predictors while reducing data gathering and storage requirements. The vanilla SIS procedure given in Algorithm 1 provides an extremely fast and efficient variable screening based on marginal regressions of each predictor with the response. While under certain independence assumptions among predictors this may prove a successful strategy in terms of estimating the true sparse model $\mathcal{M}_\star$, there are well-known issues associated with variable screening using only information from marginal regressions, such as missing important predictors from $\mathcal{M}_\star$ which happen to have low marginal correlation with the response. The vanilla ISIS procedure addresses these issues by using more thoroughly the joint covariate information through the conditional marginal regressions (10) and (11), which aim at measuring the additional contribution of a predictor x_j given the presence of the variables in $\widehat{\mathcal{M}}_1$ in the current model, all while maintaining low computational costs.

Finally, we would like to point out that the intermediate deletion steps of the vanilla ISIS procedure could be carried out with any other variable selection methods, such as the weight vector ranking with support vector machines (Rakotomamonjy 2003) or the greedy search strategies of forward selection and backward elimination (Guyon and Elisseeff 2003). In our implementation within the **SIS** package we favor the penalized likelihood criteria (12) and (13), but in principle any variable selection technique could be employed to further filter unimportant predictors.

3. Variants of ISIS

By nature of their marginal approach, sure independence screening procedures have large false selection rates, namely, many unimportant predictors in $\mathcal{M}_\star^c$ are selected after the screening steps. In order to reduce the false selection rate, Fan *et al.* (2009) suggested the idea of sample splitting. Without loss of generality, we assume the sample size n is even, and we randomly split the sample into two halves. Two variants of the ISIS methodology have been proposed in the literature; both of them relying on the idea of performing variable screening to the data in each partition separately, combining the results in a subsequent penalized likelihood step. We also revisit the approach of Fan *et al.* (2011), in which a random permutation of the observations is used to obtain a data-driven threshold for independence screening.

3.1. First variant of ISIS

Let $\widehat{\mathcal{A}}_1^{(1)}$ and $\widehat{\mathcal{A}}_1^{(2)}$ be the two sets of indices, each of size k_1, obtained after applying regular SIS to the data in each partition. As the first crude estimates of the true sparse model $\mathcal{M}_\star$, both of them should have large false selection rates. Yet, under the technical conditions given in Fan and Song (2010), the estimates should satisfy

$$\mathsf{P}(\mathcal{M}_\star \subset \widehat{\mathcal{A}}_1^{(j)}) \to 1 \quad \text{as } n \to \infty$$

for $j = 1, 2$. That is, important features should appear in both sets with probability tending to one. If we define $\widehat{\mathcal{A}}_1 = \widehat{\mathcal{A}}_1^{(1)} \cap \widehat{\mathcal{A}}_1^{(2)}$ as a new estimate of $\mathcal{M}_\star$, this new index set must also satisfy

$$\mathsf{P}(\mathcal{M}_\star \subset \widehat{\mathcal{A}}_1) \to 1 \quad \text{as } n \to \infty.$$

However, by construction, the number of unimportant predictors in $\widehat{\mathcal{A}}_1$ should be much smaller, thus reducing the false selection rate. The reason is that, in order for an unimportant predictor to appear in $\widehat{\mathcal{A}}_1$, it has to be included in both sets $\widehat{\mathcal{A}}_1^{(1)}$ and $\widehat{\mathcal{A}}_1^{(2)}$ randomly. In their theoretical support for this variant based on random splitting, Fan *et al.* (2009) provided a non-asymptotic upper bound for the probability of the event that m unimportant covariates are included in the intersection $\widehat{\mathcal{A}}_1$. The probability bound, obtained under an exchangeability condition ensuring that each unimportant feature is equally likely to be recruited by SIS, is decreasing in the dimensionality, showing an apparent "blessing of dimensionality". This is only part of the full story, since, as pointed out in Fan *et al.* (2009), the probability of missing important predictors from the true sparse model $\mathcal{M}_\star$ is expected to increase with p. However, as we show in our simulation settings of Section 4.1, and further confirm in the real data analysis of Section 4.3, the procedure is quite effective at obtaining a minimal set of features that should be included in a final model.

The remainder of the first variant of ISIS proceeds as follows. After forming the intersection $\widehat{\mathcal{A}}_1 = \widehat{\mathcal{A}}_1^{(1)} \cap \widehat{\mathcal{A}}_1^{(2)}$, we perform penalized likelihood estimation as in Algorithm 2 to obtain a refined approximation $\widehat{\mathcal{M}}_1$ to the true sparse model. We then perform the second iteration of the vanilla ISIS procedure, computing conditional marginal regressions to each partition separately to obtain sets of indices $\widehat{\mathcal{A}}_2^{(1)}$ and $\widehat{\mathcal{A}}_2^{(2)}$, each of size k_2. After taking the intersection $\widehat{\mathcal{A}}_2 = \widehat{\mathcal{A}}_2^{(1)} \cap \widehat{\mathcal{A}}_2^{(2)}$ of these two sets, we carry out sparse penalized likelihood estimation as in (12) and (13), obtaining a second approximation $\widehat{\mathcal{M}}_2$ to the true model $\mathcal{M}_\star$. As before, the iteration continues until we have an index set $\widehat{\mathcal{M}}_l$ of size d, or satisfiying $\widehat{\mathcal{M}}_l = \widehat{\mathcal{M}}_j$ for some $j < l$.

3.2. Second variant of ISIS

The variable selection performed by the first variant of ISIS could potentially lead to a very aggressive screening of predictors, reducing the overall false selection rate, but sometimes yielding undesirably minimal model sizes. The second variant of ISIS is a more conservative variable selection procedure, where we again apply regular SIS to each partition separately, but we now choose larger sets of indices $\widetilde{\mathcal{A}}_1^{(1)}$ and $\widetilde{\mathcal{A}}_1^{(2)}$ to ensure that their intersection $\widetilde{\mathcal{A}}_1 = \widetilde{\mathcal{A}}_1^{(1)} \cap \widetilde{\mathcal{A}}_1^{(2)}$ has k_1 elements. In this way, the second variant guarantees that there are at least k_1 predictors included before the penalized likelihood step, making it considerably less aggressive than the first variant.

After applying penalized likelihood to the predictors with indices in $\widetilde{\mathcal{A}}_1$, we obtain a first estimate $\widetilde{\mathcal{M}}_1$ of the true sparse model. The second iteration computes conditional marginal regressions to each partition separately, recruiting enough features in index sets $\widetilde{\mathcal{A}}_2^{(1)}$ and $\widetilde{\mathcal{A}}_2^{(2)}$ to ensure that $\widetilde{\mathcal{A}}_2 = \widetilde{\mathcal{A}}_2^{(1)} \cap \widetilde{\mathcal{A}}_2^{(2)}$ has k_2 elements. Penalized likelihood, as outlined in Section 2.2, is now applied to $\widetilde{\mathcal{M}}_1 \cup \widetilde{\mathcal{A}}_2$, yielding a second estimate $\widetilde{\mathcal{M}}_2$ of the true model $\mathcal{M}_\star$. Stopping criteria remain the same as in the first variant.

3.3. Data-driven thresholding

Motivated by a false discovery rate criterion in Fan *et al.* (2011), the following variant of ISIS determines a data-driven threshold for marginal screening. Given GLM data of the form $\{(\boldsymbol{x}_i, y_i) : i = 1, \ldots, n\}$, a random permutation π of $\{1, \ldots, n\}$ is used to decouple $\boldsymbol{x}_i$ and y_i so that the resulting data $\{(\boldsymbol{x}_{\pi(i)}, y_i) : i = 1, \ldots, n\}$ follow a null model, i.e., a model in which the features have no prediction power over the response. For the newly permuted data, we recalculate marginal regression coefficients $(\hat{\beta}_j^M)^*$ for each predictor j, with $j = 1, \ldots, p$.

The motivation behind this approach is that whenever j is the index of an important predictor in $\mathcal{M}_\star$, the MMLE $|\hat{\beta}_j^M|$ given in (4) should be larger than most of $|(\hat{\beta}_j^M)^*|$, as the random permutation is meant to eliminate the prediction power of features. For a given $q \in [0, 1]$, let $\omega_{(q)}$ be the qth quantile of $\{|(\hat{\beta}_j^M)^*| : j = 1, \ldots, p\}$. The data-driven threshold allows only a $1 - q$ proportion of inactive variables to enter the model when $\boldsymbol{x}$ and y are not related (the null model). Thus, the initial index set $\widehat{\mathcal{A}}_1$ is defined as

$$\widehat{\mathcal{A}}_1 = \{1 \leq j \leq p : |\hat{\beta}_j^M| \geq \omega_{(q)}\},$$

and the modified ISIS iteration then carries out penalized likelihood estimation in the usual way to obtain a finer approximation $\widehat{\mathcal{M}}_1$ of the true sparse model $\mathcal{M}_\star$. The complete procedure for this variant is detailed in Algorithm 3 above.

Algorithm 3 Permutation-based ISIS (Perm-ISIS)

1. Inputs: Screening model size d. Penalty $p_\lambda(\cdot)$. Quantile q. Maximum iteration number $l_{\max}$.
2. For every $j \in \{1, \ldots, p\}$, compute the MMLE $\hat{\beta}_j^M$. Obtain the randomly permuted data $\{(\boldsymbol{x}_{\pi(i)}, y_i) : i = 1, \ldots, n\}$, and let $\omega_{(q)}$ be the qth quantile of $\{|(\hat{\beta}_j^M)^*| : j = 1, \ldots, p\}$, where $(\hat{\beta}_j^M)^*$ is the second coordinate of the solution to

$$\arg\max_{\beta_0, \beta_j} \sum_{i=1}^{n} \{y_i(\beta_0 + x_{\pi(i)j}\beta_j) - b(\beta_0 + x_{\pi(i)j}\beta_j)\}.$$

   Select the variables in the index set $\widehat{\mathcal{A}}_1 = \{1 \leq j \leq p : |\hat{\beta}_j^M| \geq \omega_{(q)}\}$.
3. Apply penalized likelihood estimation on the set $\widehat{\mathcal{A}}_1$ to obtain a subset of indices $\widehat{\mathcal{M}}_1$.
4. For every $j \in \widehat{\mathcal{M}}_1^c$, solve the conditional marginal regression problem (10) and obtain $\{|\hat{\beta}_j^M| : j \in \widehat{\mathcal{M}}_1^c\}$. By randomly permuting only the variables in $\widehat{\mathcal{M}}_1^c$, let $\omega_{(q)}$ be the qth quantile of $\{|(\hat{\beta}_j^M)^*| : j \in \widehat{\mathcal{M}}_1^c\}$, where $(\hat{\beta}_j^M)^*$ is the last coordinate of the solution to

$$\arg\max_{\beta_0, \boldsymbol{\beta}_{\widehat{\mathcal{M}}_1}, \beta_j} \sum_{i=1}^{n} \{y_i(\beta_0 + \boldsymbol{x}_{i,\widehat{\mathcal{M}}_1}^\top \boldsymbol{\beta}_{\widehat{\mathcal{M}}_1} + x_{\pi(i)j}\beta_j) - b(\beta_0 + \boldsymbol{x}_{i,\widehat{\mathcal{M}}_1}^\top \boldsymbol{\beta}_{\widehat{\mathcal{M}}_1} + x_{\pi(i)j}\beta_j)\}.$$

   Select the variables in the index set $\widehat{\mathcal{A}}_2 = \{j \in \widehat{\mathcal{M}}_1^c : |\hat{\beta}_j^M| \geq \omega_{(q)}\}$, and apply penalized likelihood estimation on $\widehat{\mathcal{M}}_1 \cup \widehat{\mathcal{A}}_2$ to obtain a new subset $\widehat{\mathcal{M}}_2$.
5. Iterate the process in step 4 until we have an index set $\widehat{\mathcal{M}}_l$ such that $|\widehat{\mathcal{M}}_l| \geq d$ or $\widehat{\mathcal{M}}_l = \widehat{\mathcal{M}}_j$ for some $j < l$ or $l = l_{\max}$.
6. Outputs: $\widehat{\mathcal{M}}_l$ from step 5 along with the parameter estimate from (12) or (13).

A greedy modification of the above algorithm can be proposed to enhance variable selection performance. Specifically, we restrict the size of the sets $\widehat{\mathcal{A}}_j$ in the iterative screening steps to be at most p_0, a small positive integer. Moreover, a completely analogous algorithm can be proposed for survival data, with permutation π and data-driven threshold $\omega_{(q)}$ defined accordingly. Details of such procedure are omitted in the current presentation.

3.4. Implementation details

There are several important details in the implementation of the vanilla versions of SIS and ISIS, as well as all of the above mentioned variants.

- In order to speed up computations, and exclusively for the first screening step, all variable selection procedures use correlation learning (i.e., marginal Pearson correlations) between each predictor and the response, instead of the componentwise GLM or partial likelihood fits (4) and (5). We found no major differences in variable selection performance between this variant and the one using the MMLEs.

- Although the asymptotic theory of Fan and Song (2010) guarantees the sure screening property (7) for a sequence $\delta_n \sim n^{-\theta_*}$, with $\theta_* > 0$ a fixed constant, in practice Fan *et al.*

(2009) recommend using $d = \lfloor n/\log n \rfloor$ as a sensible choice since θ_* is typically unknown. Furthermore, Fan *et al.* (2009) also advocate using model-based choices of d, favoring smaller values in models where the response provides less information. In our numerical implementation we use $d = \lfloor n/(4\log n) \rfloor$ for logistic regression and $d = \lfloor n/(2\log n) \rfloor$ for Poisson regression, which are less informative than the real-valued response in a linear model, for which we select $d = \lfloor n/\log n \rfloor$. For the right-censored response in the survival analysis framework, we fix $d = \lfloor n/(4\log n) \rfloor$.

- Regardless of the statistical model at hand, we set $d = \lfloor n/\log n \rfloor$ for the first variant of ISIS. Note that since the selected variables for this first variant are in the intersection of two sets of size $k_l \leq d$, we typically end up with far fewer than d variables selected by this method. In any case, our procedures within the **SIS** package allow for a user-specified value of d.

- Variable selection under the traditional $p < n$ setting can also be carried out using our screening procedures, for which we fix $d = p$ as default for all variants. In this regard, practicing data analysts can view sure independence screening procedures along classical criteria for variable selection such as the AIC (Akaike 1973), BIC (Schwarz 1978), C_p (Mallows 1973) or the generalized information criterion GIC (Nishii 1984) applied directly to the full set of covariates.

- The intermediate penalized likelihood problems (12) and (13) are solved using the **glmnet** and **ncvreg** packages. Our code has an option allowing the regularization parameter $\lambda > 0$ to be selected through the AIC, BIC, EBIC or K-fold cross-validation criteria. The concavity parameter γ in the SCAD and MCP penalties can also be user-specified.

- In our permutation-based variant with data-driven thresholding, we use $q = 1$ (i.e., we take $\omega_{(q)}$ to be the maximum absolute value of the permuted estimates) and take p_0 to be 1 in the greedy modification. Note that if none of the variables is recruited, that is, exceeding the threshold for the null model, the criterion in step 5 stops the procedure.

- We can further combine the permutation-based approach of Section 3.3 with the sample splitting idea from the first two variants to define a new ISIS procedure. Concretely, we first select two subsets of indices $\widehat{\mathcal{A}}_1^{(1)}$ and $\widehat{\mathcal{A}}_1^{(2)}$, each consisting of variables whose MMLEs, or correlation with the response, exceed the data-driven thresholds $\omega_{(q)}^{(1)}$ and $\omega_{(q)}^{(2)}$ of their respective samples. If the size of their intersection is less than k_1, we define $\widehat{\mathcal{A}}_1 = \widehat{\mathcal{A}}_1^{(1)} \cap \widehat{\mathcal{A}}_1^{(2)}$; otherwise, we reduce the size of $\widehat{\mathcal{A}}_1^{(1)}$ and $\widehat{\mathcal{A}}_1^{(2)}$ to ensure their intersection has at most k_1 elements. This is done to control the size of $\widehat{\mathcal{A}}_1$ when too many variables exceed the thresholds. The rest of the iteration is carried out accordingly.

- Every variant of ISIS is coded to guarantee there will be at least two predictors at the end of the first screening step.

As the proposed ISIS variants grow more involved, the associated number of tuning parameters is bound to increase. While this may initially make some data practitioners feel uneasy, our intent here is to be as flexible as possible, providing all available tools that the powerful family of sure independence screening procedures has to offer. Additionally, it is important to clarify that there exist tuning parameters inherent to the screening procedures, which

Parameter	Description	**SIS** package options
$p_\lambda(\cdot)$	Penalty function for intermediate penalized likelihood estimation	`penalty = "SCAD"` (default) / `= "MCP"` /= `"lasso"`
λ	Method for tuning regularization parameter of penalty function $p_\lambda(\cdot)$	`tune = "bic"` (default) / `= "ebic"` /= `"aic"` /= `"cv"`
γ	Concavity parameter for SCAD and MCP penalties	`concavity.parameter = 3.7` /= `3` are defaults for SCAD/MCP. Any $\gamma > 2$ for SCAD or $\gamma > 1$ for MCP can be user-specified
d	Upper bound on the number of predictors to be selected	`nsis` default is model-based as explained in Section 3.4. It can also be user-specified.
ISIS variant	Flags which ISIS version to perform	`varISIS = "vanilla"` (default) / `= "aggr"` /= `"cons"`
Permuation variant	Flags whether to use permutation variant with data-driven thresholds	`perm = FALSE` (default) /= `TRUE`
q	Quantile used in calculating data-driven thresholds	`q = 1` (default) / can be any user-specified value in $[0, 1]$
p_0	Maximum size of active sets in greedy modification	`greedy.size = 1` (default) / can be any user-specified integer less than p

Table 1: Summary of tuning parameters for variable selection using ISIS procedures within the **SIS** package, as well as associated defaults. All ISIS variants are implemented through the `SIS` function, which we describe in Section 4.4 using a gene expression data set.

are fundamentally different from tuning parameters (e.g., driven by a K-fold cross-validation approach) needed in the intermediate penalized likelihood procedures. In any case, we detail all available options implemented in our **SIS** package in Table 1 above, where we highlight recommended default settings for practicing researchers.

4. Model selection and timings

In this section we illustrate all independence screening procedures by studying their performance on simulated data and on four popular gene expression data sets. Most of the simulation settings are adapted from the work of Fan *et al.* (2009) and Fan *et al.* (2010).

4.1. Model selection and statistical accuracy

We first conduct simulation studies comparing the runtime of the vanilla version of SIS (Van-SIS), its iterated vanilla version (Van-ISIS), the first variant (Var1-ISIS), the second variant (Var2-ISIS), the permutation-based ISIS (Perm-ISIS), its greedy modification (Perm-g-ISIS), the permutation-based variant with sample splitting (Perm-var-ISIS) and its greedy modification (Perm-var-g-ISIS), under both generalized linear models and the Cox proportional hazards model. We also demonstrate the power of ISIS and its variants, in terms of model selection and estimation, by comparing them with traditional LASSO and SCAD penalized

estimation. Our **SIS** code is tested against the competing R packages **glmnet** (Friedman *et al.* 2013) and **ncvreg** (Breheny 2013) for LASSO and SCAD penalization, respectively. All calculations were carried out on an Intel Xeon L5430 @ 2.66 GHz processor. We refer interested readers to Friedman *et al.* (2010), Breheny and Huang (2011), and Simon *et al.* (2011) for a detailed discussion in penalized likelihood estimation algorithms under generalized linear models and the Cox proportional hazards model.

Four different statistical models are considered here: Linear regression, Logistic regression, Poisson regression, and Cox proportional hazards regression. We select the configuration $(n, p) = (400, 5000)$ for all models, and generate covariates $x_1, \ldots, x_p$ as follows:

Case 1: $x_1, \ldots, x_p$ are independent and identically distributed $N(0, 1)$ random variables.

Case 2: $x_1, \ldots, x_p$ are jointly Gaussian, marginally distributed as $N(0, 1)$, and with correlation structure $\mathrm{Corr}(x_i, x_j) = 1/2$ if $i \neq j$.

Case 3: $x_1, \ldots, x_p$ are jointly Gaussian, marginally distributed as $N(0, 1)$, and with correlation structure $\mathrm{Corr}(x_i, x_4) = 1/\sqrt{2}$ for all $i \neq 4$ and $\mathrm{Corr}(x_i, x_j) = 1/2$ if i and j are distinct elements of $\{1, \ldots, p\} \setminus \{4\}$.

Case 4: $x_1, \ldots, x_p$ are jointly Gaussian, marginally distributed as $N(0, 1)$, and with correlation structure $\mathrm{Corr}(x_i, x_5) = 0$ for all $i \neq 5$, $\mathrm{Corr}(x_i, x_4) = 1/\sqrt{2}$ for all $i \notin \{4, 5\}$, and $\mathrm{Corr}(x_i, x_j) = 1/2$ if i and j are distinct elements of $\{1, \ldots, p\} \setminus \{4, 5\}$.

With independent features, Case 1 is the most straightforward for variable selection. In Cases 2–4, however, we have serial correlation among predictors such that $\mathrm{Corr}(x_i, x_j)$ does not decay as $|i - j|$ increases. We will see below that for Cases 3 and 4, the true sparse model $\mathcal{M}_\star$ is chosen such that the response is marginally independent but jointly dependent of x_4. This type of dependence is ruled out in the asymptotic theory of SIS in Fan and Song (2010), so we should expect variable selection in these settings to be more challenging for the non-iterated version of SIS.

In the Cox proportional hazards scenario, the right-censoring time is generated from the exponential distribution with mean 10. This corresponds to fixing the baseline hazard function $h_0(t) = 0.1$ for all $t \geq 0$. The true regression coefficients from the sparse model $\mathcal{M}_\star$ in each of the four settings are as follows:

Case 1: $\beta_0^\star = 0$, $\beta_1^\star = -1.5140$, $\beta_2^\star = 1.2799$, $\beta_3^\star = -1.5307$, $\beta_4^\star = 1.5164$, $\beta_5^\star = -1.3019$, $\beta_6^\star = 1.5833$, and $\beta_j^\star = 0$ for $j > 6$.

Case 2: The coefficients are the same as in Case 1.

Case 3: $\beta_0^\star = 0$, $\beta_1^\star = 0.6$, $\beta_2^\star = 0.6$, $\beta_3^\star = 0.6$, $\beta_4^\star = -0.9\sqrt{2}$, and $\beta_j^\star = 0$ for $j > 4$.

Case 4: $\beta_1^\star = 4$, $\beta_2^\star = 4$, $\beta_3^\star = 4$, $\beta_4^\star = -6\sqrt{2}$, $\beta_5^\star = 4/3$, and $\beta_j^\star = 0$ for $j > 5$. The corresponding median censoring rate is 33.5%.

For Cases 1 and 2, the coefficients were randomly generated as $(4 \log n/\sqrt{n} + |Z|/4)U$ with $Z \sim N(0, 1)$ and $U = 1$ with probability 0.5 and -1 with probability 0.5, independent of the value of Z. For Cases 3 and 4, the selected model ensures that even though $\beta_4^\star \neq 0$, the

Method	$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_1$	$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2^2$	TP	FP	Time
Van-SIS	0.24(0.10)	0.01(0.01)	6(0.00)	0(0.00)	0.26(0.02)
Van-ISIS	0.24(0.09)	0.01(0.01)	6(0.00)	0(0.00)	8.34(0.78)
Var1-ISIS	0.29(0.15)	0.02(0.02)	6(0.00)	0(0.74)	11.76(8.65)
Var2-ISIS	0.24(0.10)	0.01(0.01)	6(0.00)	0(0.00)	11.90(1.12)
Perm-ISIS	0.41(0.25)	0.05(0.05)	6(0.00)	1(1.49)	44.57(13.38)
Perm-g-ISIS	0.39(0.27)	0.04(0.05)	6(0.00)	1(1.49)	107.50(22.99)
Perm-var-ISIS	0.24(0.09)	0.01(0.01)	6(0.00)	0(0.00)	41.64(1.01)
Perm-var-g-ISIS	0.24(0.09)	0.01(0.01)	6(0.00)	0(0.00)	82.81(15.80)
SCAD	0.24(0.09)	0.01(0.01)	6(0.00)	0(0.00)	6.89(5.56)

Table 2: Linear regression, Case 1, where results are given in the form of medians and robust standard deviations (in parentheses).

Method	$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_1$	$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2^2$	TP	FP	Time
Van-SIS	2.79(2.20)	2.02(2.55)	5.5(0.75)	0(0.75)	0.36(0.05)
Van-ISIS	4.06(7.78)	2.77(7.35)	6(0.00)	2(7.46)	48.73(11.72)
Var1-ISIS	1.86(1.25)	0.79(1.07)	6(0.00)	0(0.75)	76.47(20.56)
Var2-ISIS	5.29(7.99)	4.43(7.75)	6(0.00)	5(6.72)	96.25(59.99)
Perm-ISIS	2.94(8.80)	1.87(9.03)	6(0.00)	2(7.46)	128.94(42.74)
Perm-g-ISIS	18.65(6.10)	23.38(14.40)	6(0.00)	10(0.93)	739.84(89.96)
Perm-var-ISIS	1.55(1.38)	0.53(1.43)	6(0.75)	0(0.00)	153.53(43.84)
Perm-var-g-ISIS	1.64(1.37)	0.63(1.18)	6(0.00)	0(0.75)	251.21(63.66)
SCAD	409.22(104.33)	8403.18(4844.63)	6(0.00)	20(2.24)	304.98(66.65)

Table 3: Logistic regression, Case 2, where results are given in the form of medians and robust standard deviations (in parentheses).

associated predictor x_4 and the response y are marginally independent. This is designed in order to make it challenging for the vanilla sure independence screening procedure to select this important variable. Furthermore, in Case 4, we add another important predictor x_5 with a small coefficient to make it even more challenging to identify the true sparse model.

The results are given in Tables 2–5, in which the median and robust estimate of the standard deviation (over 100 repetitions) of several performance measures are reported: ℓ_1-estimation error, squared ℓ_2-estimation error, true positives (TP), false positives (FP), and computational time in seconds (Time). In Cases 1 and 2, under the Linear and Logistic regression setups, for any type of SIS or ISIS, we employ the SCAD penalty ($\gamma = 3.7$) at the end of the screening steps; whereas LASSO is applied for Cases 3 and 4, under the Poisson and Cox proportional hazards regression frameworks. For simplicity, we exclude the performance of MCP-based screening procedures in the current analysis. Whenever necessary, for all variable selection procedures considered here, the BIC criterion is used as a fast way to select the regularization parameter $\lambda > 0$, always chosen from a path of 100 candidate λ values.

As the covariates are all independent in Case 1, it is not surprising to see that Van-SIS performs reasonably well. However, this non-iterative procedure fails in terms of identifying

Method	$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_1$	$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2^2$	TP	FP	Time
Van-SIS	3.10(0.55)	2.08(0.26)	3(0.00)	9.5(18.66)	0.14(0.04)
Van-ISIS	5.21(0.79)	2.21(2.20)	4(0.75)	29(0.00)	86.23(64.17)
Var1-ISIS	0.53(0.39)	0.09(0.11)	4(0.00)	1(0.75)	88.05(28.02)
Var2-ISIS	5.05(0.67)	2.15(0.22)	3(0.75)	29(0.75)	207.67(100.74)
Perm-ISIS	6.16(1.40)	6.56(5.54)	3(0.00)	30(0.75)	130.93(221.45)
Perm-g-ISIS	6.76(0.77)	1.70(0.74)	4(0.00)	29(0.00)	2202.81(136.08)
Perm-var-ISIS	0.26(0.21)	0.02(0.04)	4(0.00)	0(0.75)	174.90(42.50)
Perm-var-g-ISIS	0.43(0.36)	0.07(0.08)	4(0.00)	1(1.49)	231.81(89.62)
LASSO	2.97(0.07)	2.14(0.18)	3(0.00)	20(9.70)	1.47(0.36)

Table 4: Poisson regression, Case 3, where results are given in the form of medians and robust standard deviations (in parentheses).

Method	$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_1$	$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2^2$	TP	FP	Time
Van-SIS	21.27(0.49)	95.64(3.63)	3(0.75)	12(0.75)	0.15(0.04)
Van-ISIS	3.13(1.20)	1.02(1.35)	5(0.00)	11(0.00)	92.29(45.61)
Var1-ISIS	1.33(0.62)	0.38(0.42)	5(0.00)	2(2.24)	207.16(45.69)
Var2-ISIS	2.80(1.12)	0.93(1.10)	5(0.00)	11(0.00)	189.24(84.92)
Perm-ISIS	21.44(0.25)	93.93(6.42)	3(0.00)	13(0.00)	136.47(69.53)
Perm-g-ISIS	9.19(1.95)	8.26(5.53)	5(0.00)	11(0.00)	1102.28(96.11)
Perm-var-ISIS	0.95(0.76)	0.24(0.38)	5(0.00)	0(0.00)	386.87(68.42)
Perm-var-g-ISIS	1.24(0.66)	0.35(0.41)	5(0.00)	1(1.49)	509.85(147.38)
LASSO	163.09(14.17)	1035.23(173.27)	4(0.00)	313(10.63)	35.67(3.87)

Table 5: Cox proportional hazards regression, Case 4, where results are given in the form of medians and robust standard deviations (in parentheses).

the true model when correlation is present, particularly in the challenging Cases 3 and 4. When predictors are dependent, the vanilla ISIS improves significantly over SIS in terms of true positives. While the number of false positive variables may be larger in some settings, Van-ISIS provides comparable estimation errors in Cases 1–3 but significant reduction in the complicated Case 4.

In terms of further reducing the false selection rate and estimation errors, while still selecting the true model $\mathcal{M}_\star$, Var1-ISIS performs much better than Var2-ISIS. Being a more conservative variable selection approach, Var2-ISIS tends to have a higher number of false positives. This is particularly true in the Poisson regression scenario, in which the second variant even misses one important predictor.

From the permutation-based variants, the ones that combine the sample splitting approach (Perm-var-ISIS and Perm-var-g-ISIS) outperform all other ISIS procedures in terms of true positives, low false selection rates, and small estimation errors, with Var1-ISIS following closely. In particular, for Perm-var-ISIS, the number of false positives is approximately zero for all examples. The only drawback seems to be their relatively large computation cost, being at least twice as large as that of Var1-ISIS. This is to be expected considering the

amount of extra work these procedures have to perform: two rounds of marginal fits to obtain sample specific data-driven thresholds $\omega_{(q)}^{(1)}$ and $\omega_{(q)}^{(2)}$, plus two additional rounds of marginal fits to compute the corresponding index sets $\widehat{\mathcal{A}}_1^{(1)}$ and $\widehat{\mathcal{A}}_1^{(2)}$. Computational costs potentially increase further when carrying out the conditional marginal regression steps described in Section 2.2, however, the gains in statistical accuracy and model selection offset the increased timings, particularly in the equally correlated Case 2 with nonconvex penalties.

Tables 2–3 show that SCAD also enjoys the sure screening property for the relatively easy Cases 1 and 2, however, model sizes and estimation errors are significantly larger than any of the ISIS procedures in the correlated scenario. On the other hand, for the difficult Cases 3 and 4, surprisingly, LASSO rarely includes the important predictor x_4 even though it is not a marginal screening based method. While exhibiting competitive performance with some of the ISIS variants in the Poisson regression scenario, LASSO performs poorly in the Cox model setup, having prohibitively large model sizes and estimation errors.

The **ncvreg** package with SCAD outperforms all ISIS variants in terms of computational cost for the uncorrelated Case 1. Still, the vanilla SIS procedure identifies the true model faster than **ncvreg**. For the correlated Case 2, however, only the greedy modification Perm-g-ISIS is slower than SCAD. In the Poisson and Cox model setups, while the computational cost of LASSO with the **glmnet** package is smaller than any of the ISIS procedures, the vanilla SIS shows better performance in terms of timings and estimation errors.

As they become more elaborate, ISIS procedures become more computationally expensive. Yet, vanilla ISIS and most of its variants presented here, particularly Var1-ISIS and Perm-var-ISIS, are clearly competitive variable selection methods in ultrahigh dimensional statistical models, adequately using the joint covariate information, while exhibiting a very low false selection rate and competing computational cost.

4.2. Scaling in n and p with feature screening

In addition to comparing our **SIS** codes with **glmnet** and **ncvreg**, we would like to know how the timings of the vanilla SIS and ISIS procedures scale in n and p. We simulated data sets from Cases 1–3 above and, for a variety of n and p values, we took the median running time over 10 repetitions. Again, for each (n, p) pair, whenever necessary, the BIC criterion was used to select the best λ value among a path of 100 possible candidates. Figure 1 shows timings for fixed n as p grows (Cases 1–3), and for fixed p as n grows (Case 2).

From the plots we see that independence screening procedures perform uniformly faster than **ncvreg** with the SCAD penalty. For Poisson regression, vanilla SIS also outperforms **glmnet** with LASSO, particularly in the $n = p$ scenario, where **glmnet** exhibits unusually slow performance. It is worth pointing out that iterative variable screening procedures typically do not show a strictly monotone timing as n or p increase. This is due to the varying number of iterations it takes to recruit d predictors, the random splitting of the sample in the first two ISIS variants, the random permutation in the data-driven thresholding, among other factors.

4.3. Real data analysis

We now evaluate the performance of all variable screening procedures on four well-studied gene expression data sets: Leukemia (Golub *et al.* 1999), Prostate cancer (Singh *et al.* 2002), Lung cancer (Gordon *et al.* 2002) and Neuroblastoma (Oberthuer *et al.* 2006). The first

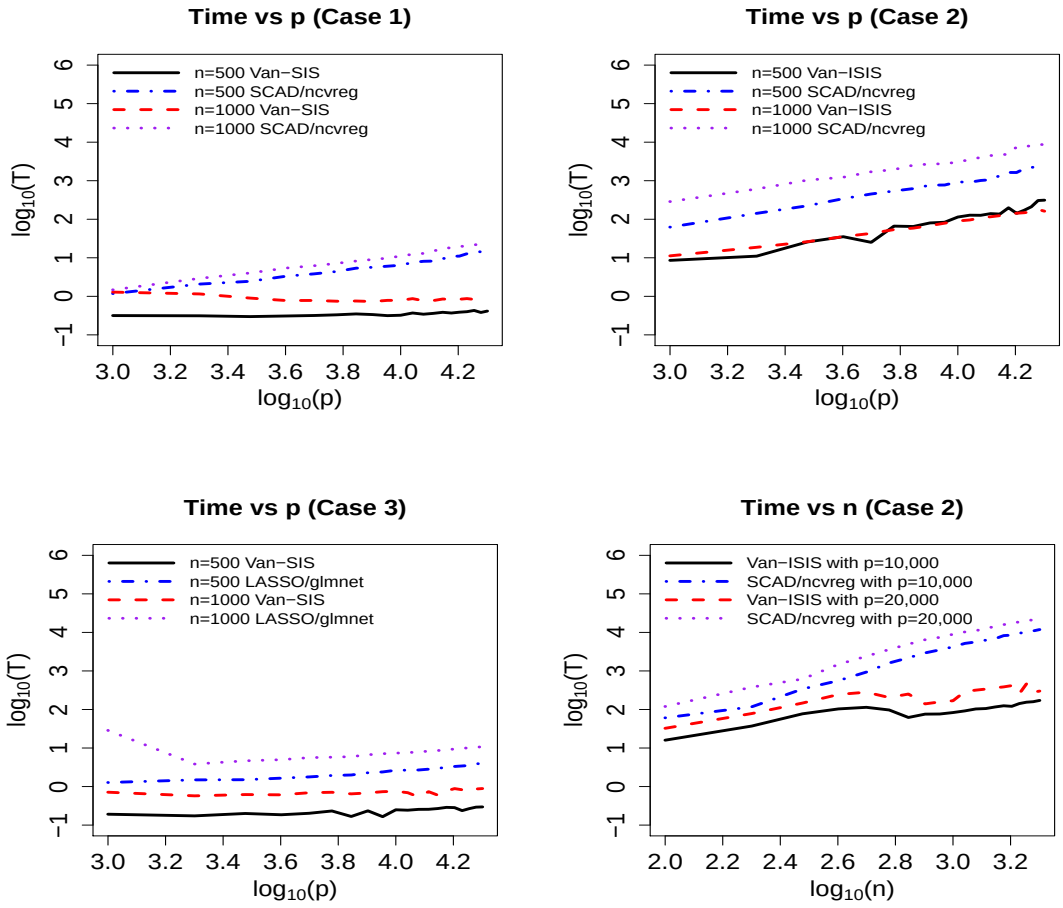

Figure 1: Median runtime in seconds taken over 10 trials (log-log scale).

three data sets come with predetermined, separate training and test sets of data vectors. The Leukemia data set contains $p = 7129$ genes for 27 acute lymphoblastic leukemia and 11 acute myeloid leukemia vectors in the training set. The test set includes 20 acute lymphoblastic leukemia and 14 acute myeloid leukemia vectors. The Prostate cancer data set consists of gene expression profiles from $p = 12600$ genes for 52 prostate "tumor" samples and 50 "normal" prostate samples in the training data set. The test set is from a different experiment and contains 25 tumor and 9 normal samples. The lung cancer data set contains 32 tissue samples in the training set (16 from malignant pleural mesothelioma and 16 from adenocarcinoma) and 149 in the test set (15 from malignant pleural mesothelioma and 134 from adenocarcinoma). Each sample consists of $p = 12533$ genes. The neuroblastoma data set consists of gene expression profiles for $p = 10707$ genes from 246 patients of the German neuroblastoma trials NB90-NB2004, diagnosed between 1989 and 2004. We analyzed the gene expression data by means of the 3-year event-free survival, indicating whether a patient survived 3 years after the diagnosis of neuroblastoma. Combining the original training and test sets, the data consists of 56 positive and 190 negative cases. For purposes of the present analysis, in each of these gene expression data sets, we initially combine the training and test samples and then perform a 50% - 50% random splitting of the observed data into new training and test data for which

Method	Leukemia training error rate	Leukemia test error rate	Leukemia model size	Prostate training error rate	Prostate test error rate	Prostate model size
Van-SIS	0.00(0.00)	0.06(0.04)	16(2.43)	0.00(0.01)	0.22(0.05)	14(3.92)
Van-ISIS	0.00(0.00)	0.06(0.04)	15(2.99)	0.00(0.01)	0.20(0.06)	14(2.43)
Var1-ISIS	0.00(0.02)	0.08(0.04)	5(1.49)	0.04(0.04)	0.19(0.09)	6(2.24)
Var2-ISIS	0.00(0.00)	0.06(0.02)	15(2.24)	0.00(0.01)	0.18(0.06)	12(2.24)
Perm-ISIS	0.00(0.00)	0.06(0.04)	15(2.99)	0.00(0.01)	0.20(0.05)	13(2.99)
Perm-g-ISIS	0.03(0.02)	0.08(0.04)	3(1.49)	0.07(0.05)	0.23(0.11)	4(1.49)
Perm-Var-ISIS	0.00(0.02)	0.08(0.04)	5(1.49)	0.04(0.04)	0.19(0.09)	5(2.24)
Perm-Var-g-ISIS	0.03(0.04)	0.08(0.04)	2(0.00)	0.08(0.04)	0.22(0.10)	4(0.93)
LASSO-CV(10)	0.00(0.00)	0.06(0.04)	17(2.99)	0.03(0.06)	0.22(0.07)	19(8.21)
NSC	0.03(0.02)	0.06(0.04)	143(456.72)	0.07(0.03)	0.20(0.12)	13(15.30)
IR	0.03(0.02)	0.14(0.10)	7129(0.00)	0.29(0.04)	0.32(0.06)	12600(0.00)

Table 6: Classification error rates and number of selected genes by various methods for the balanced Leukemia and Prostate cancer data sets. For the Leukemia data, the training and test samples are of size 36. For the Prostate cancer data, the training and test samples are of size 68. Results are given in the form of medians and robust standard deviations (in parentheses).

the number of cases remains balanced across these new samples. In this manner, for the Leukemia data, the balanced training and test samples are of size 36, for the Prostate data we have balanced training and test samples of size 68, whereas the Neuroblastoma data set has balanced training and test samples of size 123. The balanced training and test samples for the Lung cancer data are of sizes 90 and 91, respectively. Interested readers can find more details about these data sets in Golub *et al.* (1999), Singh *et al.* (2002), Gordon *et al.* (2002) and Oberthuer *et al.* (2006).

Following the approach of Dudoit *et al.* (2002), before variable screening and classification, we first standardize each sample to zero mean and unit variance. We compare the performance of all described variable screening procedures with the Nearest Shrunken Centroids (NSC) method of Tibshirani *et al.* (2002), the Independence Rule (IR) in the high-dimensional setting (Bickel and Levina 2004) and the LASSO (Tibshirani 1996), which uses ten-fold cross-validation to select its tuning parameter, applied to the full set of covariates. Under a working independence assumption in the feature space, NSC selects an important subset of variables for classification by thresholding a corresponding two-sample t statistic, whereas IR makes use of the full set of predictors.

Tables 6–7 show the median and robust standard deviation of the classification error rates and model sizes for all procedures, taken over 100 random splittings into 50% - 50% balanced training and test data. At each intermediate step of the (I)SIS procedures, we employ the LASSO with ten-fold cross-validation to further filter unimportant predictors for classification purposes. To determine a data-driven threshold for independence screening, we fix $q = 0.95$ for all permutation-based variable selection procedures. Lastly, for each data set considered, we apply all screening procedures to reduce dimensionality from the corresponding p to $d = 100$.

Method	Lung training error rate	Lung test error rate	Lung model size	NB training error rate	NB test error rate	NB model size
Van-SIS	0.00(0.00)	0.02(0.01)	14(2.43)	0.09(0.02)	0.19(0.02)	14(2.99)
Van-ISIS	0.00(0.00)	0.02(0.01)	13(1.49)	0.00(0.00)	0.22(0.03)	38(5.22)
Var1-ISIS	0.00(0.00)	0.02(0.01)	9(1.68)	0.14(0.04)	0.21(0.03)	3(2.24)
Var2-ISIS	0.00(0.00)	0.02(0.01)	13(2.24)	0.00(0.00)	0.22(0.03)	33(5.41)
Perm-ISIS	0.00(0.00)	0.02(0.01)	13(1.49)	0.00(0.00)	0.22(0.02)	38(5.97)
Perm-g-ISIS	0.00(0.01)	0.02(0.02)	2(0.75)	0.03(0.02)	0.26(0.04)	10(3.73)
Perm-Var-ISIS	0.00(0.00)	0.02(0.01)	9(2.24)	0.14(0.04)	0.21(0.03)	4(2.24)
Perm-Var-g-ISIS	0.01(0.01)	0.02(0.01)	2(0.75)	0.14(0.03)	0.21(0.04)	3(1.49)
LASSO-CV(10)	0.01(0.01)	0.02(0.02)	15(2.24)	0.14(0.04)	0.26(0.04)	23(9.70)
NSC	0.00(0.00)	0.00(0.02)	6(22.76)	0.18(0.02)	0.20(0.03)	361(5150.37)
IR	0.13(0.05)	0.14(0.06)	12533(0.00)	0.15(0.02)	0.20(0.02)	10707(0.00)

Table 7: Classification error rates and number of selected genes by various methods for the balanced Lung and Neuroblastoma (NB) cancer data sets. For the Lung data, the training and test samples are of sizes 90 and 91, respectively. For the Neuroblastoma cancer data, the training and test samples are of size 123. Results are given in the form of medians and robust standard deviations (in parentheses).

From the results in Tables 6–7, we observe that all ISIS variants perform similarly in terms of test error rates, whereas the main differences lie in the estimated model sizes. Compared with the LASSO applied to the full set of covariates, a majority of ISIS procedures select smaller models while retaining competitive classification error rates. This is in agreement with our simulation results, which highlight the benefits of variable screening over a direct high-dimensional regularized logistic regression approach. In particular, we observe the variants Var1-ISIS and Perm-var-g-ISIS provide the most parsimonious models across all four data sets, yielding optimal test error rates while using only 2 features in the case of the Lung cancer data set. Nonetheless, due to its robust performance in both the simulated data and these four gene expression data sets, and its reduced computational cost compared with all available ISIS variants, we select the vanilla ISIS of Algorithm 2 as the default variable selection procedure within our **SIS** package.

While the NSC method achieves competitive test error rates, it typically makes use of larger sets of genes which vary considerably across the different 50% - 50% training and test data splittings. The Independence Rule exhibits poor test error performance, except for the Neuroblastoma data set, where it even outperforms some of the ISIS procedures. However, this approach uses all features without performing variable selection, thus yielding models of little practical use for researchers.

4.4. Code example

All described independence screening procedures are straightforward to run using the **SIS** package. We demonstrate the `SIS` function on the Leukemia data set from the previous section. We first load the predictors and response vector from the training and test data sets.

```
R> library("SIS")
R> set.seed(9)
R> data("leukemia.train", package = "SIS")
R> data("leukemia.test", package = "SIS")
R> y1 = leukemia.train[, dim(leukemia.train)[2]]
R> x1 = as.matrix(leukemia.train[, -dim(leukemia.train)[2]])
R> y2 = leukemia.test[, dim(leukemia.test)[2]]
R> x2 = as.matrix(leukemia.test[, -dim(leukemia.test)[2]])
```

Afterwards, we carry out the balanced sample splitting as outlined above.

```
R> x = rbind(x1, x2)
R> y = c(y1, y2)
R> n = dim(x)[1]; aux = 1:n
R> ind.train1 = sample(aux[y == 0], 23, replace = FALSE)
R> ind.train2 = sample(aux[y == 1], 13, replace = FALSE)
R> ind.train = c(ind.train1, ind.train2)
R> x.train = scale(x[ind.train,])
R> y.train = y[ind.train]
R> ind.test1 = setdiff(aux[y == 0], ind.train1)
R> ind.test2 = setdiff(aux[y == 1], ind.train2)
R> ind.test = c(ind.test1, ind.test2)
R> x.test = scale(x[ind.test,])
R> y.test = y[ind.test]
```

We now perform variable selection using the Var1-ISIS and Perm-var-ISIS procedures paired with the LASSO penalty and the ten-fold cross-validation method for choosing the regularization parameter.

```
R> model1 = SIS(x.train, y.train, family = "binomial", penalty = "lasso",
+    tune = "cv", nfolds = 10, nsis = 100, varISIS = "aggr", seed = 9,
+    standardize = FALSE)
R> model2 = SIS(x.train, y.train, family = "binomial", penalty = "lasso",
+    tune = "cv", nfolds = 10, nsis = 100, varISIS = "aggr", perm = TRUE,
+    q = 0.95, seed = 9, standardize = FALSE)

R> model1$ix
[1] 1834 4377 4847 6281 6493
R> model2$ix
[1] 1834 4377 4847 6281 6493
```

Here we modified the default value $d = \lfloor n/(4 \log n) \rfloor$ to make both iterative procedures select models with at most 100 predictors. The value of $q \in [0, 1]$, from which we obtain the data-driven threshold $\omega_{(q)}$ for Perm-var-ISIS, was also customized from its default $q = 1$.

5. Discussion

Sure independence screening is a power family of methods for performing variable selection in statistical models when the dimension is much larger than the sample size, as well as in the classical setting where $p < n$. The focus of the paper is on iterative sure independence screening, which iteratively applies a large scale screening by means of conditional marginal regressions, filtering out unimportant predictors, and a moderate scale variable selection through penalized pseudo-likelihood methods, which further selects the unfiltered predictors. With the goal of providing further flexibility to the iterative screening paradigm, special attention is also paid to powerful variants which reduce the number of false positives by means of sample splitting and data-driven thresholding approaches. Compared with the versions of LASSO and SCAD we used, the iterative procedures presented in this paper are much more accurate in selecting important variables and achieving small estimation errors. In addition, computational time is also reduced, particularly in the case of nonconvex penalties, thus resulting in a robust family of procedures for model selection and estimation in ultrahigh dimensional statistical models. Extensions of the current package to more general loss-based models and nonparametric independence screening procedures, as well as implementation of conditional marginal regressions through support vector machine methods are lines of future work.

Acknowledgments

We would like to thank the Editor, the AE and the referee for constructive comments which have greatly improved the paper. This research was partially supported by NSF DMS-1308566.

References

Akaike H (1973). "Information Theory and an Extension of the Maximum Likelihood Principle." In BN Petrov, F Csáki (eds.), *Proceedings of the 2nd International Symposium on Information Theory*. Akadémiai Kiadó, Budapest.

Bernau C, Waldron L, Riester M (2014). ***survHD**: Synthesis of High-Dimensional Survival Analysis.* R package version 0.99.1, URL https://bitbucket.org/lwaldron/survhd.

Bickel PJ, Levina E (2004). "Some Theory for Fisher's Linear Discriminant Function, 'Naive Bayes', and Some Alternatives when There Are Many More Variables than Observations." *Bernoulli*, **10**(6), 989–1010.

Breheny P (2013). ***ncvreg**: Regularization Paths for SCAD- and MCP-Penalized Regression Models.* R package version 2.6-0, URL http://CRAN.R-project.org/package=ncvreg.

Breheny P, Huang J (2011). "Coordiante Descent Algorithms for Nonconvex Penalized Regression, with Applications to Biological Feature Selection." *The Annals of Applied Statistics*, **5**(1), 232–253.

Chen J, Chen Z (2008). "Extended Bayesian Information Criteria for Model Selection with Large Model Spaces." *Biometrika*, **95**(3), 759–771.

Cox DR (1975). "Partial Likelihood." *Biometrika*, **62**(2), 269–276.

Dudoit S, Fridlyand J, Speed TP (2002). "Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data." *Journal of the American Statistical Association*, **97**(457), 77–87.

Fan J, Feng Y, Saldana DF, Samworth R, Wu Y (2015). ***SIS**: Sure Independence Screening*. R package version 0.7-6, URL http://CRAN.R-project.org/package=SIS.

Fan J, Feng Y, Song R (2011). "Nonparametric Independence Screening in Sparse Ultra-High-Dimensional Additive Models." *Journal of the American Statistical Association*, **106**(494), 544–557.

Fan J, Feng Y, Wu Y (2010). "High-Dimensional Variable Selection for Cox's Proportional Hazards Model." *IMS Collections*, **6**, 70–86.

Fan J, Li R (2001). "Variable Selection Via Nonconcave Penalized Likelihood and its Oracle Properties." *Journal of the American Statistical Association*, **96**(456), 1348–1360.

Fan J, Lv J (2008). "Sure Independence Screening for Ultrahigh Dimensional Feature Space." *Journal of the Royal Statistical Society B*, **70**(5), 849–911.

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

Fan J, Song R (2010). "Sure Independence Screening in Generalized Linear Models with NP-Dimensionality." *The Annals of Statistics*, **38**(6), 3567–3604.

Feng Y, Yu Y (2013). "Consistent Cross-Validation for Tuning Parameter Selection in High-Dimensional Variable Selection." Manuscript.

Friedman J, Hastie T, Tibshirani R (2010). "Regularization Paths for Generalized Linear Models Via Coordinate Descent." *Journal of Statistical Software*, **33**(1), 1–22.

Friedman J, Hastie T, Tibshirani R (2013). ***glmnet**: Lasso and Elastic-Net Regularized Generalized Linear Models*. R package version 1.9-5, URL http://CRAN.R-project.org/package=glmnet.

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES (1999). "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring." *Science*, **286**(5439), 531–537.

Gordon GJ, Jensen RV, Hsiao LL, Gullans SR, Blumenstock JE, Ramaswamy S, Richards WG, Sugarbaker DJ, Bueno R (2002). "Translation of Microarray Data into Clinically Relevant Cancer Diagnostic Tests Using Gene Expression Ratios in Lung Cancer and Mesothelioma." *Cancer Research*, **62**, 4963–4967.

Guyon I, Elisseeff A (2003). "An Introduction to Variable and Feature Selection." *Journal of Machine Learning Research*, **3**, 1157–1182.

Kalbfleisch JD, Prentice RL (2002). *The Statistical Analysis of Failure Time Data*. 2nd edition. John Wiley & Sons, New Jersey.

Mallows CL (1973). "Some Comments on C_p." *Technometrics*, **15**(4), 661–675.

Nishii R (1984). "Asymptotic Properties of Criteria for Selection of Variables in Multiple Regression." *The Annals of Statistics*, **12**(2), 758–765.

Oberthuer A, Berthold F, Warnat P, Hero B, Kahlert Y, Spitz R, Ernestus K, König R, Haas S, Eils R, Schwab M, Brors B, Westermann F, Fischer M (2006). "Customized Oligonucleotide Microarray Gene Expression-Based Classification of Neuroblastoma Patients Outperforms Current Clinical Risk Stratification." *Journal of Clinical Oncology*, **24**(31), 5070–5078.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rakotomamonjy A (2003). "Variable Selection Using SVM-based Criteria." *Journal of Machine Learning Research*, **3**, 1357–1370.

Schwarz G (1978). "Estimating the Dimension of a Model." *The Annals of Statistics*, **6**(2), 461–464.

Simon H, Friedman J, Hastie T, Tibshirani R (2011). "Regularization Paths for Cox's Proportional Hazards Model Via Coordinate Descent." *Journal of Statistical Software*, **39**(5), 1–13.

Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, Tamayo P, Renshaw AA, D'Amico AV, Richie JP, Lander ES, Loda M, Kantoff PW, Golub TR, Sellers WR (2002). "Gene Expression Correlates of Clinical Prostate Cancer Behavior." *Cancer Cell*, **1**(2), 203–209.

Therneau TM, Lumley T (2015). ***survival**: Survival Analysis*. R package version 2.38-3, URL http://CRAN.R-project.org/package=survival.

Tibshirani R (1996). "Regression Shrinkage and Selection Via the Lasso." *Journal of the Royal Statistical Society B*, **58**(1), 267–288.

Tibshirani R (1997). "The Lasso Method for Variable Selection in the Cox Model." *Statistics in Medicine*, **16**, 385–395.

Tibshirani R, Hastie T, Narasimhan B, Chu G (2002). "Diagnosis of Multiple Cancer Types by Shrunken Centroids of Gene Expression." *Proceedings of the National Academy of Sciences of the United States of America*, **99**(10), 6567–6572.

Yu Y, Feng Y (2014a). "APPLE: Approximate Path for Penalized Likelihood Estimators." *Statistics and Computing*, **24**, 803–819.

Yu Y, Feng Y (2014b). "Modified Cross-Validation for Penalized High-Dimensional Linear Regression Models." *Journal of Computational and Graphical Statistics*, **23**(4), 1009–1027.

Yu Y, Feng Y (2015). ***apple**: Approximate Path for Penalized Likelihood Estimators*. R package version 0.3, URL http://CRAN.R-project.org/package=apple.

Zhang CH (2010). "Nearly Unbiased Variable Selection Under Minimax Concave Penalty." *The Annals of Statistics*, **38**(2), 894–942.

Zou H, Hastie T (2005). "Regularization and Variable Selection Via the Elastic Net." *Journal of the Royal Statistical Society B*, **67**(2), 301–320.

Affiliation:

Diego Franco Saldana, Yang Feng
Department of Statistics
Columbia University
New York NY, 10027, United States of America
E-mail: diego@stat.columbia.edu, yangfeng@stat.columbia.edu
URL: http://www.stat.columbia.edu/~yangfeng/

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume VV, Issue II *Submitted:* yyyy-mm-dd
MMMMMM YYYY *Accepted:* yyyy-mm-dd