Journal of the Royal Statistical Society
SERIES B Statistical Methodology

J. R. Statist. Soc. B (2008)
70, *Part* 5, *pp.* 849–911

Sure independence screening for ultrahigh dimensional feature space

Jianqing Fan

Princeton University, USA

and Jinchi Lv

University of Southern California, Los Angeles, USA

[*Read before* The Royal Statistical Society *at a meeting organized by the* Research Section *on Wednesday, April 23rd, 2008*, Professor I. L. Dryden *in the Chair*]

Summary. Variable selection plays an important role in high dimensional statistical modelling which nowadays appears in many areas and is key to various scientific discoveries. For problems of large scale or dimensionality p, accuracy of estimation and computational cost are two top concerns. Recently, Candes and Tao have proposed the Dantzig selector using L_1-regularization and showed that it achieves the ideal risk up to a logarithmic factor $\log(p)$. Their innovative procedure and remarkable result are challenged when the dimensionality is ultrahigh as the factor $\log(p)$ can be large and their uniform uncertainty principle can fail. Motivated by these concerns, we introduce the concept of sure screening and propose a sure screening method that is based on correlation learning, called sure independence screening, to reduce dimensionality from high to a moderate scale that is below the sample size. In a fairly general asymptotic framework, correlation learning is shown to have the sure screening property for even exponentially growing dimensionality. As a methodological extension, iterative sure independence screening is also proposed to enhance its finite sample performance. With dimension reduced accurately from high to below sample size, variable selection can be improved on both speed and accuracy, and can then be accomplished by a well-developed method such as smoothly clipped absolute deviation, the Dantzig selector, lasso or adaptive lasso. The connections between these penalized least squares methods are also elucidated.

Keywords: Adaptive lasso; Dantzig selector; Dimensionality reduction; Lasso; Oracle estimator; Smoothly clipped absolute deviation; Sure independence screening; Sure screening; Variable selection

1. Introduction

1.1. Background

Consider the problem of estimating a p-vector of parameters $\boldsymbol{\beta}$ from the linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ is an n-vector of responses, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^{\mathrm{T}}$ is an $n \times p$ random design matrix with independent and identically distributed (IID) $\mathbf{x}_1, \ldots, \mathbf{x}_n$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is a p-vector of parameters and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{T}}$ is an n-vector of IID random errors. When dimension p is high, it is often assumed that only a small number of predictors among $X_1, \ldots, X_p$ contribute to the response, which amounts to assuming ideally that the parameter vector $\boldsymbol{\beta}$ is sparse. With sparsity, variable selection can improve the accuracy of estimation by effectively

Address for correspondence: Jianqing Fan, Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA.
E-mail: jqfan@princeton.edu

 1369–7412/08/70849

identifying the subset of important predictors, and also enhance model interpretability with parsimonious representation.

Sparsity comes frequently with high dimensional data, which is a growing feature in many areas of contemporary statistics. The problems arise frequently in genomics such as gene expression and proteomics studies, biomedical imaging, functional magnetic resonance imaging, tomography, tumour classifications, signal processing, image analysis and finance, where the number of variables or parameters p can be much larger than sample size n. For instance, one may wish to classify tumours by using microarray gene expression or proteomics data or one may wish to associate protein concentrations with expression of genes or to predict certain clinical prognosis (e.g. injury scores or survival time) by using gene expression data. For this kind of problems, the dimensionality can be much larger than the sample size, which calls for new or extended statistical methodologies and theories. See, for example Donoho (2000) and Fan and Li (2006) for overviews of statistical challenges with high dimensionality.

Back to the problem in model (1), it is challenging to find tens of important variables out of thousands of predictors, with a number of observations usually in tens or hundreds. This is similar to finding a couple of needles in a huge haystack. A new idea in Candes and Tao (2007) is the notion of the uniform uncertainty principle on deterministic design matrices. They proposed the Dantzig selector which is the solution to an l_1-regularization problem and showed that, under the uniform uncertainty principle, this minimum l_1-estimator achieves the ideal risk, i.e. the risk of the oracle estimator with the true model known ahead of time, up to a logarithmic factor $\log(p)$. Appealing features of the Dantzig selector include that

(a) it is easy to implement because the convex optimization that the Dantzig selector solves can easily be recast as a linear program and
(b) it has the oracle property in the sense of Donoho and Johnstone (1994).

Despite their remarkable achievement, we still have four concerns when the Dantzig selector is applied to high or ultrahigh dimensional problems. First, a potential hurdle is the computational cost for large or huge scale problems such as implementing linear programs in dimension of tens or hundreds of thousands. Second, the factor $\log(p)$ can become large and may not be negligible when dimension p grows rapidly with sample size n. Third, as dimensionality grows, their uniform uncertainty principle condition may be difficult to satisfy, which will be illustrated later by using a simulated example. Finally, there is no guarantee that the Dantzig selector picks up the right model though it has the oracle property. These four concerns inspire our work.

1.2. Dimensionality reduction

Dimension reduction or feature selection is an effective strategy to deal with high dimensionality. With dimensionality reduced from high to low, the computational burden can be reduced drastically. Meanwhile, accurate estimation can be obtained by using some well-developed lower dimensional method. Motivated by this along with those concerns on the Dantzig selector, we have the following main goal in our paper:

> reduce dimensionality p from a large or huge scale (say, $\exp\{O(n^{\xi})\}$ for some $\xi > 0$) to a relatively large scale d (e.g. $o(n)$) by a fast and efficient method.

We achieve this by introducing the concept of sure screening and proposing a sure screening method which is based on correlation learning which filters out the features that have weak correlation with the response. Such correlation screening is called sure independence screening (SIS). Here and below, by sure screening we mean a property that all the important variables

survive after variable screening with probability tending to 1. This dramatically narrows down the search for important predictors. In particular, applying the Dantzig selector to the much smaller submodel relaxes our first concern on the computational cost. In fact, this not only speeds up the Dantzig selector but also reduces the logarithmic factor in mimicking the ideal risk from $\log(p)$ to $\log(d)$, which is smaller than $\log(n)$ and hence relaxes our second concern above.

Oracle properties in a stronger sense, say, mimicking the oracle in not only selecting the right model, but also estimating the parameters efficiently, give a positive answer to our third and fourth concerns above. Theories on oracle properties in this sense have been developed in the literature. Fan and Li (2001) laid down groundwork on variable selection problems in the finite parameter setting. They discussed a family of variable selection methods that adopt a penalized likelihood approach, which includes well-established methods such as the Akaike information criterion and Bayes information criterion, as well as more recent methods like the bridge regression in Frank and Friedman (1993), the lasso in Tibshirani (1996), and the smoothly clipped absolute deviation (SCAD) method in Fan (1997) and Antoniadis and Fan (2001), and established oracle properties for non-concave penalized likelihood estimators. Later on, Fan and Peng (2004) extended the results to the setting of $p = o(n^{1/3})$ and showed that the oracle properties continue to hold. An effective algorithm for optimizing penalized likelihood, the local quadratic approximation, was proposed in Fan and Li (2001) and well studied in Hunter and Li (2005). Zou (2006) introduced an adaptive lasso in a finite parameter setting and showed that the lasso does not have oracle properties as conjectured in Fan and Li (2001), whereas the adaptive lasso does. Zou and Li (2008) propose a local linear approximation algorithm that recasts the computation of non-concave penalized likelihood problems into a sequence of penalized L_1-likelihood problems. They also proposed and studied the one-step sparse estimators for non-concave penalized likelihood models.

There is a huge literature on the problem of variable selection. To name a few in addition to those mentioned above, Fan and Li (2002) studied variable selection for Cox's proportional hazards model and frailty model, Efron *et al.* (2004) proposed the least angle regression algorithm LARS, Hunter and Li (2005) proposed a new class of algorithms, minorization–maximization algorithms, for variable selection, Meinshausen and Bühlmann (2006) looked at the problem of variable selection with the lasso for high dimensional graphs and Zhao and Yu (2006) gave an almost necessary and sufficient condition on model selection consistency of the lasso. Meier *et al.* (2008) proposed a fast implementation for the group lasso. More recent studies include Huang *et al.* (2008), Paul *et al.* (2008), Zhang (2007) and Zhang and Huang (2008), which signficantly advance the theory and methods of the penalized least squares (PLS) approaches. It is worth mentioning that in variable selection there is a weaker concept than consistency, called persistency, which was introduced by Greenshtein and Ritov (2004). Motivation of this concept lies in the fact that, in machine learning such as tumour classification, primary interest centres on the misclassification errors or more generally expected losses, not the accuracy of estimated parameters. Greenshtein and Ritov (2004) studied the persistence of lasso-type procedures in high dimensional linear predictor selection, and Greenshtein (2006) extended the results to more general loss functions. Meinshausen (2007) considered a case with finite non-sparsity and showed that, under quadratic loss, the lasso is persistent, but the rate of persistency is slower than that of a relaxed lasso.

1.3. Some insight into high dimensionality

To gain some insight into challenges of high dimensionality in variable selection, let us look at a situation where all the predictors $X_1, \ldots, X_p$ are standardized and the distribution of $\mathbf{z} = \boldsymbol{\Sigma}^{-1/2}\mathbf{x}$

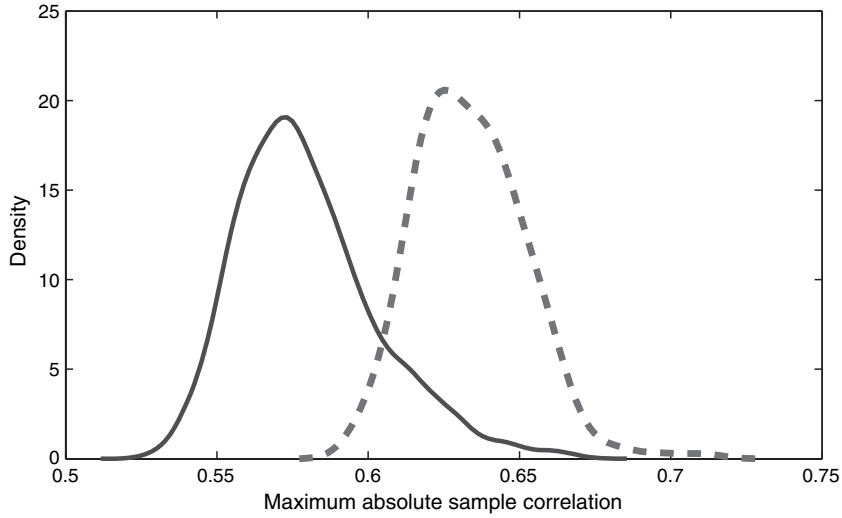

Fig. 1. Distributions of the maximum absolute sample correlation coefficient when $n = 60$ and $p = 1000$ (———) and $n = 60$ and $p = 5000$ (- - - -), based on 500 simulations

is spherically symmetric, where $\mathbf{x} = (X_1, \ldots, X_p)^{\mathrm{T}}$ and $\boldsymbol{\Sigma} = \mathrm{cov}(\mathbf{x})$. Clearly, the transformed predictor vector $\mathbf{z}$ has covariance matrix I_p. Our way of study in this paper is to separate the effects of the covariance matrix $\boldsymbol{\Sigma}$ and the distribution of $\mathbf{z}$, which gives us a better understanding of difficulties of high dimensionality in variable selection.

The real difficulty when dimension p is larger than sample size n comes from four facts. First, the design matrix $\mathbf{X}$ is rectangular, having more columns than rows. In this case, the matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is huge and singular. The maximum spurious correlation between a covariate and the response can be large (see, for example, Fig. 1) because of the dimensionality and the fact that an unimportant predictor can be highly correlated with the response variable owing to the presence of important predictors associated with the predictor. These make variable selection difficult. Second, the population covariance matrix $\boldsymbol{\Sigma}$ may become ill conditioned as n grows, which adds difficulty to variable selection. Third, the minimum non-zero absolute coefficient $|\beta_i|$ may decay with n and fall close to the noise level, say, the order $\{\log(p)/n\}^{-1/2}$. Fourth, the distribution of $\mathbf{z}$ may have heavy tails. Therefore, in general, it is challenging to estimate the sparse parameter vector $\boldsymbol{\beta}$ accurately when $p \gg n$.

When dimension p is large, some of the intuition might not be accurate. This is exemplified by the data piling problems in high dimensional space that were observed in Hall *et al.* (2005). A challenge with high dimensionality is that important predictors can be highly correlated with some unimportant ones, which usually increases with dimensionality. The maximum spurious correlation also grows with dimensionality. We illustrate this by using a simple example. Suppose that the predictors $X_1, \ldots, X_p$ are independent and follow the standard normal distribution. Then, the design matrix is an $n \times p$ random matrix, each entry an independent realization from $\mathcal{N}(0, 1)$. The maximum absolute sample correlation coefficient between predictors can be very large. This is indeed against our intuition, as the predictors are independent. To show this, we simulated 500 data sets with $n = 60$ and $p = 1000$ and $p = 5000$. Fig. 1 shows the distributions of the maximum absolute sample correlation of predictors. The multiple canonical correlation between two groups of predictors (e.g. 2 in one group and 3 in another) can even be much larger, as there are already

$$\binom{p}{2}\binom{p-2}{3} = O(p^5)$$

choices of the two groups in our example. Hence, sure screening when p is large is very challenging.

The paper is organized as follows. In the next section we propose a sure screening method, SIS, and discuss its rationale as well as its connection with other methods of dimensionality reduction. In Section 3 we review several known techniques for model selection in the reduced feature space and present two simulations and one real data example to study the performance of SIS-based model selection methods. In Section 4 we discuss some extensions of SIS and, in particular, an iterative SIS is proposed and illustrated by three simulated examples. Section 5 is devoted to the asymptotic analysis of SIS and an iteratively thresholded ridge regression screener as well as two SIS-based model selection methods. Some concluding remarks are given in Section 6. Technical details are provided in Appendix A.

2. Sure independence screening

2.1. A sure screening method: correlation learning

By sure screening we mean a property that all the important variables survive after applying a variable screening procedure with probability tending to 1. A dimensionality reduction method is desirable if it has the sure screening property. Below we introduce a simple sure screening method using componentwise regression or equivalently correlation learning. Throughout the paper we centre each input variable so that the observed mean is 0, and we scale each predictor so that the sample standard deviation is 1. Let $\mathcal{M}_* = \{1 \leqslant i \leqslant p : \beta_i \neq 0\}$ be the true sparse model with non-sparsity size $s = |\mathcal{M}_*|$. The other $p - s$ variables can also be correlated with the response variable via linkage to the predictors that are contained in the model. Let $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_p)^{\mathrm{T}}$ be a p-vector that is obtained by componentwise regression, i.e.

$$\boldsymbol{\omega} = \mathbf{X}^{\mathrm{T}}\mathbf{y}, \tag{2}$$

where the $n \times p$ data matrix $\mathbf{X}$ is first standardized columnwise as mentioned before. Hence, $\boldsymbol{\omega}$ is really a vector of marginal correlations of predictors with the response variable, rescaled by the standard deviation of the response.

For any given $\gamma \in (0, 1)$, we sort the p componentwise magnitudes of the vector $\boldsymbol{\omega}$ in a decreasing order and define a submodel

$$\mathcal{M}_\gamma = \{1 \leqslant i \leqslant p : |\omega_i| \text{ is among the first } [\gamma n] \text{ largest of all}\}, \tag{3}$$

where $[\gamma n]$ denotes the integer part of γn. This is a straightforward way to shrink the full model $\{1, \ldots, p\}$ down to a submodel $\mathcal{M}_\gamma$ with size $d = [\gamma n] < n$. Such correlation learning ranks the importance of features according to their marginal correlation with the response variable and filters out those that have weak marginal correlations with the response variable. We call this correlation screening method SIS, since each feature is used independently as a predictor to decide how useful it is for predicting the response variable. This concept is broader than correlation screening and is applicable to generalized linear models, classification problems under various loss functions and non-parametric learning under sparse additive models (Ravikumar *et al.*, 2007).

The computational cost of SIS or correlation learning is that of multiplying a $p \times n$ matrix by an n-vector plus obtaining the largest d components of a p-vector, so SIS has computational complexity $O(np)$.

It is worth mentioning that SIS uses only the order of componentwise magnitudes of $\boldsymbol{\omega}$, so it is indeed invariant under scaling. Thus the idea of SIS is identical to selecting predictors by using their correlations with the response. To implement SIS, we note that linear models with more than n parameters are not identifiable with only n data points. Hence, we may choose $d = [\gamma n]$ to be conservative, for instance, $n-1$ or $n/\log(n)$ depending on the order of sample size n. Although SIS is proposed to reduce dimensionality p from high to below sample size n, nothing can stop us applying it with final model size $d \geqslant n$, say $\gamma \geqslant 1$. It is obvious that larger d means larger probability of including the true model $\mathcal{M}_*$ in the final model $\mathcal{M}_\gamma$.

SIS is a hard-thresholding-type method. For orthogonal design matrices, it is well understood. But, for general design matrices, there is no theoretical support for it, though this kind of idea is frequently used in applications. It is important to identify the conditions under which the sure screening property holds for SIS, i.e.

$$P(\mathcal{M}_* \subset \mathcal{M}_\gamma) \to 1 \qquad \text{as } n \to \infty \tag{4}$$

for some given γ. This question as well as how the sequence $\gamma = \gamma_n \to 0$ should be chosen in theory will be answered by theorem 1 in Section 5. We would like to point out that the simple thresholding algorithm (see, for example, Baron *et al.* (2005) and Gribonval *et al.* (2007)) that is used in sparse approximation or compressed sensing is a one-step greedy algorithm and is related to SIS. In particular, our asymptotic analysis in Section 5 helps to understand the performance of the simple thresholding algorithm.

2.2. Rationale of correlation learning

To understand better the rationale of correlation learning, we now introduce an iteratively thresholded ridge regression screener (ITRRS), which is an extension of the dimensionality reduction method SIS. But, for practical implementation, only correlation learning is needed. The ITRRS also provides a very nice technical tool for our understanding of the sure screening property of correlation screening and other methods.

When there are more predictors than observations, it is well known that the least squares estimator $\hat{\boldsymbol{\beta}}_{\mathrm{LS}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{+}\mathbf{X}^{\mathrm{T}}\mathbf{y}$ is noisy, where $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{+}$ denotes the Moore–Penrose generalized inverse of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$. We therefore consider ridge regression, namely linear regression with l_2-regularization to reduce the variance. Let $\boldsymbol{\omega}^\lambda = (\omega_1^\lambda, \ldots, \omega_p^\lambda)^{\mathrm{T}}$ be a p-vector that is obtained by ridge regression, i.e.

$$\boldsymbol{\omega}^\lambda = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda I_p)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}, \tag{5}$$

where $\lambda > 0$ is a regularization parameter. It is obvious that

$$\boldsymbol{\omega}^\lambda \to \hat{\boldsymbol{\beta}}_{\mathrm{LS}} \qquad \text{as } \lambda \to 0, \tag{6}$$

and the scaled ridge regression estimator tends to the componentwise regression estimator:

$$\lambda\boldsymbol{\omega}^\lambda \to \boldsymbol{\omega} \qquad \text{as } \lambda \to \infty. \tag{7}$$

In view of property (6) to make $\boldsymbol{\omega}^\lambda$ less noisy we should choose a large regularization parameter λ to reduce the variance in the estimation. Note that the ranking of the absolute components of $\boldsymbol{\omega}^\lambda$ is the same as that of $\lambda\boldsymbol{\omega}^\lambda$. In light of property (7) the componentwise regression estimator is a specific case of ridge regression with regularization parameter $\lambda = \infty$, namely, it makes the resulting estimator as little noisy as possible.

For any given $\delta \in (0, 1)$, we sort the p componentwise magnitudes of the vector $\boldsymbol{\omega}^\lambda$ in a descending order and define a submodel

$$\mathcal{M}_{\delta,\lambda}^1 = \{1 \leqslant i \leqslant p : |\omega_i^\lambda| \text{ is among the first } [\delta p] \text{ largest of all}\}. \tag{8}$$

This procedure reduces the model size by a factor of $1-\delta$. The idea of the ITRRS to be introduced below is to perform dimensionality reduction as above successively until the number of remaining variables drops to below sample size n.

It will be shown in theorem 2 in Section 5 that, under some regularity conditions and when the tuning parameters λ and δ are chosen appropriately, with overwhelming probability the submodel $\mathcal{M}_{\delta,\lambda}^1$ will contain the true model $\mathcal{M}_*$ and its size is of order n^θ for some $\theta > 0$ lower than the original one p. This property stimulates us to propose the ITRRS as follows.

(a) First, carry out the procedure in submodel (8) to the full model $\{1, \ldots, p\}$ and obtain a submodel $\mathcal{M}_{\delta,\lambda}^1$ with size $[\delta p]$.
(b) Then, apply a similar procedure to the model $\mathcal{M}_{\delta,\lambda}^1$ and again obtain a submodel $\mathcal{M}_{\delta,\lambda}^2 \subset \mathcal{M}_{\delta,\lambda}^1$ with size $[\delta^2 p]$, and so on.
(c) Finally, obtain a submodel $\mathcal{M}_{\delta,\lambda} = \mathcal{M}_{\delta,\lambda}^k$ with size $d = [\delta^k p] < n$, where $[\delta^{k-1} p] \geqslant n$.

We point out that this procedure is different from thresholded ridge regression, as the submodels and estimated parameters change over the course of iterations. The only exception is the case $\lambda = \infty$, in which the rank of variables does not vary with iterations.

Now we are ready to see that the correlation learning which was introduced in Section 2.1 is a specific case of the ITRRS since componentwise regression is a specific case of ridge regression with an infinite regularization parameter. The ITRRS provides a very nice technical tool for understanding how fast the dimension p can grow comparing with sample size n and how the final model size d can be chosen while the sure screening property still holds for correlation learning. The question of whether the ITRRS has the sure screening property as well as how the tuning parameters γ and δ should be chosen will be answered by theorem 3 in Section 5.

The number of steps in the ITRRS depends on the choice of $\delta \in (0, 1)$. We shall see in theorem 3 that δ cannot be chosen too small, which means that there should not be too many iteration steps in the ITRRS. This is due to the cumulation of the probability errors of missing some important variables over the iterations. In particular, the backward stepwise deletion regression which deletes one variable each time in the ITRRS until the number of remaining variables drops to below the sample size might not work in general as it requires $p-d$ iterations. When p is of exponential order, even though the probability of deleting some important predictors in each step of deletion is exponentially small, the cumulative error in exponential order of operations may not be negligible.

2.3. Connections with other dimensionality reduction methods

As pointed out before, SIS uses the marginal information of correlation to perform dimensionality reduction. The idea of using marginal information to deal with high dimensionality has also appeared independently in Huang *et al.* (2008) who proposed the use of marginal bridge estimators to select variables for sparse high dimensional regression models. We now look at SIS in the context of classification, in which the idea of independent screening appears natural and has been widely used.

The problem of classification can be regarded as a specific case of the regression problem with response variable taking discrete values such as ± 1. For high dimensional problems like tumour classification using gene expression or proteomics data, it is not wise to classify the data by using the full feature space because of accumulation of noise and interpretability. This is well demonstrated both theoretically and numerically in Fan and Fan (2008). In addition, many of the features come into play through linkage to the important predictors (see, for

example, Fig. 1). Therefore feature selection is important for high dimensional classification. How to select important features effectively and how many of them to include are two tricky questions to answer. Various feature selection procedures have been proposed in the literature to improve the classification power in the presence of high dimensionality. For example, Tibshirani *et al.* (2002) introduced the nearest shrunken centroids method, and Fan and Fan (2008) propose the features annealed independence rules procedure. Theoretical justifications for these methods are given in Fan and Fan (2008).

SIS can readily be used to reduce the feature space. Now suppose that we have n_1 samples from class 1 and n_2 samples from class -1. Then the componentwise regression estimator (2) becomes

$$\boldsymbol{\omega} = \sum_{Y_i=1} \mathbf{x}_i - \sum_{Y_i=-1} \mathbf{x}_i. \tag{9}$$

Written more explicitly, the jth component of the p-vector $\boldsymbol{\omega}$ is

$$\omega_j = (n_1 \bar{X}_{j,1} - n_2 \bar{X}_{j,2})/(\text{standard deviation of the } j\text{th feature}),$$

by recalling that each covariate in equation (9) has been normalized marginally, where $\bar{X}_{j,1}$ is the sample average of the jth feature with class label '1' and $\bar{X}_{j,2}$ is the sample average of the jth feature with class label '-1'. When $n_1 = n_2$, ω_j is simply a version of the two-sample t-statistic except for a scaling constant. In this case, feature selection using SIS is the same as that using two-sample t-statistics. See Fan and Fan (2008) for a theoretical study of the sure screening property in this context.

Two-sample t-statistics are commonly used in feature selection for high dimensional classification problems such as in a significance analysis of gene selection in microarray data analysis (see, for example, Storey and Tibshirani (2003) and Fan and Ren (2006)) as well as in the nearest shrunken centroids method of Tibshirani *et al.* (2002). Therefore SIS is an insightful and natural extension of this widely used technique. Although not directly applicable, the sure screening property of SIS in theorem 1 after some adaptation gives theoretical justification for the nearest shrunken centroids method. See Fan and Fan (2008) for a sure screening property.

By using SIS we can single out the important features and thus reduce significantly the feature space to a much lower dimensional space. From this point on, many methods such as the linear discrimination rule or the naive Bayes rule can be applied to conduct the classification in the reduced feature space. This idea will be illustrated on a leukaemia data set in Section 3.3.3.

3. Sure independence screening based model selection techniques

3.1. Estimation and model selection in the reduced feature space

As shown later in theorem 1 in Section 5, with correlation learning, we can shrink the full model $\{1, \ldots, p\}$ straightforwardly and accurately down to a submodel $\mathcal{M} = \mathcal{M}_\gamma$ with size $d = [\gamma n] = o(n)$. Thus the original problem of estimating the sparse p-vector $\boldsymbol{\beta}$ in model (1) reduces to estimating a sparse d-vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^{\mathrm{T}}$ that is based on the now much smaller submodel $\mathcal{M}$, namely,

$$\mathbf{y} = \mathbf{X}_{\mathcal{M}} \boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{10}$$

where $\mathbf{X}_{\mathcal{M}} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^{\mathrm{T}}$ denotes an $n \times d$ submatrix of $\mathbf{X}$ that is obtained by extracting its columns corresponding to the indices in $\mathcal{M}$. Apparently SIS can speed up variable selection dramatically when the original dimension p is ultrahigh.

Now we briefly review several well-developed moderate dimensional techniques that can be applied to estimate the d-vector $\boldsymbol{\beta}$ in equation (10) at the scale of d that is comparable with n. These methods include the SCAD method in Fan and Li (2001) and Fan and Peng (2004), the adaptive lasso in Zou (2006) and the Dantzig selector in Candes and Tao (2007), among others.

3.1.1. Penalized least squares and smoothly clipped absolute deviation

Penalization is commonly used in variable selection. Fan and Li (2001, 2006) have given a comprehensive overview of feature selection and a unified framework based on the penalized likelihood approach to the problem of variable selection. They considered the PLS problem

$$l(\boldsymbol{\beta}) = \frac{1}{2n}\sum_{i=1}^{n}(Y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})^2 + \sum_{j=1}^{d} p_{\lambda_j}(|\beta_j|), \tag{11}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^{\mathrm{T}} \in \mathbf{R}^d$ and $p_{\lambda_j}(\cdot)$ is a penalty function indexed by a regularization parameter λ_j. Variation of the regularization parameters across the predictors allows us to incorporate some prior information. For example, we may want to keep certain important predictors in the model and to choose not to penalize their coefficients. The regularization parameters λ_j can be chosen, for instance, by cross-validation (see, for example, Breiman (1996) and Tibshirani (1996)). A unified and effective algorithm for optimizing penalized likelihood, which is called the local quadratic approximation, was proposed in Fan and Li (2001) and well studied in Hunter and Li (2005). In particular, the local quadratic approximation can be employed to minimize the above PLS problem. In our implementation, we choose $\lambda_j = \lambda$ and select λ by the Bayesian information criterion.

An alternative and effective algorithm to minimize the PLS problem (11) is the local linear approximation that was proposed by Zou and Li (2008). With the local linear approximation, problem (11) can be cast as a sequence of penalized L_1 regression problems so that the LARS (Efron *et al.*, 2004) or other algorithms can be employed. More explicitly, given the estimate $\{\hat{\beta}_j^{(k)}, j = 1, \ldots, d\}$ at the kth iteration, instead of minimizing probelm (11), we minimize

$$\frac{1}{2n}\sum_{i=1}^{n}(Y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})^2 + \sum_{j=1}^{d} w_j^{(k)}|\beta_j|, \tag{12}$$

which after adding the constant term $\Sigma_{j=1}^{d} p_{\lambda_j}(|\hat{\beta}_j^{(k)}|)$ is a local linear approximation to $l(\boldsymbol{\beta})$ in problem (11), where $w_j^{(k)} = |p'_{\lambda_j}(|\hat{\beta}_j^{(k)}|)|$. Problem (12) is a convex problem and can be solved by LARS and other algorithms such as those in Friedman *et al.* (2007) and Meier *et al.* (2008). In this sense, the PLS problem (11) can be regarded as a family of iteratively reweighted penalized L_1-problems and the function $p'_{\lambda}(\cdot)$ dictates the amount of penalty at each location. The emphasis on non-concave penalty functions by Fan and Li (2001) is to ensure that the penalty decreases to zero as $|\hat{\beta}_j^{(k)}|$ becomes large. This reduces unnecessary biases of the penalized likelihood estimator, leading to the oracle property in Fan and Li (2001). Fig. 2 depicts how the SCAD function is approximated locally by a linear or quadratic function and the derivative functions $p'_{\lambda}(\cdot)$ for some commonly used penalty functions. When the initial value $\boldsymbol{\beta} = 0$, the first-step estimator is indeed the lasso so the implementation of SCAD can be regulated as an iteratively reweighted penalized penalized L_1-estimator with the lasso as an initial estimator. An advantage of the SCAD penalty is that zero is not an absorbing state. See Section 6 for further discussion of the choice of initial values $\{\hat{\beta}_j^{(0)}, j = 1, \ldots, d\}$.

The PLS problem (11) depends on the choice of penalty function $p_{\lambda_j}(\cdot)$. Commonly used penalty functions include the l_p-penalty, $0 \leqslant p \leqslant 2$, the non-negative garrotte in Breiman (1995),

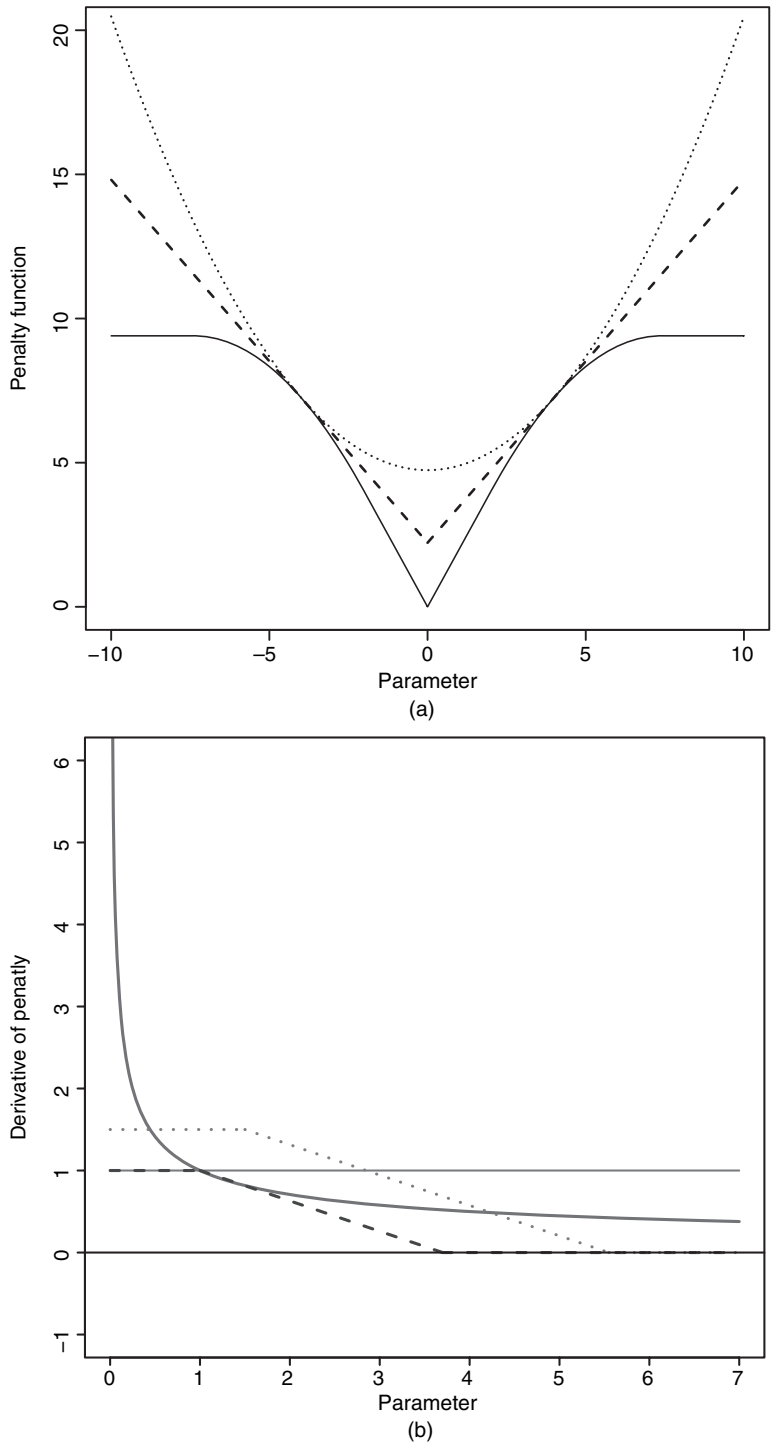

Fig. 2. (a) SCAD penalty (———) and its local linear (-------) and quadratic (·····) approximations and (b) $p'_\lambda(\cdot)$ for penalized L_1 (———), SCAD with $\lambda = 1$ (-------) and $\lambda = 1.5$ (·····) and adaptive lasso (———) with $\gamma = 0.5$

the SCAD penalty in Fan (1997) and the minimax concave penalty in Zhang (2007) (see below for a definition). In particular, l_1-penalized least squares is called the lasso in Tibshirani (1996). In seminal papers, Donoho and Huo (2001) and Donoho and Elad (2003) showed that the penalized l_0-solution can be found by the penalized l_1-method when the problem is sufficiently sparse, which implies that the best subset regression can be found by using penalized l_1-regression. Antoniadis and Fan (2001) proposed PLS for wavelets denoising with irregular designs. Fan and Li (2001) advocated penalty functions with three properties: sparsity, unbiasedness and continuity. More details on characterization of these three properties can be found in Fan and Li (2001) and Antoniadis and Fan (2001). For penalty functions, they showed that singularity at the origin is a necessary condition to generate sparsity and folded concavity is required to reduce the estimation bias. It is well known that the l_p-penalty with $0 \leqslant p < 1$ does not satisfy the continuity condition, the l_p-penalty with $p > 1$ does not satisfy the sparsity condition and the l_1-penalty (lasso) has sparsity and continuity but generates estimation bias, as demonstrated in Fan and Li (2001), Zou (2006) and Meinshausen (2007).

Fan (1997) proposed a continuously differentiable penalty function called the SCAD penalty, which is defined by

$$p'_\lambda(|\beta|) = \lambda \left\{ I(|\beta| \leqslant \lambda) + \frac{(a\lambda - |\beta|)_+}{(a-1)\lambda} I(|\beta| > \lambda) \right\} \qquad \text{for some } a > 2. \tag{13}$$

Fan and Li (2001) suggested using $a = 3.7$. This function has similar features to the penalty function $\lambda|\beta|/(1+|\beta|)$ that was advocated in Nikolova (2000). The minimum concave penalty in Zhang (2007) translates the flat part of the derivative of the SCAD to the origin and is given by

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

which minimizes the maximum of the concavity. The SCAD penalty and minimum concave penalty satisfy the above three conditions simultaneously. We shall show in theorem 5 in Section 5 that SIS followed by SCAD enjoys oracle properties.

3.1.2. Adaptive lasso

The lasso in Tibshirani (1996) has been widely used because of its convexity. It, however, generates estimation bias. This problem was pointed out in Fan and Li (2001) and formally shown in Zou (2006) even in a finite parameter setting. To overcome this bias problem, Zou (2006) proposed an adaptive lasso and Meinshausen (2007) proposed a relaxed lasso.

The idea in Zou (2006) is to use an adaptively weighted l_1-penalty in the PLS problem (11). Specifically, he introduced the penalization term

$$\lambda \sum_{j=1}^{d} \omega_j |\beta_j|,$$

where $\lambda \geqslant 0$ is a regularization parameter and $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_d)^{\mathrm{T}}$ is a known weight vector. He further suggested the use of the weight vector $\hat{\boldsymbol{\omega}} = 1/|\hat{\boldsymbol{\beta}}|^\gamma$, where $\gamma \geqslant 0$, the power is understood componentwise and $\hat{\boldsymbol{\beta}}$ is a root n consistent estimator. Its connections with the family of folded concavity PLS is apparent from problem (12) and Fig. 2. However, zero is an absorbing state of the adaptive lasso.

The case $\gamma = 1$ is closely related to the non-negative garrotte in Breiman (1995). Zou (2006) also showed that the adaptive lasso can be solved by the LARS algorithm, which was proposed in Efron *et al.* (2004). Using the same finite parameter set-up as that in Knight and Fu (2000), Zou

(2006) established that the adaptive lasso has oracle properties as long as the tuning parameter is chosen in a way such that $\lambda/n^{1/2} \to 0$ and $\lambda n^{(\gamma-1)/2} \to \infty$ as $n \to \infty$.

3.1.3. *Dantzig selector*

The Dantzig selector was proposed in Candes and Tao (2007) to recover a sparse high dimensional parameter vector in the linear model. Adapted to the setting in equation (10), it is the solution $\hat{\boldsymbol{\beta}}_{\mathrm{DS}}$ to the l_1-regularization problem

$$\min_{\boldsymbol{\zeta} \in \mathbf{R}^d} (\|\boldsymbol{\zeta}\|_1) \qquad \text{subject to } \|\mathbf{X}_{\mathcal{M}}^{\mathrm{T}} \mathbf{r}\|_\infty \leqslant \lambda_d \sigma, \tag{14}$$

where $\lambda_d > 0$ is a tuning parameter, $\mathbf{r} = \mathbf{y} - \mathbf{X}_{\mathcal{M}} \boldsymbol{\zeta}$ is an n-vector of the residuals and $\|\cdot\|_1$ and $\|\cdot\|_\infty$ denote the l_1- and l_∞-norms respectively. They pointed out that the above convex optimization problem can easily be recast as a linear program:

$$\min\Big(\sum_{i=1}^{d} u_i\Big) \qquad \text{subject to } -\mathbf{u} \leqslant \boldsymbol{\zeta} \leqslant \mathbf{u} \text{ and } -\lambda_d \sigma \mathbf{1} \leqslant \mathbf{X}_{\mathcal{M}}^{\mathrm{T}}(\mathbf{y} - \mathbf{X}_{\mathcal{M}} \boldsymbol{\zeta}) \leqslant \lambda_d \sigma \mathbf{1},$$

where the optimization variables are $\mathbf{u} = (u_1, \ldots, u_d)^{\mathrm{T}}$ and $\boldsymbol{\zeta} \in \mathbf{R}^d$, and $\mathbf{1}$ is a d-vector of 1s.

We shall show in theorem 4 in Section 5 that an application of SIS followed by the Dantzig selector can achieve the ideal risk up to a factor of $\log(d)$ with $d < n$, rather than the original $\log(p)$. In particular, if dimension p is growing exponentially fast, i.e. $p = \exp\{O(n^\xi)\}$ for some $\xi > 0$, then a direct application of the Dantzig selector results in a loss of a factor $O(n^\xi)$ which could be too large to be acceptable. In contrast, with the dimensionality first reduced by SIS the loss is now merely of a factor $\log(d)$, which is less than $\log(n)$.

3.2. *Sure independence screening based model selection methods*

For the problem of ultrahigh dimensional variable selection, we propose first to apply a sure screening method such as SIS to reduce the dimensionality from p to a relatively large scale d, say, below sample size n. Then we use a lower dimensional model selection method such as SCAD, the Dantzig selector, lasso, or adaptive lasso. We call SIS followed by SCAD and the Dantzig selector SIS–SCAD and SIS–DS respectively for short in the paper. In some situations, we may want to reduce further the model size down to $d' < d$ by using a method such as the Dantzig selector along with hard thresholding or the lasso with suitable tuning, and finally to choose a model with a more refined method such as SCAD or the adaptive lasso. In the paper these two methods will be referred to as SIS–DS–SCAD and SIS–DS–AdaLasso respectively for simplicity. Fig. 3 shows a schematic diagram of these approaches.

The idea of SIS makes it feasible to do model selection with ultrahigh dimensionality and speeds up variable selection drastically. It also makes the model selection problem efficient and modular. SIS can be used in conjunction with any model selection technique including Bayesian methods (see, for example, George and McCulloch (1997)) and the lasso. We did not include

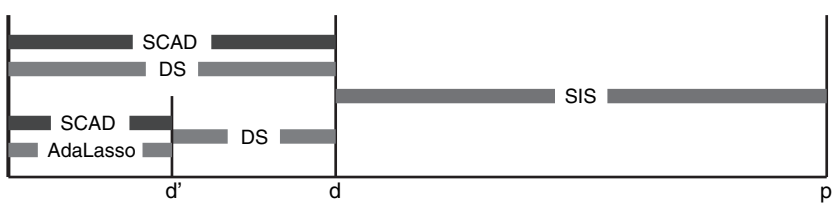

Fig. 3. Methods of model selection with ultrahigh dimensionality

an SIS–lasso method for numerical studies because of the approximate equivalence between the Dantzig selector and the lasso (Bickel *et al.*, 2008; Meinshausen *et al.*, 2007).

3.3. Numerical studies

To study the performance of the SIS-based model selection methods that were proposed above, we now present two simulations and one real data example.

3.3.1. Simulation I: 'independent features'

For the first simulation, we used the linear model (1) with IID standard Gaussian predictors and Gaussian noise with standard deviation $\sigma = 1.5$. We considered two such models with $(n, p) = (200, 1000)$ and $(n, p) = (800, 20000)$. The sizes s of the true models, i.e. the numbers of non-zero coefficients, were chosen to be 8 and 18, and the non-zero components of the p-vectors β were randomly chosen as follows. We set $a = 4\ \log(n)/n^{1/2}$ and $5\ \log(n)/n^{1/2}$ respectively and picked non-zero coefficients of the form $(-1)^u(a + |z|)$ for each model, where u was drawn from a Bernoulli distribution with parameter 0.4 and z was drawn from the standard Gaussian distribution. In particular, the l_2-norms $\|\beta\|$ of the two simulated models are 6.795 and 8.908. For each model we simulated 200 data sets. Even with IID standard Gaussian predictors, these settings are non-trivial since there is non-negligible sample correlation between the predictors, which reflects the difficulty of high dimensional variable selection. As evidence, we report in Fig. 4 the distributions of the maximum absolute sample correlation when $n = 200$ and $p = 1000$ and $p = 5000$. It reveals significant sample correlation between the predictors. The multiple canonical correlation between two groups of predictors can be much larger.

To estimate the sparse p-vectors β, we employed six methods: the Dantzig selector using a primal dual algorithm, the lasso using the LARS algorithm and the SIS–SCAD, SIS–DS, SIS–DS–SCAD and SIS–DS–AdaLasso methods (see Fig. 3). For the SIS–SCAD and SIS–DS methods, we chose $d = [n/\log(n)]$ and, for the last two methods, we chose $d = n - 1$ and $d' = [n/\log(n)]$ and in the middle step the Dantzig selector was used to reduce the model size further from d to d' by choosing variables with the d' largest componentwise magnitudes of the estimated d-vector (see Fig. 3).

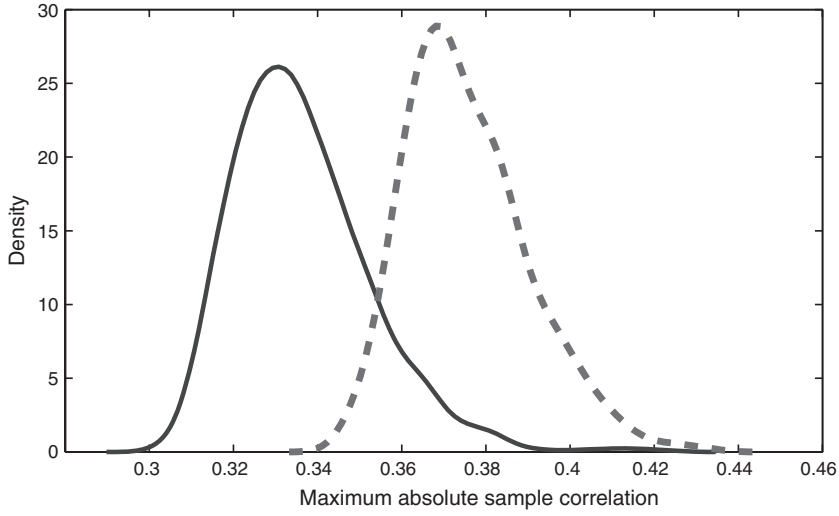

Fig. 4. Distributions of the maximum absolute sample correlation when $n = 200$ and $p = 1000$ (———) and $n = 200$ and $p = 5000$ (- - - - -)

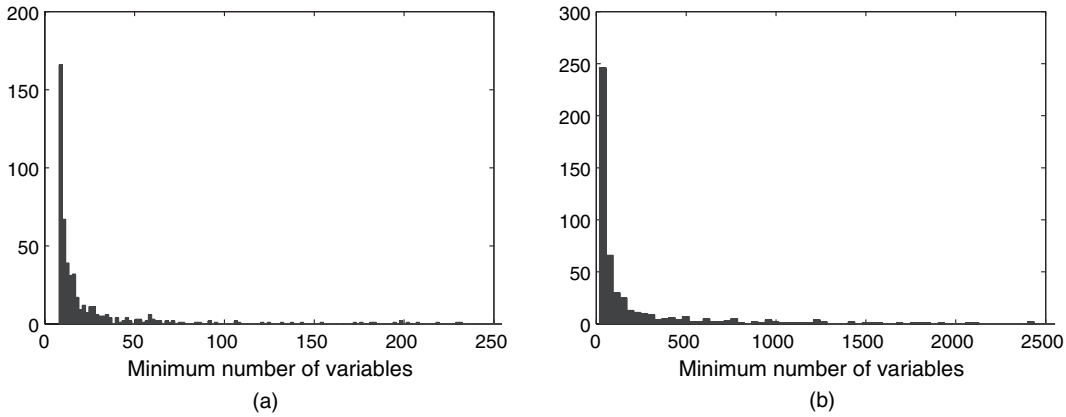

Fig. 5. Distribution of the minimum number of selected variables that is required to include the true model by using SIS when (a) $n = 200$ and $p = 1000$ and (b) $n = 800$ and $p = 20\,000$ in simulation I

Table 1. Results of simulation I: medians of the selected model sizes and estimation errors (in parentheses)

p	*Results for the following methods:*					
	Dantzig selector	*Lasso*	*SIS–SCAD*	*SIS–DS*	*SIS–DS–SCAD*	*SIS–DS–AdaLasso*
1000	10^3	62.5	15	37	27	34
	(1.381)	(0.895)	(0.374)	(0.795)	(0.614)	(1.269)
20000	—	—	37	119	60.5	99
	—	—	(0.288)	(0.732)	(0.372)	(1.014)

The simulation results are summarized in Fig. 5 and Table 1. Fig. 5, which was produced on the basis of 500 simulations, depicts the distribution of the minimum number of selected variables, i.e. the selected model size, that is required for sure screening by using SIS. It shows clearly that in both settings it is safe to shrink the full model down to a submodel of size $[n/\log(n)]$ with SIS, which is consistent with the sure screening property of SIS that is shown in theorem 1 in Section 5. For example, for the case $n = 200$ and $p = 1000$, reducing the model size to 50 includes the variables in the true model with high probability and, for the case $n = 800$ and $p = 20\,000$, it is safe to reduce the dimension to about 500. For each of the above six methods, we report in Table 1 the median of the selected model sizes and the median of the estimation errors $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|$ in l_2-norm. Four entries of Table 1 are missing owing to the limited computing power and software that were used. In comparison, SIS reduces the computational burden significantly.

From Table 1 we see that the Dantzig selector gives non-sparse solutions and the lasso using cross-validation for selecting its tuning parameter produces large models. This can be because the biases in the lasso require a small bandwidth in cross-validation, whereas a small bandwidth results in lack of 'sparsistency' in the terminology of Ravikumar *et al.* (2007). This has also been observed and demonstrated in work by Lam and Fan (2007) in the context of estimating sparse covariance or precision matrices. We should point out here that a variation of the Dantzig selector, the Gauss–Dantzig selector in Candes and Tao (2007), should yield much smaller

models, but for simplicity we did not include it in our simulation. Of all the methods, SIS–SCAD performs the best and generates much smaller and more accurate models. It is clear to see that SCAD gives more accurate estimates than the adaptive lasso in view of the estimation errors. Also, SIS followed by the Dantzig selector improves the accuracy of estimation over using the Dantzig selector alone, which is in line with our theoretical result.

3.3.2. Simulation II: 'dependent' features

For the second simulation, we used similar models to those in simulation I except that the predictors are now correlated with each other. We considered three models with $(n, p, s) = (200, 1000, 5)$, (200, 1000, 8) and (800, 20000, 14), where s denotes the size of the true model, i.e. the number of non-zero coefficients. The three p-vectors $\boldsymbol{\beta}$ were generated in the same way as in simulation I. We set $(\sigma, a) = (1, 2\log(n)/n^{1/2})$, $(1.5, 4\log(n)/n^{1/2})$, $(2, 4\log(n)/n^{1/2})$. In particular, the l_2-norms $\|\boldsymbol{\beta}\|$ of the three simulated models are 3.304, 6.795 and 7.257. To introduce correlation between predictors, we first used MATLAB function `sprandsym` to generate randomly an $s \times s$ symmetric positive definite matrix $\mathbf{A}$ with condition number $n^{1/2}/\log(n)$ and drew samples of s predictors $X_1, \ldots, X_s$ from $\mathcal{N}(\mathbf{0}, \mathbf{A})$. Then we took $Z_{s+1}, \ldots, Z_p \sim \mathcal{N}(\mathbf{0}, I_{p-s})$ and defined the remaining predictors as $X_i = Z_i + rX_{i-s}$, $i = s+1, \ldots, 2s$, and $X_i = Z_i + (1-r)X_1$, $i = 2s+1, \ldots, p$, with $r = 1 - 4\log(n)/p$, $1 - 5\log(n)/p$ and $1 - 5\log(n)/p$. For each model we simulated 200 data sets.

We applied the same six methods as those in simulation I to estimate the sparse p-vectors $\boldsymbol{\beta}$. For the SIS–SCAD and SIS–DS methods, we chose $d = [\frac{3}{2}n/\log(n)]$, $[\frac{3}{2}n/\log(n)]$ and $[n/\log(n)]$, and, for the last two methods, we chose $d = n - 1$ and $d' = [\frac{3}{2}n/\log(n)]$, $[\frac{3}{2}n/\log(n)]$ and $[n/\log(n)]$. The simulation results are similarly summarized in Fig. 6 (which is based on 500 simulations) and Table 2. Similar conclusions to those from simulation I can be drawn. As in simulation I, for simplicity we did not include the Gauss–Dantzig selector. It is interesting to observe that, in the first setting here, the lasso gives large models and its estimation errors are noticeable compared with the norm of the true coefficient vector $\boldsymbol{\beta}$.

3.3.3. Leukaemia data analysis

We also applied SIS to select features for the classification of a leukaemia data set. The leukaemia data from high density Affymetrix oligonucleotide arrays have previously been analysed in Golub *et al.* (1999) and are available from `http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi`. There are 7129 genes and 72 samples from two classes: 47 in class ALL (acute lymphocytic leukaemia) and 25 in class AML (acute mylogenous leukaemia).

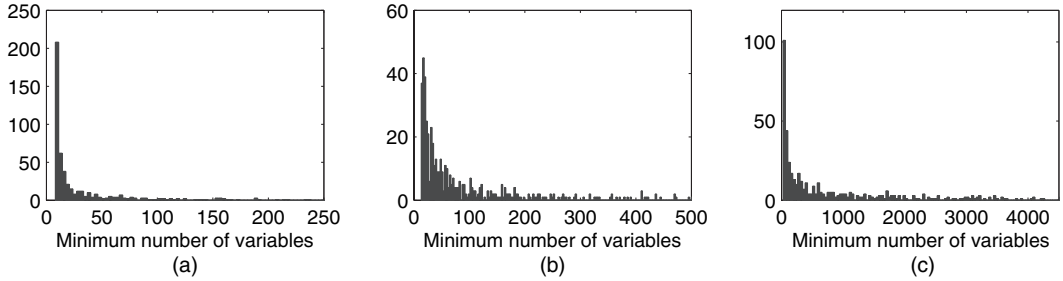

Fig. 6. Distribution of the minimum number of selected variables that is required to include the true model by using SIS when (a) $n = 200$, $p = 1000$ and $s = 5$, (b) $n = 200$, $p = 1000$ and $s = 8$ and (c) $n = 800$, $p = 20\,000$ and $s = 8$ in simulation II

Table 2. Results of simulation II: medians of the selected model sizes and estimation errors (in parentheses)

p	*Results for the following methods:*					
	Dantzig selector	*Lasso*	*SIS–SCAD*	*SIS–DS*	*SIS–DS–SCAD*	*SIS–DS–AdaLasso*
1000	10^3	91	21	56	27	52
($s=5$)	(1.256)	(1.257)	(0.331)	(0.727)	(0.476)	(1.204)
($s=8$)	10^3	74	18	56	31.5	51
	(1.465)	(1.257)	(0.458)	(1.014)	(0.787)	(1.824)
20000	—	—	36	119	54	86
	—	—	(0.367)	(0.986)	(0.743)	(1.762)

Table 3. Classification errors in the leukaemia data set

Method	*Training error*	*Test error*	*Number of genes*
SIS–SCAD–LD	0/38	1/34	16
SIS–SCAD–NB	4/38	1/34	16
Nearest shrunken centroids	1/38	2/34	21

Among those 72 samples, 38 (27 in class ALL and 11 in class AML) of them were set as the training sample and the remaining 34 (20 in class ALL and 14 in class AML) of them were set to be the test sample.

We used the two methods SIS–SCAD–LD and SIS–SCAD–NB that will be introduced below to carry out the classification. For each method, we first applied SIS to select $d=[2n/\log(n)]$ genes with $n=38$ the training sample size that was chosen above and then used SCAD to obtain a family of models indexed by the regularization parameter λ. Here, we should point out that our classification results are not very sensitive to the choice of d as long as it is not too small. There are certainly many ways to tune the regularization parameter λ. For simplicity, we chose a λ that produces a model with size equal to the optimal number of features that was determined by the features annealed independence rules procedure in Fan and Fan (2008). 16 genes were picked up by their approach. Now we selected 16 genes and obtained a linear model with size 16 by using the SIS–SCAD method. Finally, the SIS–SCAD–LD method directly used the above linear discrimination rule to do classification, and the SIS–SCAD–NB method applied the naive Bayes rule to the resulting 16-dimensional feature space.

The classification results of the SIS–SCAD–LD, SIS–SCAD–NB and nearest shrunken centroids method in Tibshirani *et al.* (2002) are shown in Table 3. The results of the nearest shrunken centroids method were extracted from Tibshirani *et al.* (2002). The SIS–SCAD–LD and SIS–SCAD–NB methods both chose 16 genes and made one test error with training errors 0 and 4 respectively, whereas the nearest shrunken centroids method picked up 21 genes and made one training error and two test errors.

4. Extensions of sure independence screening

Like model building in linear regression, there are many variations of the implementation of correlation learning. This section discusses some extensions of SIS to enhance its methodological power. In particular, iterative SIS (ISIS) is proposed to overcome some weak points of SIS. The methodological power of ISIS is illustrated by three simulated examples.

4.1. Some extensions of correlation learning

The key idea of SIS is to apply a single componentwise regression. Three potential issues, however, might arise with this approach. First, some unimportant predictors that are highly correlated with the important predictors can have higher priority for being selected by SIS than other important predictors that are relatively weakly related to the response. Second, an important predictor that is marginally uncorrelated but jointly correlated with the response cannot be picked by SIS and thus will not enter the estimated model. Third, the issue of collinearity between predictors adds difficulty to the problem of variable selection. These three issues will be addressed in the extensions of SIS below, which allow us to use more fully the joint information of the covariates rather than just the marginal information in variable selection.

4.1.1. Iterative sure independence screening: iterative correlation learning

It will be shown that when the model assumptions are satisfied, which excludes basically the three aforementioned problems, SIS can accurately reduce the dimensionality from ultrahigh to a moderate scale, say, below the sample size. But, when those assumptions fail, it could happen that SIS would miss some important predictors. To overcome this problem, we propose below ISIS to enhance the methodological power. It is an iterative application of the SIS approach to variable selection. The essence is to apply iteratively a large-scale variable screening followed by a moderate-scale careful variable selection.

ISIS works as follows. In the first step, we select a subset of k_1 variables $\mathcal{A}_1 = \{X_{i_1}, \ldots, X_{i_{k_1}}\}$ using an SIS-based model selection method such as the SIS–SCAD or SIS–lasso methods. These variables were selected, using SCAD or the lasso, on the basis of the joint information of $[n/\log(n)]$ variables that survive after correlation screening. Then we have an n-vector of residuals from regressing the response Y over $X_{i_1}, \ldots, X_{i_{k_1}}$. In the next step, we treat those residuals as the new responses and apply the same method as in the previous step to the remaining $p - k_1$ variables, which results in a subset of k_2 variables $\mathcal{A}_2 = \{X_{j_1}, \ldots, X_{j_{k_2}}\}$. We remark that fitting the residuals from the previous step on $\{X_1, \ldots, X_p\} \backslash \mathcal{A}_1$ can significantly weaken the priority of those unimportant variables that are highly correlated with the response through their associations with $X_{i_1}, \ldots, X_{i_{k_1}}$, since the residuals are uncorrelated with those selected variables in $\mathcal{A}_1$. This helps to solve the first issue. It also makes those important predictors that are missed in the previous step possible to survive, which addresses the second issue above. In fact, after variables in $\mathcal{A}_1$ enter the model, those that are marginally weakly correlated with Y purely due to the presence of variables in $\mathcal{A}_1$ should now be correlated with the residuals. We can keep on doing this until we obtain l disjoint subsets $\mathcal{A}_1, \ldots, \mathcal{A}_l$ whose union $\mathcal{A} = \cup_{i=1}^{l} \mathcal{A}_i$ has a size d, which is less than n. In practical implementation, we can choose, for example, the largest l such that $|\mathcal{A}| < n$. From the selected features in $\mathcal{A}$, we can choose the features by using a moderate scale method such as SCAD, the lasso or the Dantzig selector.

For the problem of ultrahigh dimensional variable selection, we now have ISIS-based model selection methods which are extensions of SIS-based model selection methods. Applying a moderate dimensional method such as SCAD, the Dantzig selector, lasso or adaptive lasso to $\mathcal{A}$ will produce a model that is very close to the true sparse model $\mathcal{M}_*$. The idea of ISIS is somewhat

related to the boosting algorithm (Freund and Schapire, 1997). In particular, if SIS is used to select only one variable at each iteration, i.e. $|\mathcal{A}_i| = 1$, ISIS is equivalent to a form of matching pursuit or a greedy algorithm for variable selection (Barron *et al.*, 2008).

4.1.2. Grouping and transformation of the input variables

Grouping the input variables is often used in various problems. For instance, we can divide the pool of p variables into disjoint groups each with five variables. The idea of variable screening via SIS can be applied to select a small number of groups. In this way there is less chance of missing the important variables by taking advantage of the joint information among the predictors. Therefore a more reliable model can be constructed.

A notorious difficulty of variable selection lies in the collinearity between the covariates. Effective ways to rule out those unimportant variables that are highly correlated with the important variables are being sought after. A good idea is to transform the input variables. Two possible ways stand out in this regard. One is subject-related transformation and the other is statistical transformation.

Subject-related transformation is a useful tool. In some cases, a simple linear transformation of the input variables can help to weaken correlation between the covariates. For example, in somatotype studies common sense tells us that predictors such as the weights w_1, w_2 and w_3 at 2, 9 and 18 years are positively correlated. We could directly use w_1, w_2 and w_3 as the input variables in a linear regression model, but a better way of model selection in this case is to use less correlated predictors such as $(w_1, w_2 - w_1, w_3 - w_2)^{\mathrm{T}}$, which is a linear transformation of $(w_1, w_2, w_3)^{\mathrm{T}}$ that specifies the changes of the weights instead of the weights themselves. Another important example is financial time series such as the prices of the stocks or interest rates. Differencing can significantly weaken the correlation between those variables.

Methods of statistical transformation include an application of a clustering algorithm such as the hierarchical clustering or k-means algorithm using the correlation metrics first to group variables into highly correlated groups and then apply sparse principal components analysis to construct weakly correlated predictors. Now those weakly correlated predictors from each group can be regarded as the new covariates and an SIS-based model selection method can be employed to select them.

The statistical techniques that we introduced above can help to identify the important features and thus to improve the effectiveness of the SIS-based model selection strategy. Introduction of non-linear terms and transformation of variables can also be used to reduce the modelling biases of linear models. Ravikumar *et al.* (2007) introduced sparse additive models to deal with non-linear feature selection.

4.2. Numerical evidence

To study the performance of the ISIS method that was proposed above, we now present three simulated examples. The aim is to examine the extent to which ISIS can improve SIS in the situation where the conditions of SIS fail. We evaluate the methods by counting the frequencies that the selected models include all the variables in the true model, namely the ability of correctly screening unimportant variables.

4.2.1. Simulated example I

For the first simulated example, we used a linear model

$$Y = 5X_1 + 5X_2 + 5X_3 + \varepsilon,$$

Table 4. Results of simulated example I: accuracy of SIS, the lasso and ISIS in including the true model $\{X_1, X_2, X_3\}$

p	n	*Method*	*Results for the following values of* ρ*:*			
			$\rho=0$	$\rho=0.1$	$\rho=0.5$	$\rho=0.9$
100	20	SIS	0.755	0.855	0.690	0.670
		Lasso	0.970	0.990	0.985	0.870
		ISIS	1	1	1	1
	50	SIS	1	1	1	1
		Lasso	1	1	1	1
		ISIS	1	1	1	1
1000	20	SIS	0.205	0.255	0.145	0.085
		Lasso	0.340	0.555	0.556	0.220
		ISIS	1	1	1	1
	50	SIS	0.990	0.960	0.870	0.860
		Lasso	1	1	1	1
		ISIS	1	1	1	1
	70	SIS	1	0.995	0.97	0.97
		Lasso	1	1	1	1
		ISIS	1	1	1	1

where $X_1, \ldots, X_p$ are p predictors and $\varepsilon \sim N(0, 1)$ is noise that is independent of the predictors. In the simulation, a sample of $(X_1, \ldots, X_p)$ with size n was drawn from a multivariate normal distribution $N(0, \Sigma)$ whose covariance matrix $\Sigma = (\sigma_{ij})_{p\times p}$ has entries $\sigma_{ii} = 1$, $i = 1, \ldots, p$ and $\sigma_{ij} = \rho$, $i \neq j$. We considered 20 such models characterized by (p, n, ρ) with $p = 100$, 1000, $n = 20, 50, 70$ and $\rho = 0$, 0.1, 0.5, 0.9, and for each model we simulated 200 data sets.

For each model, we applied SIS and ISIS to select n variables and tested their accuracy in including the true model $\{X_1, X_2, X_3\}$. For ISIS, the SIS–SCAD method with $d = [n/\log(n)]$ was used at each step and we kept on collecting variables in those disjoint $\mathcal{A}_j$s until we obtained n variables (if there were more variables than needed in the final step, we included only those with the largest absolute coefficients). In Table 4, we report the percentages of SIS, lasso and ISIS that include the true model. All these three methods select $n-1$ variables, to make fair comparisons. It is clear that the collinearity (large value of ρ) and high dimensionality deteriorate the performance of SIS and the lasso, and the lasso outperforms SIS somewhat. However, when the sample size is 50 or more, the difference in performance is very small, but SIS has much less computational cost. In contrast, ISIS improves dramatically the performance of this simple SIS and lasso. Indeed, in this simulation, ISIS always picks all true variables. It can even have less computational cost than the lasso when the lasso is used in the implementation of ISIS.

4.2.2. Simulated example II

For the second simulated example, we used the same set-up as in example I except that ρ was fixed to be 0.5 for simplicity. In addition, we added a fourth variable X_4 to the model and the linear model is now

$$Y = 5X_1 + 5X_2 + 5X_3 - 15\rho^{1/2}X_4 + \varepsilon,$$

where $X_4 \sim N(0, 1)$ and has correlation $\rho^{1/2}$ with all the other $p-1$ variables. The way that X_4 was introduced is to make it uncorrelated with the response Y. Therefore, SIS cannot pick up the true model except by chance.

Table 5. Results of simulated example II: accuracy of SIS, the lasso and ISIS in including the true model $\{X_1, X_2, X_3, X_4\}$†

p	Method	Results for the following values of n:		
		$n=20$	$n=50$	$n=70$
100	SIS	0.025	0.490	0.740
	Lasso	0.000	0.360	0.915
	ISIS	1	1	1
1000	SIS	0.000	0.000	0.000
	Lasso	0.000	0.000	0.000
	ISIS	1	1	1

†$\rho = 0.5$.

Again we simulated 200 data sets for each model. In Table 5, we report the percentages of SIS, lasso and ISIS that include the true model of four variables. In this simulation example, SIS performs somewhat better than the lasso in variable screening, and ISIS outperforms significantly the simple SIS and lasso. In this simulation it always picks all true variables. This demonstrates that ISIS can effectively handle the second problem that was mentioned at the beginning of Section 4.1.

4.2.3. *Simulated example III*

For the third simulated example, we used the same set-up as in example II except that we added a fifth variable X_5 to the model and the linear model is now

$$Y = 5X_1 + 5X_2 + 5X_3 - 15\rho^{1/2}X_4 + X_5 + \varepsilon,$$

where $X_5 \sim N(0,1)$ and is uncorrelated with all the other $p-1$ variables. Again X_4 is uncorrelated with the response Y. The way that X_5 was introduced was to make it have a very small correlation with the response and in fact the variable X_5 has the same proportion of contribution to the response as the noise ε does. For this particular example, X_5 has weaker marginal correlation with Y than $X_6, \ldots, X_p$ and hence has a lower priority of being selected by SIS.

For each model we simulated 200 data sets. In Table 6, we report the accuracy in percentage of SIS, lasso and ISIS in including the true model. It is clear to see that ISIS can improve significantly over the simple SIS and lasso and always picks all true variables. This shows again that ISIS can pick up two difficult variables X_4 and X_5, which addresses simultaneously the second and third problem at the beginning of Section 4.

4.2.4. *Simulations I and II in Section 3.3 revisited*

Now let us go back to the two simulation studies that were presented in Section 3.3. For each of them, we applied the technique of ISIS with SCAD and $d = [n/\log(n)]$ to select $q = [n/\log(n)]$ variables. After that, we estimated the q-vector $\boldsymbol{\beta}$ by using SCAD. This method is referred to as ISIS–SCAD. We report in Table 7 the median of the model sizes selected and the median of the estimation errors $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|$ in l_2-norm. We can see clearly that ISIS improves over simple SIS. The improvements are more drastic for simulation II in which covariates are more correlated and the variable selections are more challenging.

Table 6. Results of simulated example III: accuracy of SIS, the lasso and ISIS in including the true model $\{X_1, X_2, X_3, X_4, X_5\}$†

p	*Method*	*Results for the following values of n:*		
		$n = 20$	$n = 50$	$n = 70$
100	SIS	0.000	0.285	0.645
	Lasso	0.000	0.310	0.890
	ISIS	1	1	1
1000	SIS	0.000	0.000	0.000
	Lasso	0.000	0.000	0.000
	ISIS	1	1	1

†$\rho = 0.5$.

Table 7. Simulations I and II in Section 3.3 revisited: medians of the model sizes selected and the estimation errors (in parentheses) for the ISIS–SCAD method

p	*Results for simulation I*		*Results for simulation II*
1000	13 (0.329)	($s = 5$)	11 (0.223)
		($s = 8$)	13.5 (0.366)
20000	31 (0.246)		27 (0.315)

5. Asymptotic analysis

We introduce an asymptotic framework below and present the sure screening property for both SIS and ITRRS as well as the consistency of the SIS-based model selection methods SIS–DS and SIS–SCAD.

5.1. Assumptions

Recall from model (1) that $Y = \Sigma_{i=1}^{p} \beta_i X_i + \varepsilon$. Throughout the paper we let $\mathcal{M}_* = \{1 \leqslant i \leqslant p : \beta_i \neq 0\}$ be the true sparse model with non-sparsity size $s = |\mathcal{M}_*|$ and define

$$\begin{aligned} \mathbf{z} &= \boldsymbol{\Sigma}^{-1/2}\mathbf{x}, \\ \mathbf{Z} &= \mathbf{X}\boldsymbol{\Sigma}^{-1/2}, \end{aligned} \tag{15}$$

where $\mathbf{x} = (X_1, \ldots, X_p)^{\mathrm{T}}$ and $\boldsymbol{\Sigma} = \mathrm{cov}(\mathbf{x})$. Clearly, the n rows of the transformed design matrix $\mathbf{Z}$ are IID copies of $\mathbf{z}$ which now has covariance matrix I_p. For simplicity, all the predictors $X_1, \ldots, X_p$ are assumed to be standardized to have mean 0 and standard deviation 1. Note that the design matrix $\mathbf{X}$ can be factored into $\mathbf{Z}\boldsymbol{\Sigma}^{1/2}$. Below we shall make assumptions on $\mathbf{Z}$ and $\boldsymbol{\Sigma}$ separately.

We denote by $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ the largest and smallest eigenvalues of a matrix respectively. For $\mathbf{Z}$, we are concerned with a concentration property of its extreme singular values as follows.

The random matrix $\mathbf{Z}$ is said to have the concentration property if there are some $c, c_1 > 1$ and $C_1 > 0$ such that the deviation inequality

$$P\{\lambda_{\max}(\tilde{p}^{-1}\tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^{\mathrm{T}}) > c_1 \text{ or } \lambda_{\min}(\tilde{p}^{-1}\tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^{\mathrm{T}}) < 1/c_1\} \leqslant \exp(-C_1 n) \tag{16}$$

holds for any $n \times \tilde{p}$ submatrix $\tilde{\mathbf{Z}}$ of $\mathbf{Z}$ with $cn < \tilde{p} \leqslant p$. We shall call it property C for short. Property C amounts to a distributional constraint on $\mathbf{z}$. Intuitively, it means that with large probability the n non-zero singular values of the $n \times \tilde{p}$ matrix $\tilde{\mathbf{Z}}$ are of the same order, which is reasonable since $\tilde{p}^{-1}\tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^{\mathrm{T}}$ will approach I_n as $\tilde{p} \to \infty$: the larger the $\tilde{p}$, the closer to I_n. It relies on random-matrix theory to derive the deviation inequality (16). In particular, property C holds when $\mathbf{x}$ has a p-variate Gaussian distribution (see Appendix A.7). We conjecture that it should be shared by a wide class of spherically symmetric distributions. For studies on the extreme eigenvalues and limiting spectral distributions, see, for example, Silverstein (1985), Bai and Yin (1993), Bai (1999), Johnstone (2001) and Ledoux (2001, 2005).

Some of the assumptions below are purely technical and serve only to provide theoretical understanding of the newly proposed methodology. We have no intent to make our assumptions the weakest possible.

Condition 1. $p > n$ and $\log(p) = O(n^{\xi})$ for some $\xi \in (0, 1-2\kappa)$, where κ is given by condition 3.

Condition 2. $\mathbf{z}$ has a spherically symmetric distribution and property C. Also, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ for some $\sigma > 0$.

Condition 3. $\operatorname{var}(Y) = O(1)$ and, for some $\kappa \geqslant 0$ and $c_2, c_3 > 0$,

$$\min_{i \in \mathcal{M}_*} |\beta_i| \geqslant \frac{c_2}{n^{\kappa}} \quad \text{and} \quad \min_{i \in \mathcal{M}_*} |\operatorname{cov}(\beta_i^{-1} Y, X_i)| \geqslant c_3.$$

As seen later, κ controls the rate of probability error in recovering the true sparse model. Although $b = \min_{i \in \mathcal{M}_*} |\operatorname{cov}(\beta_i^{-1} Y, X_i)|$ is assumed here to be bounded away from 0, our asymptotic study applies as well to the case where $b \to 0$ as $n \to \infty$. In particular, when the variables in $\mathcal{M}_*$ are uncorrelated, $b = 1$. This condition rules out the situation in which an important variable is marginally uncorrelated with Y, but jointly correlated with Y.

Condition 4. There are some $\tau \geqslant 0$ and $c_4 > 0$ such that

$$\lambda_{\max}(\boldsymbol{\Sigma}) \leqslant c_4 n^{\tau}.$$

This condition rules out the case of strong collinearity.

The largest eigenvalue of the population covariance matrix $\boldsymbol{\Sigma}$ is allowed to diverge as n grows. When there are many predictors, often their covariance matrix is block diagonal or nearly block diagonal under a suitable permutation of the variables. Therefore $\lambda_{\max}(\boldsymbol{\Sigma})$ usually does not grow too fast with n. In addition, condition 4 holds for the covariance matrix of a stationary time series (see Bickel and Levina (2004, 2008)). See also Grenander and Szegö (1984) for more details on the characterization of extreme eigenvalues of the covariance matrix of a stationary process in terms of its spectral density.

5.2. Sure screening property

Analysing the p-vector ω in equation (2) when $p > n$ is essentially difficult. The approach that we took is first to study the specific case with $\Sigma = I_p$ and then to relate the general case to the specific case.

Theorem 1 (accuracy of SIS). Under conditions 1–4, if $2\kappa + \tau < 1$ then there is some $\theta < 1 - 2\kappa - \tau$ such that, when $\gamma \sim cn^{-\theta}$ with $c > 0$, we have, for some $C > 0$,

$$P(\mathcal{M}_* \subset \mathcal{M}_\gamma) = 1 - O[\exp\{-Cn^{1-2\kappa}/\log(n)\}].$$

We should point out here that $s \leqslant [\gamma n]$ is implied by our assumptions as demonstrated in the technical proof. Theorem 1 shows that SIS has the sure screening property and can reduce from exponentially growing dimension p down to a relatively large scale $d = [\gamma n] = O(n^{1-\theta}) < n$ for some $\theta > 0$, where the reduced model $\mathcal{M} = \mathcal{M}_\gamma$ still contains all the variables in the true model with an overwhelming probability. In particular, we can choose the submodel size d to be $n - 1$ or $n/\log(n)$ for SIS if conditions 1–4 are satisfied.

Another interpretation of theorem 1 is that it requires the model size $d = [\gamma n] = n^{\theta^*}$ with $\theta^* > 2\kappa + \tau$ in order to have the sure screening property. The weaker the signal, the larger the κ and hence the larger the required model size is. Similarly, the more severe the collinearity, the larger the τ and the larger the required model size is. In this sense, the restriction that $2\kappa + \tau < 1$ is not needed, but $\kappa < \frac{1}{2}$ is needed since we cannot detect signals that are of smaller order than root n consistent. In the former case, there is no guarantee that θ^* can be taken to be smaller than 1.

The proof of theorem 1 depends on the iterative application of the following theorem, which demonstrates the accuracy of each step of ITRRS. We first describe the result of the first step of ITRRS. It shows that, as long as the ridge parameter λ is sufficiently large and the percentage of remaining variables δ is sufficiently large, the sure screening property is ensured with overwhelming probability.

Theorem 2 (asymptotic sure screening). Under conditions 1–4, if $2\kappa + \tau < 1$, $\lambda(p^{3/2}n)^{-1} \to \infty$, and $\delta n^{1-2\kappa-\tau} \to \infty$ as $n \to \infty$, then we have, for some $C > 0$,

$$P(\mathcal{M}_* \subset \mathcal{M}^1_{\delta,\lambda}) = 1 - O[\exp\{-Cn^{1-2\kappa}/\log(n)\}].$$

Theorem 2 reveals that, when the tuning parameters are chosen appropriately, with an overwhelming probability the submodel $\mathcal{M}^1_{\delta,\lambda}$ will contain the true model $\mathcal{M}_*$ and its size is an order n^θ (for some $\theta > 0$) lower than the original one. This property stimulated us to propose ITRRS.

Theorem 3 (accuracy of ITRRS). Let the assumptions of theorem 2 be satisfied. If $\delta n^\theta \to \infty$ as $n \to \infty$ for some $\theta < 1 - 2\kappa - \tau$, then successive applications of procedure (8) for k times results in a submodel $\mathcal{M}_{\delta,\lambda}$ with size $d = [\delta^k p] < n$ such that, for some $C > 0$,

$$P(\mathcal{M}_* \subset \mathcal{M}_{\delta,\lambda}) = 1 - O[\exp\{-Cn^{1-2\kappa}/\log(n)\}].$$

Theorem 3 follows from iterative application of theorem 2 k times, where k is the first integer such that $[\delta^k p] < n$. This implies that $k = O\{\log(p)/\log(n)\} = O(n^\xi)$. Therefore, the accumulated error probability, from the union bound, is still exponentially small with a possibility of a different constant C.

ITRRS has now been shown to have the sure screening property. As mentioned before, SIS is a specific case of ITRRS with an infinite regularization parameter and hence enjoys also the sure screening property.

Note that the number of steps in ITRRS depends on the choice of $\delta \in (0,1)$. In particular, δ cannot be too small or, equivalently, the number of iteration steps in ITRRS cannot be too large, owing to the accumulation of the probability errors of missing some important variables over the iterations. In particular, the stepwise deletion method which deletes one variable each time in ITRRS might not work since it requires $p-d$ steps of iterations, which may exceed the error bound in theorem 2.

5.3. Consistency of methods SIS–DS and SIS–SCAD

To study the property of the Dantzig selector, Candes and Tao (2007) introduced the notion of the uniform uncertainty principle on deterministic design matrices which essentially states that the design matrix obeys a 'restricted isometry hypothesis'. Specifically, let $\mathbf{A}$ be an $n \times d$ deterministic design matrix and for any subset $T \subset \{1, \ldots, d\}$. Denote by $\mathbf{A}_T$ the $n \times |T|$ submatrix of $\mathbf{A}$ that is obtained by extracting its columns corresponding to the indices in T. For any positive integer $S \leqslant d$, the S-restricted isometry constant $\delta_S = \delta_S(\mathbf{A})$ of $\mathbf{A}$ is defined to be the smallest quantity such that

$$(1-\delta_S)\|\mathbf{v}\|^2 \leqslant \|\mathbf{A}_T\mathbf{v}\|^2 \leqslant (1+\delta_S)\|\mathbf{v}\|^2$$

holds for all subsets T with $|T| \leqslant S$ and $\mathbf{v} \in \mathbf{R}^{|T|}$. For any pair of positive integers S and S' with $S + S' \leqslant d$, the (S, S')-restricted orthogonality constant $\theta_{S,S'} = \theta_{S,S'}(\mathbf{A})$ of $\mathbf{A}$ is defined to be the smallest quantity such that

$$|\langle \mathbf{A}_T\mathbf{v}, \mathbf{A}_{T'}\mathbf{v}' \rangle| \leqslant \theta_{S,S'}\|\mathbf{v}\|\,\|\mathbf{v}'\|$$

holds for all disjoint subsets T and T' of cardinalities $|T| \leqslant S$ and $|T'| \leqslant S'$, $\mathbf{v} \in \mathbf{R}^{|T|}$ and $\mathbf{v}' \in \mathbf{R}^{|T'|}$.

The following theorem is obtained by the sure screening property of SIS in theorem 1 along with theorem 1.1 in Candes and Tao (2007), where $\varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$ for some $\sigma > 0$. To avoid the selection bias in the screening step, we can split the sample into two halves: the first half is used to screen variables and the second half is used to construct the Dantzig estimator. The same technique applies to SCAD, but we avoid this step to facilitate the presentation.

Theorem 4 (consistency of method SIS–DS). Assume, with large probability, that $\delta_{2s}(\mathbf{X}_{\mathcal{M}}) + \theta_{s,2s}(\mathbf{X}_{\mathcal{M}}) \leqslant t < 1$ and choose $\lambda_d = \{2\log(d)\}^{1/2}$ in problem (14). Then, with large probability, we have

$$\|\hat{\boldsymbol{\beta}}_{\mathrm{DS}} - \boldsymbol{\beta}\|^2 \leqslant C\,\{\log(d)\}^{1/2}\, s\sigma^2,$$

where $C = 32/(1-t)^2$ and s is the number of non-zero components of $\boldsymbol{\beta}$.

Theorem 4 shows that method SIS–DS, i.e. SIS followed by the Dantzig selector, can now achieve the ideal risk up to a factor of $\log(d)$ with $d < n$, rather than the original $\log(p)$.

Now let us look at method SIS–SCAD, i.e. SIS followed by SCAD. For simplicity, a common regularization parameter λ is used for the SCAD penalty function. Let $\hat{\boldsymbol{\beta}}_{\mathrm{SCAD}} = (\hat{\beta}_1, \ldots, \hat{\beta}_d)^{\mathrm{T}}$ be a minimizer of the SCAD–PLS method in problem (11). The following theorem is obtained by the sure screening property of SIS in theorem 1 along with theorems 1 and 2 in Fan and Peng (2004).

Theorem 5 (oracle properties of method SIS–SCAD). If $d = o(n^{1/3})$ and the assumptions of theorem 2 in Fan and Peng (2004) are satisfied, then, with probability tending to 1, the SCAD–PLS estimator $\hat{\boldsymbol{\beta}}_{\mathrm{SCAD}}$ satisfies

(a) $\hat{\beta}_i = 0$ for any $i \notin \mathcal{M}_*$ and

(b) the components of $\hat{\beta}_{\mathrm{SCAD}}$ in $\mathcal{M}_*$ perform as well as if the true model $\mathcal{M}_*$ were known.

The SIS–SCAD method has been shown to enjoy oracle properties.

6. Concluding remarks

This paper studies the problem of high dimensional variable selection for the linear model. The concept of sure screening is introduced and a sure screening method which is based on correlation learning that we call SIS is proposed. SIS has been shown to be capable of reducing from exponentially growing dimensionality to below sample size accurately. It speeds up variable selection dramatically and can also improve the accuracy of estimation when dimensionality is ultrahigh. SIS combined with well-developed variable selection techniques including SCAD, the Dantzig selector, lasso and adaptive lasso provides a powerful tool for high dimensional variable selection. The tuning parameter d can be taken as $d = [n/\log(n)]$ or $d = n - 1$, depending on which model selector is used in the second stage. For non-concave PLS problem (12), when we directly apply the local linear approximation algorithm to the original problem with $d = p$, we need initial values that are not available. SIS provides a method that makes this feasible by screening many variables and furnishing the corresponding coefficients with 0s. The initial value in problem (12) can be taken as the ordinary least squares estimate if $d = [n/\log(n)]$ and 0 (corresponding to $w_j^{(0)} \equiv p'_\lambda(0+)$) when $d = n - 1$. The latter corresponds to the lasso.

Some extensions of SIS have also been discussed. In particular, ISIS was proposed to enhance the finite sample performance of SIS, particularly in the situations where the technical conditions fail. This raises a challenging question: to what extent does ISIS relax the conditions for SIS to have the sure screening property? An ITRRS has been introduced to understand better the rationale of SIS and serves as a technical device for proving the sure screening property. As a by-product, it is demonstrated that the stepwise deletion method may have no sure screening property when the dimensionality is of an exponential order. This raises another interesting question whether the sure screening property holds for a greedy algorithm such as the stepwise addition or matching pursuit and how large the selected model must be if it does.

The paper leaves open the problem of extending the SIS and ISIS methods that were introduced for the linear models to the family of generalized linear models and other general loss functions such as the hinge loss and the loss that is associated with the support vector machine. Questions including how to define associated residuals to extend ISIS and whether the sure screening property continues to hold naturally arise. The paper focuses only on random designs which commonly appear in statistical problems, whereas for many problems in fields such as image analysis and signal processing the design matrices are often deterministic. It remains open how to impose a set of conditions that ensure the sure screening property. It also remains open whether the sure screening property can be extended to the sparse additive model in non-parametric learning as studied by Ravikumar *et al.* (2007). These questions are beyond the scope of the current paper and are interesting topics for future research.

Acknowledgements

Financial support from National Science Foundation grants DMS-0354223, DMS-0704337 and DMS-0714554, and National Institutes of Health grant R01-GM072611 is gratefully acknowledged. Lv's research was partially supported by National Science Foundation grant DMS-0806030 and the 2008 Zumberge Individual Award from the James H. Zumberge Faculty Research and Innovation Fund at the University of Southern California. We are grateful to the referees for their constructive and helpful comments.

Appendix A

Hereafter we use both C and c to denote generic positive constants for notational convenience.

A.1. Proof of theorem 1

Motivated by the results in theorems 2 and 3, the idea is to apply successively dimensionality reduction in a way that is described in expression (17) below. To enhance readability, we split the whole proof into two main steps and multiple substeps.

Step 1: let $\delta \in (0, 1)$. Similarly to expression (8), we define a submodel

$$\tilde{\mathcal{M}}_\delta^1 = \{1 \leqslant i \leqslant p : |\omega_i| \text{ is among the first } [\delta p] \text{ largest of all}\}. \tag{17}$$

We aim to show that, if $\delta \to 0$ in such a way that $\delta n^{1-2\kappa-\tau} \to \infty$ as $n \to \infty$, we have, for some $C > 0$,

$$P(\mathcal{M}_* \subset \tilde{\mathcal{M}}_\delta^1) = 1 - O[\exp\{-Cn^{1-2\kappa}/\log(n)\}]. \tag{18}$$

The main idea is to relate the general case to the specific case with $\boldsymbol{\Sigma} = I_p$, which is separately studied in Appendices A.4–A.6 below. A key ingredient is the representation (19) below of the $p \times p$ random matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}$. Throughout, let $\mathbf{S} = (\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{+}\mathbf{Z}^{\mathrm{T}}\mathbf{Z}$ and $\mathbf{e}_i = (0, \ldots, 1, \ldots, 0)^{\mathrm{T}}$ be a unit vector in $\mathbf{R}^p$ with the ith entry 1 and 0 elsewhere, $i = 1, \ldots, p$.

Since $\mathbf{X} = \mathbf{Z}\boldsymbol{\Sigma}^{1/2}$, it follows from equation (45) in Appendix A.4 that

$$\mathbf{X}^{\mathrm{T}}\mathbf{X} = p\boldsymbol{\Sigma}^{1/2}\tilde{\mathbf{U}}^{\mathrm{T}} \operatorname{diag}(\mu_1, \ldots, \mu_n)\tilde{\mathbf{U}}\boldsymbol{\Sigma}^{1/2}, \tag{19}$$

where $\mu_1, \ldots, \mu_n$ are n eigenvalues of $p^{-1}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}$, $\tilde{\mathbf{U}} = (I_n, \mathbf{0})_{n\times p}\mathbf{U}$, and $\mathbf{U}$ is uniformly distributed on the orthogonal group $\mathcal{O}(p)$. By expressions (1) and (2), we have

$$\boldsymbol{\omega} = \mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} + \mathbf{X}^{\mathrm{T}}\boldsymbol{\varepsilon} \triangleq \boldsymbol{\xi} + \boldsymbol{\eta}. \tag{20}$$

We shall study the above two random vectors $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ separately.

Step 1.1: first, we consider term $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_p)^{\mathrm{T}} = \mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}$.

Step 1.1.1 (bounding $\|\boldsymbol{\xi}\|$ from above): it is obvious that

$$\operatorname{diag}(\mu_1^2, \ldots, \mu_n^2) \leqslant \lambda_{\max}(p^{-1}\mathbf{Z}\mathbf{Z}^{\mathrm{T}})^2 I_n$$

and $\tilde{\mathbf{U}}\boldsymbol{\Sigma}\tilde{\mathbf{U}}^{\mathrm{T}} \leqslant \lambda_{\max}(\boldsymbol{\Sigma})I_n$. These and equation (19) lead to

$$\|\boldsymbol{\xi}\|^2 \leqslant p^2\, \lambda_{\max}(\boldsymbol{\Sigma})\, \lambda_{\max}(p^{-1}\mathbf{Z}\mathbf{Z}^{\mathrm{T}})^2 \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}^{1/2}\tilde{\mathbf{U}}^{\mathrm{T}}\tilde{\mathbf{U}}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}. \tag{21}$$

Let $Q \in \mathcal{O}(p)$ such that $\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta} = \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\| Q\mathbf{e}_1$. Then, it follows from lemma 1 that

$$\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}^{1/2}\tilde{\mathbf{U}}^{\mathrm{T}}\tilde{\mathbf{U}}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta} = \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|^2\langle Q^{\mathrm{T}}\mathbf{S}Q\mathbf{e}_1, \mathbf{e}_1\rangle \overset{(\mathrm{d})}{=} \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|^2\langle \mathbf{S}\mathbf{e}_1, \mathbf{e}_1\rangle,$$

where we use the symbol $\overset{(\mathrm{d})}{=}$ to denote being identical in distribution for brevity. By condition 3, $\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|^2 = \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\beta} \leqslant \operatorname{var}(Y) = O(1)$ and thus, by lemma 4, we have, for some $C > 0$,

$$P\left\{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}^{1/2}\tilde{\mathbf{U}}^{\mathrm{T}}\tilde{\mathbf{U}}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta} > O\left(\frac{n}{p}\right)\right\} \leqslant O\{\exp(-Cn)\}. \tag{22}$$

Since $\lambda_{\max}(\boldsymbol{\Sigma}) = O(n^\tau)$ and $P\{\lambda_{\max}(p^{-1}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}) > c_1\} \leqslant \exp(-C_1 n)$ by conditions 2 and 4, (21) and (22) along with Bonferroni's inequality yield

$$P\{\|\boldsymbol{\xi}\|^2 > O(n^{1+\tau}p)\} \leqslant O\{\exp(-C_n)\}. \tag{23}$$

Step 1.1.2 (bounding $|\xi_i|$, $i \in \mathcal{M}_*$, from below): this needs a delicate analysis. Now fix an arbitrary $i \in \mathcal{M}_*$. By equation (19), we have

$$\xi_i = p\mathbf{e}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{1/2}\tilde{\mathbf{U}}^{\mathrm{T}} \operatorname{diag}(\mu_1, \ldots, \mu_n)\tilde{\mathbf{U}}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}.$$

Note that $\|\boldsymbol{\Sigma}^{1/2}\mathbf{e}_i\| = \operatorname{var}(X_i)^{1/2} = 1$ and $\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\| = O(1)$. By condition 3, there is some $c > 0$ such that

$$|\langle \boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}, \boldsymbol{\Sigma}^{1/2}\mathbf{e}_i\rangle| = |\beta_i||\mathrm{cov}(\beta_i^{-1}Y, X_i)| \geqslant c/n^{\kappa}. \tag{24}$$

Thus, there is a $Q \in \mathcal{O}(p)$ such that $\boldsymbol{\Sigma}^{1/2}\mathbf{e}_i = Q\mathbf{e}_1$ and

$$\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta} = \langle \boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}, \boldsymbol{\Sigma}^{1/2}\mathbf{e}_i\rangle Q\mathbf{e}_1 + O(1)Q\mathbf{e}_2.$$

Since $(\mu_1, \ldots, \mu_n)^{\mathrm{T}}$ is independent of $\tilde{\mathbf{U}}$ by lemma 1 and the uniform distribution on the orthogonal group $\mathcal{O}(p)$ is invariant under itself, it follows that

$$\xi_i \overset{(\mathrm{d})}{=} p\langle \boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}, \boldsymbol{\Sigma}^{1/2}\mathbf{e}_i\rangle R_1 + O(p)R_2 \overset{\triangle}{=} \xi_{i,1} + \xi_{i,2}, \tag{25}$$

where $\mathbf{R} = (R_1, R_2, \ldots, R_p)^{\mathrm{T}} = \tilde{\mathbf{U}}^{\mathrm{T}}\mathrm{diag}(\mu_1, \ldots, \mu_n)\tilde{\mathbf{U}}\mathbf{e}_1$. We shall examine the above two terms $\xi_{i,1}$ and $\xi_{i,2}$ separately. Clearly,

$$R_1 \geqslant \mathbf{e}_1^{\mathrm{T}}\tilde{\mathbf{U}}^{\mathrm{T}}\lambda_{\min}(p^{-1}\mathbf{Z}\mathbf{Z}^{\mathrm{T}})I_n\tilde{\mathbf{U}}\mathbf{e}_1 = \lambda_{\min}(p^{-1}\mathbf{Z}\mathbf{Z}^{\mathrm{T}})\langle \mathbf{S}\mathbf{e}_1, \mathbf{e}_1\rangle,$$

and thus, by condition 2, lemma 4 in Appendix A.5 and Bonferroni's inequality, we have, for some $c > 0$ and $C > 0$,

$$P(R_1 < cn/p) \leqslant O\{\exp(-Cn)\}.$$

This, along with expression (24), gives, for some $c > 0$,

$$P(|\xi_{i,1}| < cn^{1-\kappa}) \leqslant O\{\exp(-Cn)\}. \tag{26}$$

Similarly to step 1.1.1, it can be shown that

$$P\{\|\mathbf{R}\|^2 > O(n/p)\} \leqslant O\{\exp(-Cn)\}. \tag{27}$$

Since $(\mu_1, \ldots, \mu_n)^{\mathrm{T}}$ is independent of $\tilde{\mathbf{U}}$ by lemma 1, the argument in the proof of lemma 5 in Appendix A.6 applies to show that the distribution of $\tilde{\mathbf{R}} = (R_2, \ldots, R_p)^{\mathrm{T}}$ is invariant under the orthogonal group $\mathcal{O}(p-1)$. Then, it follows that $\tilde{\mathbf{R}} \overset{(\mathrm{d})}{=} \|\tilde{\mathbf{R}}\|\,\mathbf{W}/\|\mathbf{W}\|$, where $\mathbf{W} = (W_1, \ldots, W_{p-1})^{\mathrm{T}} \sim \mathcal{N}(0, I_{p-1})$, independent of $\|\tilde{\mathbf{R}}\|$. Thus, we have

$$R_2 \overset{(\mathrm{d})}{=} \|\tilde{\mathbf{R}}\|W_1/\|\mathbf{W}\|. \tag{28}$$

In view of expressions (27), (28) and $\xi_{i,2} = O(pR_2)$, applying the argument in the proof of lemma 5 gives, for some $c > 0$,

$$P(|\xi_{i,2}| > cn^{1/2}|W|) \leqslant O\{\exp(-Cn)\}, \tag{29}$$

where W is an $\mathcal{N}(0, 1)$-distributed random variable.

Let $x_n = c\sqrt{(2C)}n^{1-\kappa}/\sqrt{\log(n)}$. Then, by the classical Gaussian tail bound, we have

$$P(cn^{1/2}|W| > x_n) \leqslant \sqrt{(2/\pi)}\frac{\exp\{-Cn^{1-2\kappa}/\log(n)\}}{\sqrt{(2C)}n^{1/2-\kappa}/\sqrt{\log(n)}} = O[\exp\{-Cn^{1-2\kappa}/\log(n)\}],$$

which, along with inequality (29) and Bonferroni's inequality, shows that

$$P(|\xi_{i,2}| > x_n) \leqslant O[\exp\{-Cn^{1-2\kappa}/\log(n)\}]. \tag{30}$$

Therefore, by Bonferroni's inequality, combining expressions (25), (26) and (30) gives, for some $c > 0$,

$$P(|\xi_i| < cn^{1-\kappa}) \leqslant O[\exp\{-Cn^{1-2\kappa}/\log(n)\}], \qquad i \in \mathcal{M}_*. \tag{31}$$

Step 1.2: then, we examine term $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_p)^{\mathrm{T}} = \mathbf{X}^{\mathrm{T}}\boldsymbol{\varepsilon}$.

Step 1.2.1 (bounding $\|\boldsymbol{\eta}\|$ from above): clearly, we have

$$\mathbf{X}\mathbf{X}^{\mathrm{T}} = \mathbf{Z}\boldsymbol{\Sigma}\mathbf{Z}^{\mathrm{T}} \leqslant \mathbf{Z}\,\lambda_{\max}(\boldsymbol{\Sigma})I_p\mathbf{Z}^{\mathrm{T}} = p\,\lambda_{\max}(\boldsymbol{\Sigma})\,\lambda_{\max}(p^{-1}\mathbf{Z}\mathbf{Z}^{\mathrm{T}})I_n.$$

Then, it follows that

$$\|\boldsymbol{\eta}\|^2 = \boldsymbol{\varepsilon}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\boldsymbol{\varepsilon} \leqslant p\,\lambda_{\max}(\boldsymbol{\Sigma})\,\lambda_{\max}(p^{-1}\mathbf{Z}\mathbf{Z}^{\mathrm{T}})\|\boldsymbol{\varepsilon}\|^2. \tag{32}$$

From condition 2, we know that $\varepsilon_1^2/\sigma^2, \ldots, \varepsilon_n^2/\sigma^2$ are IID χ_1^2-distributed random variables. Thus, by inequality (47) in lemma 3 in Appendix A.5, there are some $c > 0$ and $C > 0$ such that

$$P(\|\varepsilon\|^2 > cn\sigma^2) \leqslant \exp(-Cn),$$

which along with inequality (32), conditions 2 and 4, and Bonferroni's inequality yield

$$P\{\|\boldsymbol{\eta}\|^2 > O(n^{1+\tau}p)\} \leqslant O\{\exp(-Cn)\}. \tag{33}$$

Step 1.2.2 (bounding $|\eta_i|$ from above): given that $\mathbf{X} = X$, $\boldsymbol{\eta} = X^{\mathrm{T}}\varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 X^{\mathrm{T}}X)$. Hence, $(\eta_i|\mathbf{X} = X) \sim \mathcal{N}\{0, \mathrm{var}(\eta_i|\mathbf{X} = X)\}$ with

$$\mathrm{var}(\eta_i|\mathbf{X} = X) = \sigma^2 \mathbf{e}_i^{\mathrm{T}} X^{\mathrm{T}} X \mathbf{e}_i. \tag{34}$$

Let $\mathcal{E}$ be the event $\{\mathrm{var}(\eta_i|\mathbf{X}) \leqslant cn\}$ for some $c > 0$. Then, using the same argument as that in step 1.1.1, we can easily show that, for some $C > 0$,

$$P(\mathcal{E}^c) \leqslant O\{\exp(-Cn)\}. \tag{35}$$

On the event $\mathcal{E}$, we have

$$P(|\eta_i| > x|\mathbf{X}) \leqslant P\{\sqrt{(cn)}|W| > x\} \qquad \text{for any } x > 0, \tag{36}$$

where W is an $\mathcal{N}(0, 1)$-distributed random variable. Thus, it follows from inequalities (35) and (36) that

$$P(|\eta_i| > x) \leqslant O\{\exp(-Cn)\} + P\{\sqrt{(cn)}|W| > x\}. \tag{37}$$

Let $x_n' = \sqrt{(2cC)}n^{1-\kappa}/\sqrt{\log(n)}$. Then, invoking the classical Gaussian tail bound again, we have

$$P\{\sqrt{(cn)}|W| > x_n'\} = O[\exp\{-Cn^{1-2\kappa}/\log(n)\}],$$

which, along with inequality (37) and condition 1, shows that

$$P\{\max_i |\eta_i| > o(n^{1-\kappa})\} \leqslant O[p\exp\{-Cn^{1-2\kappa}/\log(n)\}] = O[\exp\{-Cn^{1-2\kappa}/\log(n)\}]. \tag{38}$$

Step 1.3: finally, we combine the results that were obtained in steps 1.1 and 1.2. By Bonferroni's inequality, it follows from expressions (20), (23), (31), (33) and (38) that, for some constants $c_1, c_2, C > 0$,

$$P(\min_{i \in \mathcal{M}_*} |\omega_i| < c_1 n^{1-\kappa} \text{ or } \|\boldsymbol{\omega}\|^2 > c_2 n^{1+\tau}p) \leqslant O[s\exp\{-Cn^{1-2\kappa}/\log(n)\}]. \tag{39}$$

This shows that, with overwhelming probability $1 - O[s\exp\{-Cn^{1-2\kappa}/\log(n)\}]$, the magnitudes of ω_i, $i \in \mathcal{M}_*$, are uniformly at least of order $n^{1-\kappa}$ and more importantly, for some $c > 0$,

$$\#\{1 \leqslant k \leqslant p : |\omega_k| \geqslant \min_{i \in \mathcal{M}_*} |\omega_i|\} \leqslant c\frac{n^{1+\tau}p}{(n^{1-\kappa})^2} = \frac{cp}{n^{1-2\kappa-\tau}}, \tag{40}$$

where $\#\{\cdot\}$ denotes the number of elements in a set.

Now, we are ready to see from inequality (40) that, if δ satisfies $\delta n^{1-2\kappa-\tau} \to \infty$ as $n \to \infty$, then equation (18) holds for some constant $C > 0$ that is larger than that in inequality (39).

Step 2: fix an arbitrary $r \in (0, 1)$ and choose a shrinking factor δ of the form $(n/p)^{1/(\kappa-r)}$, for some integer $k \geqslant 1$. We successively perform dimensionality reduction until the number of remaining variables drops to below sample size n.

(a) First, carry out procedure (17) to the full model $\tilde{\mathcal{M}}_\delta^0 \triangleq \{1, \ldots, p\}$ and obtain a submodel $\tilde{\mathcal{M}}_\delta^1$ with size $[\delta p]$.
(b) Then, apply a similar procedure to the model $\tilde{\mathcal{M}}_\delta^1$ and again obtain a submodel $\tilde{\mathcal{M}}_\delta^2 \subset \tilde{\mathcal{M}}_\delta^1$ with size $[\delta^2 p]$, and so on.
(c) Finally, obtain a submodel $\tilde{\mathcal{M}}_\delta \triangleq \tilde{\mathcal{M}}_\delta^k$ with size $d = [\delta^k p] = [\delta^r n] < n$, where $[\delta^{k-1}p] = [\delta^{r-1}n] > n$.

It is obvious that $\tilde{\mathcal{M}}_\delta = \mathcal{M}_\gamma$, where $\gamma = \delta^r < 1$.

Now fix an arbitrary $\theta_1 \in (0, 1-2\kappa-\tau)$ and pick some $r < 1$ very close to 1 such that $\theta_0 = \theta_1/r < 1-2\kappa-\tau$. We choose a sequence of integers $k \geqslant 1$ in a way such that

$$\delta n^{1-2\kappa-\tau} \to \infty \quad \text{and} \quad \delta n^{\theta_0} \to 0 \qquad \text{as } n \to \infty, \tag{41}$$

where $\delta = (n/p)^{1/(k-r)}$. Then, applying the above scheme of dimensionality reduction results in a submodel $\tilde{\mathcal{M}}_\delta = \mathcal{M}_\gamma$, where $\gamma = \delta^r$ satisfies

$$\gamma n^{r(1-2\kappa-\tau)} \to \infty \quad \text{and} \quad \gamma n^{\theta_1} \to 0 \qquad \text{as } n \to \infty. \tag{42}$$

Before going further, let us make two important observations. First, for any principal submatrix $\mathbf{\Sigma}^0$ of $\mathbf{\Sigma}$ corresponding to a subset of variables, condition 4 ensures that

$$\lambda_{\max}(\mathbf{\Sigma}^0) \leqslant \lambda_{\max}(\mathbf{\Sigma}) \leqslant c_4 n^{\tau}.$$

Second, by definition, property (16) holds for any $n \times \tilde{p}$ submatrix $\tilde{\mathbf{Z}}$ of $\mathbf{Z}$ with $cn < \tilde{p} \leqslant p$, where $c > 1$ is some constant. Thus, the probability bound in equation (18) is uniform over dimension $\tilde{p} \in (cn, p]$. Therefore, for some $C > 0$, by expressions (41) and (18) we have, in each step $1 \leqslant i \leqslant k$ of the above dimensionality reduction,

$$P(\mathcal{M}_* \subset \tilde{\mathcal{M}}_\delta^i | \mathcal{M}_* \subset \tilde{\mathcal{M}}_\delta^{i-1}) = 1 - O[\exp\{-Cn^{1-2\kappa}/\log(n)\}],$$

which along with Bonferroni's inequality gives

$$P(\mathcal{M}_* \subset \mathcal{M}_\gamma) = 1 - O[k \exp\{-Cn^{1-2\kappa}/\log(n)\}]. \tag{43}$$

It follows from expression (41) that $k = O\{\log(p)/\log(n)\}$, which is of order $O\{n^\xi/\log(n)\}$ by condition 1. Thus, a suitable increase of the constant $C > 0$ in equation (43) yields

$$P(\mathcal{M}_* \subset \mathcal{M}_\gamma) = 1 - O[\exp\{-Cn^{1-2\kappa}/\log(n)\}].$$

Finally, in view of expression (42), the above probability bound holds for any $\gamma \sim cn^{-\theta}$, with $\theta < 1 - 2\kappa - \tau$ and $c > 0$. This completes the proof.

A.2. Proof of theorem 2

We observe that expression (8) uses only the order of componentwise magnitudes of $\boldsymbol{\omega}^\lambda$, so it is invariant under scaling. Therefore, in view of expression (7) we see from step 1 of the proof of theorem 1 that theorem 2 holds for sufficiently large regularization parameter λ.

It remains to specify a lower bound on λ. Now we rewrite the p-vector $\lambda\boldsymbol{\omega}^\lambda$ as

$$\lambda\boldsymbol{\omega}^\lambda = \boldsymbol{\omega} - \{I_p - (I_p + \lambda^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\}\boldsymbol{\omega}.$$

Let $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_p)^{\mathrm{T}} = \{I_p - (I_p + \lambda^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\}\boldsymbol{\omega}$. It follows easily from $\mathbf{X}^{\mathrm{T}}\mathbf{X} = \mathbf{\Sigma}^{1/2}\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\mathbf{\Sigma}^{1/2}$ that

$$\lambda_{\max}(\mathbf{X}^{\mathrm{T}}\mathbf{X}) \leqslant p\, \lambda_{\max}(p^{-1}\mathbf{Z}\mathbf{Z}^{\mathrm{T}})\, \lambda_{\max}(\mathbf{\Sigma}),$$

and thus

$$\begin{aligned}\|\boldsymbol{\zeta}\|^2 &\leqslant \lambda_{\max}\{I_p - (I_p + \lambda^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\}^2 \|\boldsymbol{\omega}\|^2 \\ &\leqslant \lambda_{\max}(\lambda^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X})^2 \|\boldsymbol{\omega}\|^2 \\ &\leqslant \lambda^{-2} p^2\, \lambda_{\max}(p^{-1}\mathbf{Z}\mathbf{Z}^{\mathrm{T}})^2\, \lambda_{\max}(\mathbf{\Sigma})^2 \|\boldsymbol{\omega}\|^2,\end{aligned}$$

which along with inequality (39), conditions 2 and 4, and Bonferroni's inequality show that

$$P\{\|\boldsymbol{\zeta}\| > O(\lambda^{-1} n^{(1+3\tau)/2} p^{3/2})\} \leqslant O[s \exp\{-Cn^{1-2\kappa}/\log(n)\}].$$

Again, by Bonferroni's inequality and inequality (39), any λ satisfying $\lambda^{-1} n^{(1+3\tau)/2} p^{3/2} = o(n^{1-\kappa})$ can be used. Note that $\kappa + \tau/2 < \frac{1}{2}$ by assumption. So in particular, we can choose any λ satisfying $\lambda(p^{3/2}n)^{-1} \to \infty$ as $n \to \infty$.

A.3. Proof of theorem 3

Theorem 3 is a straightforward corollary to theorem 2 by the argument in step 2 of the proof of theorem 1.

Throughout Appendices A.4–A.6 below, we assume that $p > n$ and the distribution of $\mathbf{z}$ is continuous and spherically symmetric, i.e. invariant under the orthogonal group $\mathcal{O}(p)$. For brevity, we use $\mathcal{L}(\cdot)$ to denote the probability law or distribution of the random variable indicated. Let $S^{q-1}(r) = \{x \in \mathbf{R}^q : \|x\| = r\}$ be the centred sphere with radius r in q-dimensional Euclidean space $\mathbf{R}^q$. In particular, S^{q-1} is referred to as the unit sphere in $\mathbf{R}^q$.

A.4. The distribution of $\mathbf{S} = (\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^+\mathbf{Z}^{\mathrm{T}}\mathbf{Z}$

It is a classical fact that the orthogonal group $\mathcal{O}(p)$ is compact and admits a probability measure that is invariant under the action of itself, say,

$$Q \cdot g \overset{\triangle}{=} Qg, \qquad g \in \mathcal{O}(p), \quad Q \in \mathcal{O}(p).$$

This invariant distribution is referred to as the uniform distribution on the orthogonal group $\mathcal{O}(p)$. We often encounter projection matrices in multivariate statistical analysis. In fact, the set of all $p \times p$ projection matrices of rank n can equivalently be regarded as the Grassmann manifold $\mathcal{G}_{p,n}$ of all n-dimensional subspaces of the Euclidean space $\mathbf{R}^p$; throughout, we do not distinguish them and write

$$\mathcal{G}_{p,n} = \{U^{\mathrm{T}} \operatorname{diag}(I_n, 0)U : U \in \mathcal{O}(p)\}.$$

It is well known that the Grassmann manifold $\mathcal{G}_{p,n}$ is compact and there is a natural $\mathcal{O}(p)$ action on it, say,

$$Q \cdot g \overset{\triangle}{=} Q^{\mathrm{T}} g Q, \qquad g \in \mathcal{G}_{p,n}, \quad Q \in \mathcal{O}(p).$$

Clearly, this group action is transitive, i.e., for any $g_1, g_2 \in \mathcal{G}_{p,n}$, there is some $Q \in \mathcal{O}(p)$ such that $Q \cdot g_1 = g_2$. Moreover, $\mathcal{G}_{p,n}$ admits a probability measure that is invariant under the $\mathcal{O}(p)$ action that was defined above. This invariant distribution is referred to as the uniform distribution on the Grassmann manifold $\mathcal{G}_{p,n}$. For more on group action and invariant measures on special manifolds, see Eaton (1989) and Chikuse (2003).

The uniform distribution on the Grassmann manifold is not easy to deal with directly. A useful fact is that the uniform distribution on $\mathcal{G}_{p,n}$ is the image measure of the uniform distribution on $\mathcal{O}(p)$ under the mapping

$$\varphi : \mathcal{O}(p) \to \mathcal{G}_{p,n}, \qquad \varphi(U) = U^{\mathrm{T}} \operatorname{diag}(I_n, 0)U, \quad U \in \mathcal{O}(p).$$

By the assumption that $\mathbf{z}$ has a continuous distribution, we can easily see that, with probability 1, the $n \times p$ matrix $\mathbf{Z}$ has full rank n. Let $\mu_1^{1/2}, \ldots, \mu_n^{1/2}$ be its n singular values. Then, $\mathbf{Z}$ admits a singular value decomposition

$$\mathbf{Z} = \mathbf{V}\mathbf{D}_1\mathbf{U}, \tag{44}$$

where $\mathbf{V} \in \mathcal{O}(n)$, $\mathbf{U} \in \mathcal{O}(p)$ and $\mathbf{D}_1$ is an $n \times p$ diagonal matrix whose diagonal elements are $\mu_1^{1/2}, \ldots, \mu_n^{1/2}$. Thus,

$$\mathbf{Z}^{\mathrm{T}}\mathbf{Z} = \mathbf{U}^{\mathrm{T}} \operatorname{diag}(\mu_1, \ldots, \mu_n, 0, \ldots, 0)\mathbf{U} \tag{45}$$

and its Moore–Penrose generalized inverse is

$$(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{+} = \sum_{i=1}^{n} \frac{1}{\mu_i} \mathbf{u}_i \mathbf{u}_i^{\mathrm{T}},$$

where $\mathbf{U}^{\mathrm{T}} = (\mathbf{u}_1, \ldots, \mathbf{u}_p)$. Therefore, we have the decomposition

$$\mathbf{S} = (\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{+}\mathbf{Z}^{\mathrm{T}}\mathbf{Z} = \mathbf{U}^{\mathrm{T}} \operatorname{diag}(I_n, 0)\mathbf{U}, \qquad \mathbf{U} \in \mathcal{O}(p). \tag{46}$$

From equation (44), we know that $\mathbf{Z} = \mathbf{V} \operatorname{diag}(\mu_1^{1/2}, \ldots, \mu_n^{1/2})(I_n, \mathbf{0})_{n \times p}\mathbf{U}$, and thus

$$(I_n, \mathbf{0})_{n \times p}\mathbf{U} = \operatorname{diag}(1/\mu_1^{1/2}, \ldots, 1/\mu_n^{1/2})\mathbf{V}^{\mathrm{T}}\mathbf{Z}.$$

By the assumption that $\mathcal{L}(\mathbf{z})$ is invariant under the orthogonal group $\mathcal{O}(p)$, the distribution of $\mathbf{Z}$ is also invariant under $\mathcal{O}(p)$, i.e.

$$\mathbf{Z}Q \overset{(\mathrm{d})}{=} \mathbf{Z} \qquad \text{for any } Q \in \mathcal{O}(p).$$

Thus, conditional on $\mathbf{V}$ and $(\mu_1, \ldots, \mu_n)^{\mathrm{T}}$, the conditional distribution of $(I_n, \mathbf{0})_{n \times p}\mathbf{U}$ is invariant under $\mathcal{O}(p)$, which entails that

$$(I_n, \mathbf{0})_{n \times p}\mathbf{U} \overset{(\mathrm{d})}{=} (I_n, \mathbf{0})_{n \times p}\tilde{\mathbf{U}},$$

where $\tilde{\mathbf{U}}$ is uniformly distributed on the orthogonal group $\mathcal{O}(p)$. In particular, we see that $(\mu_1, \ldots, \mu_n)^{\mathrm{T}}$ is independent of $(I_n, \mathbf{0})_{n \times p}\mathbf{U}$. Therefore, these facts along with equation (46) yield the following lemma.

Lemma 1. $\mathcal{L}\{(I_n, \mathbf{0})_{n \times p}\mathbf{U}\} = \mathcal{L}\{(I_n, \mathbf{0})_{n \times p}\tilde{\mathbf{U}}\}$ and $(\mu_1, \ldots, \mu_n)^{\mathrm{T}}$ is independent of $(I_n, \mathbf{0})_{n \times p}\mathbf{U}$, where $\tilde{\mathbf{U}}$ is

uniformly distributed on the orthogonal group $\mathcal{O}(p)$ and $\mu_1, \ldots, \mu_n$ are n eigenvalues of $\mathbf{Z}\mathbf{Z}^{\mathrm{T}}$. Moreover, $\mathbf{S}$ is uniformly distributed on the Grassmann manifold $\mathcal{G}_{p,n}$.

For simplicity, we do not distinguish $\tilde{\mathbf{U}}$ and $\mathbf{U}$ in singular value decomposition (44).

A.5. Deviation inequality on $\langle \mathbf{S}\mathbf{e}_1, \mathbf{e}_1 \rangle$

Lemma 2. $\mathcal{L}(\langle \mathbf{S}\mathbf{e}_1, \mathbf{e}_1 \rangle) = \chi_n^2/(\chi_n^2 + \chi_{p-n}^2)$, where χ_n^2 and χ_{p-n}^2 are two independent χ^2-distributed random variables with degrees of freedom n and $p-n$ respectively, i.e. $\langle \mathbf{S}\mathbf{e}_1, \mathbf{e}_1 \rangle$ has a beta distribution with parameters $n/2$ and $(p-n)/2$.

Proof. Lemma 1 gives $\mathcal{L}(\mathbf{S}) = \mathcal{L}\{\mathbf{U}^{\mathrm{T}} \operatorname{diag}(I_n, 0)\mathbf{U}\}$, where $\mathbf{U}$ is uniformly distributed on $\mathcal{O}(p)$. Clearly, $(\mathbf{U}\mathbf{e}_1)$ is a random vector on the unit sphere S^{p-1}. It can be shown that $\mathbf{U}\mathbf{e}_1$ is uniformly distributed on the unit sphere S^{p-1}.

Let $\mathbf{W} = (W_1, \ldots, W_p)^{\mathrm{T}} \sim \mathcal{N}(\mathbf{0}, I_p)$. Then, we have $\mathbf{U}\mathbf{e}_1 =^{(\mathrm{d})} \mathbf{W}/\|\mathbf{W}^{\mathrm{T}}\|$ and

$$\langle \mathbf{S}\mathbf{e}_1, \mathbf{e}_1 \rangle = (\mathbf{U}\mathbf{e}_1)^{\mathrm{T}} \operatorname{diag}(I_n, 0)\mathbf{U}\mathbf{e}_1 \overset{(\mathrm{d})}{=} \frac{W_1^2 + \ldots + W_n^2}{W_1^2 + \ldots + W_p^2}.$$

This proves lemma 2. □

Lemmas 3 and 4 below give sharp deviation bounds on the beta distribution.

Lemma 3 (moderate deviation). Let $\xi_1, \ldots, \xi_n$ be IID χ_1^2-distributed random variables. Then,

(a) for any $\varepsilon > 0$, we have

$$P\{n^{-1}(\xi_1 + \ldots + \xi_n) > 1 + \varepsilon\} \leqslant \exp(-A_\varepsilon n), \tag{47}$$

where $A_\varepsilon = \{\varepsilon - \log(1+\varepsilon)\}/2 > 0$, and

(b) for any $\varepsilon \in (0, 1)$, we have

$$P\{n^{-1}(\xi_1 + \ldots + \xi_n) < 1 - \varepsilon\} \leqslant \exp(-B_\varepsilon n), \tag{48}$$

where $B_\varepsilon = \{-\varepsilon - \log(1+\varepsilon)\}/2 > 0$.

Proof.

(a) Recall that the moment-generating function of a χ_1^2-distributed random variable ξ is

$$M(t) = E\exp(t\xi) = (1-2t)^{-1/2}, \qquad t \in (-\infty, \tfrac{1}{2}). \tag{49}$$

Thus, for any $\varepsilon > 0$ and $0 < t < \frac{1}{2}$, by Chebyshev's inequality (see, for example, van der Vaart and Wellner (1996)), we have

$$P\left(\frac{\xi_1 + \ldots + \xi_n}{n} > 1 + \varepsilon\right) \leqslant \frac{1}{\exp\{(1+\varepsilon)nt\}} E\exp\{t(\xi_1 + \ldots + \xi_n)\} = \exp\{-n\, f_\varepsilon(t)\},$$

where $f_\varepsilon(t) = \frac{1}{2}\log(1-2t) + (1+\varepsilon)t$. Setting the derivative $f_\varepsilon'(t)$ to zero gives $t = \varepsilon/2(1+\varepsilon)$, where f_ε attains the maximum $A_\varepsilon = \{\varepsilon - \log(1+\varepsilon)\}/2$, $\varepsilon > 0$. Therefore, we have

$$P\{n^{-1}(\xi_1 + \ldots + \xi_n) > 1 + \varepsilon\} \leqslant \exp(-A_\varepsilon n).$$

This proves inequality (47).

(b) For any $0 < \varepsilon < 1$ and $t > 0$, by Chebyshev's inequality and equation (49), we have

$$P\{n^{-1}(\xi_1 + \ldots + \xi_n) < 1 - \varepsilon\} \leqslant \frac{1}{\exp(tn\varepsilon)} E\exp\{t(1-\xi_1) + \ldots + t(1-\xi_n)\} = \exp\{-n\, g_\varepsilon(t)\},$$

where $g_\varepsilon(t) = \frac{1}{2}\log(1+2t) - (1-\varepsilon)t$. Taking $t = \varepsilon/2(1-\varepsilon)$ yields inequality (48).

Lemma 4 (moderate deviation). For any $C > 0$, there are constants c_1 and c_2 with $0 < c_1 < 1 < c_2$ such that

$$P\left(\langle \mathbf{S}\mathbf{e}_1, \mathbf{e}_1 \rangle < c_1 \frac{n}{p} \text{ or } > c_2 \frac{n}{p}\right) \leqslant 4\exp(-Cn). \tag{50}$$

Proof. From lemma 2, we know that $\langle \mathbf{Se}_1, \mathbf{e}_1 \rangle =^{(\mathrm{d})} \xi/\eta$, where ξ is χ^2_n distributed and η is χ^2_p distributed. Note that A_ε and B_ε are increasing in ε and have the same range $(0, \infty)$. For any $C > 0$, it follows from the proof of lemma 3 that there are $\tilde{c}_1$ and $\tilde{c}_2$ with $0 < \tilde{c}_1 < 1 < \tilde{c}_2$, such that $B_{1-\tilde{c}_1} = C$ and $A_{\tilde{c}_2 - 1} = C$. Now define

$$\mathcal{A} = \left\{ \frac{\xi}{n} < \tilde{c}_1 \text{or} > \tilde{c}_2 \right\},$$
$$\mathcal{B} = \left\{ \frac{\eta}{p} < \tilde{c}_1 \text{or} > \tilde{c}_2 \right\}.$$

Let $c_1 = \tilde{c}_1/\tilde{c}_2$ and $c_2 = \tilde{c}_2/\tilde{c}_1$. Then, it can easily be shown that

$$\left\{ \langle \mathbf{Se}_1, \mathbf{e}_1 \rangle < c_1 \frac{n}{p} \text{or} > c_2 \frac{n}{p} \right\} \subset \mathcal{A} \cup \mathcal{B}. \tag{51}$$

It follows from inequalities (47) and (48) and the choice of $\tilde{c}_1$ and $\tilde{c}_2$ above that

$$\begin{aligned} P(\mathcal{A}) &\leqslant 2\exp(-Cn), \\ P(\mathcal{B}) &\leqslant 2\exp(-Cp). \end{aligned} \tag{52}$$

Therefore, by $p \geqslant n$ and Bonferroni's inequality, the results follow from expressions (51) and (52).

A.6. Deviation inequality on $\langle \mathbf{Se}_1, \mathbf{e}_2 \rangle$

Lemma 5. Let $\mathbf{Se}_1 = (V_1, V_2, \ldots, V_p)^{\mathrm{T}}$. Then, given that the first co-ordinate $V_1 = v$, the random vector $(V_2, \ldots, V_p)^{\mathrm{T}}$ is uniformly distributed on the sphere $S^{p-2}\{\sqrt{(v - v^2)}\}$. Moreover, for any $C > 0$, there is some $c > 1$ such that

$$P(|V_2| > cn^{1/2} p^{-1} |W|) \leqslant 3\exp(-Cn), \tag{53}$$

where W is an independent $\mathcal{N}(0, 1)$-distributed random variable.

Proof. In view of equation (46), it follows that

$$\|b\mathbf{V}\|^2 = \mathbf{e}_1^{\mathrm{T}} \mathbf{Se}_1 = V_1,$$

where $\mathbf{V} = (V_1, \ldots, V_p)^{\mathrm{T}}$. For any $Q \in \mathcal{O}(p-1)$, let $\tilde{Q} = \mathrm{diag}(1, Q) \in \mathcal{O}(p)$. Thus, by lemma 1, we have

$$\begin{aligned} \tilde{Q}\mathbf{V} &\overset{(\mathrm{d})}{=} (\mathbf{U}\tilde{Q}^{\mathrm{T}})^{\mathrm{T}} \,\mathrm{diag}(I_n, 0)(\mathbf{U}\tilde{Q}^{\mathrm{T}})\tilde{Q}\mathbf{e}_1 \\ &\overset{(\mathrm{d})}{=} \mathbf{U}^{\mathrm{T}} \,\mathrm{diag}(I_n, 0)\mathbf{U}\mathbf{e}_1 \overset{(\mathrm{d})}{=} \mathbf{V}. \end{aligned}$$

This shows that, given $V_1 = v$, the conditional distribution of $(V_2, \ldots, V_p)^{\mathrm{T}}$ is invariant under the orthogonal group $\mathcal{O}(p-1)$. Therefore, given $V_1 = v$, the random vector $(V_2, \ldots, V_p)^{\mathrm{T}}$ is uniformly distributed on the sphere $S^{p-2}\{\sqrt{(v - v^2)}\}$.

Let $W_1, \ldots, W_{p-1}$ be IID $\mathcal{N}(0, 1)$-distributed random variables, independent of V_1. Conditioning on V_1, we have

$$V_2 \overset{(\mathrm{d})}{=} \sqrt{(V_1 - V_1^2)} \frac{W_1}{\sqrt{(W_1^2 + \ldots + W_{p-1}^2)}}. \tag{54}$$

Let $C > 0$ be a constant. From the proof of lemma 4, we know that there is some $c_2 > 1$ such that

$$P(V_1 > c_2 n/p) \leqslant 2\exp(-Cn). \tag{55}$$

It follows from inequality (48) that there is some $0 < c_1 < 1$ such that

$$P\{W_1^2 + \ldots + W_{p-1}^2 < c_1(p-1)\} \leqslant \exp\{-C(p-1)\} \leqslant \exp(-Cn), \tag{56}$$

since $p > n$. Let $c = \sqrt{(c_2/c_1)}$. Then, by $V_1 - V_1^2 \leqslant V_1$ and Bonferroni's inequality, inequality (53) follows immediately from expressions (54)–(56).

A.7. Verifying property C for Gaussian distributions

In this section, we check property (16) for Gaussian distributions. Assume that $\mathbf{x}$ has a p-variate Gaussian distribution. Then, the $n \times p$ design matrix $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, I_n \otimes \mathbf{\Sigma})$ and

$$\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, I_n \otimes I_p) = \mathcal{N}(\mathbf{0}, I_{n \times p}),$$

i.e. all the entries of $\mathbf{Z}$ are IID $\mathcal{N}(0, 1)$ random variables, where the symbol '$\otimes$' denotes the Kronecker product of two matrices. We shall invoke results in the random matrix theory on extreme eigenvalues of random matrices in Gaussian ensemble.

Before proceeding, let us make two simple observations. First, in studying singular values of $\mathbf{Z}$, the role of n and p is symmetric. Second, when $p > n$, by letting $\mathbf{W} \sim \mathcal{N}(\mathbf{0}, I_{m \times p})$, independent of $\mathbf{Z}$, and

$$\tilde{\mathbf{Z}}_{(n+m) \times p} = \begin{pmatrix} \mathbf{Z} \\ \mathbf{W} \end{pmatrix},$$

then the extreme singular values of $\mathbf{Z}$ are sandwiched by those of $\tilde{\mathbf{Z}}$. Therefore, a combination of lemmas 6 and 7 below immediately implies property (16).

Lemma 6. Let $p \geqslant n$ and $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, I_{n \times p})$. Then, there is some $C > 0$ such that, for any eigenvalue λ of $p^{-1}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}$ and any $r > 0$,

$$P\{|\lambda^{1/2} - E(\lambda^{1/2})| > r\} \leqslant C \exp(-pr^2/C).$$

Moreover, for each λ, the same inequality holds for a median of $\lambda^{1/2}$ instead of the mean.

Proof. See proposition 3.2 in Ledoux (2005) and note that Gaussian measures satisfy the dimension-free concentration inequality (3.6) in Ledoux (2005).

Lemma 7. Let $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, I_{n \times p})$. If $p/n \to \gamma > 1$ as $n \to \infty$, then we have

$$\lim_{n \to \infty} (\text{median}[\sqrt{\{\lambda_{\max}(p^{-1}\mathbf{Z}\mathbf{Z}^{\mathrm{T}})\}}]) = 1 + \gamma^{-1/2}$$

and

$$\liminf_{n \to \infty} (E[\sqrt{\{\lambda_{\min}(p^{-1}\mathbf{Z}\mathbf{Z}^{\mathrm{T}})\}}]) \geqslant 1 - \gamma^{-1/2}.$$

Proof. The first result follows directly from Geman (1980):

$$\lambda_{\max}(p^{-1}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}) \to (1 + \gamma^{-1/2})^2 \qquad \text{almost surely as } n \to \infty.$$

For the smallest eigenvalue, it is well known that (see, for example, Silverstein (1985) or Bai (1999))

$$\lambda_{\min}(p^{-1}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}) \to (1 - \gamma^{-1/2})^2 \qquad \text{almost surely as } n \to \infty.$$

This and Fatou's lemma entail the second result.

References

Antoniadis, A. and Fan, J. (2001) Regularization of wavelets approximations (with discussion). *J. Am. Statist. Ass.*, **96**, 939–967.

Bai, Z. D. (1999) Methodologies in spectral analysis of large dimensional random matrices, a review. *Statist. Sin.*, **9**, 611–677.

Bai, Z. D. and Yin, Y. Q. (1993) Limit of smallest eigenvalue of a large dimensional sample covariance matrix. *Ann. Probab.*, **21**, 1275–1294.

Baron, D., Wakin, M. B., Duarte, M. F., Sarvotham, S. and Baraniuk, R. G. (2005) Distributed compressed sensing. *Manuscript*.

Barron, A., Cohen, A., Dahmen, W. and DeVore, R. (2008) Approximation and learning by greedy algorithms. *Ann. Statist.*, **36**, 64–94.

Bickel, P. J. and 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**, 989–1010.

Bickel, P. J. and Levina, E. (2008) Regularized estimation of large covariance matrices. *Ann. Statist.*, **36**, 199–227.
Bickel, P. J., Ritov, Y. and Tsybakov, A. (2008) Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.*, **36**, in the press.
Breiman, L. (1995) Better subset regression using the nonnegative garrote. *Technometrics*, **37**, 373–384.
Breiman, L. (1996) Heuristics of instability and stabilization in model selection. *Ann. Statist.*, **24**, 2350–2383.
Candes, E. and Tao, T. (2007) The Dantzig selector: statistical estimation when p is much larger than n (with discussion). *Ann. Statist.*, **35**, 2313–2404.
Chikuse, Y. (2003) Statistics on special manifolds. *Lect. Notes Statist*, **174**.
Donoho, D. L. (2000) High-dimensional data analysis: the curses and blessings of dimensionality. *American Mathematical Society Conf. Math Challenges of the 21st Century*.
Donoho, D. L. and Elad, M. (2003) Maximal sparsity representation via l_1 minimization. *Proc. Natn. Acad. Sci. USA*, **100**, 2197–2202.
Donoho, D. L. and Huo, X. (2001) Uncertainty principles and ideal atomic decomposition. *IEEE Trans. Inform. Theory*, **47**, 2845–2862.
Donoho, D. L. and Johnstone, I. M. (1994) Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**, 425–455.
Eaton, M. L. (1989) *Group Invariance Applications in Statistics*. Hayward: Institute of Mathematical Statistics.
Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004) Least angle regression (with discussion). *Ann. Statist.*, **32**, 407–499.
Fan, J. (1997) Comments on "Wavelets in statistics: a review," by A. Antoniadis. *J. Ital. Statist. Ass.*, **6**, 131–138.
Fan, J. and Fan, Y. (2008) High dimensional classification using features annealed independence rules. *Ann. Statist.*, to be published.
Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Ass.*, **96**, 1348–1360.
Fan, J. and Li, R. (2002) Variable selection for Cox's proportional hazards model and frailty model. *Ann. Statist.*, **30**, 74–99.
Fan, J. and Li, R. (2006) Statistical challenges with high dimensionality: feature selection in knowledge discovery. In *Proc. Int. Congr. Mathematicians* (eds M. Sanz-Sole, J. Soria, J. L. Varona and J. Verdera), vol. III, pp. 595–622. Freiburg: European Mathematical Society.
Fan, J. and Peng, H. (2004) Nonconcave penalized likelihood with diverging number of parameters. *Ann. Statist.*, **32**, 928–961.
Fan, J. and Ren, Y. (2006) Statistical analysis of DNA microarray data. *Clin. Cancer Res.*, **12**, 4469–4473.
Frank, I. E. and Friedman, J. H. (1993) A statistical view of some chemometrics regression tools (with discussion). *Technometrics*, **35**, 109–148.
Freund, Y. and Schapire, R. E. (1997) A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, **55**, 119–139.
Friedman, J., Hastie, T., Höfling, H. and Tibshirani, R. (2007) Pathwise coordinate optimization. *Ann. Appl. Statist.*, **1**, 302–332.
Geman, S. (1980) A limit theorem for the norm of random matrices. *Ann. Probab.*, **8**, 252–261.
George, E. I. and McCulloch, R. E. (1997) Approaches for Bayesian variable selection. *Statist. Sin.*, **7**, 339–373.
Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. and Lander, E. S. (1999) Molecular classification of cancer: class discovery and class prediction by expression monitoring. *Science*, **286**, 531–537.
Greenshtein, E. (2006) Best subset selection, persistence in high dimensional statistical learning and optimization under l_1 constraint. *Ann. Statist.*, **34**, 2367–2386.
Greenshtein, E. and Ritov, Y. (2004) Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, **10**, 971–988.
Grenander, U. and Szegö, G. (1984) *Toeplitz Forms and Their Applications*. New York: Chelsea.
Gribonval, R., Mailhe, B., Rauhut, H., Schnass, K. and Vandergheynst, P. (2007) Average case analysis of multichannel thresholding. In *Proc. Int. Conf. Acoustic and Speech Signal Processing*. New York: Institute of Electrical and Electronics Engineers.
Hall, P., Marron, J. S. and Neeman, A. (2005) Geometric representation of high dimension, low sample size data. *J. R. Statist. Soc.* B, **67**, 427–444.
Huang, J., Horowitz, J. and Ma, S. (2008) Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann. Statist.*, **36**, 587–613.
Hunter, D. and Li, R. (2005) Variable selection using MM algorithms. *Ann. Statist.*, **33**, 1617–1642.
Johnstone, I. M. (2001) On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.*, **29**, 295–327.
Knight, K. and Fu, W. (2000) Asymptotics for Lasso-type estimators. *Ann. Statist.*, **28**, 1356–1378.
Lam, C. and Fan, J. (2007) Sparsistency and rates of convergence in large covariance matrices estimation. *Manuscript*.
Ledoux, M. (2001) *The Concentration of Measure Phenomenon*. Cambridge: American Mathematical Society.
Ledoux, M. (2005) Deviation inequalities on largest eigenvalues. *Manuscript*.

Meier, L., van de Geer, S. and Bühlmann, P. (2008) The group lasso for logistic regression. *J. R. Statist. Soc.* B, **70**, 53-71.
Meinshausen, N. (2007) Relaxed Lasso. *Computnl Statist. Data Anal.*, **52**, 374–393.
Meinshausen, N. and Bühlmann, P. (2006) High dimensional graphs and variable selection with the Lasso. *Ann. Statist.*, **34**, 1436–1462.
Meinshausen, N., Rocha, G. and Yu, B. (2007) Discussion of "The Dantzig selector: statistical estimation when p is much larger than n". *Ann. Statist.*, **35**, 2373–2384.
Nikolova, M. (2000) Local strong homogeneity of a regularized estimator. *SIAM J. Appl. Math.*, **61**, 633–658.
Paul, D., Bair, E., Hastie, T. and Tibshirani, R. (2008) "Pre-conditioning" for feature selection and regression in high-dimensional problems. *Ann. Statist.*, to be published.
Ravikumar, P., Lafferty, J., Liu, H. and Wasserman, L. (2007) Sparse additive models. *Manuscript*.
Silverstein, J. W. (1985) The smallest eigenvalue of a large dimensional Wishart matrix. *Ann. Probab.*, **13**, 1364–1368.
Storey, J. D. and Tibshirani, R. (2003) Statistical significance for genome-wide studies. *Proc. Natn. Acad. Sci. USA*, **100**, 9440–9445.
Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B, **58**, 267–288.
Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natn. Acad. Sci. USA*, **99**, 6567–6572.
van der Vaart, A. W. and Wellner, J. A. (1996) *Weak Convergence and Empirical Processes*. New York: Springer.
Zhang, C.-H. (2007) Penalized linear unbiased selection. *Technical Report 2007-003*. Department of Statistics, Rutgers University, Piscataway.
Zhang, C.-H. and Huang, J. (2008) The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Ann. Statist.*, **36**, 1567–1594.
Zhao, P. and Yu, B. (2006) On model selection consistency of Lasso. *J. Mach. Learn. Res.*, **7**, 2541–2567.
Zou, H. (2006) The adaptive Lasso and its oracle properties. *J. Am. Statist. Ass.*, **101**, 1418–1429.
Zou, H. and Li, R. (2008) One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.*, to be published.

Discussion on the Paper by Fan and Lv

Peter Bickel (*University of California at Berkeley*)

Professor Fan and Professor Lv are to be congratulated on this timely paper. A paradigm of much statistical activity is the generalized regression setting. We observe repeatedly a large number of covariates $\mathbf{X}$ and an outcome Y: what our machine learning colleagues call a training sample. Our goals, in a stylized form, are twofold:

(a) to construct as effective a method as possible for predicting a new Y given its $\mathbf{X}$;
(b) to gain insight into the relationships between $\mathbf{X}$ and Y for scientific purposes, as well as, hopefully, to construct an improved prediction method.

Fan and Lv's focus is very much on this second aspect, which is also known as model selection.

We live in an era of massive and complex data arising in many fields of endeavours ranging from the hard sciences such as physics, astronomy and biology through the social sciences such as economics to critical applications such as medical science and various aspects of engineering. All these situations are marked by

(i) a very large number p of predictors and
(ii) a much more modest number n of observations.

The paper's timeliness comes from its dealing head on with this environment.

Formally, Fan and Lv consider the linear regression model

$$\mathbf{Y}_{n\times 1} = (\mathbf{X}^{(1)}_{n\times 1}, \ldots, \mathbf{X}^{(p)}_{n\times 1})\boldsymbol{\beta}_{p\times 1} + \varepsilon_{n\times 1},$$

where $p \gg n$.

Let $\mathcal{M} \equiv \{j : \beta_j \neq 0\}$ and $\tilde{\mathcal{M}}$ be the smallest such set, $|\tilde{\mathcal{M}}| < n$, where $|\cdot|$ is cardinality. Fan and Lv propose *sure* independence screening and sequential and other refinements to obtain an estimate $\mathcal{M}^*$ of $\tilde{\mathcal{M}}$ with $\mathcal{M}^* \lesssim n$.

They focus on the important property, which was also introduced by Meinshausen and Yu (2008), that, as $p, n \to \infty$,

$$\mathbb{P}(\mathcal{M}^* \supset \tilde{\mathcal{M}}) \to 1.$$

Their paper stimulated me to ask three questions.

Question 1

Sure independence screening corresponds roughly to testing

$$H_j : \beta_j = 0, \qquad j = 1, \ldots, p,$$

where both for H_j and its alternative all $\beta_k, k \neq j$, are 0, and then screening out the $100(1-\alpha)\%$ largest p-values.

Do methods such as Benjamini and Hochberg's (1995) and later methods attempting to keep the false discovery rate at α have consistency properties under the same conditions?

Question 2

'All models are false but some models are useful'

G. E. P. Box

Suppose that there are a number of linear models which fit equally well or, even, a linear model holds, but $\mathcal{M}$ is not unique so $\boldsymbol{\beta}$ is not identifiable. The natural question is, what variables should be presented as important since the R-representations or approximations to $\mathbb{E}[Y|\mathbf{X}]$ that are given by $\Sigma\{\beta_j \mathbf{X}^{(j)} : j \in \mathcal{M}_m\}, m = 1, \ldots, R, |\mathcal{M}_m \leqslant K$ small, are equally good. What we would like is something like

$$\mathbb{P}(\mathcal{M}^* \supset \text{all potentially important factors}) \to 1.$$

I give a possible definition of importance in my discussion to Candes and Tao (2007) (Bickel, 2007).

Question 3

Fan and Lv propose 'screen first; fit after, whereas others (Bickel *et al.*, 2008; Meinshausen and Bühlmann, 2006; Bühlmann and Meier, 2008; Zou and Li, 2008) propose 'fit first; screen after'.

(a) How do these compare in terms of consistency and oracle properties?
(b) Is there sequentially no real difference?

As my questions indicate, I found this paper very stimulating and it is my pleasure to propose the vote of thanks.

Peter Bühlmann (*Eidgenössiche Technische Hochschule, Zürich*)

I congratulate Fan and Lv for their stimulating and thought-provoking paper. Variable screening is among the primary goals in high dimensional data analysis. Having a computationally efficient and statistically accurate method for retaining relevant and deleting thousands of irrelevant variables is highly desirable.

Sure independence screening (SIS) is a marginal method. This makes it very easy to use. To understand the properties of a marginal view, consider the well-known relationship for a linear model of the form $Y = \Sigma_{j=1}^{p} \beta_j X^{(j)} + \varepsilon$:

$$\beta_j \neq 0 \Leftrightarrow \text{Parcorr}(Y, X^{(j)} | \{X^{(j)}; k \neq j\}) \neq 0.$$

Of course, if $\text{corr}(X^{(j)}, X^{(k)}) = 0$ for $j \neq k$ there is an exact correspondence to the marginal view:

$$\beta_j \neq 0 \Leftrightarrow \text{corr}(Y, X^{(j)}) \neq 0.$$

And, in fact, Fan and Lv justify SIS for the situation with fairly uncorrelated variables: their discussion of condition 4 in Section 5.1 implies (for large p) that the correlation matrix among the X-variables is 'not too far away' from the identity. In contrast with the purely marginal view, it is possible to start with the marginal approach and then gradually to consider partial correlations from low to higher order. This can be achieved within the framework of so-called faithful distributions, a concept which is mainly used in the literature about graphical modelling. For linear models, Bühlmann and Kalisch (2008) introduce partial faithfulness which holds if and only if for every j

$$\text{Parcorr}(Y, X^{(j)} | X^{(\mathcal{S})}) = 0 \text{ for some } \mathcal{S} \subseteq \{1, \ldots, p\} \setminus j \Rightarrow \beta_j = 0.$$

Bühlmann and Kalisch (2008) argue that the class of linear models satisfying this condition is quite broad. Roughly speaking, the partial faithfulness assumption implies that a large (in absolute value) marginal or partial correlation does not tell us much, but a zero (partial) correlation says plenty. The idea of SIS is the other way round: a large marginal correlation is interpreted as importance for the corresponding variable

Table 8. Summary of computationally tractable methods for variable screening

Method	*Assumption*	*Computational complexity*
SIS	'Fairly' uncorrelated covariates	$O(np)$
Lasso	Coherence condition for design	$O\{np\min(n, p)\}$
PC algorithm	Partial faithfulness	$O(np^{\gamma})$ $(1 \leqslant \gamma \leqslant C)$

whereas no decision is taken for small correlations. The PC algorithm (Spirtes *et al.*, 2000) exploits the partial faithfulness assumption. Instead of assuming fairly uncorrelated covariates (as in SIS) or partial faithfulness, the lasso (Tibshirani, 1996) is another alternative which requires some coherence assumptions for the design matrix ruling out cases with too strong linear dependence (of certain design submatrices). The methods are summarized in Table 8.

The exponent γ in the computational complexity of the PC algorithm (see Table 8) depends on the underlying sparsity. Asymptotic theory for high dimensional settings include, for the lasso, Meinshausen and Bühlmann (2006), van de Geer (2008), Meinshausen and Yu (2008), Zhang and Huang (2008) and Bickel *et al.* (2008), and, for the PC algorithm, Kalisch and Bühlmann (2007) and Bühlmann and Kalisch (2008). For finite samples, we consider two simulation models: model 1,

$$\text{example III from Section 4.2.3,} \qquad p = 1000, n = 50, \rho = 0.5$$

model 2,

$$Y = \sum_{j=1}^{5} X^{(j)} + \varepsilon, \qquad p = 1000, n = 50, \rho = 0.5.$$

For model 1, Fan and Lv report that $\mathbb{P}(\mathcal{M}_{\text{true}} \subseteq \hat{\mathcal{M}}) = 0$ for SIS and the lasso when using $|\hat{\mathcal{M}}| = n - 1$. But some differences between the methods can be easily detected in Figs 7 and 8. In addition to the single number $\mathbb{P}(\mathcal{M}_{\text{true}} \subseteq \hat{\mathcal{M}})$, it is important to report also performance measures such as the number of true positive results or receiver operating characteristic curves to obtain a more complete picture. In our small simulation study we see that the lasso (and also the PC algorithm) has a better 'global' accuracy than SIS. The price to pay for this higher accuracy is a more complicated procedure although we should note that

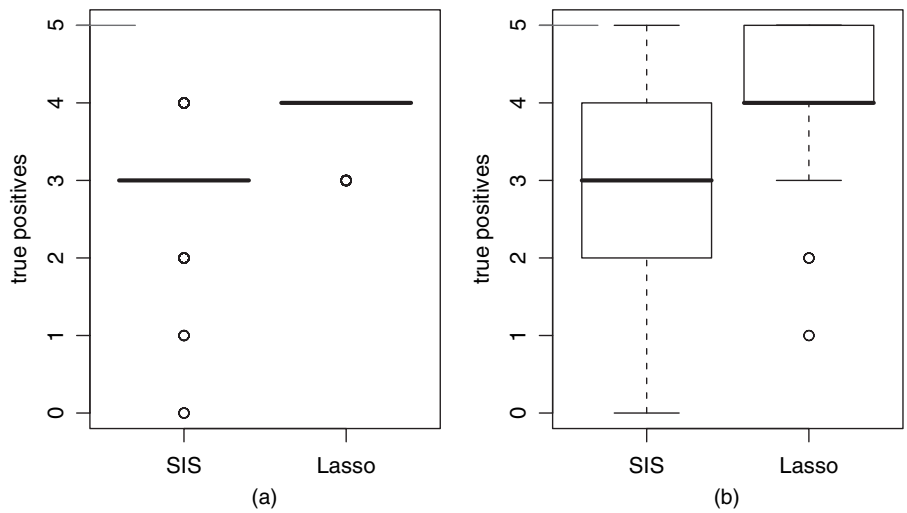

Fig. 7. Boxplots of the number of true positives $\Sigma_{j=1}^{p} I(\hat{\beta}_j \neq 0, \beta_j \neq 0)$ (with $\hat{\mathcal{M}} = n - 1$) for 100 simulations from (a) model 1 and (b) model 2: the number of effective variables equals 5

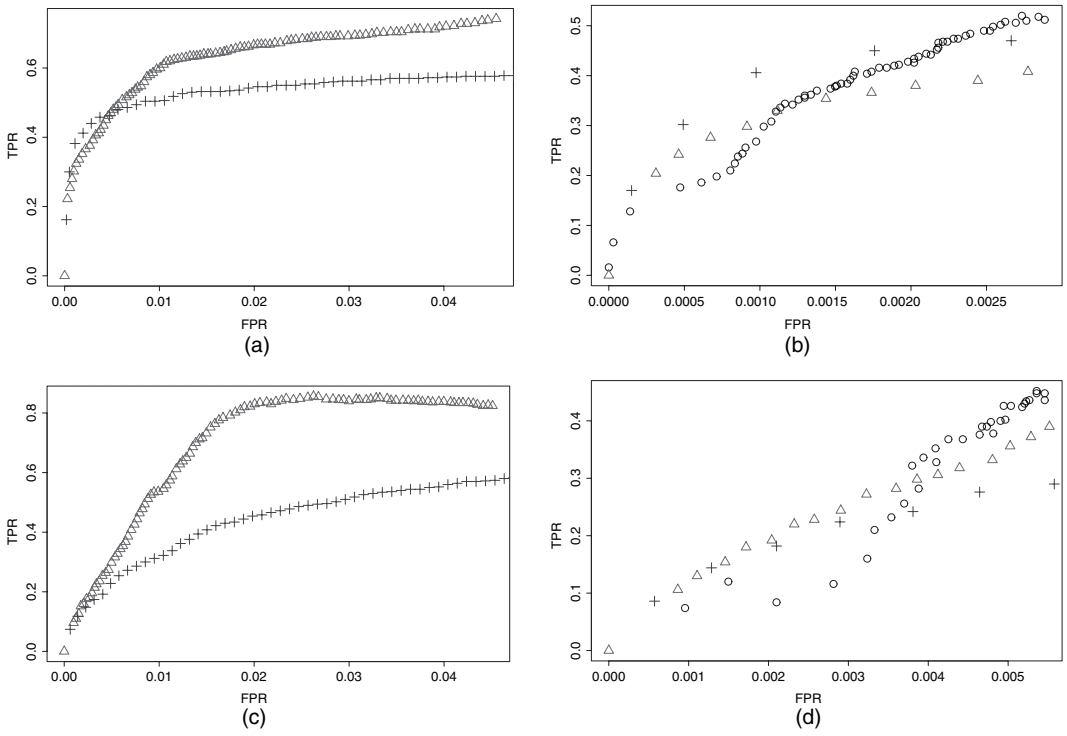

Fig. 8. Receiver operating characteristic curves for (a), (b) model 1 and (c), (d) model 2 (Δ, lasso; +, SIS; O, PC algorithm): (b), (d) enlargement for the domain with small false positive rate

the lasso has also linear computational complexity in dimensionality p if $p \gg n$. Interestingly, we note that SIS does well in the conservative domain where the false positive rate is very low. I do not know whether we can expect such behaviour in a wide variety of scenarios: if such findings would be true in general, this would indeed be a strong argument in favour of the simple SIS method for detecting very few but most relevant variables among say thousands of others. A (presumably difficult) theory which would support such a finding is lacking though.

I agree with the authors that iterative SIS (ISIS) mitigates many of the problems with the marginal approach of SIS. However, we need to choose a tuning parameter k (or denoted in the paper by $k_1, k_2, \ldots, k_l$) which is really unpleasant: ideally, for some rough sort of variable screening, no other tuning parameter should be involved except the number of variables which are to be selected from screening. When using $k = 1$ in ISIS, we end up with a procedure which is somewhere between orthogonal matching pursuit (which is almost identical to forward variable selection) and matching pursuit, which is the same as L_2-boosting with componentwise linear least squares (Friedman, 2001; Bühlmann and Yu, 2003; Bühlmann, 2006). In particular, in the high dimensional setting with fairly low signal-to-noise ratio, the boosting approach is in our experience often better than orthogonal matching pursuit or forward variable selection. Why should we use ISIS? Why should we not use the established boosting approach for variable screening (which is presumably not so different from the lasso; see Efron *et al.* (2004))? And, if there are strong reasons for ISIS, how should we select the tuning parameter k for screening whose optimal choice may be in conflict with accurate prediction?

Finally, the authors stress the fact about ultrahigh dimensionality. In their framework, the dimensionality $p = p_n$ is a function of sample size such that

$$\log(p_n) = O(n^{\xi}) \qquad \text{for some } \xi > 0.$$

The usual approaches in asymptotic analysis (exponential inequalities and entropy arguments) would require that $\xi < 1$, which is equivalent to $\log(p_n)/n \to 0$ $(n \to \infty)$. Fan and Lv write in Section 5.1 (in the discussion of condition 1) that 'the concentration property (16) makes restrictions on ξ'. What is the upper

bound for $\xi > 0$, e.g. in the Gaussian case? Do we see here another range of high dimensionality, or is ultra-high dimensionality the same as high dimensionality where $\log(p_n)/n \to 0$?

It is my pleasure to second the vote of thanks: this paper will stimulate plenty of future research.

The vote of thanks was passed by acclamation

Qiwei Yao (*London School of Economics and Political Science*)

The authors are congratulated for tackling a challenging statistical problem with a simple and effective procedure—sure independence screening. Directly motivated by their work, An *et al.* (2008) have revisited some conventional stepwise regression procedures coupled with information criteria. The method adopted consists of two stages. In stage 1, forward addition is performed to grow the regression model until a modified Bayes information criterion BIC stops to decrease. In stage 2, backward deletion is employed to delete redundant variables again according to the modified BIC. The computation is carried out by using the standard sweep operation.

The conventional BIC must be modified to cope with the cases $p > n$ or $p \gg n$. More precisely, we have proposed the two modified versions of BIC as follows:

$$\mathrm{BICP}_k = \log(\hat{\sigma}_k^2) + \frac{2k}{n}\log(p),$$

$$\mathrm{BICC}_k = \log(\hat{\sigma}_k^2 + c_0) + \frac{k}{n}\log(n),$$

where k denotes the number of selected regression variables, and $c_0 > 0$ is a fixed constant. BICP uses the penalty weight $2\log(p)/n$ (instead of $\log(n)/n$) to penalize the models with large k heavily. It has been proved that the above two-stage procedure using BICP leads to a consistent estimator for the true sparse

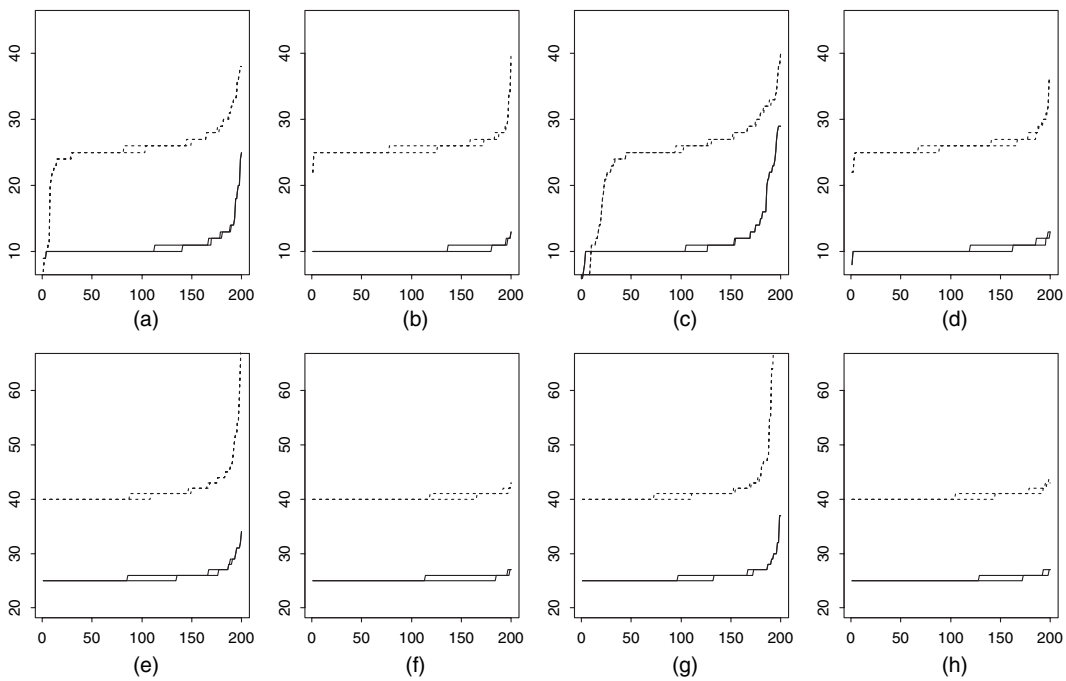

Fig. 9. Plots of the numbers of selected regression variables by the forward search and backward search in 200 replications: (a) BICP, $n = 200$, $p = 1000$, $s = 10$ (———) and $s = 25$ (-------); (b) BICC, $n = 200$, $p = 1000$, $s = 10$ (———) and $s = 25$ (-------); (c) BICP, $n = 200$, $p = 2000$, $s = 10$ (———) and $s = 25$ (-------); (d) BICC, $n = 200$, $p = 2000$, $s = 10$ (———) and $s = 25$ (-------); (e) BICP, $n = 800$, $p = 10000$, $s = 25$ (———) and $s = 40$ (-------); (f) BICC, $n = 800$, $p = 10000$, $s = 25$ (———) and $s = 40$ (-------); (g) BICP, $n = 800$, $p = 20000$, $s = 25$ (———) and $s = 40$ (-------); (h) BICC, $n = 800$, $f = 20000$, $s = 25$ (———) and $s = 40$ (-------)

model. In contrast, BICC uses the same penalty weight $\log(n)/n$ as the standard BIC. But it inserts a positive constant c_0 in the logarithmic function. In fact BICC also applies when p is fixed.

Fig. 9 plots the selected numbers of regression variables in ascending order from a simulation study with 200 replications. We use model (1) with the first $s\,\beta$ non-zero, all $x_{ij} \sim N(0,1)$, and ε_i independent $N(0,1)$. The non-zero β are of the form $(-1)^u(b+|v|)$, where $b = 2.5\sqrt{\{2\log(p)/n\}}$, u is a Bernoulli random variable with $P(u=1) = P(u=0) = 0.5$ and $v \sim N(0,1)$. The dependences between different x_{ij} are set as follows: for any $1 \leqslant k \leqslant n$ and $1 \leqslant i \neq j \leqslant s$, $\mathrm{corr}(X_{ki}, X_{kj}) = (-1)^{u_1} \times 0.5^{|i-j|}$, $\mathrm{corr}(X_{ki}, X_{k,i+s}) = (-1)^{u_2}\rho$ and $\mathrm{corr}(X_{ki}, X_{k,i+2s}) = (-1)^{u_3}(1-p^2)^{1/2}$, where $\rho \sim U[0.2, 0.8]$, and u_1, u_2 and u_3 are independent and are of the same distribution as u. The numerical results indicate that both criteria work fine, although the performance of BICC is better than that of BICP.

Richard Samworth (*University of Cambridge*)

I congratulate the authors for a very interesting and timely contribution to an important problem. The power of the methodology is well demonstrated and there are many opportunities to explore its possible

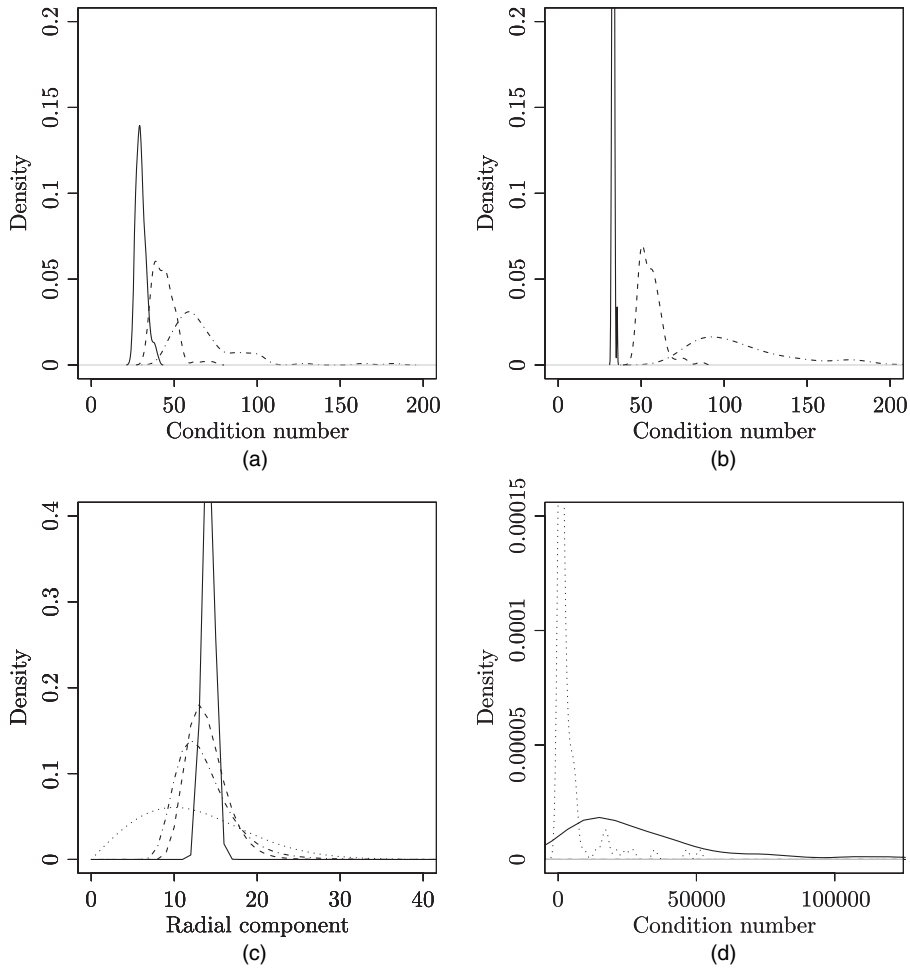

Fig. 10. (a), (b) Estimated densities of the condition number of $\tilde{p}^{-1}\tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^{\mathsf{T}}$, based on 100 simulations, when $\mathbf{Y}$ is Gaussian (———) and multivariate t with $\nu = 20$ (-------) and $\nu = 10$ (· - - - ·) degrees of freedom ((a) $n = 100$, $\tilde{p} = 200$; (b) $n = 1000$, $\tilde{p} = 2000$), (c) corresponding densities of the radial components of these spherically symmetric densities when $\tilde{p} = 200$, as well as the density of a scaled Weibull distribution with shape parameter 2 (·····), and (d) estimated condition number density of $\tilde{p}^{-1}\tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^{\mathsf{T}}$ in the Weibull case, with $n = 100$, $\tilde{p} = 200$ (·····) and $n = 1000$, $\tilde{p} = 2000$ (———)

extensions beyond the linear model. Most of the technical conditions required for their theoretical results are natural and interpretable. An exception is the concentration property that is imposed on the $n \times p$ matrix $\mathbf{Z}$ in expression (16), namely that there exist $c, c_1 > 1$ and $C_1 > 0$ such that

$$P\{\lambda_{\max}(\tilde{p}^{-1}\tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^{\mathrm{T}}) > c_1 \text{ or } \lambda_{\min}(\tilde{p}^{-1}\tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^{\mathrm{T}}) < 1/c_1\} \leqslant \exp(-C_1 n)$$

for any $n \times \tilde{p}$ submatrix $\tilde{\mathbf{Z}}$ of $\mathbf{Z}$ with $cn \leqslant \tilde{p} \leqslant p$. The authors prove that this condition is satisfied when the entries of $\mathbf{Z}$ are independent standard normal random variables and we now study their conjecture that it holds for a wide class of spherically symmetric distributions.

The condition is expected to be most restrictive when $\tilde{p}$ is at the lower end of the range, so in the simulations below I took $\tilde{p} = cn$ with $c = 2$. To compare with the case of independent Gaussian entries, the rows of $\tilde{\mathbf{Z}}$ were taken to be independent and each row was generated as $a\mathbf{Y}$, where the distribution of $\mathbf{Y}$ was multivariate t with $\nu = 10$ and $\nu = 20$ degrees of freedom, and a was such that each component had unit variance. Figs 10(a) and 10(b) plot estimated densities of the condition number of $\tilde{p}^{-1}\tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^{\mathrm{T}}$ (i.e. the ratio of the largest and smallest eigenvalues) when $n = 100$ and $n = 1000$ respectively. Note that, in the Gaussian case, the condition number density becomes more concentrated as n increases. For the multivariate t case with $\nu = 20$, the distribution is quite similar at both sample sizes, whereas when $\nu = 10$ it is more dispersed with a long right-hand tail at the larger sample size, suggesting that the concentration property may fail to hold there.

We can generate the rows of $\tilde{\mathbf{Z}}$ having a spherically symmetric distribution as $aR\mathbf{U}$, where the direction $\mathbf{U}$ is uniform on the unit sphere in $\mathbb{R}^{\tilde{p}}$, where R is independent of $\mathbf{U}$ and is supported on $[0,\infty)$, and a is chosen so that each component has unit variance. However, it is important to remember (see Hall *et al.* (2005), for instance) that, in high dimensions, the distribution of the radial component aR tends to be quite highly concentrated—see Fig. 10(c). Even when R has a distribution that is traditionally thought of as light tailed in a univariate context, such as a Weibull distribution with shape parameter 2, the distribution of aR is much more dispersed (Fig. 10(c)), and Fig. 10(d) shows that in such a context the corresponding matrix $\tilde{\mathbf{Z}}$ can become badly ill conditioned as n becomes large.

Overall, then, we conclude that although condition (16) appears reasonable for the authors' purposes, it remains of interest to describe the theoretical properties of independence screening when the rows of $\mathbf{Z}$ have heavier tails and condition (16) may fail.

Peter Hall (*University of Melbourne*) **and D. M. Titterington and Jing-Hao Xue** (*University of Glasgow*)

We thank Fan and Lv for their innovative and stimulating paper. Here we propose a regression-oriented version of the classification-based approach that was developed in Hall *et al.* (2008) and raise the general issue of using a predictive model as a prelude to variable selection.

Our tilting procedure boiled down to the quadratic programming problem of minimizing

$$d_{L2}(q, q_0) = \sum_{k=1}^{p} (q_k - p^{-1})^2,$$

subject to

$$\sum_{k=1}^{p} q_k \Delta_k = c_1, \tag{57}$$

$\Sigma_{k=1}^{p} q_k = 1$, and $q_k \geqslant 0$, for each k, where Δ_k is a cross-validated measure of the performance of the classifier provided by the kth feature variable X_k. Variables corresponding to $q_k = 0$ are deselected from the classifier.

In the linear regression analogue, Δ_k is very close to being a linear function of $\hat{\rho}_k^2$, the squared sample correlation coefficient between Y and X_k, so, as with the method of Fan and Lv, variables exhibiting low sample correlation with the response are deselected, the cut-off dictating or being dictated by c_1 in equation (57).

However, it is not always appropriate to attack a variable selection problem via one of linear prediction, and an advantage of the method of Hall *et al.* (2008) is that it does not require a predictive model as an intermediary for variable selection. To appreciate the dangers, note that, if one of the significant components of X influences Y in a non-linear, indeed non-monotone, way, then it might be overlooked by a linear model approach; see Segal *et al.* (2003) for a real gene expression example of this and Hall and Miller (2008) for a more general discussion.

Finally, a few details: first, if in equation (9) Y_i is scored as n_1^{-1} for class 1 and $-n_2^{-1}$ for class 2, w_j corresponds to the t-statistic for any sample sizes. Secondly, correlation-based variable selection frequently

appears in the machine learning literature (Guyon and Elisseeff, 2003), although typically without the theoretical detail of the current paper. Thirdly, unless $\mathcal{A}_1$ in Section 4.1.1 excludes any unimportant predictor that is highly correlated with the important predictors, iterative sure independence screening will not deal with the first issue that is referred to in Section 4.1.

C. Anagnostopoulos and D. K. Tasoulis (*Imperial College London*)

Professor Fan and Professor Lv address the increasingly important problem of ultrahigh dimensional variable selection by considering the utility of marginal correlations, rather than regression coefficients. Modern applications, including bioinformatics and text mining, often generate such data, and developing methods that are both computationally efficient and able to handle extreme dimensionality is a contemporary challenge.

Much work on variable selection concerns a 'true model', which is defined in terms of regression coefficients rather than marginal correlations. Selection via regression coefficients tends to overfit, so screening with marginal correlations is certainly a good idea. Fan and Lv are to be congratulated for developing a theoretical framework for screening based on marginal correlations. This framework readily accommodates a second-stage selection procedure, using other approaches. The proof that, under suitable conditions, the consistency of smoothly clipped absolute deviation is preserved by sure independence screening is a particularly useful and promising result. This approach is computationally attractive also, since the screening does not require matrix inversion. Notably, in contexts where data points arrive sequentially, the computations necessary for sure independence screening may be implemented in an incremental fashion. In conjunction with recursive implementations of the lasso and other estimators (e.g. Anagnostopoulos *et al.* (2008)) this could allow for on-line variable selection in ultrahigh dimensional data streams.

Although Fan and Lv are content with their theoretical assumptions, it would be very interesting to weaken them, in particular

(a) that the true model is linear in the sample size and
(b) that variables in the true model are marginally correlated with the response.

Investigating the extent to which these assumptions can be relaxed may enhance the intuitive appeal of the theorems, as well as yield insights into the applicability of sure independence screening to real data contexts where the assumption of a small true model may be unwarranted.

We have a small comment about the leukaemia data analysis. This data set has been extensively studied and used as a benchmark for various data reduction methods. The results that are reported in the paper indicate that the sure independence screening–smoothly clipped absolute deviation combination can yield very high classification rates that are comparable with other methods (e.g. Tassoulis *et al.* (2006)) by selecting only 16 variates. It would be interesting to examine the degree of overlap with those genes which are identified by domain experts as important predictors (Golub *et al.*, 1999).

Wenyang Zhang (*University of Bath*) **and Yingcun Xia** (*National University of Singapore*)

We congratulate Professor Fan and Professor Lv for such a brilliant paper. We believe that this paper will have a huge influence on high dimensional inference and will stimulate many further researches.

Componentwise regression is an easy way to select variables; however, it may suffer from collinearity. In this paper, Professor Fan and Professor Lv cleverly avoid this problem by iteratively using componentwise regression, which is quite interesting. It is also very interesting to see that such a simple and easy-to-implement method enjoys so many good asymptotic properties.

If we understand the paper correctly, the core idea is to select the variables by two stages. In the first stage, a simple, easy-to-implement and quick method is used to remove the least important variables. In the second stage, a more delicate, sophisticated and accurate method is applied to reduce the variables further. It is very important for the method in the first stage to be unbiased; otherwise, the whole variable selection procedure would fail. The proposed sure independence screening (SIS) may not be unbiased under some circumstances. We appreciate that asymptotically we know when SIS is biased; it is just a matter of checking whether the technical conditions hold or not. However, practically, how do we know when SIS is biased? The proposed iterative SIS (ISIS) seems more safe to use to guard against bias. Is it sensible always to use ISIS?

The selection of tuning parameter d is not an issue in SIS; it can be taken as either $n-1$ or $[n/\log(n)]$. However, in ISIS, it could be an issue, because the selected variables are cumulated through the iterations. The number of selected variables may be above n if the tuning parameter is not carefully selected. Is there any way to select the tuning parameter optimally? Also, is there any stopping rule for ISIS?

Intuitively, traditional forward selection in model selection can almost overcome the collinearity problem. It would be interesting to see a comparison between ISIS and forward selection or stepwise regression.

Iain M. Johnstone (*Stanford University*)

I join the discussants in thanking the authors for a fascinating paper. My comments expand on the authors' remarks about multiple correlation, specifically in their introduction on how large they can be even in the absence of any dependence. For example, when the number of variables p and q is proportional to sample size n, Wachter (1980) showed that, for Gaussian data, the largest sample canonical correlation $R_{p,q;n}$ converges to $\rho_{p,q;n} = \sqrt{\{p/n(1-q/n)\}} + \sqrt{\{q/n(1-p/n)\}}$. For $p = 10, q = 20$ and $n = 100$, values that are quite plausible for a small empirical study, this is as large as 0.71.

An asymptotic approximation to the null distribution of the largest sample canonical correlation is now available (Johnstone, 2008a). It states that $\text{logit}(R_{p,q})$ is approximately distributed as $\mu + \sigma W_1$, where simple explicit formulae are available for μ and σ; and W_1 follows the Tracy-Widom distribution for real-valued variates. This approximation is quite accurate at conventional percentiles (90th, 95th and 99th) for even the smallest values of $p = q = 2$—the relative error in the percentiles is less than 10% and becomes much smaller as p and q grow even a little (Johnstone, 2008b).

Finally, the authors make use of the Dantzig selector that was introduced by Candes and Tao (2007) and establish its consistency in theorem 4 under an assumption on restricted orthogonality of the columns of the design matrix **A**. We note that there is a concentration of measure property for the null distribution of the largest canonical correlation. It states that, for $t > 0$,

$$P(R_{p,q;n} \geqslant ER_{p,q;n} + t) \leqslant \exp\{-(n-2)t^2/8\}.$$

It follows from the asymptotic approximation result that was referred to above that, for p sufficiently large, $ER_{p,q;n} < \rho_{p,q;n}$. Work in progress shows that in turn this leads to concrete bounds for a slightly modified notion of restricted orthogonality.

Sylvia Richardson and Leonardo Bottolo (*Imperial College London*)

This discussion relates to the applicability of sure independence screening (SIS) and its extension iterative SIS (ISIS) to analyse realistic data sets. Most of the theoretical developments presented assume that the covariates are quasi-independent: a favourable case for asymptotics but an unlikely situation in most applications, in particular in genetics and genomics. Acknowledging that the conditions imposed in Section 5.1 to limit the correlations are restrictive, the authors propose the extension ISIS but, in doing so, relinquish sure screening properties. If ISIS has no theoretical foundation, its use relies on arbitrary choices which could give misleading results.

In all the examples shown, correlations between predictors are simulated in a symmetric fashion. Such symmetry is rarely encountered but could be the key behind the apparent good performance of ISIS. In genetics and genomics, we typically see block-like patterns of correlated variables, in line with biological pathways or linkage disequilibrium. To investigate the performance of ISIS under realistic conditions, we built a test-case using as predictors 1421 non-redundant markers across the whole genome of rats recombinant inbred lines ($n = 29$), publicly available from the STAR Consortium (`www.snp-star.eu`). The predictors have complex correlation patterns created by population structure (Fig. 11(a) (bottom)). We simulated three true effects at position (90, 721, 1300) with size (-1, 2.5, 1) and $N(0, 0.75)$ noise. This mimics a typical situation in genetics where, besides a large effect, there could be a few others with smaller magnitude.

Applying the ISIS–lasso method fails to recover the true effects (one out of three is totally ignored; Figs 11(a), middle, and 11(b), bottom), highlighting that additional tuning would be necessary, and that 'sure screening' properties in correlated cases will be elusive. A more useful formulation of the problem would acknowledge the inherent uncertainty of variable selection in high dimensional space and would aim, instead, to find a well-supported set of solutions by exploring the large model space. Such exploration is possible using, for example, powerful Bayesian variable selection algorithms that cope with arbitrary correlation structure among the predictors. Our own evolutionary stochastic search algorithm ESS (Bottolo and Richardson, 2008) gives posterior support to the marginal inclusion of the three simulated markers (Fig. 11(b)); the top three models visited by ESS are the true model ($R^2 = 0.962$), another three-marker model involving close-by marker 98 instead of 90 ($R^2 = 0.957$) and a single marker (721) model ($R^2 = 0.879$), giving a coherent picture of model uncertainty.

We have built several such examples, giving us serious misgivings about the use of ISIS for model choice and feature selection in cases where structured correlation exists among the predictors.

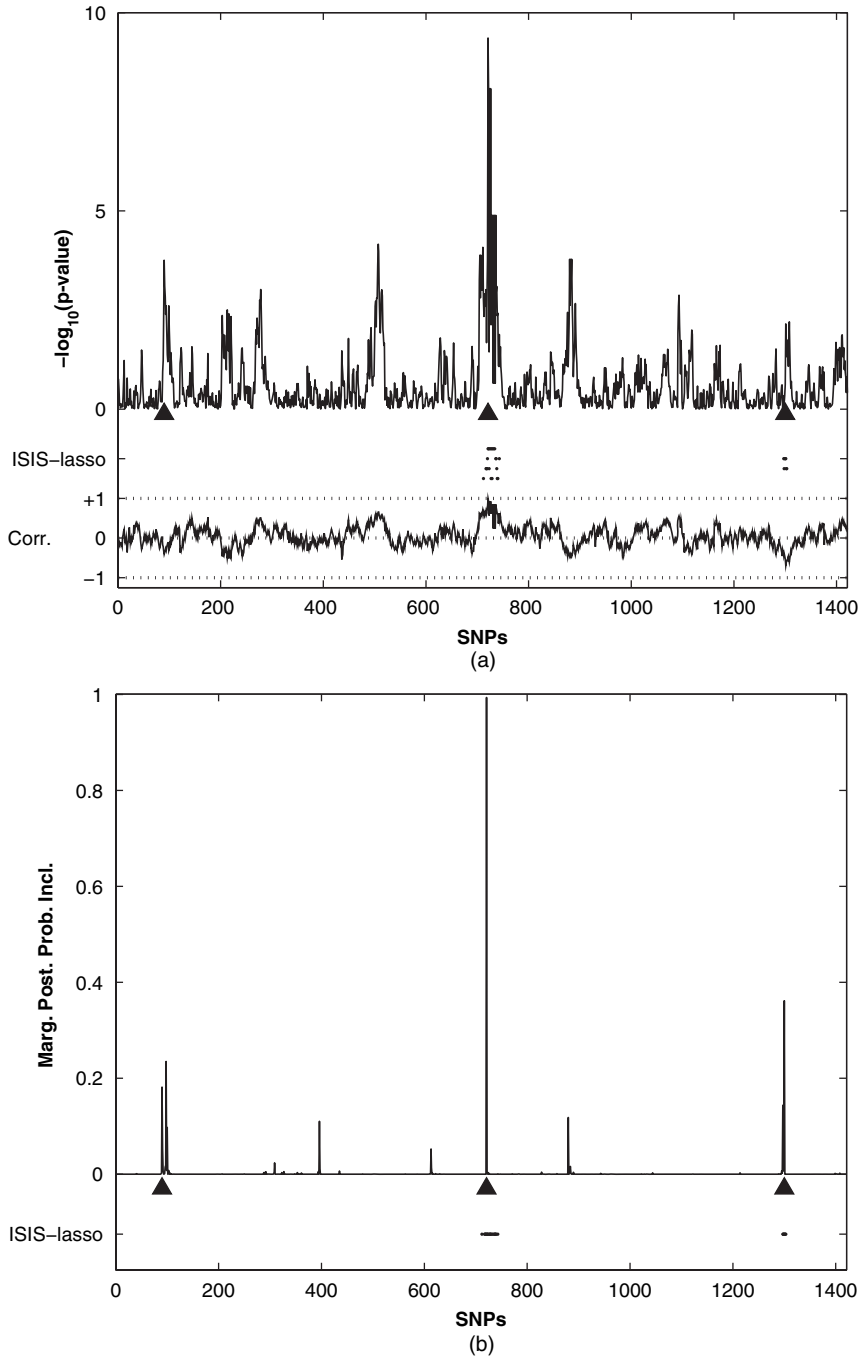

Fig. 11. (a) $-\log_{10}$-transformation of p-values obtained from a t-test (top), markers selected from running the ISIS–lasso (output from four successive iterations with $[n/\log(n)] = 8$) (middle), sample correlation between the marker at position 721 and the set of remaining predictors (bottom) and (b) marginal posterior probability of inclusion for the simulated data set applying evolutionary stochastic search (top) and the final set of 28 selected markers from running the ISIS–lasso (bottom): ▲, position where the effects have been simulated

John T. Kent (*University of Leeds*)

This paper falls under the general heading of model simplification in regression analysis. There are two main approaches to this problem:

(a) variable selection or sparsity, as in this paper, and
(b) subspace selection, focusing on selected linear combinations of the x-variables.

The latter point of view forms the basis of several techniques including principal component regression, partial least squares and also the inverse regression methods that have been developed by Ker-Chau Li, Dennis Cook and others. In particular, Li *et al.* (2007) developed links between partial least squares and inverse regression.

Although there has been some interaction between traditions (a) and (b) (e.g. Li (2007)), the two traditions have largely been developed independently. However, one point of contact is componentwise regression, which appears in this paper as a starting point in Section 2.1 and also lies behind the simplest version of partial least squares. After standardizing the x-variables to have variance 1, this regression estimate can be phrased either in terms of the forward regression of y on each component of x in turn or in terms of the backward regression of each component of x on y. Of course, if the x-variables are uncorrelated, this estimate coincides with the ordinary least squares regression estimate. However, there seems to be some hidden regularity in practical problems which makes it a sensible estimate in a much wider range of settings.

The following contributions were received in writing after the meeting.

Kofi Adragni and R. Dennis Cook (*University of Minnesota, Minneapolis*)

Fan and Lv give a compelling case for predictor screening based essentially on marginal linear relationships and penalization. We have been working on methodology called screening by principal fitted components (SPFC) (Cook, 2007) in the $p \gg n$ context. Instead of using marginal relations, or a forward regression of Y on $\mathbf{X}$ as in penalized least squares, we adopt an inverse regression approach regressing $\mathbf{X}$ on Y. Consider the relatively simple inverse regression model

$$\mathbf{X} = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\lambda}\mathbf{f}_y + \boldsymbol{\varepsilon}. \tag{58}$$

The term $\mathbf{f}_y \in \mathbb{R}^r$ is a user-selected function of y, $\boldsymbol{\mu} \in \mathbb{R}^p$, $\boldsymbol{\Gamma} \in \mathbb{R}^{p \times d}$, $\boldsymbol{\lambda} \in \mathbb{R}^{d \times r}$ and $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \mathbf{I})$. Cook (2007) showed that $Y \perp\!\!\!\perp \mathbf{X} | \boldsymbol{\Gamma}^{\mathrm{T}} \mathbf{X}$. With $d = 1$, estimating the sparse subspace span($\boldsymbol{\Gamma}$) is equivalent to estimating span($\boldsymbol{\beta}$) in equation (1) of the paper, the zero elements of $\boldsymbol{\Gamma}$ identifying the predictors to be screened.

Consider the special case with $\boldsymbol{\Gamma} \in \mathbb{R}^p$ and $\mathbf{f}_y = y - \bar{y}$. Let $\boldsymbol{\Phi} = \boldsymbol{\Gamma}\boldsymbol{\lambda}$ and $\mathbb{X}^{\mathrm{T}} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$. The maximum likelihood estimator under the inverse model (58) of the $p \times 1$ vector $\Phi = (\phi_1, \phi_2, \ldots, \phi_p)^{\mathrm{T}}$ is $\mathbb{X}^{\mathrm{T}}(y - \bar{y})$. After normalization this corresponds to the $\boldsymbol{\omega}$ of equation (2) in the paper. Consequently, SPFC reduces to sure independence screening (SIS) with $\mathbf{f}_y = y - \bar{y}$. Following Fan and Lv, we can select predictors by taking the first $[\gamma n]$ with the largest normalized $|\phi_i|$. But we can also tie the selection to test statistics for $\phi_i = 0$, which automatically introduces an equivalent scaling. Because of these connections we expect that SIS will work best when var($\mathbf{X}|Y$) is a diagonal matrix, which is not a case represented in Fan and Lv's simulations.

The inverse regression approach is more flexible than forward regressions. Unlike SIS or penalized least squares, inverse regression models can easily accommodate a categorical response, non-linearities and non-constant variance var($Y|\mathbf{X}$) and still perform well. As an example, we generated $n = 70$ observations on $p = 500$ independent predictors $\mathbf{X} = (X_1, \ldots, X_p)^{\mathrm{T}}$ with $X_1 \sim$ Unif(l, 10) and $X_i \sim N(0, 4)$, $i = 2, \ldots, p$. The response was generated as $y = (5X_1)\varepsilon$ where $\varepsilon \sim N(0, 1)$. We used SIS and generated 200 data sets to estimate the frequency that the only active predictor X_1 is among the first 35 (Fan and Lv's $\gamma = 0.5$) with the largest normalized $|\phi_i|$. The result was as expected under random selection: SIS included X_1 in the first 35 predictors 12% of time. In contrast, SPFC with a piecewise linear basis $\mathbf{f}_y$ captured X_1 among the first *two* predictors 98% of the time. We have obtained similar results with non-linear mean functions and a constant variance. Our results so far suggest that SPFC is an effective generalization of SIS, and that it can be enhanced to produce a generalization of ISIS.

Ursula Gather and Charlotte Guddat (*Dortmund University of Technology*)

Firstly, we congratulate the authors on their fine paper. It not only provides a very useful approach to variable selection for ultrahigh dimensional predictor spaces but also is an approach which is as simple as it can be—an aspect often undervalued.

Our interest is in the robustness of sure independence screening (SIS). As SIS employs the componentwise correlations ω_j, $j=1,\ldots,p$, rescaled by the standard deviation of Y it can be conjectured that the variable selection based on this non-robust measure suffers from outliers and contaminations of the data.

A straightforward robustification of SIS which we suggest replaces the classical estimators by their robust counterparts, i.e. the median and the median absolute deviation are used instead of the mean and the standard deviation respectively. For estimating the correlation one can use for example the Gnanadesikan–Kettenring estimator (Gnanadesikan and Kettenring, 1972). We call the resulting method robust SIS (ROSIS). This is a similar robustification to the dimension adjustment method for the non-robust sliced inverse regression method of dimension reduction (Li, 1991; Gather *et al.*, 2001, 2002).

We compare SIS and ROSIS under the two simple models

$$Y = X_1 + 0.1\varepsilon \tag{59}$$

and

$$Y = 5X_1 + 5X_2 + 5X_3 + \varepsilon \tag{60}$$

where the predictor vector $\mathbf{X} = (X_1, \ldots, X_p)^{\mathrm{T}} \sim \mathcal{N}^p(0,1)$ and the error term $\varepsilon \sim \mathcal{N}(0,1)$ are independent.

Table 9. Accuracy of SIS and ROSIS in including the true model $\{X_1\}$ or $\{X_1, X_2, X_3\}$ respectively

p	Model	Results for the following methods:					
		Clean		10% outlier		20% outlier	
		SIS	ROSIS	SIS	ROSIS	SIS	ROSIS
100	(59)	1.000	1.000	0.828	1.000	0.788	1.000
	(60)	1.000	0.996	0.714	0.944	0.722	0.942
1000	(59)	1.000	1.000	0.072	1.000	0.068	1.000
	(60)	1.000	0.824	0.054	0.796	0.052	0.740

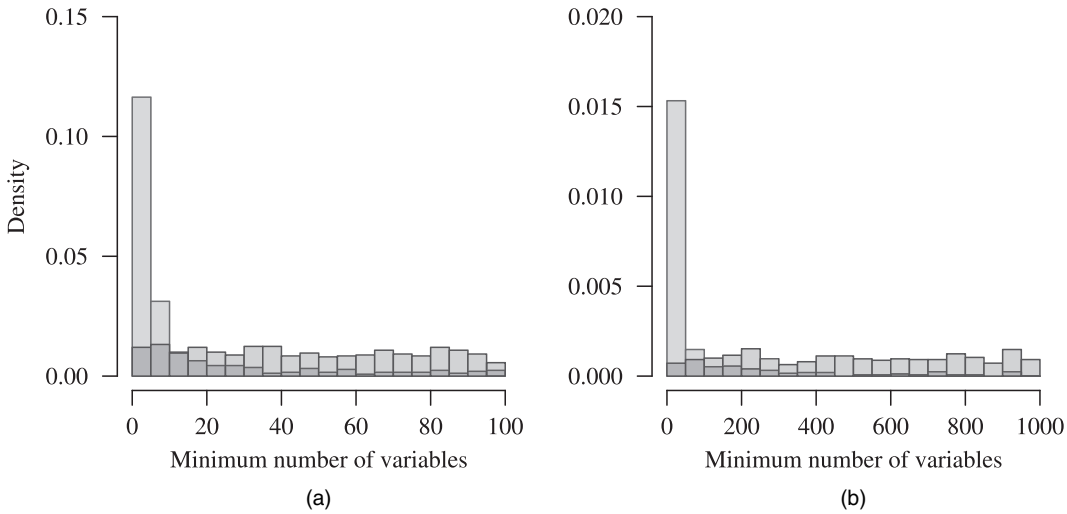

Fig. 12. Distribution of the minimum number of selected variables required to include the true model (60) by using SIS (■) and ROSIS (□) for samples of size $n = 70$ contaminated by seven outliers: (a) $p = 100$; (b) $p = 1000$

In each case we generate $n = 20, 50, 70$ data points with dimension $p = 100, 1000$. To analyse the sensitivity of SIS against outliers we contaminate the samples by adding $2\sqrt{\chi^2_{p;0.999}}$ to 0%, 10% or 20% of the observations of X_1 (see also Gather *et al.* (2001)). As performance criterion, we calculate the percentage of cases where the true submodel is included when we reduce to the dimension $n - 1$; see Table 9 for $n = 70$. Also, we count the minimum number of variables such that all true predictors are selected; see Fig. 12 for 10% contamination and $n = 70$.

The simulation results show clearly that ROSIS outperforms SIS when the data are contaminated by outliers, whereas in the clean situation SIS works only slightly better.

Though being aware of iterative SIS (ISIS), which performs better, we have only concentrated on SIS so far. Certainly, ISIS will yield better results than SIS under contamination but its performance will also suffer from outliers. We then suggest replacing SIS by ROSIS and using a robust model selection procedure in the ISIS algorithm.

Eitan Greenshtein (*Duke University, Durham*)

This paper of Fan and Lv provides further understanding of fundamentals in regression with $p \gg n$, and even with $p \gg\gg n$.

In what follows I shall question how crucial variable selection is. This is in spite of the very impressive advantages of the screening that is suggested by Fan and Lv.

When the ultimate goal is to find a parsimonious model, variable selection is essential by definition. I shall consider cases where the ultimate goal is to provide a good prediction. In such cases, variable selection is still performed, as a way of regularization. It is a convention that, given n observations, we should avoid using procedures which depend on more than n variables. I wonder to what extent this convention is helpful.

In Section 2.3 the authors write that 'classification can be regarded as a specific case of regression problem'. I shall stretch this analogy.

Consider an example where $\mathbf{X} = (X_1, \ldots, X_p)$ is a random vector of explanatory variables and Y is a response variable $Y = 0, 1$. Suppose that $\mathbf{X} \sim N(\boldsymbol{\mu}^i, I)$ when $Y = i$, $i = 0, 1$. Assume equal prior probabilities for the events $Y = i$, $i = 0, 1$. Suppose that there are $n^i = 25$ examples for which $Y = i$, $i = 0, 1$. Thus, there are total of $n = n^0 + n^1 = 50$ observations.

The optimal classifier is Fisher's rule, which requires knowledge of $\boldsymbol{\mu}^i$, $i = 0, 1$. As demonstrated by Fan and Fan (2008), estimating high dimensional $\boldsymbol{\mu}^i$ by the corresponding maximum likelihood estimator and plugging into Fisher's rule could lead to a very weak classifier.

Consider a case where $p = 100\,000$ and $\boldsymbol{\mu}^0 = (0, \ldots, 0)$, while $\boldsymbol{\mu}^1$ has 50 000 zero entries and its remaining 50 000 entries are all equal to 0.03. Even the optimal classifier, based on an optimal selection of (say) $d = 50$ variables, would have a misclassification rate 0.46. Simulations, in which the maximum likelihood estimator for $\boldsymbol{\mu}^i$, $i = 0, 1$, is plugged into Fisher's rule, resulted in an average misclassification rate of 0.41. Let $\boldsymbol{\nu} = \boldsymbol{\mu}^1 - \boldsymbol{\mu}^0$. In Greenshtein and Park (2008) it is demonstrated by simulations that a classifier which is based on estimating $\boldsymbol{\nu}$ by a non-parametric empirical Bayes method and plugged into (a slightly modified) Fisher's rule would have an average misclassification rate 0.11. This classifier depends virtually on *all* the $p = 100\,000$ variables, though the weight of each variable is very small.

Fan and Lv's paper is devoted to sparse situations, unlike our above configuration. However, the above empirical Bayes method performs very well also in sparse situations, compared with various classification methods especially designed for such situations.

Gareth M. James and Peter Radchenko (*University of Southern California, Los Angeles*)

We congratulate the authors on introducing a powerful new methodology for addressing an increasingly important problem. Although the theoretical aspects of this work are impressive we have concentrated our discussion on the practical behaviour of the authors' methodology. The basic moral of this paper is that when dealing with extremely large numbers of predictors one should use an iterative two-step approach. At each iteration, one first uses a simple bivariate criterion to rank the predictors and hence to obtain a 'moderate' number of variables. Then a multivariate variable selection method is used to obtain the final set of predictors. The authors present convincing evidence that this approach can produce considerable improvements, in terms of both computational cost as well as statistical accuracy, over directly working with the full data set. The idea that a large number of variables can be discarded with little risk of eliminating important variables seems reasonable.

The authors work primarily with smoothly clipped absolute deviation (SCAD) when implementing the iterative sure independence screening (ISIS) approach. We were interested in the robustness of ISIS to different plug-in methods. Hence we reran the simulation results from Section 4.2.1 using two alternatives

Table 10. Simulation comparison†

p	n	ρ	*Results for the following methods:*					
			ISIS	*Lasso*	*Iterative lasso*	*Forward*$_1$	*Forward*$_{n/4}$	*Forward*$_{n/2}$
100	20	0	1.000	0.970	0.885	0.730	0.850	0.895
		0.5	1.000	0.985	0.820	0.515	0.790	0.865
	50	0	1.000	1.000	1.000	1.000	1.000	1.000
		0.5	1.000	1.000	1.000	1.000	1.000	1.000
1000	20	0	1.000	0.340	0.305	0.250	0.275	0.235
		0.5	1.000	0.556	0.180	0.025	0.130	0.165
	50	0	1.000	1.000	1.000	1.000	1.000	1.000
		0.5	1.000	1.000	0.985	0.940	0.990	1.000

†The ISIS and lasso results are taken from Table 4. The iterative lasso, Forward$_1$, Forward$_{n/4}$ and Forward$_{n/2}$ methods respectively replace SCAD in the ISIS method with the lasso and forward selection using $K=1$, $K=n/4$ and $K=n/2$.

to the SCAD plug-in. The first replaced SCAD with the lasso. The second replaced SCAD with a version of forward selection which selected the K variables with largest correlations to the response. We utilized three values: $K=1$, $K=n/4$ and $K=n/2$. In all other respects the set-up was the same as for Section 4.2.1. Our results are provided in Table 10.

For the $n=50$ scenarios all methods gave almost perfect predictions. For the $p=100, n=20$, scenario we found that the iterative forward method improved as K grew with $K=n/2$ giving slightly superior results to those of the iterative lasso approach. For the $p=1000, n=20$, scenario the iterative lasso outperformed the iterative forward methods. However, interestingly, the iterative lasso either gave the same performance as the standard lasso or performed worse. In addition the iterative lasso and forward selection methods both substantially underperformed compared with the iterative SCAD results reported by Fan and Lv. We drew the following conclusions from these results.

First, applying the iterative approach does not always cause an improvement, as demonstrated by the inferior performance of the iterative lasso over the standard lasso. Second, at least in certain scenarios, the iterative approach seems to be sensitive to the plug-in method with SCAD providing significantly superior results to those of the lasso and forward selection methods.

Chenlei Leng (*National University of Singapore*) **and Hansheng Wang** (*Peking University, Beijing*)

We congratulate Professor Fan and Professor Lv for a thought-provoking paper, which provides us with deep understanding about variable selection in an ultrahigh dimensional set-up. We would like to comment as follows.

The important work of Breiman (1996) and Tibshirani (1996) demonstrated clearly that shrinkage estimation is a promising solution for variable selection. The first paper on the asymptotic results of the lasso was Knight and Fu (2000). However, the important question regarding whether those shrinkage methods are consistent in model selection (Shao, 1997) was not clear. In a seminal paper, Fan and Li (2001) developed smoothly clipped absolute deviation (SCAD) and, more importantly, introduced a general theoretical framework to understand the asymptotic behaviour of various shrinkage methods. As a consequence, Fan and Li (2001) is also partially responsible for the recent development of the adaptive lasso methods (Zou, 2006; Wang *et al.*, 2007a; Zhang and Lu, 2007).

Note that the oracle properties defined in Fan and Li (2001) depend on an appropriate selection of tuning parameters, for which prediction-based criteria such as generalized cross-validation have been commonly used in practice. Nevertheless, Leng *et al.* (2006) and Wang *et al.* (2007b) showed that this practice leads to seriously overfitted models. For model selection consistency, a Bayes information type of criterion is a justifiable alternative. Results were established for SCAD (Wang *et al.*, 2007b) and the adaptive lasso (Wang and Leng, 2007) with a fixed dimension, and also for these two methods with diverging model dimensions (Wang *et al.*, 2008).

It is very natural to ask whether similar results can be established in an ultrahigh dimensional set-up. In particular, we are very interested in knowing the answers to the following questions.

(a) How can the parameter γ in the first stage of sure independence screening (SIS) be automatically tuned? The authors' numerical studies suggest that $[kn/\log(n)]$ might be a good choice, with a reasonably range of κ (e.g. $\kappa = 1, 2, \ldots$). However, we still believe that a completely data-driven choice can make SIS more attractive for real practitioners.
(b) What is known about the stochastic error involved in SIS's first-stage screening? Is it ignorable in its second-stage shrinkage estimation? Are the Bayes information criteria developed in the existing literature still applicable? We believe that research along those directions will further enhance the applicability of SIS in an ultrahigh dimensional setting.

Lastly, we conclude by congratulating the authors again for such a wonderful piece of work!

Elizaveta Levina and Ji Zhu (*University of Michigan, Ann Arbor*)

We congratulate the authors on developing an attractive and practical method for high dimensional variable selection with many potential applications. One area where this method may have applications is genomewide association studies, where one is interested in identifying common genetic factors that influence health and disease. For complex diseases and traits, the genetic contribution of a true association is often expected to be moderate. In terms of the model in the paper, this would result in a low signal-to-noise ratio (SNR) $\mathrm{var}(X\beta)/\mathrm{var}(\varepsilon)$ but, in the simulations in Section 4, the SNRs are all relatively high, ranging from 40 to 200. Thus we decided to investigate briefly the behaviour of iterative sure independence screening (ISIS) under lower SNR levels.

To illustrate the point, we mimicked the simple simulated example I from Section 4.2.1. Specifically, we considered the linear model

$$Y = 5X_1 + 5X_2 + 5X_3 + \varepsilon,$$

where $X_1, \ldots, X_p$ ($p = 1000$) have a multivariate normal distribution $N(0, \Sigma)$ and $\varepsilon \sim N(0, \sigma^2)$ is independent of the predictors. The covariance matrix Σ has diagonal elements equal to 1 and off-diagonal elements equal to $\rho = 0.5$. Instead of setting $\sigma = 1$, which corresponds to an SNR of 150, we considered several different values of σ, i.e. $\sigma = 1, 2, 4, 8, 12, 16$. The last scenario $\sigma = 16$ corresponds to an SNR of 0.6.

We considered $n = 25$,50,100 and ran ISIS exactly as described in Section 4.1.1 using the lasso at the second stage of variable selection, with the tuning parameter selected by the Bayes information criterion. For each simulation, we recorded how many important variables (out of X_1, X_2 and X_3) were selected. The results over 100 replications are summarized in Table 11.

Not surprisingly, as the SNR decreases, the performance of ISIS degrades. For example, when $n = 50$, $p = 1000$ and $\sigma = 12$, which corresponds to an SNR of about 1 (which is considered relatively high in genomewide association studies), the ISIS could still identify some important variables (identifying at least one important variable in 61 out of 100 simulations) but could not identify all three important variables.

Table 11. Simulation results for $p = 1000$ and $\rho = 0.5$†

σ	*Results for the following values of n:*											
	$n = 25$				$n = 50$				$n = 100$			
	#0	*#1*	*#2*	*#3*	*#0*	*#1*	*#2*	*#3*	*#0*	*#1*	*#2*	*#3*
1	16	32	18	34	0	0	0	100	0	0	0	100
2	21	35	20	24	1	1	1	96	0	0	0	100
4	23	37	25	15	2	4	22	72	0	0	1	99
8	62	35	2	1	9	42	38	11	1	0	26	73
12	73	25	2	0	39	39	22	0	12	20	45	23
16	84	16	0	0	63	32	5	0	25	51	17	7

†The true model contains three important variables. #3 corresponds to the number of times that all three important variables were selected by ISIS out of 100 simulations, and similarly for #2, #1 and #0.

A theoretical question to consider is whether the asymptotic results in the paper could be extended to incorporate the SNR or σ^2 into the rate explicitly. Modifications of the method that would allow it to be applied to low SNR large-scale problems may also be an interesting topic for further investigation.

Runze Li (*Pennsylvania State University, University Park*)

Fan and Lv are to be congratulated for their inspiring work. I have some comments on both screening and post-screen variable selection.

Screening

Consider a regression model, $E(y|\mathbf{x}) = \eta(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$, and assume that $\mathbf{x}$ follows an elliptical distribution with mean $\boldsymbol{\mu}_x$ and covariance Σ_x (Fang *et al.*, 1990). It can be shown (see, for example, Li (2008)) that

$$E(\mathbf{x} - \boldsymbol{\mu}_x)y = k\Sigma_x\boldsymbol{\beta} \tag{61}$$

for some constant k. Equation (61) implies that, assuming elliptical symmetry on $\mathbf{x}$, the sure independence screening procedure may be directly applied to generalized linear models, single-index models, Cox's model and various regression models in the literature.

Post-screen variable selection

The penalized least squares problem (PLS) (11) in the paper provides a unified framework for post-screen variable selection. I would like to comment on two fundamental issues in the implementation of the PLS: tuning parameter (λ) selection and optimization of PLS.

Selection of λ plays a crucial role in PLS. This issue has been carefully studied in Wang *et al.* (2007), who suggested using a Bayes information criterion type of λ-selector to achieve the oracle property (Fan and Li, 2001). Zhang *et al.* (2008) further proposed generalized information criteria for λ-selection in penalized likelihood and studied its asymptotic behaviour.

In the same spirit as the lasso, Fan and Li (2001) advocated non-convex penalties, such as the smoothly clipped absolute deviation penalty. It is challenging to minimize PLS with non-convex penalties. This optimization problem has been studied by several researchers. With the local quadratic approximation (LQA), the minimization of PLS can be carried out by iterative ridge regression, which can be easily implemented. The LQA algorithm also provides a robust standard error formula for the resulting estimate. With the local linear approximation, one may obtain a solution to PLS by iteratively reweighed penalized L_1-regression, and the least angle regression algorithm LARS can be used to solve a weighed penalized L_1-regression. For some penalty functions, such as smoothly clipped absolute deviation, iteratively conditional (co-ordinate) minimization (ICM) provides another alternative to minimize PLS. As demonstrated in Friedman *et al.* (2007), ICM for penalized L_1-regression may be much faster than the LARS algorithm for a large-scale linear regression problem. Dziak (2004) and Zhang (2006) conducted some numerical comparisons between ICM and LQA for PLS and penalized likelihood respectively. From their numerical results, the ICM algorithm performs well.

Yufeng Liu (*University of North Carolina, Chapel Hill*)

The authors are to be congratulated for their stimulating and path breaking paper. Variable selection is an extremely important aspect in the model building process, especially for high dimensional problems. Sure independence screening (SIS) is a simple yet powerful procedure. With the solid theoretical justification, SIS allows researchers to prescreen the variables marginally before applying well-established statistical methods. It is certain that this work will have a great influence in the field of high dimensional variable screening and selection.

Various feature selection procedures for classification have been proposed to improve classification accuracy and interpretability. SIS provides another promising technique to rank features and to reduce the dimension of data for binary classification. In particular, if two classes are labelled as $y \in \{\pm 1\}$ and have n_1 and n_2 samples, the SIS criterion becomes $w_j = (n_1\bar{X}_{j,1} - n_2\bar{X}_{j,2})/(\text{standard deviation of the } j\text{th feature})$. When the sampling proportions are not balanced, i.e. $n_1 \gg n_2$ or $n_2 \gg n_1$, w_j may be affected accordingly; see Qiao and Liu (2008). One possible modification is $w_j^* = (\bar{X}_{j,1} - \bar{X}_{j,2})/(\text{standard deviation of the } j\text{th feature})$. It will be interesting to compare them.

For multiclass problems with $y \in \{1, \ldots, k\}$ and $k \geqslant 3$, feature ranking can be more challenging. There are several properties that a good ranking criterion for multiclass problems should have. First of all, the ranking criterion should be invariant of the coding system for y. For example, the popular between-groups to within-groups sum of squares ratio BW criterion (Dudoit *et al.*, 2000) satisfies this property. The BW criterion measures the relevance of each feature by the ratio of the between-class to the within-class sums

of squares. Secondly, the ranking should be robust against unbalanced sample sizes, i.e. features which discriminate smaller classes should be protected. One such example is the 'smarter' BW ratio (Dudoit *et al.*, 2000). Despite the popular use of BW ratios, there are some drawbacks. In particular, the BW ratio tends to select highly correlated features and it does not reveal interactions between features. Natural questions include how to generalize SIS for multiclass problems, and how to resolve the issues of the BW ratio.

The authors demonstrated the performance of the SIS–SCAD–LD and SIS–SCAD–NB methods. It would be interesting to see how SIS may improve the classification accuracy of other classifiers such as the support vector machine (Vapnik, 1998). Furthermore, there is some work using L_1-regularization to perform simultaneous classification and variable selection (e.g. Wang and Shen (2007) and Wu and Liu (2007)). In particular, Wang and Shen (2007) developed convergence rates of the generalization error with dimension $p = O\{\exp(n)\}$. Comparisons of SIS with L_1-regularization in the context of classification remain to be seen.

N. T. Longford (*SNTL, Reading, and Universitat Pompeu Fabra, Barcelona*)

The problem that is studied in the paper is truly formidable and the solutions proposed match it with their ingenuity. The problem is understated by the description of looking for 'a couple needles in a huge haystack', because a satisfactory solution must locate *all* the needles, even if they come with a little hay. Maybe the two kinds of error that can be committed, omission of a needle and inclusion of hay, could be associated with losses (penalties). If these could be stated (elicited from the relevant experts), then we could clearly see or be able to assess (even) more realistically the value of a method such as sure independence screening or iterative sure independence screening.

The fundamental assumption made is that there is a small number of important covariates (needles) and the remaining variables are useless (hay). A more realistic assumption is that there is a pyramid of variables; a select few of them (at the top) are very important, and then for every lower level of importance there are increasingly more variables; at the bottom of the pyramid, there is the vast majority of variables that are nearly or completely irrelevant. I appreciate that working with such a weaker assumption would be difficult. However, the methods that are considered are not subjected to a stern test in the simulations in Section 3.3 by assuming the needles and hay instead of the pyramid or a similar scenario.

With the 'pyramid' assumption, the problem is not that of identifying the minimal valid model (the select list of variables), but of finding an invalid model which yields efficient inferences. Technically, such a model is not valid because it cannot include all the covariates that are associated with non-zero regression parameters. I think that we should prefer efficiency (the best trade-off of variance and bias) to validity (no bias as an imperative), even if it makes asymptotics less relevant.

Weiqi Luo, Paul D. Baxter and Charles C. Taylor (*University of Leeds*)

Being able to select variables reliably, and automatically, in high dimensional models is a notoriously difficult problem. The paper tackles this question, introducing not only a computationally efficient method (sure independence screening (SIS)), but at the same time introducing comprehensive theory explaining the details of the procedure as well as theory to guide its practical implementation. We congratulate the authors on their excellent work.

The numerical simulations that are illustrated in Section 4.2 show the performance of the iterative SIS (ISIS) method. When n variables are selected from candidate predictors, ISIS always picks out all the true variables, but the false discovery rate is quite high. We would like to point out another relevant approach to dimension reduction. The jack-knife partial least squares regression (JKPLSR) algorithm, which was proposed by Westad and Martens (2000), is based on significance tests of the regression coefficients estimated in a PLSR model. Analogous to ISIS, JKPLSR is effective as a pre-step to reduce dimensionality.

We used the simulation study (as described in Section 4.2.1) to compare ISIS and JKPLSR. For simplicity, we considered models only with dimension $p = 100$. The performance of both methods was evaluated by correct hit and false discovery rate statistics. Table 12 shows the comparative results. Similar performance was observed when $n = 20$; however, JKPLSR produced a simpler model when $n = 50$.

A second, more specific, comment is with respect to the reliability and robustness of ISIS. According to the settings of the simulated linear models (Sections 4.2.1–4.2.3), less than 1% model error is included (comparing noise variance with signal variance). It is of interest to experiment with more simulations under different configurations, e.g.

(a) varying the variance of model error,
(b) using skew instead of symmetric distributions for modelled predictors or
(c) increasing the number of non-zero coefficients in the model.

Table 12. Results of a simulated study†

n	*Method*	*Hit rate*				*False discovery rate*			
		$\rho=0$	$\rho=0.1$	$\rho=0.5$	$\rho=0.9$	$\rho=0$	$\rho=0.1$	$\rho=0.5$	$\rho=0.9$
20	ISIS	1	1	1	1	0.175	0.175	0.175	0.175
	JKPLSR	0.991	0.997	0.995	0.970	0.123	0.100	0.088	0.101
50	ISIS	1	1	1	1	0.485	0.485	0.485	0.485
	JKPLSR	1	1	1	1	0.001	0.001	0.001	0.002

†The average correct hit and false discovery rate was calculated across 200 simulation runs; $p=100$.

It is possible that these configurations will have a strong influence on the performance of ISIS. We wonder whether the sure screening property still holds for ISIS under these general conditions, and we would welcome the authors' comments on this.

J. S. Marron (*University of North Carolina, Chapel Hill*)
The model selection issues that are addressed here are deep and the viewpoints are novel. The approaches taken are simple and straightforward, so it is the models used, the *sure screening* criterion and the style of analysis that are keenly interesting. For many real life high dimensional data contexts, I feel that the assumption (that might be critical to the whole notion of sure screening) that the number of 'important parameters' is less than the sample size is a serious limitation. For example in genetic applications it seems easily conceivable that the actual number of 'active genes' could be much larger than the number of data points. Nevertheless, it is still valuable to understand the asymptotic behaviour in this domain.

In the definition of sure screening, it sounds sensible to insist that the probability of finding important variables tends to 1, but it seems that it should make sense also to control the false discovery rate. For example one could consider making some sort of statement about the number of unimportant variables. This false discovery rate could also be used to compare methods which share the sure screening property.

In situations where it is sensible to consider a dimension which is an exponentially large function of the sample size, why does it make sense to index the asymptotics by using the sample size? It is certainly traditional to do so, but traditionally the dimension is fixed, which is certainly not relevant here. With such large dimensions, it seems more sensible to index the asymptotics by the dimension, and then to express the sample size as a function of that. In addition, this type of framework allows natural interface with fixed sample size asymptotics, as done by Hall *et al.* (2005). See Ahn *et al.* (2007) for recent results of that type.

Jeffrey S. Morris (*University of Texas M.D. Anderson Cancer Center, Houston*)
I congratulate the authors on an interesting and thought-provoking paper. Working at a cancer centre, I have been involved with various projects involving high throughput genomic data yielding extremely high dimensional data (large p) and very small sample sizes (small n). A key problem of interest in cancer research is to build models by using subsets of genomic factors to predict clinical response, with the eventual goal of developing personalized therapy strategies.

In this setting, it is common to perform separate single-variable regressions on each of the p factors (equivalent to sure independence screening (SIS) in this paper) to reduce the number of factors to a more manageable number before applying formal variable selection methods. Considering how widely used this approach is in practice, it is insightful to see a formal study of this approach, investigating its properties through theoretical exploration and simulation. It is encouraging to see positive results, suggesting that it may not be a bad idea.

However, I am concerned about how well it would perform for typical genomic data, which tend to be characterized by

(a) small sample sizes and
(b) between-gene correlation.

The smallest sample size that is considered in simulation studies is $n=200$, larger than typical microarray studies. It is more typical to have several dozen arrays and $p=30\,000$ or so features. I wonder whether SIS would perform very well in that setting. If the number of features d were chosen by $n/\log(n)$ as in

this paper, then we would pick only $d = 7$ features when $n = 20$, $d = 9$ features when $n = 30$ and $d = 13$ features when $n = 50$. Do I really expect that I should be able to pick out reliably the seven, nine or 13 crucial features from 30000 by using a univariate screening method? This is especially difficult, given that we know from biology that the expression levels of different genes are not independent, but have correlations that are induced by complex biological pathways. The set of most predictive single-factor predictors may be highly correlated and may leave out factors with moderate marginal effects but strong joint effects.

The authors acknowledge this problem and discuss an approach, iterative SIS (ISIS), that applies the principles of boosting to alleviate it. I expect ISIS to perform better, given that it focuses on partial instead of marginal effects. Further theoretical investigation of this method would be interesting and insightful.

Again, I congratulate the authors on their contribution to this important and challenging problem of variable selection in high dimensional spaces.

Christian P. Robert (*Université Paris Dauphine and Centre de Recherche en Economie et Statistique, Malakoff*)

Although I appreciate the *tour de force* involved in the paper and in particular by the proof that $\mathbb{P}(\mathcal{M}_* \subset \mathcal{M}_\gamma)$ goes to 1, I can only get an overall feeling of slight disbelief about the statistical consequences of the results contained in the paper: in short, I basically question the pertinence of assuming a 'true' model in settings when $p \gg n$.

Indeed, when constructing a statistical model like the regression model at the core of the paper, it is highly improbable that there is a *single* model, e.g. a *single* subset of regressors that explains the data. Therefore, to assume, as the authors do,

(a) that there is such a subset and
(b) that a statistical procedure will pick the 'right' regressors when applied in a context where $p \gg n$

strikes me as implausible or only applicable in formalized settings such as orthogonal regressors. If confronted by the opposite, as when reading this paper, my natural reaction is to question the final relevance of the asymptotic results in terms of statistical meaning. Once again, far from the idea of casting doubt about the mathematical validity of those asymptotic results, but they seem to be orthogonal to the purposes of statistical modelling. In most if not all practical settings, considering a large number p of potential regressors implies that a wide range of alternative submodels will enjoy the same *predictive* properties, especially if $n \ll p$ because, in this setting, an *explicative* model is in my opinion statistically meaningless. Significant variables may be identified in such cases but not a single monolithic collection of those, I am afraid.

It thus seems to me that a decisional approach that focuses on the decisional consequences of model selection rather than assuming the existence of a single true model would be more appropriate, especially because it naturally accounts for correlation between covariates. In addition, using a loss function on the βs or on the models allows for a rational definition of 'important variables', instead of the 0–1 dichotomy that is found in the paper. That traditional model choice procedures suffer from computational difficulties and are in practice producing suboptimal solutions is a recognized problem, even though more efficient exploration techniques are under development (Hans *et al.*, 2007a, b; Liang *et al.*, 2008; Bottolo and Richardson, 2008). In addition, adopting a more sensible predictive perspective means that missing the exploration of the full submodel space is only relevant if better fitting models are omitted. I also think that this is more than a mere philosophical difference of perspectives, since it has direct consequences on the way that inference is conducted and since the overall simplicity of the hard threshold is more convincing for practitioners than more elaborate modelling.

Keming Yu (*Brunel University, Uxbridge*)

Although some newly proposed variable selection methods for high dimensional statistical modelling typically fall under the rule of l_p-penalty least squares ($p = 0, 1, 2$), this paper is welcomed to attract one's eyes to the combination of simple correlation learning with other methods for the aim. However, the correlation learning rule may not work or exclude the following cases:

(a) componentwise regression mainly measures the 'linear' correlation whereas many dependent structures such as neural network training follow 'non-linear' dependence;
(b) some of the variables in regression analysis are dummy variables, and the componentwise magnitudes that are associated with these variables largely depend on the numerical values assigned, so it is likely that an unimportant dummy variable but with big assigned values is selected, and this issue

may not be easily avoided even using the extension of correlation learning that is proposed in the paper.

For this we may propose an alternative variable selection rule named 'score learning'. On the basis of a neural network, for example, we simply use a 'keep-one-in' (or 'keep-two-in') rule to evaluate the associated score function such as squared error for all p ($p \gg n$) variables, than we rank or order the output scores and take the first d ($< n$) variables with large scores (or maybe the other way round).

Along the lines of neural networks, we further point out that Hinton and Salakhutdinov (2006) have recently described a different way of training a multilayer neural network with a small central layer to reconstruct high dimensional input vectors.

Cun-Hui Zhang (*Rutgers University, Piscataway*)

We congratulate the authors for their correct call for attention to the utility of screening and great effort in studying its effectiveness.

Computational issues

In addition to the minimax concave penalty, we introduced a PLUS algorithm (Zhang, 2007a, 2008) to compute an entire path of local minimizers of concave penalized loss. Moreover, we proved the selection consistency $P\{\mathrm{sgn}(\hat{\beta}_j) = \mathrm{sgn}(\beta_j)\, \forall\, j\} \to 1$ for exactly the PLUS solutions MC+ and SCAD+ at the universal penalty level $\lambda_* = \sigma\sqrt{\{2\log(p)/n\}}$ (Donoho and Johnstone, 1994). Since iterative approximations are not guaranteed to converge to the 'right' local minimizer, selection consistency holds for the PLUS algorithm after consistent screening.

The computational cost of the PLUS algorithm is the same as that of the least angle regression algorithm LARS (Efron *et al.*, 2004) per step. Heuristics provide the computational cost of the PLUS algorithm as $O(snp)$ up to λ_*, higher than $O(np)$ for correlation screening, the same as iterative sure independence screening with $O(s)$ iterations, but lower than $O(p^3 + np^2)$ for the iteratively thresholded ridge regression screener. As Qiwei Yao discusses, we prove selection consistency of stepwise regression, intuitively also costing $O(snp)$ to compute.

Consistency

The consistency (4) of correlation screening hinges on condition 3, although the rankings of $|E(X_j Y)|$ are non-trivial. In Huang *et al.* (2006), ω_j provide consistent weights for the adaptive lasso under a partial orthogonality condition. We report simulation results without screening, at λ_*, with estimated σ at $p = 20\,000$, in the settings of simulation I (Table 13). An additional example is included with reduced signal. Compared with Tables 1 and 7, our simulations underscore the scalability and consistency of the PLUS algorithm and the importance of picking a proper method after consistent screening.

Information limits

Consider $\mathbf{X}$ with independent and identically distributed $N(0, \Sigma)$ rows, where the eigenvalues of Σ are all in $[c_*, c^*] \subseteq (0, \infty)$. Selection consistency requires $\min_{\beta_j \neq 0} |\beta_j| \geqslant M\sigma\sqrt{\{\log(p)/n\}}$ according to Wainwright

Table 13. Simulation results for the lasso MC+ and SCAD+ methods

Variable	*Results for (n, p, s) = (200, 1000, 8) and the following minimum signals a:*						*Results for (800, 20000, 18) and minimum signal a = 5 log(n)/√(n)*	
	$4\log(n)/\sqrt{n}$			$\log(n)/\sqrt{n}$			MC+	MC+($\hat{\sigma}$)
	Lasso	*MC+*	*SCAD+*	*Lasso*	*MC+*	*SCAD+*		
median($\lvert\hat{\mathcal{M}}\rvert$)	10	8	8	10	8	8	18	18
median($\lVert\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\rVert^2$)	1.52	0.18	0.18	1.34	0.34	0.76	0.09	0.09
%($\hat{\mathcal{M}} = \mathcal{M}_*$)	0.08	0.86	0.87	0.07	0.54	0.30	0.91	0.91
mean(steps)	11	17	25	11	13	17	37	37

(2007) for $\Sigma = I_p$ and Zhang (2007b) for general Σ. This information bound is achieved by the lasso for $\Sigma = I_p$ (Wainwright, 2007) and by the PLUS algorithm for general Σ (Zhang, 2007a, b). The same should hold post correlation screening under the additional condition 3. Since condition (16) holds for $\tilde{p} > 2n$, the answer seems to lie in equation (18) and $\delta = (n/p)^{1/(k-r)}$ in expression (41), after dealing with the dependence between the rows of $\mathbf{X}$ at the beginning of step 2(b).

Hao Helen Zhang (*North Carolina State University, Raleigh*)

We congratulate the authors for their thought-provoking and fascinating work on a fundamental yet challenging topic in variable selection. Driven by the pressing need of high dimensional data analysis in many fields, the problem of dimension reduction without losing relevant information becomes increasingly important. Fan and Lv successfully tackled the extremely challenging case, where $\log(p) = O(n^{\xi}), \xi > 0$. The proposed sure independence screening (SIS) is a state of the art method for high dimensional variable screening: simple, powerful and having optimal properties. This work is a substantial contribution to the area of variable selection and will also have a significant effect in other scientific fields.

Extension to non-parametric models

In linear models, marginal correlation coefficients between linear predictors and the response are effective measures to capture the strength of their linear relationship. However, correlation coefficients generally do not work for ranking non-linear effects. Consider the additive model

$$Y_i = \sum_{j=1}^{p} f_i(X_{ij}) + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where f_j takes an arbitrary non-linear function form. Motivated by the ranking idea of SIS, one could first fit a univariate smoother for each predictor and then use some marginal statistics to rank the covariates. Many interesting questions arise in this approach. Firstly, what are good measures to characterize the strength of the non-linear relationship fully? Possible choices include non-parametric test statistics, p-values and goodness-of-fit statistics like R^2. But which is best? Also, how do we develop consistent selection theory for the procedure of screening non-linear effects? All these questions are challenging because of the complicated estimation that is involved in non-parametric modelling. It would be interesting to explore whether and how the SIS can be extended to this context.

Connection to multiple hypotheses testing and false discovery rate control

The variable selection problem can be regarded as the problem of testing multiple hypotheses: $H_1 : \beta_1 = 0, \ldots, H_p : \beta_p = 0$. Screening important variables is hence equivalent to identifying the hypotheses to be rejected. The false discovery rate (Benjamini and Hochberg, 1995) has been developed to control the proportion of false rejections. Some consistent procedures based on individual tests of each parameter have been developed (Potscher, 1983; Bauer *et al.*, 1988). Recently, Bunea *et al.* (2006) considered the case when p increases with n, and showed that the false discovery rate or Bernoulli adjustment can lead to consistent selection of variables under certain conditions. Their method is based on the ordered p-values of individual t-statistics for testing $H_j : \beta_j = 0$, $j = 1, \ldots, p$. It would be interesting to compare SIS with these adjusted multiple hypotheses testing approaches.

Harrison H. Zhou (*Yale University, New Haven*) **and Xihong Lin** (*Harvard University, Cambridge*)

An important finding of this paper is that the method proposed can identify the true model with a high probability in ultrahigh dimensional variable selection settings such as $p = \exp(n^{\xi})$, with $\xi > 0$ arbitrarily large. To understand when this result can be applied in practice, we consider the following special linear model. Denote by $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_p)$ an $n \times p$ design matrix. Let $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with

(a) each of the p $\mathbf{X}$s being independent $N(0, 1)$,
(b) $p = \exp(n^{\xi})$ with $\xi > 0$,
(c) $\boldsymbol{\varepsilon} \sim N(0, \mathbf{I}_{n \times n})$ and
(d) $\beta_1 = n^{-\kappa}$ for some $\kappa < \frac{1}{2}$, and $\beta_j = 0$ for all $j \geqslant 2$.

Let $\lambda = \infty$, and write

$$\hat{\beta}_1 = \frac{1}{n}\mathbf{X}_1^{\mathrm{T}}\mathbf{y} = \frac{1}{n}\|\mathbf{X}_1\|^2\beta_1 + \frac{1}{n}\mathbf{X}_1^{\mathrm{T}}\boldsymbol{\varepsilon},$$

$$\hat{\beta}_j = \frac{1}{n}\mathbf{X}_j^{\mathrm{T}}\mathbf{y} = \frac{1}{n}\mathbf{X}_j^{\mathrm{T}}(\mathbf{X}_1\beta_1 + \boldsymbol{\varepsilon}) \qquad \text{for } j \geqslant 2.$$

Note that $\hat{\beta}_1 = \{1 + o(1)\}\beta_1 = \{1 + o(1)\}n^{-\kappa}$. Given $\mathbf{X}_1$ and ε, $\hat{\beta}_j$ are independent and identically distributed normal. Hence the maximum noise

$$\max_{2 \leqslant j \leqslant p} (\hat{\beta}_j) = \{1 + o(1)\}\sqrt{\{2\log(p)\}}\sqrt{\left(\frac{1 + \beta_1^2}{n}\right)} = \{1 + o(1)\}\sqrt{\{2(1 + \beta_1^2)\}}n^{(\xi - 1)/2}.$$

This calculation suggests that the true model can be identified with a high probability when $\kappa < (1 - \xi)/2$, i.e. $\xi < 1 - 2\kappa$. However, it is difficult to identify the true model when $\xi > 1 - 2\kappa$, as the maximum noise dominates the true signal. Can the authors' method be applied in this example to identify the true model when $\xi > 1 - 2\kappa$?

This example is related to the scenario in which some predictors are highly correlated. For example, when p is large, it is expected that there is a predictor X_j with $j \geqslant 2$ such that the sample correlation coefficient between X_j and true predictor X_1 is arbitrarily close to 1. The authors proposed a useful iterative sure independence screening procedure to deal with such a correlated X case. Can the authors provide some guidelines on how to choose k_i in each step and when to stop the iteration to ensure that the true model can be identified with a high probability, or β can be estimated with some nice risk property? It seems that the procedure proposed assumes that, if a variable is selected in the previous steps, it cannot be deleted in later steps. Intuitively, it would be desirable to let variables in and out at each step. Can the authors' procedure be modified to allow for this? We realize that the problem might become more complicated for a general covariance matrix of $\mathbf{x}$. For example, when the covariance is non-stationary, such as a constant exchangeable correlation among the $\mathbf{X}$s, the concentration property may not hold. Is the method still applicable in this case, and what are the required assumptions about signals relative to noise for the true model to be identified?

We would like to make one minor comment. We think, under *condition 3*, the term $\log(d)$ for the risk of method SIS–DS in theorem 4 may not be necessary given the other assumed conditions. Hence the result might be more general.

Hui Zou (*University of Minnesota, Minneapolis*)

I congratulate Professor Fan and Professor Lv for an excellent and stimulating paper which discusses several fundamental issues in high dimensional data analysis. My comments will focus on the oracle properties after sure independence screening.

The size of reduced dimension

Fan and Peng (2004) extended the oracle properties of non-concave penalized likelihood estimators in a finite dimension setting (Fan and Li, 2001) to the diverging dimension setting with $p = o(n^{1/3})$. Combined with sure independence screening (SIS), this result allows us to reduce the dimension from $p \gg n$ to $d = o(n^{1/3})$ and we still have an oracle-like estimator. In real applications, we wish to use SIS to screen out noise features and also want to do so conservatively. Thus, it is of interest to know whether theorem 5 holds for larger d, i.e. $d = o(n^{\nu})$ and $\nu > \frac{1}{3}$. Some positive answers are reported in Zou and Zhang (2008) which show that, under reasonably weak conditions, the adaptive elastic net estimator enjoys the oracle properties for $p = O(n^{\nu})$ as long as $0 \leqslant \nu < 1$. Hence, we believe that SIS can be used in a more conservative way without sacrificing any theoretical optimality.

Which oracle estimator should be mimicked?

In Fan and Li (2001) and Fan and Peng (2004) the likelihood estimators are considered; thus the oracle estimator should be the maximum likelihood estimator (MLE) and the non-concavely penalized likelihood estimator mimics the MLE oracle estimator. However, in linear regression models, we often do not wish to impose the error distribution assumption unless there is strong evidence for doing so. Thus the MLE estimator cannot be used to construct the oracle estimator. A popular oracle estimator is the least squares estimator, and the non-concavely penalized least squares estimator mimics the least squares oracle estimator. When the error distribution is non-normal, then the least squares oracle can be inefficient. Zou and Yuan (2008) discuss the issues with the oracle in the oracle model selection theory and propose the composite quantile regression (CQR) oracle estimator. Zou and Yuan (2008) show that the relative efficiency of the CQR oracle compared with the least squares oracle is greater than 70% regardless of the error distribution. Kai *et al.* (2008) further show that the efficiency lower bound can be as high as 86.4%. In the Gaussian model the relative efficiency is 95.5%. In a wide class of non-normal error models, the CQR oracle could be much more efficient and sometimes arbitrarily more efficient than the least squares oracle. Therefore, the CQR oracle is a safe and efficient alternative to the least square oracle.

The **authors** replied later, in writing, as follows.

We are very grateful to all the contributors for their stimulating comments and questions on the role of variable screening and selection on high dimensional statistical modelling. This paper would not have been in the current form without the benefit of private communications with Professor Peter Bickle, Professor Peter Bühlmann, Professor Eitan Greenshtein, Professor Qiwei Yao, Professor Cun-Hui Zhang and Dr Wenyang Zhang at various stages of this research. We shall not be able to resolve all the points in a brief rejoinder—indeed, the discussion can be seen as a collective research agenda for the future and some of the agendas have already been undertaken by the discussants.

Independent learning

We would like to point out that correlation learning is a specific case of independent learning that we advocate, which ranks the features according to the marginal utility of each feature. Correlation ranking is the same as feature ranking according to the reduction of the residual sum of squares in the least squares setting. In general, the marginal utility can be the quasi-likelihood or classification margin, contributed by each individual feature. This has.been made more explicit in Fan (2007) and Fan *et al.* (2008). We do not claim that independent learning can solve all high dimensional problems, but we indicate its power for some class of problems with ultrahigh dimensionality. The computational expediency and stability are prominently featured in independent learning.

We are very pleased to see that independent learning can indeed be derived from an empirical likelihood viewpoint as elucidated by Hall, Titterington and Xue. An added feature of the empirical likelihood approach is that the classifier is automatically built on the basis of the selected features. The idea of independent learning is also applicable to generalized additive models as discussed by Helen Zhang. The critical aspect is that the degrees of freedom for each component should be comparable or adjusted as elaborated in the generalized likelihood ratio tests by Fan *et al.* (2001) and Fan and Jiang (2007). This also partially responds to the question that was raised by Keming Yu on non-linear regression and categorical covariates. Although our theory does not cover the case with categorical variables, our method does. The discussion by Runze Li suggested further that correlation learning is applicable to the non-linear single-index model, which is more general than a single-layer neural network model. This also answers partially the question raised by Keming Yu.

Relationship to multiple testing and selection consistency

Several discussants (Bickel, Bühlmann, Marron, Luo, Baxter and Taylor, and Helen Zhang) link independent learning with multiple testing. Bickel raises several important theoretical questions from different perspectives and Bühlmann provides nice receiver operating characteristic curves, both for further understanding of sure independence screening (SIS). As Bickel correctly points out, our procedure is similar to multiple-testing problems to see whether each feature is correlated with the response variable. Translating the test statistics into P-values puts them into the same scale, as Adragni and Cook, Richardson and other discussants correctly point out. Helen Zhang mentions some existing work that answers the selection consistency question that was raised by Bickel. However, sure screening and multiple testing have a different philosophy and evaluation criterion. Multiple testing aims at controlling the false discovery rate (FDR) whereas screening focuses on missed discoveries. In the simulated example II in Section 4.2.2, for example, failing to discover the variable X_4, which is uncorrelated with Y having marginal regression coefficient 0, is regarded as a serious mistake, whereas in the multiple-testing problem this would even be regarded as a correct decision. Hence, the evaluation criterion is also different from the multiple-testing problem.

Several contributors (Bühlmann, Qiwei Yao, Cun-Hui Zhang, Leng and Wang) address the issue of selection consistency. This corresponds to no false and missed discoveries in variable selection, if the evaluation criterion for model selection is used. Although this is a very nice property, the selection consistency in ultrahigh dimensional space is a stringent requirement. The selection consistency is usually achieved by more complicated procedures than independent learning. For example, Bühlmann explores the idea of partial faithfulness, Qiwei Yao suggests stepwise regression procedures using modified information criteria, Leng and Wang discuss the penalized likelihood methods (Fan and Li, 2001), and so does Cun-Hui Zhang. However, in high dimensional statistical endeavours, a procedure with low FDR and no missed variables is already remarkable, if the procedure is computationally expedient and stable. Such a procedure can indeed be constructed by using SIS below.

As discussed above, SIS is not designed to control the FDR. But, it can easily be used to reduce the FDR with no missed variables. The idea is very simple. Split the data randomly into two halves and apply

SIS separately to both subsets of the data to obtain two submodels. Since we assume that the method has a sure screening property, both submodels contain all relevant variables in the model. Therefore, we take the common variables in these two submodels as the selected model. This selected model should have a low FDR, as a falsely discovered variable must appear independently twice in the selected model. The probability of such an event is merely $(n/p)^2$ under the mild condition on exchangeability, thanks to the 'blessing of dimensionality'. See Fan *et al.* (2008) for details and extensions. In particular, they showed that the probability of choosing r extra variables is bounded by $(n^2/p)^r/r!$.

Tuning parameters

Several contributors (Bühlmann, James and Radchenko, Leng and Wang, Runze Li, Wenyang Zhang, Zhou and Lin) discuss the need for a data-driven method for choosing the tuning parameters in both stage 1 and stage 2. We agree wholeheartedly. In the first stage, our preference is to select sufficiently many features, such as $d = n$ or $d = n/\log(n)$, though one can easily use twofold cross-validation to choose the number of features d. This answers partially the question that was raised by Zhang and Xia. In the second or final stage, Leng and Wang suggest using a Bayes information type of criterion and provide related references. Bühlmann comments correctly that a tuning scheme is needed for the second stage of iterative SIS (ISIS). Since ISIS means to be a simple screening procedure, a simple selection scheme suffices in many situations. The predetermined parameters $k_1, \ldots, k_l$ in the second stage should be a decreasing sequence and the geometric sequence is more appropriate. Suppose, for example, that we wish to run ISIS for five iterations ($l = 5$) and to decrease the number of selected variables at each stage by a factor of θ (0.75, say). Then, the first iteration should choose $k_1 = n(1-\theta)/(1-\theta^5)$, the second iteration picks θk_1 variables and the third stage selects $\theta^2 k_1$ variables, and so on. This avoids the ambiguity of variable selection of ISIS.

Clarifications

Some of our concepts were poorly presented and cause confusion as seen in some of discussions. First of all, the paper stressed the simplicity and utility of independent learning rules in high dimensional feature screening. Although correlation learning is an important specific example, we stress in fact independent screening. This is why we chose the title with sure independent learning. Secondly, we would like to clarify that condition 4 in the paper indicates a constraint on population collinearity, whereas usual conditions on the design talk about sample collinearity. The difference between these two types of collinearity could be severe when the dimensionality is much larger than the sample size, as illustrated by Figs 1 and 4 in the paper. Condition 4 accommodates the situation in which the features can be divided into several uncorrelated groups, each satisfying condition 4. Thirdly, although $d = n - 1$ is our default in the screening stage, we do not rule out the possibility of selecting more features in the first stage. This partially answers the concerns by Morris, and Richardson and Bottolo that $d = n/\log(n)$ or $d = n - 1$ can be too small in the first stage for some applications. In other words, we do not disagree with the comments that were made by Greenshtein and Marron, who have in mind to construct as effective a method as possible to predict future observations (the first goal in Bickel's comment). However, for the second goal in Bickel's comment, to gain insight into the relationship between features and response for scientific purposes, as well as, hopefully, to construct an improved prediction method, selection of smaller numbers of features without too much compromise of prediction errors is also an important object and hence the default that $d = n - 1$ makes sense. Lastly, our goal is feature screening. The number of selected features is an order of magnitude larger than the number of active features. This makes sure screening much more easy and feasible.

Questions

Various contributors raise excellent questions that can be seen as a research agenda. For brevity, we cannot respond to most of them.

Bickel's questions 2 and 3 touch the foundation of feature selection. We are pleased that he extends the concept of sure screening to the situation when more than one parsimonious models fit approximately equally well. In such a situation, the independence screening would be more likely to obtain all important factors than more sophisticated variable selection approaches, as the former is more likely to retain highly correlated variables that would fit approximately equally well in the final model. In addition, among those parsimonious models, the variables with a large marginal utility are usually preferred. Apparently much work is needed to compare screen first–fit after types of method with fit first–screen after types of method in terms of consistency and oracle properties, but the former would be faster and can deal with higher dimensionality.

Bühlmann questions the advantages of ISIS in comparison with the boosting algorithm. With the predetermined variable selection schemes at stage 2 that were shown above, ISIS chooses multiple features in

the second stage using the joint information of important covariates. It is less greedy than the boosting algorithm. At the same time, it avoids solving large optimization problems as smoothly clipped absolute deviation or the lasso would have done without prescreening. In other words, it bridges the gap between these two extreme methods. It would be interesting to study and compare consistency properties of ISIS and the boosting approach, as commented by Bühlmann. The concentration property (16) always holds for the Gaussian case, as shown in our theoretical study. Bühlmann, and Zhou and Lin both raise an excellent question about the upper bound of dimensionality for our theoretical results. Inspecting our technical proofs more carefully yielded the need for such an upper bound, which is stated in condition 1.

Leng and Wang, and Zhang and Xia raise the question of how the stochastic error in the screening stage impacts on the second stage of estimation. For many applications, this would not be severe. As commented in the paper, to avoid the selection bias in the screening stage, we can split the sample into two portions, where the first portion is used to screen variables and the second portion is used for shrinkage estimation.

Comments

Johnstone provides beautiful theoretical results on the distribution of the largest sample canonical correlation which gives us a better idea of how the problem of collinearity becomes severe when we have only a modest sample size compared with the dimensionality. We agree with Samworth that the concentration property should hold for a broader class of spherically symmetric distributions, as he shows via careful simulations, and that it is important to derive the theoretical properties of independence screening for heavier-tailed distributions that may have no concentration property.

We appreciate the remark by Hall, Titterington and Xue that the correlation coefficient for binary data becomes a t-statistic for any sample size provided that class 1 is scored by n_1^{-1} and class 2 is scored with $-n_2^{-1}$. The idea is related to assigning an empirical prior to the class labels in a reverse order. This also provides some insights to the question that was raised by Yufeng Liu, who would like to know the relative merits between the correlation ranking and t-statistic ranking. Yufeng Liu discusses several scenarios in classification problems in which SIS deserves further development. One possible method of feature ranking in multiclass problems is to rank them according to the F-statistics or their variants. An alternative is to regard it as a sequence of two-class problems and to use ISIS to select all relevant features.

We appreciate the connections between SIS and screening by principal fitted components by using the inverse regression made by Adragni and Cook. In their simulation, the response is uncorrelated with all predictors, and this explains why SIS performs poorly. It is unclear to us how the piecewise linear basis $\mathbf{f}_y$ was constructed, but screening by principal fitted components in this simulated example is the same as correlation learning based on a non-linear transform $\mathbf{f}_y$ of the response, which is now correlated with the relevant predictor. This gives advantages over plain SIS, which does not use any transform, for this simulated model.

Greenshtein remarks that the non-parametric empirical Bayes method performs very well for classifications and works well also for sparse situations. As the method does not explicitly explore the sparsity, we would not expect it to adapt to the sparse setting as well as the methods that are tailored for this setting. The non-parametric empirical Bayes method is useful for constructing an effective method for prediction class labels, without selecting features. This would not be suitable for achieving aim (b) of Bickel's discussion.

Levina and Zhu look at the performance of ISIS with the lasso plug-in under lower signal-to-noise levels through a simulation study. We agree that incorporating the signal-to-noise ratio the convergence rate will give us a better picture of its effect. Of importance is to develop extensions of ISIS that are robust to low signal-to-noise ratio. Some related questions have also been addressed by Luo, Baxter and Taylor.

Several discussants, including Anagnostopoulos and Tasoulis, bring up questions of relaxing the technical assumptions in the paper to give more insights into the applicability of SIS. We believe that these questions will certainly stimulate much new research on variable screening and selection. For the leukaemia data analysis, we cannot compare the overlap of the genes selected since we do not have the keys to check that.

Robustness

Several contributors, including Gather and Guddat, Hui Zou, and Luo, Baxter and Taylor, bring up the issue of robustness to outliers and model assumptions. We appreciate their efforts to make the procedure more robust to those assumptions. In particular, Gather and Guddat, and Hui Zou both propose more robust procedures to the outliers. We agree with all discussants that robustness to outliers and to model assumptions are important issues and they have addressed some of those. Independent learning is still in the infancy and certainly needs more researchers to nurture and understand it.

Criticisms

Many discussants give very critical scrutiny of independent learning in high dimensional modelling. SIS and ISIS are simple procedures and cannot expect to address all the needs.

Robert casts doubts on the assumption of the existence of a single true model when $p \gg n$, as Bickel does. We acknowledge that there are many models that are statistically indifferentiable given the limited amount of information, but some are more useful. SIS and ISIS are procedures to pick some submodels that have large marginal contributions. The asymptotic results provide merely an ideal situation under which our common sense of independent screening works. However, Bayesian methods are viable tools for selecting a family of submodels that have similar performance.

Longford comes up with a pyramid view of the importance of the variables which leads to a weaker assumption than sparsity in the narrow sense. In a sense, the classical best subset selection provides such a pyramid view on the most important k-variable models. However, such a best set selection is an NP-hard problem and classical stepwise addition or stepwise algorithms provide a useful proxy. In high dimensional endeavours, however, the accumulation of noise and computational cost make these methods more challenging to use and to understand. The penalized least squares methods provide an alternative solution to these traditional methods with more efficient computation and easier structure to understand its statistical properties. The 'SCAD+' and 'MCP+' (Zhang, 2007) or lasso solution paths provide, in a sense, such a pyramid view. SIS and ISIS purely assist in reducing the dimensionality so that a more efficient solution path can be constructed.

Richardson and Bottolo speculate that sure screening can be elusive in some correlated cases. The poor performance of SIS in their simulation is related to the selection of the tuning parameter and leakage effect from the peaks. If a larger d such as $d = n - 1$ is used in the first stage of screening and if the leakage issue is addressed, then the results of SIS can be significantly improved, by looking at their figures. In other words, the sure screening property still holds in their simulated example. In searching for quantitative trait loci or other similar biological endeavours, the leakage issue should be addressed. The peak locations are often of interest and large values around the peak are regarded as the leakage from the peaks and correspond to the same genetic locus.

Computation

Many contributors touch on the issue of computation, including Bühlmann, Runze Li and Cun-Hui Zhang, who address different computation algorithms and their computational complexity. We agree with those discussants that SIS has the smallest computational cost at the screening stage. The PC algorithm for exploring partial faithfulness is certainly very stimulating and useful. The PLUS algorithm that was proposed by Zhang (2007) is creative for effectively finding the solution paths to the folded concave penalized least squares problems and is backed by asymptotic theory. Alternative algorithms are iteratively reweighted penalized L_1-regression proposed by Zou and Li (2008) and elaborated further in the paper, the iterative co-ordinatewise minimization that was discussed by Runze Li and the local quadratic approximation (Fan and Li, 2001). With these, we agree with Cun-Hui Zhang and Runze Li that the implementation of folded concave penalized least squares problems are not much harder or slower to compute than the lasso. However, the gains in bias reduction can be substantial in the high dimensional setting. We believe that, with better understanding and implementation, the folded concave penalized likelihood (Fan and Li, 2001) will play even more important roles in high dimensional statistical modelling and feature selection.

Conclusion

Taken together, the discussants cover a wide range of topics, from foundations, philosophy and theory to methods, computation and applications. The wide interest in high dimensional learning and related methods in many fields, from bioinformatics and genetics to climatology and finance, clearly presents exciting opportunities for interdisciplinary collaborations and expanded exchanges in ideas and tools between statistics and other disciplines. We are very pleased to conclude by reiterating our thanks to all the contributors, and to the Royal Statistical Society and the journal for hosting this forum.

References in the discussion

Ahn, J., Marron, J. S., Muller, K. M. and Chi, Y.-Y. (2007) The high-dimension, low-sample-size geometric representation holds under mild conditions. *Biometrika*, **94**, 760–766.

An, H.-Z., Huang, D., Yao, Q. and Zhang, C.-H. (2008) Adjusted information criteria based stepwise searching for feature variables in high-dimensional linear regression. To be published.

Anagnostopoulos, C., Tasoulis, D., Hand, D. J. and Adams, N. (2008) Optimisation for variable selection in data streams. In *Proc. 18th Eur. Conf. Artificial Intelligence*. To be published.
Bauer, P., Potscher, B. M. and Hackl, P. (1988) Model selection by multiple test procedures. *Statistics*, **19**, 39–44.
Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple hypotheses testing. *J. R. Statist. Soc.* B, **57**, 289–300.
Bickel, P. (2007) Discussion of Candes and Tao. *Ann. Statist.*, **35**, 2313–2404.
Bickel, P., Ritov, Y. and Tsybakov, A. (2008) Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.*, **36**, in the press.
Bottolo, L. and Richardson, S. (2008) Evolutionary stochastic search. *Technical Report*. Centre for Biostatistics, Imperial College London, London. (Available from `www.bgx.org.uk`.)
Breiman, L. (1996) Heuristics of instability and stabilization in model selection. *Ann. Statist.*, **24**, 2350–2383.
Bühlmann, P. (2006) Boosting for high-dimensional linear models. *Ann. Statist.*, **34**, 559–583.
Bühlmann, P. and Kalisch, M. (2008) Variable selection for high-dimensional models: partial faithful distributions, strong associations and the PC-algorithm. *Technical Report*. Eidgenössiche Technische Hochschule, Zürich.
Bühlmann, P. and Meier, L. (2008) Discussion on 'One-step sparse estimates in nonconcave penalized likelihood models' (by H. Zou and R. Li). *Ann. Statist.*, **36**, 1534–1541.
Bühlmann, P. and Yu, B. (2003) Boosting with the L_2 loss: regression and classification. *J. Am. Statist. Ass.*, **98**, 324–339.
Bunea, F., Wegkamp, M. and Auguste, A. (2006) Consistent variable selection in high dimensional regression via multiple testing. *J. Statist. Planng Inf.*, **136**, 4349–4364.
Candes, E. and Tao, T. (2007) The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, **35**, 2313–2351.
Cook, R. D. (2007) Fisher Lecture: Dimension reduction in regression. *Statist. Sci.*, **22**, 1–26.
Donoho, D. L. and Johnstone, I. (1994) Minimax risk over l_p-balls for l_q-error. *Probab. Theory Reltd Flds*, **99**, 277–303.
Dudoit, S., Fridlyand, J. and Speed, T. P. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Statist. Ass.*, **97**, 77–87.
Dziak, J. (2004) Penalized likelihood and quasi-likelihood for variable selection in linear models. *Master's Thesis*. Department of Statistics, Pennsylvania State University, University Park.
Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004) Least angle regression (with discussion). *Ann. Statist.*, **32**, 407–499.
Fan, J. (2007) Variable screening in high-dimensional feature space. In *Proc. 4th Int. Congr. Chinese Mathematicians*, vol. II (eds L. Ji, K. Liu, L. Yang and S. T. Yau), pp. 735–747. Beijing: High Education Press.
Fan, J. and Fan, Y. (2008) High dimensional classification using features annealed independence rules. *Ann. Statist.*, to be published.
Fan, J. and Jiang, J. (2007) Nonparametric inference with generalized likelihood ratio tests (with discussion). *Test*, **16**, 409–478.
Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Ass.*, **96**, 1348–1360.
Fan, J. and Peng, H. (2004) On non-concave penalized likelihood with diverging number of parameters. *Ann. Statist.*, **32**, 928–961.
Fan, J., Samworth, R. and Wu, Y. (2008) Ultra-dimensional variable selection via independent learning: beyond the linear model. To be published.
Fan, J., Zhang, C. and Zhang, J. (2001) Generalized likelihood ratio statistics and Wilks phenomenon. *Ann. Statist.*, **29**, 153–193.
Fang, K. T., Kotz, S. and Ng, K. W. (1990) *Symmetric Multivariate and Related Distributions*. London: Chapman and Hall.
Friedman, J. (2001) Greedy function approximation: a gradient boosting machine. *Ann. Statist.*, **29**, 1189–1232.
Friedman, J., Hastie, T., Höfling, H. and Tibshirani, R. (2007) Pathwise coordinate optimization. *Ann. Appl. Statist.*, **1**, 302–332.
Gather, U., Hilker, T. and Becker, C. (2001) A robustified version of sliced inverse regression. In *Statistics in Genetics and in the Environmental Sciences: Proc. Wrkshp Statistical Methodology for the Sciences: Environmetrics and Genetics, Ascona, May 23rd–28th* (eds L. T. Fernholz, S. Morgenthaler and W. Stahel), pp. 147–157. Basel: Birkhäuser.
Gather, U., Hilker, T. and Becker, C. (2002) A note on outlier sensitivity of sliced inverse regression. *Statistics*, **13**, 271–281.
van de Geer, S. (2008) High-dimensional generalized linear models and the Lasso. *Ann. Statist.*, **36**, 614–645.
Gnanadesikan, R. and Kettenring, J. (1972) Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, **28**, 81–124.
Golub, T. R., Slomin, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M. L., 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**, 531–537.

Greenshtein, E. and Park, J. (2008) Application on nonparametric empirical Bayes estimation to high dimensional classification. To be published.
Guyon, I. and Elisseeff, A. (2003) An introduction to variable and feature selection. *J. Mach. Learn. Res.*, **3**, 1157–1182.
Hall, P., Marron, J. S. and Neeman, A. (2005) Geometric representation of high dimension, low sample size data. *J. R. Statist. Soc.* B, **67**, 427–444.
Hall, P. and Miller, H. (2008) Using generalised correlation to effect variable selection in very high dimensional problems. To be published.
Hall, P., Titterington, D. M. and Xue, J.-H. (2008) Tilting methods for assessing the influence of components in a classifier. To be published.
Hans, C., Dobra, A. and West, M. (2007a) Shotgun stochastic search in regression with many predictors. *J. Am. Statist. Ass.*, **102**, 507–516.
Hans, C., Wang, Q., Dobra, A. and West, M. (2007b) SSS: high-dimensional Bayesian regression model search. *Bull. Int. Soc. Bayes. Anal.*, **24**, 8–9.
Hinton, G. E. and Salakhutdinov, R. R. (2006) Reducing the dimensionality of data with neural networks. *Science*, **313**, 504–507.
Huang, J., Ma, S. and Zhang, C.-H. (2006) Adaptive LASSO for sparse high-dimensional regression models. *Statist. Sin.*, to be published.
Johnstone, I. M. (2008a) Multivariate analysis and Jacobi ensembles: largest eigenvalue, Tracy-Widom limits and rates of convergence. *Ann. Statist.*, to be published.
Johnstone, I. M. (2008b) Approximate null distribution of the largest characteristic root in multivariate analysis. *Technical Report*. Department of Statistics, Stanford University, Stanford.
Kai, B., Li, R. and Zou, H. (2008) Local polynomial composite quantile regression. *Technical Report 08-86*. Methodology Center, Pennsylvania State University, University Park.
Kalisch, M. and Bühlmann, P. (2007) Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *J. Mach. Learn. Res.*, **8**, 613–636.
Knight, K. and Fu, W. (2000) Asymptotics for lasso-type estimators. *Ann. Statist.*, **28**, 1356–1378.
Leng, C., Lin, Y. and Wahba, G. (2006) A note on lasso and related procedures in model selection. *Statist. Sin.*, **16**, 1273–1284.
Li, K.-C. (1991) Sliced inverse regression for dimension reduction (with discussion). *J. Am. Statist. Ass.*, **86**, 316–342.
Li, L. (2007) Sparse sufficient dimension reduction. *Biometrika*, **94**, 603–613.
Li, R. (2008) Discussion of "Sure independence screening for ultra-high dimensional feature Space" by J. Fan and J. Lv: a detailed version. (Available from `http://www.stat.psu.edu/~rli/discussion.pdf`.)
Li, L., Cook, R. D. and Tsai, C.-L. (2007) Partial inverse regression. *Biometrika*, **94**, 615–625.
Liang, F., Paulo, R., Molina, G., Clyde, M. A. and Berger, J. O. (2008) Mixtures of g priors for Bayesian variable selection. *J. Am. Statist. Ass.*, **103**, 410–423.
Meinshausen, N. and Bühlmann, P. (2006) High-dimensional graphs and variable selection with the Lasso. *Ann. Statist.*, **34**, 1436–1462.
Meinshausen, N. and Yu, B. (2008) Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.*, **36**, in the press.
Potscher, B. (1983) Order estimation in ARMA models by Lagrange multiplier tests. *Ann. Statist.*, **11**, 872–885.
Qiao, X. and Liu, Y. (2008) Adaptive weighted learning for unbalanced multicategory classification. *Biometrics*, to be published, doi:10.1111/j.1541-0420.2008.01017.x.
Segal, M. R., Dahlquist, K. D. and Conklin, B. R. (2003) Regression approach for microarray data analysis. *J. Computnl Biol.*, **10**, 961–980.
Shao, J. (1997) An asymptotic theory for linear model selection. *Statist. Sin.*, **7**, 221–264.
Spirtes, P., Glymour, C. and Scheines, R. (2000) *Causation, Prediction, and Search*, 2nd edn. Cambridge: MIT Press.
Tasoulis, D. K., Plagianakos, V. P. and Vrahatis, M. N. (2006) Unsupervised clustering in mRNA expression profiles. *Comput. Biol. Med.*, **36**, 1126–1142.
Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B, **58**, 267–288.
Vapnik, V. (1998) *Statistical Learning Theory*. Chichester: Wiley.
Wachter, K. W. (1980) The limiting empirical measure of multiple discriminant ratios. *Ann. Statist.*, **8**, 937–957.
Wainwright, M. J. (2007) Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *Technical Report*. Department of Statistics, University of California, Berkeley.
Wang, H. and Leng, C. (2007) Unified lasso estimation via least squares approximation. *J. Am. Statist. Ass.*, **101**, 1418–1429.
Wang, H., Li, B. and Leng, C. (2008) Shrinkage tuning parameter selection with a diverging number of parameters. *Technical Report*. (Available from `http://hansheng.gsm.pku.edu.cn/work.htm/DivergeBIC.pdf`.)
Wang, H., Li, G. and Tsai, C. L. (2007a) Regression coefficient and autoregressive order shrinkage and selection via the lasso. *J. R. Statist. Soc.* B, **69**, 63–78.

Wang, H., Li, R. and Tsai, C. L. (2007b) On the consistency of scad tuning parameter selector. *Biometrika*, **94**, 553–558.
Wang, L. and Shen, X. (2007) On L_1-norm multi-class support vector machines: methodology and theory. *J. Am. Statist. Ass.*, **102**, 583–594.
Westad, F. and Martens, H. (2000) Variable selection in NIR based on significance testing in partial least squares regression. *J. Near Infrared Spectrosc.*, **8**, 117–124.
Wu, Y. and Liu, Y. (2007) Robust truncated-hinge-loss support vector machines. *J. Am. Statist. Ass.*, **102**, 974–983.
Zhang, C.-H. (2007a) Penalized linear unbiased selection. *Technical Report 2007-003*. Department of Statistics, Rutgers University, Piscataway.
Zhang, C.-H. (2007b) Information-theoretic optimality of variable selection with concave penalty. *Technical Report 2007-008*. Department of Statistics, Rutgers University, Piscataway.
Zhang, C.-H. (2008) Discussion: One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.*, **36**, 1553–1560.
Zhang, Y. (2006) Variable selection via penalized likelihood and iterative conditional minimization algorithm. *Master's Thesis*. Department of Statistics, Pennsylvania State University, University Park.
Zhang, C.-H. and Huang, J. (2008) The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Ann. Statist.*, **36**, 1567–1594.
Zhang, Y., Li, R. and Tsai, C.-L. (2008) Regularization parameter selection for penalized likelihood function of GLIM and Cox models. *Technical Report 08-87*. Methodology Center, Pennsylvania State University, University Park.
Zhang, H. H. and Lu, W. (2007) Adaptive lasso for cox's proportional hazard model. *Biometrika*, **94**, 691–703.
Zou, H. (2006) The adaptive lasso and its oracle properties. *J. Am. Statist. Ass.*, **101**, 1418–1429.
Zou, H. and Li, R. (2008) One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.*, to be published.
Zou, H. and Yuan, M. (2008) Composite quantile regression and the oracle model selection theory. *Ann. Statist.*, **36**, 1108–1126.
Zou, H. and Zhang, H. H. (2008) On the adaptive elastic-net with a diverging number of parameters. *Ann. Statist*., to be published.