

Sure Independence Screening for Ultra-High Dimensional Feature Space *

Jianqing Fan

Department of Operations Research and Financial Engineering
Princeton University

Jinchi Lv

Information and Operations Management Department
Marshall School of Business
University of Southern California

August 8, 2007

Abstract

Variable selection plays an important role in high dimensional statistical modeling which nowadays appears in many areas of scientific discoveries. With large scale or dimensionality p , computational cost and estimation accuracy are two top concerns for any statistical procedure. In a recent paper, Candès and Tao (2007) propose a minimum ℓ_1 estimator, the Dantzig selector, and show that it mimics the ideal risk within a factor of $\log p$. Their innovative procedure and remarkable result are challenged when the dimensionality is ultra high. The factor $\log p$ can be large and their uniform uncertainty principle can fail.

Motivated by those concerns, we introduce the concept of sure screening and propose a sure screening method Sure Independence Screening (SIS) to reduce dimensionality from high to a relatively large scale d that is below sample size. In a fairly general asymptotic framework, SIS is shown to have the sure screening property for even exponentially growing dimension. As a methodological extension, an iterate SIS (ISIS) is also proposed to enhance the finite sample performance. With high dimension reduced accurately to below sample size, variable selection can be accomplished by some refined lower dimensional method such as the SCAD, Dantzig selector, Lasso, or adaptive Lasso.

Short title: Sure Independence Screening

AMS 2000 subject classifications: Primary 62J99; secondary 62F12

Keywords: Variable selection, dimensionality reduction, SIS, sure screening, oracle estimator, SCAD, Dantzig selector, Lasso, adaptive Lasso

*Financial support from the NSF grants DMS-0354223 and DMS-0704337 and NIH grant R01-GM072611 is gratefully acknowledged. We are grateful to the anonymous referees for their constructive and helpful comments. Address for correspondence: Jianqing Fan, Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544. Phone: (609) 258-7924. E-mail: jqfan@princeton.edu.

1 Introduction

1.1 Background

Consider the problem of estimating a p -vector of parameters β from the linear model

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

where $\mathbf{y} = (Y_1, \dots, Y_n)^T$ is an n -vector of responses, $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^T$ is an $n \times p$ random design matrix with i.i.d. $\mathbf{x}_1, \dots, \mathbf{x}_n$, $\beta = (\beta_1, \dots, \beta_p)^T$ is a p -vector of parameters, and $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T$ is an n -vector of i.i.d. random errors. When dimension p is high, it is often assumed that only a small number of predictors among X_1, \dots, X_p contribute to the response, which amounts to assuming ideally that the parameter vector β is sparse. With sparsity, variable selection can improve estimation accuracy by effectively 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 MRI, tomography, tumor 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 tumors using microarray gene expression or proteomics data; one may wish to associate protein concentrations with expression of genes or predict certain clinical prognosis (e.g., injury scores or survival time) 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, e.g., Donoho (2000) and Fan and Li (2006) for overviews of statistical challenges with high dimensionality.

Back to the problem in (1), it is challenging to find tens of important variables out of thousands of predictors, with 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 Candès and Tao (2007) is the notion of uniform uncertainty principle (UUP) on deterministic design matrices. They proposed the Dantzig selector, which is solution to an ℓ_1 -regularization problem, and showed that under UUP, 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: 1) it is easy to implement because the convex optimization the Dantzig selector solves can easily be recast as a linear program; and 2) 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 ultra-high dimensional problems. First, a potential hurdle is the computational cost for large or huge scale problems such as implementing linear programs in dimension tens of thousands. Second, the factor $\log p$ can become large and may not be negligible when dimension p grows fast with sample size n . Third, as dimensionality grows, their UUP condition may be hard to satisfy, which will be illustrated later using a simulated example. Finally, there is no guarantee 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, 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 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 one. 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) lay 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 AIC and BIC, as well as more recent methods like the bridge regression in Frank and Friedman (1993), Lasso in Tibshirani (1996), and SCAD in Fan (1997) and Antoniadis and Fan (2001), and established oracle properties for nonconcave penalized likelihood estimators. Later on, Fan and Peng (2004) extend the results to the setting of $p = o(n^{1/3})$ and show that the oracle properties continue to hold. An effective algorithm for optimizing penalized likelihood, local quadratic approximation (LQA), was proposed in Fan and Li (2001) and well studied in Hunter and Li (2005). Zou (2006) introduces an adaptive Lasso in a finite parameter setting and shows that Lasso does not have oracle properties as conjectured in Fan and Li (2001), whereas the adaptive Lasso does. Zou and Li (2007) propose a local linear approximation algorithm and one-step sparse estimates for nonconcave 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) study variable selection for Cox's proportional hazards model and frailty model; Efron, Hastie, Johnstone and Tibshirani (2004) propose LARS; Hunter and Li (2005) propose a new class of algorithms, MM algorithms, for variable selection; Meinshausen and Bühlmann (2006) look at the problem of variable selection with the Lasso for high dimensional graphs, and Zhao and Yu (2006) give an almost necessary and sufficient condition on model selection consistency of Lasso. More recent studies include Huang, Horowitz and Ma (2007), Paul *et al.* (2007), Zhang (2007), and Zhang and Huang (2007), which significantly advances the theory and methods of the penalized

least-squares approaches. It is worth to mention that in variable selection, there is a weaker concept than consistency, called persistency, introduced by Greenshtein and Ritov (2004). Motivation of this concept lies in the fact that in machine learning such as tumor classifications, the primary interest centers on the misclassification errors or more generally expected losses, not the accuracy of estimated parameters. Greenshtein and Ritov (2004) study the persistence of Lasso-type procedures in high dimensional linear predictor selection, and Greenshtein (2006) extends the results to more general loss functions. Meinshausen (2006) considers a case with finite nonsparsity and shows that under quadratic loss, Lasso is persistent, but the rate of persistency is slower than that of a relaxed Lasso.

1.3 Some insight on high dimensionality

To gain some insight on challenges of high dimensionality in variable selection, let us look at a situation where all the predictors X_1, \dots, X_p are standardized and the distribution of $\mathbf{z} = \Sigma^{-1/2}\mathbf{x}$ is spherically symmetric, where $\mathbf{x} = (X_1, \dots, X_p)^T$ and $\Sigma = \text{cov}(\mathbf{x})$. Clearly, the transformed predictor vector \mathbf{z} has covariance matrix I_p . Our way of study in this paper is to separate the impacts of the covariance matrix Σ and the distribution of \mathbf{z} , which gives us a better understanding on 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 $\mathbf{X}^T\mathbf{X}$ is huge and singular. The maximum spurious correlation between a covariate and the response can be large (see, e.g., Figure 1) because of the dimensionality and the fact that an unimportant predictor can be highly correlated with the response variable due to the presence of important predictors associated with the predictor. These make variable selection difficult. Second, the population covariance matrix Σ may become ill-conditioned as n grows, which adds difficulty to variable selection. Third, the minimum nonzero absolute coefficient $|\beta_i|$ may decay with n and get 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 β 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 observed in Hall, Marron and Neeman (2005). A challenge with high dimensionality is that important predictors can be highly correlated with some unimportant predictors, which usually increases with dimensionality. The maximum spurious correlation also grows with dimensionality. We illustrate this using a simple example. Suppose the predictors X_1, \dots, 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 among 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 5000, respectively. Figure 1 shows the distributions of the maximum absolute sample correlation. 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.

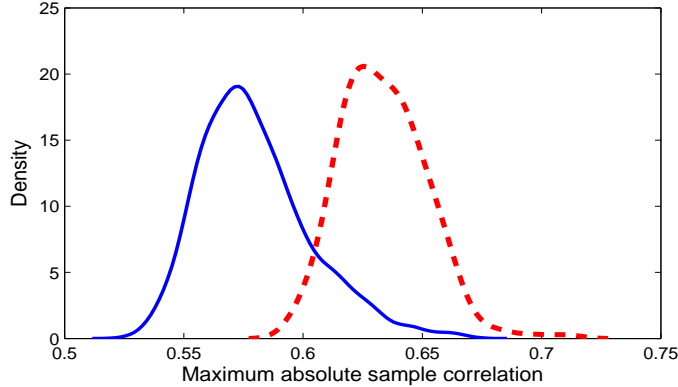


Figure 1: Distributions of the maximum absolute sample correlation coefficient when $n = 60, p = 1000$ (solid curve) and $n = 60, p = 5000$ (dashed curve).

The paper is organized as follows. In the next section we propose a sure screening method Sure Independence Screening (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, 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 the Appendix.

2 Sure Independence Screening

2.1 A sure screening method: SIS

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

$$\boldsymbol{\omega} = \mathbf{X}^T \mathbf{y}, \quad (2)$$

where the $n \times p$ data matrix \mathbf{X} is first standardized columnwise as mentioned before.

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

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

where $[\gamma n]$ denotes the integer part of γn . This is a straightforward way to shrink the full model $\{1, \dots, p\}$ down to a submodel \mathcal{M}_γ with size $d = [\gamma n] < n$. We call this method Sure Independence Screening (SIS). Its computational cost is that of multiplying a $p \times n$ matrix with an n -vector plus getting the largest d components of a p -vector, so SIS has computational complexity $O(np)$. We remark here that d is below sample size n .

It is worth to mention that SIS uses only the order of componentwise magnitudes of ω , so it is indeed invariant under scaling. Thus the idea of SIS is identical to selecting predictors using their correlations with the response. To implement SIS, 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 $d \geq n$, say, $\gamma \geq 1$. It is obvious that larger d means larger probability to include the true model \mathcal{M}_* in the final model \mathcal{M}_γ .

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.,

$$P(\mathcal{M}_* \subset \mathcal{M}_\gamma) \rightarrow 1 \quad \text{as } n \rightarrow \infty \quad (4)$$

for some given γ . This question as well as how the sequence $\gamma = \gamma_n \rightarrow 0$ should be chosen in theory will be answered by Theorem 1 in Section 5.

2.2 Rationale of SIS

To better understand the rationale of SIS, 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 the simple SIS is needed. ITRRS also provides a very nice technical tool for our understanding of the sure property of SIS and other methods.

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

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

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

$$\omega^\lambda \rightarrow \hat{\beta}_{\text{LS}} \quad \text{as } \lambda \rightarrow 0, \quad (6)$$

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

$$\lambda \omega^\lambda \rightarrow \omega \quad \text{as } \lambda \rightarrow \infty. \quad (7)$$

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

For any given $\delta \in (0, 1)$, we sort the p componentwise magnitudes of the vector ω^λ in a descending order and define a submodel

$$\mathcal{M}_{\delta,\lambda}^1 = \left\{ 1 \leq i \leq p : |\omega_i^\lambda| \text{ is among the first } [\delta p] \text{ largest of all} \right\}. \quad (8)$$

This procedure reduces the model size by a factor of $1 - \delta$. The idea of ITRRS that will 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 an order n^θ for some $\theta > 0$ lower than the original one p . This property stimulates us to propose ITRRS as follows:

- First, carry out the procedure in (8) to the full model $\{1, \dots, p\}$ and get a submodel $\mathcal{M}_{\delta,\lambda}^1$ with size $[\delta p]$;
- 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;
- Finally, get a submodel $\mathcal{M}_{\delta,\lambda}^k = \mathcal{M}_{\delta,\lambda}^k$ with size $d = [\delta^k p] < n$, where $[\delta^{k-1} p] \geq n$.

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

Now we are ready to see that the simple SIS introduced in Section 2.1 is a specific case of ITRRS since the componentwise regression is a specific case of the ridge regression with an infinite regularization parameter and ranking is invariant under scaling. 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 SIS. The question of whether 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. We would like to point out that the Simple Thresholding Algorithm (see, e.g., 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 related to the ITRRS. In particular, our asymptotic analysis in Section 5 sheds the lights of understanding the performance of the Simple Thresholding Algorithm.

The number of steps in ITRRS depends on the choice of $\delta \in (0, 1)$. We will see in Theorem 3 that δ can not be chosen too small which means that there should not be too many iteration steps in 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 ITRRS until the number of remaining variables drops to below 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 is often adopted. For example, Huang, Horowitz and Ma (2007) propose independently to use 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 tumor classification using gene expression or proteomics data, it might not be wise to classify the data using the full feature space due to the noise accumulation and interpretability. This is well demonstrated both theoretically and numerically in Fan and Fan (2007). In addition, many of the features come into play through linkage to the important ones (see, e.g., Figure 1). Therefore feature selection is important for high dimensional classification. How to effectively select important features 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 presence of high dimensionality. For example, Tibshirani *et al.* (2002) introduce the nearest shrunken centroids method, and Fan and Fan (2007) propose the Features Annealed Independence Rules (FAIR) procedure. Theoretical justification for the nearest shrunken centroids method is still missing.

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

$$\omega = \sum_{Y_i=1} \mathbf{x}_i - \sum_{Y_i=-1} \mathbf{x}_i, \quad (9)$$

Written more explicitly, the j -th component of the p -vector ω is

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

by recalling that each component in (9) has been normalized marginally, where $\bar{X}_{j,1}$ is the sample average of the j -th feature with class label “1” and $\bar{X}_{j,2}$ is the sample average of the j -th 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 the two-sample t -statistics. See Fan and Fan (2007) for a theoretical study.

Two-sample t -statistics are commonly used in feature selection for high dimensional classification problems such as in the significance analysis of gene selection in microarray data analysis (see, e.g., Storey and Tibshirani, 2003; 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.

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

3 SIS 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 SIS we can shrink the full model $\{1, \dots, p\}$ straightforward and accurately down to a submodel $\mathcal{M} = \mathcal{M}_\gamma$ with size $d = \lceil \gamma n \rceil = o(n)$. Thus the original problem of estimating the sparse p -vector β in (1) reduces to estimating a sparse d -vector $\beta = (\beta_1, \dots, \beta_d)^T$ based on the now much smaller submodel \mathcal{M} , namely,

$$\mathbf{y} = \mathbf{X}_{\mathcal{M}}\beta + \varepsilon, \quad (10)$$

where $\mathbf{X}_{\mathcal{M}} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^T$ denotes an $n \times d$ submatrix of \mathbf{X} 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 ultra high.

Now we briefly review several well-developed lower dimensional techniques that can be applied to estimate the d -vector β in (10) at the scale of $d < n$. Those methods include the Dantzig selector in Candès and Tao (2007), SCAD in Fan and Li (2001) and Fan and Peng (2004), adaptive Lasso in Zou (2006), among others.

3.1.1 Dantzig selector

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

$$\min_{\zeta \in \mathbf{R}^d} \|\zeta\|_1 \quad \text{subject to} \quad \|(\mathbf{X}_{\mathcal{M}})^T \mathbf{r}\|_\infty \leq \lambda_d \sigma, \quad (11)$$

where $\lambda_d > 0$ is a tuning parameter, $\mathbf{r} = \mathbf{y} - \mathbf{X}_{\mathcal{M}}\zeta$ 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:

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

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

We will 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. On the other hand, with the dimensionality first reduced by SIS the loss is now merely of a factor $\log d$, which is less than $\log n$.

3.1.2 Penalized least-squares and SCAD

Penalization is commonly used in variable selection. Fan and Li (2006) give a comprehensive overview of feature selection and a unified framework based on penalized likelihood approach to the problem of variable selection. They consider the penalized least squares (PLS)

$$\ell(\boldsymbol{\beta}) = \frac{1}{2n} \sum_{i=1}^n (Y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \sum_{j=1}^d p_{\lambda_j}(|\beta_j|), \quad (12)$$

where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_d)^T \in \mathbf{R}^d$ 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 choose not to penalize their coefficients. The regularization parameters λ_j can be chosen, for instance, by cross-validation (see, e.g., Breiman, 1996 and Tibshirani, 1996). A unified and effective algorithm for optimizing penalized likelihood, called local quadratic approximation (LQA), was proposed in Fan and Li (2001) and well studied in Hunter and Li (2005). In particular, LQA can be employed to minimize the above PLS. In our implementation, we choose $\lambda_j = \lambda$ and select λ by BIC.

An alternative and effective algorithm to minimize the penalized least-squares problem (12) is the local linear approximation (LLA) proposed by Zou and Li (2007). With the local linear approximation, the problem (12) can be cast as a sequence of penalized L_1 regression problems so that the LARS algorithm (Efron, *et al.*, 2004) can be employed.

The PLS (12) depends on the choice of penalty function $p_{\lambda_j}(\cdot)$. Commonly used penalty functions include the ℓ_p -penalty, $0 \leq p \leq 2$, nonnegative garrote in Breiman (1995), and smoothly clipped absolute deviation (SCAD) penalty, in Fan (1997) and a minimax concave penalty (MCP) in Zhang (2007) (see below for definition). In particular, the ℓ_1 -penalized least squares is called Lasso in Tibshirani (1996). In seminal papers, Donoho and Huo (2001) and Donoho and Elad (2003) show that penalized ℓ_0 -solution can be found by penalized ℓ_1 -method when the problem is sparse enough, which implies that the best subset regression can be found by using the penalized ℓ_1 -regression. Antoniadis and Fan (2001) propose the PLS for wavelets denoising with irregular designs. Fan and Li (2001) advocate 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 nonconvexity is required to reduce the estimation bias. It is well known that ℓ_p -penalty with $0 \leq p < 1$ does not satisfy the continuity condition, ℓ_p -penalty with $p > 1$ does not satisfy the sparsity condition, and ℓ_1 -penalty (Lasso) possesses the sparsity and continuity, but generates estimation bias. For example, studies in Fan and Li (2001), Meinshausen (2006), and Zou (2006) show some bias problems in the Lasso.

Fan (1997) proposes a continuously differentiable penalty function called the smoothly clipped absolute deviation (SCAD) penalty, which is defined by

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

Fan and Li (2001) suggest using $a = 3.7$. This function has similar feature to the penalty function $\lambda|\beta|/(1+|\beta|)$ advocated in Nikolova (2000). The MCP 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 MCP satisfy the above three conditions simultaneously. We will show in Theorem 5 in Section 5 that SIS followed by the SCAD enjoys the oracle properties.

3.1.3 Adaptive Lasso

The Lasso in Tibshirani (1996) has been widely used due to its convexity. Although it is easy to implement, an intrinsic bias problem comes with it. This 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) proposes an adaptive Lasso and Meinshausen (2006) proposes a relaxed Lasso.

The idea in Zou (2006) is to use an adaptively weighted ℓ_1 penalty in the PLS (12). Specifically, he introduced the following penalization term

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

where $\lambda \geq 0$ is a regularization parameter and $\boldsymbol{\omega} = (\omega_1, \dots, \omega_d)^T$ is a known weight vector. He further suggested using the weight vector $\hat{\boldsymbol{\omega}} = 1/|\hat{\boldsymbol{\beta}}|^\gamma$, where $\gamma \geq 0$, the power is understood componentwise, and $\hat{\boldsymbol{\beta}}$ is a root- n consistent estimator. The case of $\gamma = 1$ is closely related to the nonnegative garrote in Breiman (1995). He also showed that the adaptive Lasso can be solved by the LARS algorithm, which was proposed in Efron, Hastie, Johnstone and Tibshirani (2004). Using the same finite parameter setup as that in Knight and Fu (2000), Zou (2006) establishes that the adaptive Lasso has the oracle properties as long as the tuning parameter is chosen in a way such that $\lambda/\sqrt{n} \rightarrow 0$ and $\lambda n^{\frac{\gamma-1}{2}} \rightarrow \infty$ as $n \rightarrow \infty$.

3.2 SIS based model selection methods

For the problem of ultra-high dimensional variable selection, we propose first to apply a sure screening method such as SIS to reduce 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 the SCAD, Dantzig selector, Lasso, or adaptive Lasso. We call SIS followed by the SCAD and Dantzig selector SIS-SCAD and SIS-DS, respectively for short in the paper. In some situations, we may want to further reduce the model size down to $d' < d$ using a method such as the Dantzig selector along with the hard thresholding or the

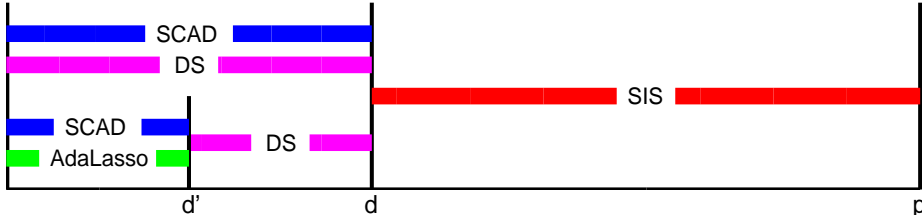


Figure 2: Methods of model selection with ultra high dimensionality.

Lasso with a suitable tuning, and finally choose a model with a more refined method such as the SCAD or adaptive Lasso. In the paper these two methods will be referred to as SIS-DS-SCAD and SIS-DS-AdaLasso, respectively for simplicity. Figure 2 shows a schematic diagram of these approaches.

The idea of SIS makes it feasible to do model selection with ultra high 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 the Bayesian methods (see, e.g., George and McCulloch, 1997).

3.3 Numerical studies

To study the performance of SIS based model selection methods 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 i.i.d. standard Gaussian predictors and Gaussian noise with standard deviation $\sigma = 1.5$. We considered two such models with $(n, p) = (200, 1000)$ and $(800, 20000)$, respectively. The sizes s of the true models, i.e., the numbers of nonzero coefficients, were chosen to be 8 and 18, respectively, and the nonzero components of the p -vectors β were randomly chosen as follows. We set $a = 4 \log n / \sqrt{n}$ and $5 \log n / \sqrt{n}$, respectively, and picked nonzero 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, respectively. For each model we simulated 200 data sets. Even with i.i.d. standard Gaussian predictors, the above settings are nontrivial since there is nonnegligible sample correlation among the predictors, which reflects the difficulty of high dimensional variable selection. As an evidence, we report in Figure 3 the distributions of the maximum absolute sample correlation when $n = 200$ and $p = 1000$ and 5000, respectively. It reveals significant sample correlation among 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 (DS) using a primal-dual algorithm, Lasso using the LARS algorithm, SIS-SCAD, SIS-DS, SIS-DS-SCAD, and SIS-DS-AdaLasso. For SIS-SCAD and SIS-DS, we chose $d = \lceil n / \log n \rceil$ and for the last two methods, we

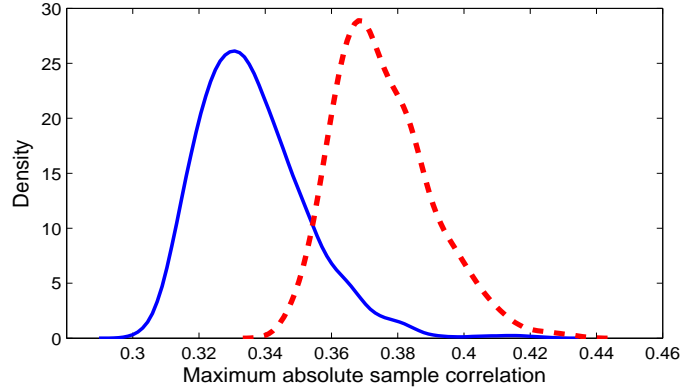


Figure 3: Distributions of the maximum absolute sample correlation when $n = 200, p = 1000$ (solid curve) and $n = 200, p = 5000$ (dashed curve).

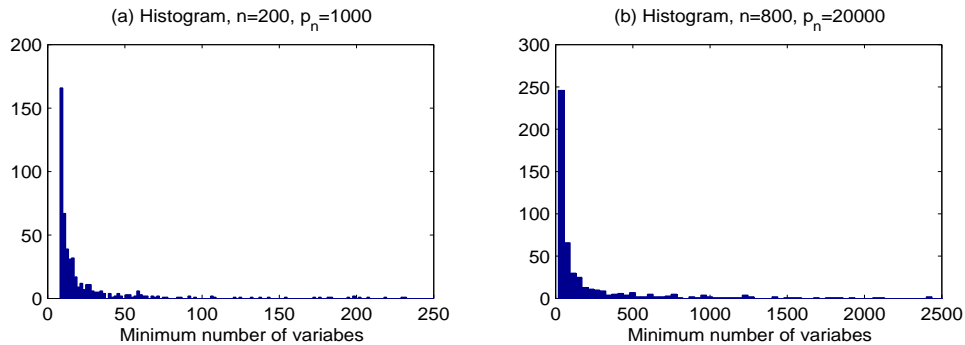


Figure 4: (a) Distribution of the minimum number of selected variables required to include the true model by using SIS when $n = 200, p = 1000$ in simulation I. (b) The same plot when $n = 800, p = 20000$.

chose $d = n - 1$ and $d' = \lceil n / \log n \rceil$ and in the middle step the Dantzig selector was used to further reduce the model size from d to d' by choosing variables with the d' largest componentwise magnitudes of the estimated d -vector.

The simulation results are summarized in Figure 4 and Table 1. Figure 4 depicts the distribution of the minimum number of selected variables, i.e., the selected model size, that is required to include the true model by using SIS. It shows clearly that in both settings it is safe to shrink the full model down to a submodel of size $\lceil n / \log n \rceil$ with SIS, which is consistent with the sure screening property of SIS shown in Theorem 1 in Section 5. For example, for the case of $n = 200$ and $p = 1000$, reducing the model size to 50 includes the variables in the true model with high probability, and for the case of $n = 800$ and $p = 20000$, 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 median of the estimation errors $\|\widehat{\beta} - \beta\|$ in ℓ_2 -norm. When the dimensionality is very high, both Lasso and Dantzig selector become computationally very expensive, so four entries of Table 1 are missing. In comparison, SIS reduces the computational

Table 1: Results of simulation I

p	Medians of the selected model sizes (upper entry) and the estimation errors (lower entry)					
	DS	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

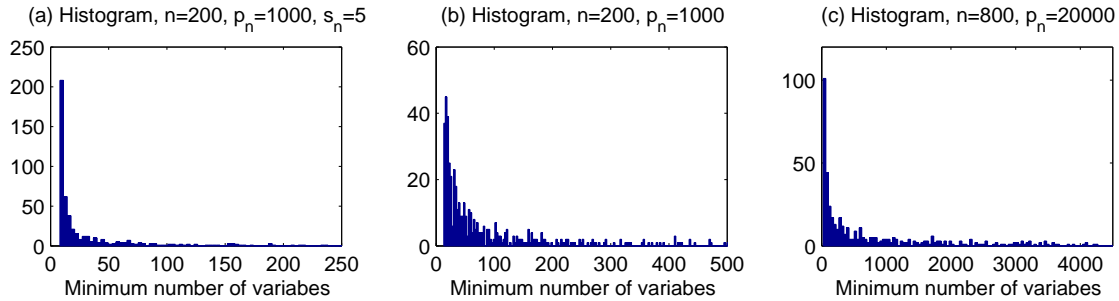


Figure 5: (a) Distribution of the minimum number of selected variables required to include the true model by using SIS when $n = 200, p = 1000, s = 5$ in simulation II. (b) The same plot when $n = 200, p = 1000, s = 8$. (c) The same plot when $n = 800, p = 20000$.

burden significantly.

From Table 1 we see that the Dantzig selector gives nonsparse solutions and the Lasso using the cross-validation for selecting its tuning parameter produces large models. We should point out here that a variation of the Dantzig selector, the Gauss-Dantzig selector in Candès and Tao (2007), should yield much smaller models, but for simplicity we did not include it in our simulation. Among all 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 estimation accuracy 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 $(800, 20000, 14)$, respectively, where s denotes the size of the true model, i.e., the number of nonzero coefficients. The three p -vectors β were generated in the same way as in simulation I. We set $(\sigma, a) = (1, 2 \log n / \sqrt{n})$, $(1.5, 4 \log n / \sqrt{n})$, and $(2, 4 \log n / \sqrt{n})$, respectively. In particular, the ℓ_2 -norms $\|\beta\|$ of the three simulated models are 3.304, 6.795, and 7.257, respectively. To introduce correlation between predictors, we first used a Matlab function `sprandsym` to randomly generate an

Table 2: Results of simulation II

p	Medians of the selected model sizes (upper entry) and the estimation errors (lower entry)					
	DS	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
	10^3	74	18	56	31.5	51
$(s = 8)$	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

$s \times s$ symmetric positive definite matrix \mathbf{A} with condition number $\sqrt{n}/\log n$, and drew samples of s predictors X_1, \dots, X_s from $\mathcal{N}(\mathbf{0}, \mathbf{A})$. Then we took $Z_{s+1}, \dots, Z_p \sim \mathcal{N}(\mathbf{0}, I_{p-s})$ and defined the remaining predictors as $X_i = Z_i + rX_{i-s}, i = s+1, \dots, 2s$ and $X_i = Z_i + (1-r)X_1, i = 2s+1, \dots, p$ with $r = 1 - 4 \log n/p, 1 - 5 \log n/p$, and $1 - 5 \log n/p$, respectively. 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 β . For SIS-SCAD and SIS-DS, we chose $d = \lfloor \frac{3}{2}n/\log n \rfloor, \lceil \frac{3}{2}n/\log n \rceil$, and $\lfloor n/\log n \rfloor$, respectively, and for the last two methods, we chose $d = n - 1$ and $d' = \lfloor \frac{3}{2}n/\log n \rfloor, \lceil \frac{3}{2}n/\log n \rceil$, and $\lfloor n/\log n \rfloor$, respectively. The simulation results are similarly summarized in Figure 5 and Table 2. Similar conclusions as those from simulation I can be drawn. As in simulation I, we did not include the Gauss-Dantzig selector for simplicity. It is interesting to observe that in the first setting here, the Lasso gives large models and its estimation errors are noticeable compare to the norm of the true coefficient vector β .

3.3.3 Leukemia data analysis

We also applied SIS to select features for the classification of a Leukemia data set. The Leukemia data from high-density Affymetrix oligonucleotide arrays were previously analyzed in Golub *et al.* (1999) and are available at <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 leukemia) and 25 in class AML (acute myelogenous leukemia). 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 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 $d = \lfloor 2n/\log n \rfloor$ genes with $n = 38$ the training sample size chosen above and then used the SCAD to get 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 determined by the Features Annealed Independence Rules (FAIR) procedure

in Fan and Fan (2007). 16 genes were picked up by their approach. Now we selected 16 genes and got a linear model with size 16 by using SIS-SCAD. 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 (NB) rule to the resulted 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 both chose 16 genes and made 1 test error with training errors 0 and 4, respectively, while the nearest shrunken centroids method picked up 21 genes and made 1 training error and 2 test errors.

Table 3: Classification errors on the Leukemia 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

4 Extensions of SIS

Like modeling building in linear regression, there are many variations in the implementation of SIS. This section discusses some extensions of SIS to enhance its methodological power. In particular, an 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 SIS

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 to be 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 can not 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 An ISIS

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 ultra high to a relatively large scale, say, below sample size. But when those assumptions fail, it could happen that SIS would

miss some important predictors. To overcome this problem, we propose below an ISIS to enhance the methodological power. It is an iterative applications of the SIS approach to variable selection.

The ISIS works as follows. In the first step, we select a subset of k_1 variables $\mathcal{A}_1 = \{X_{i_1}, \dots, X_{i_{k_1}}\}$ using an SIS based model selection method such as the SIS-SCAD or SIS-LASSO. Then we have an n -vector of the residuals from regressing the response Y over $X_{i_1}, \dots, X_{i_{k_1}}$. 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 $\mathcal{A}_2 = \{X_{j_1}, \dots, X_{j_{k_2}}\}$. We remark that fitting the residuals from the previous step on $\{X_1, \dots, X_p\} \setminus \mathcal{A}_1$ can significantly weaken the priority of those unimportant variables that are highly correlated with the response through their associations with $X_{i_1}, \dots, X_{i_{k_1}}$, since the residuals are uncorrelated with those selected variables in \mathcal{A}_1 . This helps solving 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 entering into 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 get ℓ disjoint subsets $\mathcal{A}_1, \dots, \mathcal{A}_\ell$ whose union $\mathcal{A} = \cup_{i=1}^{\ell} \mathcal{A}_i$ has a size d , which is less than n . In practical implementation, we can choose, for example, the largest l such that $|\mathcal{A}| < n$.

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

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 5 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 ones 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 weaken correlation among the covariates. For example, in somatotype studies the 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 $(w_1, w_2 - w_1, w_3 - w_2)^T$, which is a linear transformation of $(w_1, w_2, w_3)^T$ that specifies the changes of the weights instead of the weights themselves. Another important example is the financial time series such as the prices of the stocks or interest rates. Differencing can significantly weaken the correlation

among those variables.

Methods of statistical transformation include an application of a clustering algorithm such as the hierarchical clustering or k -mean algorithm using the correlation metrics to first group variables into highly correlated groups and then apply the sparse principal components analysis (PCA) 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.

Those statistical techniques we introduced above can help identify the important features and thus improve the effectiveness of the vanilla SIS based model selection strategy.

4.2 Numerical evidence

To study the performance of the ISIS 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

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

where X_1, \dots, X_p are p predictors and $\varepsilon \sim N(0, 1)$ is a noise that is independent of the predictors. In the simulation, a sample of (X_1, \dots, 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 $\sigma_{ii} = 1, i = 1, \dots, p$ and $\sigma_{ij} = \rho, i \neq j$. We considered 18 such models characterized by (p, n, ρ) with $p = 100, 1000, n = 20, 50, 70$, and $\rho = 0, 0.1, 0.5, 0.9$, respectively, and for each model we simulated 200 data sets.

For each model, we applied SIS and the ISIS to select n variables and tested their accuracy in including the true model $\{X_1, X_2, X_3\}$. For the ISIS, the SIS-SCAD with $d = \lceil n / \log n \rceil$ was used at each step and we kept on collecting variables in those disjoint \mathcal{A}_j 's until we got n variables (if there were more variables than needed in the final step, we only included those with the largest absolute coefficients). In Table 4, we report the percentages of SIS, LASSO and ISIS that include the true model. It is clear that the collinearity (large value of ρ) and high-dimensionality deteriorate the performance of SIS and LASSO, and 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. On the other hand, 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 LASSO when LASSO is used in the implementation of ISIS.

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

p	n		$\rho = 0$	$\rho = 0.1$	$\rho = 0.5$	$\rho = 0.9$
100	20	SIS	.755	.855	.690	.670
		LASSO	.970	.990	.985	.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	.205	.255	.145	.085
		LASSO	.340	.555	.556	.220
		ISIS	1	1	1	1
	50	SIS	.990	.960	.870	.860
		LASSO	1	1	1	1
		ISIS	1	1	1	1
	70	SIS	1	.995	.97	.97
		LASSO	1	1	1	1
		ISIS	1	1	1	1

4.2.2 Simulated example II

For the second simulated example, we used the same setup 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

$$Y = 5X_1 + 5X_2 + 5X_3 - 15\sqrt{\rho}X_4 + \varepsilon,$$

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

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 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 mentioned at the beginning of Section 4.1.

4.2.3 Simulated example III

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

$$Y = 5X_1 + 5X_2 + 5X_3 - 15\sqrt{\rho}X_4 + X_5 + \varepsilon,$$

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 X_5 was introduced is to make it have a very small correlation with the

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

p	$\rho = 0.5$	$n = 20$	$n = 50$	$n = 70$
100	SIS	.025	.490	.740
	LASSO	.000	.360	.915
	ISIS	1	1	1
1000	SIS	.000	.000	.000
	LASSO	.000	.000	.000
	ISIS	1	1	1

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, \dots, X_p and hence has a lower priority to be 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 the ISIS can improve significantly over the simple SIS and LASSO and always picks all true variables. This shows again that the ISIS is able to pick up two difficult variables X_4 and X_5 , which addresses simultaneously the second and third problem at the beginning of Section 4.

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

p	$\rho = 0.5$	$n = 20$	$n = 50$	$n = 70$
100	SIS	.000	.285	.645
	LASSO	.000	.310	.890
	ISIS	1	1	1
1000	SIS	.000	.000	.000
	LASSO	.000	.000	.000
	ISIS	1	1	1

4.2.4 Simulations I and II in Section 3.3 revisited

Now let us go back to the two simulation studies 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 β by using SCAD. This method is referred to as ISIS-SCAD. We report in Table 7 the median of the selected model sizes and median of the estimation errors $\|\widehat{\beta} - \beta\|$ in ℓ_2 -norm. We can see clearly that ISIS improves over the simple SIS. The improvements are more drastic for simulation II in which covariates are more correlated and the variable selections are more challenging.

Table 7: Simulations I and II in Section 3.3 revisited: Medians of the selected model sizes (upper entry) and the estimation errors (lower entry)

p	Simulation I	Simulation II	
	ISIS-SCAD		ISIS-SCAD
1000	13	$(s = 5)$	11
	0.329		0.223
		$(s = 8)$	13.5
			0.366
20000	31		27
	0.246		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 (1) that $Y = \sum_{i=1}^p \beta_i X_i + \varepsilon$. Throughout the paper we let $\mathcal{M}_* = \{1 \leq i \leq p : \beta_i \neq 0\}$ be the true sparse model with nonsparsity rate $s = |\mathcal{M}_*|$ and define

$$\mathbf{z} = \Sigma^{-1/2} \mathbf{x} \quad \text{and} \quad \mathbf{Z} = \mathbf{X} \Sigma^{-1/2}, \quad (14)$$

where $\mathbf{x} = (X_1, \dots, X_p)^T$ and $\Sigma = \text{cov}(\mathbf{x})$. Clearly, the n rows of the transformed design matrix \mathbf{Z} are i.i.d. copies of \mathbf{z} which now has covariance matrix I_p . For simplicity, all the predictors X_1, \dots, 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 $\mathbf{Z} \Sigma^{1/2}$. Below we will make assumptions on \mathbf{Z} and Σ 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:

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

$$P\left(\lambda_{\max}(\tilde{p}^{-1} \tilde{\mathbf{Z}} \tilde{\mathbf{Z}}^T) > c_1 \text{ and } \lambda_{\min}(\tilde{p}^{-1} \tilde{\mathbf{Z}} \tilde{\mathbf{Z}}^T) < 1/c_1\right) \leq e^{-C_1 n} \quad (15)$$

holds for any $n \times \tilde{p}$ submatrix $\tilde{\mathbf{Z}}$ of \mathbf{Z} with $cn < \tilde{p} \leq p$. We will 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 nonzero singular values of the $n \times \tilde{p}$ matrix $\tilde{\mathbf{Z}}$ are of the same order, which is reasonable since $\tilde{p}^{-1} \tilde{\mathbf{Z}} \tilde{\mathbf{Z}}^T$ will approach I_n as $\tilde{p} \rightarrow \infty$: the larger the \tilde{p} , the closer to I_n . It relies on the random matrix theory (RMT) to derive the deviation inequality in (15). 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, e.g., Silverstein (1985), Bai and Yin (1993), Bai (1999), Johnstone (2001), and Ledoux (2001, 2005).

Some of the assumptions below are purely technical and only serve 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 $\xi > 0$.

Even though we do not explicit restriction on ξ , the concentration property (15) makes restrictions on ξ .

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. $\text{var}(Y) = O(1)$ and for some $\kappa \geq 0$ and $c_2, c_3 > 0$,

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

As seen later, κ controls the rate of probability error in recovering the true sparse model. Although $b = \min_{i \in \mathcal{M}_*} |\text{cov}(\beta_i^{-1}Y, X_i)|$ is assumed here to be bounded away from zero, our asymptotic study applies as well to the case where b tends to zero as $n \rightarrow \infty$. 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 exist some $\tau \geq 0$ and $c_4 > 0$ such that

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

This condition rules the case of strong collinearity.

The largest eigenvalue of the population covariance matrix $\mathbf{\Sigma}$ is allowed to diverge as n grows. When there are many predictors, it is often the case that their covariance matrix is block diagonal or nearly block diagonal under a suitable permutation of the variables. Therefore $\lambda_{\max}(\mathbf{\Sigma})$ 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, 2007). 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

Analyzing the p -vector $\boldsymbol{\omega}$ in (2) when $p > n$ is essentially difficult. The approach we took is to first study the specific case with $\mathbf{\Sigma} = I_p$ and then relate the general case to the specific case.

Theorem 1. (Accuracy of SIS). *Under Conditions 1–4, if $2\kappa + \tau < 1$ then there exists 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 $s \leq [\gamma n]$ is implied by our assumptions as demonstrated in the technical proof. The above theorem shows that SIS has the sure screening property and can reduce up to 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.

The proof of Theorem 1 depends on the iterative application of the following theorem, which demonstrates the accuracy of the 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 large enough and the percentage of remaining variables δ is large enough, the sure screening property is ensured with the overwhelming probability.

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

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

The above theorem reveals that when the tuning parameters are chosen appropriately, with an overwhelming probability the submodel $\mathcal{M}_{\delta,\lambda}^1$ 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 $\delta n^\theta \rightarrow \infty$ as $n \rightarrow \infty$ for some $\theta < 1 - 2\kappa - \tau$, then successive applications of the procedure in (8) for k times results in a submodel $\mathcal{M}_{\delta,\lambda}$ with size $d = \lceil \delta^k p \rceil < n$ 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 $\lceil \delta^k p \rceil < n$. This implies that $k = O(\log p / \log \delta) = O(n^\xi)$. Therefore, the accumulated error probability, from the union bound, is still of exponentially small with a possibility of a different constant C .

ITRRS has now been shown to possess 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 can not be too small, or equivalently, the number of iteration steps in ITRRS can not be too large, due to the cumulation 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 may not work since it requires $p - d$ steps of iterations, which may exceed the error bound in Theorem 2.

5.3 Consistency of SIS-DS and SIS-SCAD

To study the property of the Dantzig selector, Candès and Tao (2007) introduce the notion of uniform uncertainty principle (UUP) 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 $T \subset \{1, \dots, d\}$, denote by \mathbf{A}_T the $n \times |T|$ submatrix of \mathbf{A} obtained by

extracting its columns corresponding to the indices in T . For any positive integer $S \leq d$, the S -restricted isometry constant $\delta_S = \delta_S(\mathbf{A})$ of \mathbf{A} is defined to be the smallest quantity such that

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

holds for all subsets T with $|T| \leq S$ and $\mathbf{v} \in \mathbf{R}^{|T|}$. For any pair of positive integers S, S' with $S+S' \leq d$, 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

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

holds for all disjoint subsets T, T' of cardinalities $|T| \leq S$ and $|T'| \leq S'$, $\mathbf{v} \in \mathbf{R}^{|T|}$, 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 Candès and Tao (2007), where $\varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$ for some $\sigma > 0$.

Theorem 4. (Consistency of SIS-DS). *Assume with large probability, $\delta_{2s}(\mathbf{X}_{\mathcal{M}}) + \theta_{s,2s}(\mathbf{X}_{\mathcal{M}}) \leq t < 1$ and choose $\lambda_d = \sqrt{2 \log d}$ in (11). Then with large probability, we have*

$$\left\| \widehat{\boldsymbol{\beta}}_{\text{DS}} - \boldsymbol{\beta} \right\|^2 \leq C (\log d) s \sigma^2,$$

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

This theorem shows that 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 SIS-SCAD, that is, SIS followed by the SCAD. For simplicity, a common regularization parameter λ is used for the SCAD penalty function. Let $\widehat{\boldsymbol{\beta}}_{\text{SCAD}} = (\widehat{\beta}_1, \dots, \widehat{\beta}_d)^T$ be a minimizer of the SCAD-PLS in (12). 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 SIS-SCAD). *If $d = o(n^{1/3})$ and the assumptions of Theorem 2 in Fan and Peng (2004) be satisfied, then, with probability tending to one, the SCAD-PLS estimator $\widehat{\boldsymbol{\beta}}_{\text{SCAD}}$ satisfies: (i) $\widehat{\beta}_i = 0$ for any $i \notin \mathcal{M}_*$; (ii) the components of $\widehat{\boldsymbol{\beta}}_{\text{SCAD}}$ in \mathcal{M}_* perform as well as if the true model \mathcal{M}_* were known.*

The SIS-SCAD has been shown to enjoy the oracle properties.

6 Concluding remarks

This paper studies the problem of high dimensional variable selection for the linear model. The concept of sure screening was introduced and a fast and efficient sure screening method Sure Independence Screening (SIS) was proposed. SIS has been shown to be able to reduce dimensionality from high up to an exponential growth to a relatively large scale that is below sample size. It can not only speed up variable selection drastically but also improve the estimation accuracy when dimensionality is ultra high. SIS combined with well-developed lower dimensional technique such as the SCAD, Dantzig selector, Lasso, or adaptive Lasso provides a powerful tool for high dimensional variable selection.

Some extensions of SIS were also discussed. In particular, the ISIS was proposed to enhance the performance of SIS in practice, particular in the situations where technical conditions fail. This raises a challenging question: To what extent the ISIS relaxes the technical conditions of SIS to possess the sure screening properties.

ITRRS was introduced for better understanding SIS and was used as a technical device for establishing the sure screening property. As a by-product, we demonstrated that the stepwise deletion estimator might not have the sure screening property when the dimensionality is of exponential order. This raises an interesting question if the stepwise addition or its related greedy algorithms such as matching pursuit would have the sure screening property. If so, how large the selected model has to be. This is an interesting future topic.

Another interesting problem is to extend the SIS and ISIS to the family of generalized linear models or other more general loss function such as the hinge loss and the loss associated with the support vector machine. How to define the associated 'residuals' for iterative SIS? Will sure screening property continue to hold?

The study of the paper focuses only on random designs that are commonly seen in statistical problems mentioned in the introduction. In some problems on signal processing or image analysis, the design matrix can be deterministic. It remains open how to impose a set of conditions such that the sure screening property holds.

A Appendix

For notational convenience, we use both C and c to denote generic positive constants throughout.

A.1 Proof of Theorem 1

Motivated by the results in Theorems 2 and 3, the idea is to successively apply dimensionality reduction in a way described in (16) below. To enhance the readability, we split the whole proof into two main steps and multiple substeps.

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

$$\widetilde{\mathcal{M}}_\delta^1 = \{1 \leq i \leq p : |\omega_i| \text{ is among the first } [\delta p] \text{ largest of all}\}. \quad (16)$$

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

$$P\left(\mathcal{M}_* \subset \widetilde{\mathcal{M}}_\delta^1\right) = 1 - O(\exp(-Cn^{1-2\kappa}/\log n)). \quad (17)$$

The main idea is to relate the general case to the specific case with $\Sigma = I_p$, which is separately studied in Sections A.4–A.6 below. A key ingredient is the representation (18) below of the $p \times p$ random matrix $\mathbf{X}^T \mathbf{X}$. Throughout, let $\mathbf{S} = (\mathbf{Z}^T \mathbf{Z})^+ \mathbf{Z}^T \mathbf{Z}$ and $\mathbf{e}_i = (0, \dots, 1, \dots, 0)^T$ be a unit vector in \mathbf{R}^p with the i -th entry 1 and 0 elsewhere, $i = 1, \dots, p$.

Since $\mathbf{X} = \mathbf{Z}\Sigma^{1/2}$, it follows from (44) that

$$\mathbf{X}^T \mathbf{X} = p\Sigma^{1/2} \widetilde{\mathbf{U}}^T \text{diag}(\mu_1, \dots, \mu_n) \widetilde{\mathbf{U}} \Sigma^{1/2}, \quad (18)$$

where μ_1, \dots, μ_n are n eigenvalues of $p^{-1}\mathbf{Z}\mathbf{Z}^T$, $\tilde{\mathbf{U}} = (I_n, \mathbf{0})_{n \times p}$, and \mathbf{U} is uniformly distributed on the orthogonal group $\mathcal{O}(p)$. By (1) and (2), we have

$$\boldsymbol{\omega} = \mathbf{X}^T \mathbf{X} \boldsymbol{\beta} + \mathbf{X}^T \boldsymbol{\varepsilon} \hat{=} \boldsymbol{\xi} + \boldsymbol{\eta}. \quad (19)$$

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

Step 1.1. First, we consider term $\boldsymbol{\xi} = (\xi_1, \dots, \xi_p)^T = \mathbf{X}^T \mathbf{X} \boldsymbol{\beta}$.

Step 1.1.1. Bounding $\|\boldsymbol{\xi}\|$ from above. It is obvious that

$$\text{diag}(\mu_1^2, \dots, \mu_n^2) \leq [\lambda_{\max}(p^{-1}\mathbf{Z}\mathbf{Z}^T)]^2 I_n$$

and $\tilde{\mathbf{U}}\boldsymbol{\Sigma}\tilde{\mathbf{U}}^T \leq \lambda_{\max}(\boldsymbol{\Sigma})I_n$. These and (18) lead to

$$\|\boldsymbol{\xi}\|^2 \leq p^2 \lambda_{\max}(\boldsymbol{\Sigma}) [\lambda_{\max}(p^{-1}\mathbf{Z}\mathbf{Z}^T)]^2 \boldsymbol{\beta}^T \boldsymbol{\Sigma}^{1/2} \tilde{\mathbf{U}}^T \tilde{\mathbf{U}} \boldsymbol{\Sigma}^{1/2} \boldsymbol{\beta}. \quad (20)$$

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

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

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

$$P\left(\boldsymbol{\beta}^T \boldsymbol{\Sigma}^{1/2} \tilde{\mathbf{U}}^T \tilde{\mathbf{U}} \boldsymbol{\Sigma}^{1/2} \boldsymbol{\beta} > O\left(\frac{n}{p}\right)\right) \leq O(e^{-Cn}). \quad (21)$$

Since $\lambda_{\max}(\boldsymbol{\Sigma}) = O(n^\tau)$ and $P(\lambda_{\max}(p^{-1}\mathbf{Z}\mathbf{Z}^T) > c_1) \leq e^{-C_1 n}$ by Conditions 2 and 4, (20) and (21) along with Bonferroni's inequality yield

$$P\left(\|\boldsymbol{\xi}\|^2 > O(n^{1+\tau} p)\right) \leq O(e^{-Cn}). \quad (22)$$

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 (18), we have

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

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

$$\left| \langle \boldsymbol{\Sigma}^{1/2} \boldsymbol{\beta}, \boldsymbol{\Sigma}^{1/2} \mathbf{e}_i \rangle \right| = |\beta_i| |\text{cov}(\beta_i^{-1} Y, X_i)| \geq c/n^\kappa. \quad (23)$$

Thus, there exists $Q \in \mathcal{O}(p)$ 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 $(\mu_1, \dots, \mu_n)^T$ 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

$$\xi_i \stackrel{(d)}{=} p \langle \boldsymbol{\Sigma}^{1/2} \boldsymbol{\beta}, \boldsymbol{\Sigma}^{1/2} \mathbf{e}_i \rangle R_1 + O(p) R_2 \hat{=} \xi_{i,1} + \xi_{i,2}, \quad (24)$$

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

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

and thus by Condition 2, Lemma 4, and Bonferroni's inequality, we have for some $c > 0$ and $C > 0$,

$$P(R_1 < cn/p) \leq O(e^{-Cn}).$$

This along with (23) gives for some $c > 0$,

$$P(|\xi_{i,1}| < cn^{1-\kappa}) \leq O(e^{-Cn}). \quad (25)$$

Similarly to Step 1.1.1, it can be shown that

$$P(\|\mathbf{R}\|^2 > O(n/p)) \leq O(e^{-Cn}). \quad (26)$$

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

$$R_2 \stackrel{(d)}{=} \|\tilde{\mathbf{R}}\| W_1 / \|\mathbf{W}\|. \quad (27)$$

In view of (26), (27), and $\xi_{i,2} = O(pR_2)$, applying the argument in the proof of Lemma 5 gives for some $c > 0$,

$$P(|\xi_{i,2}| > c\sqrt{n}|W|) \leq O(e^{-Cn}), \quad (28)$$

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

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

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

which along with (28) and Bonferroni's inequality shows that

$$P(|\xi_{i,2}| > x) \leq O(\exp(-Cn^{1-2\kappa}/\log n)). \quad (29)$$

Therefore, by Bonferroni's inequality, combining (24), (25), and (29) together gives for some $c > 0$,

$$P(|\xi_i| < cn^{1-\kappa}) \leq O(\exp(-Cn^{1-2\kappa}/\log n)), \quad i \in \mathcal{M}_*. \quad (30)$$

Step 1.2. Then, we examine term $\boldsymbol{\eta} = (\eta_1, \dots, \eta_p)^T = \mathbf{X}^T \boldsymbol{\varepsilon}$.

Step 1.2.1. Bounding $\|\boldsymbol{\eta}\|$ from above. Clearly, we have

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

Then, it follows that

$$\|\boldsymbol{\eta}\|^2 = \boldsymbol{\varepsilon}^T \mathbf{X} \mathbf{X}^T \boldsymbol{\varepsilon} \leq p \lambda_{\max}(\boldsymbol{\Sigma}) \lambda_{\max}(p^{-1} \mathbf{Z} \mathbf{Z}^T) \|\boldsymbol{\varepsilon}\|^2. \quad (31)$$

From Condition 2, we know that $\varepsilon_1^2/\sigma^2, \dots, \varepsilon_n^2/\sigma^2$ are i.i.d. χ_1^2 -distributed random variables. Thus, by (46) in Lemma 3, there exist some $c > 0$ and $C > 0$ such that

$$P\left(\|\boldsymbol{\varepsilon}\|^2 > cn\sigma^2\right) \leq e^{-Cn},$$

which along with (31), Conditions 2 and 4, and Bonferroni's inequality yields

$$P\left(\|\boldsymbol{\eta}\|^2 > O(n^{1+\tau}p)\right) \leq O(e^{-Cn}). \quad (32)$$

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

$$\text{var}(\eta_i | \mathbf{X} = X) = \sigma^2 \mathbf{e}_i^T X^T X \mathbf{e}_i. \quad (33)$$

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

$$P(\mathcal{E}^c) \leq O(e^{-Cn}). \quad (34)$$

On the event \mathcal{E} , we have

$$P(|\eta_i| > x | \mathbf{X}) \leq P(\sqrt{cn}|W| > x) \text{ for any } x > 0, \quad (35)$$

where W is a $\mathcal{N}(0, 1)$ -distributed random variable. Thus, it follows from (34) and (35) that

$$P(|\eta_i| > x) \leq O(e^{-Cn}) + P(\sqrt{cn}|W| > x). \quad (36)$$

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 (36) shows that

$$P(|\eta_i| > o(n^{1-\kappa})) \leq O(\exp(-Cn^{1-2\kappa}/\log n)). \quad (37)$$

Step 1.3. Finally, we combine the results obtained in Steps 1.1 and 1.2 together. By Bonferroni's inequality, it follows from (19), (22), (30), (32), and (37) that for some constants $c_1, c_2, C > 0$,

$$P\left(\min_{i \in \mathcal{M}_*} |\omega_i| < c_1 n^{1-\kappa} \text{ or } \|\boldsymbol{\omega}\|^2 > c_2 n^{1+\tau} p\right) \leq O(s \exp(-Cn^{1-2\kappa}/\log n)). \quad (38)$$

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 $c > 0$,

$$\#\left\{1 \leq k \leq p : |\omega_k| \geq \min_{i \in \mathcal{M}_*} |\omega_i|\right\} \leq c \frac{n^{1+\tau} p}{(n^{1-\kappa})^2} = \frac{cp}{n^{1-2\kappa-\tau}}, \quad (39)$$

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

Now, we are ready to see from (39) that if δ satisfies $\delta n^{1-2\kappa-\tau} \rightarrow \infty$ as $n \rightarrow \infty$, then (17) holds for some constant $C > 0$ larger than that in (38).

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

- First, carry out the procedure in (16) to the full model $\widetilde{\mathcal{M}}_\delta^0 \hat{=} \{1, \dots, p\}$ and get a submodel $\widetilde{\mathcal{M}}_\delta^1$ with size $[\delta p]$;
- Then, apply a similar procedure to the model $\widetilde{\mathcal{M}}_\delta^1$ and again obtain a submodel $\widetilde{\mathcal{M}}_\delta^2 \subset \widetilde{\mathcal{M}}_\delta^1$ with size $[\delta^2 p]$, and so on;
- Finally, get a submodel $\widetilde{\mathcal{M}}_\delta \hat{=} \widetilde{\mathcal{M}}_\delta^k$ with size $d = [\delta^k p] = [\delta^r n] < n$, where $[\delta^{k-1} p] = [\delta^{r-1} n] > n$.

It is obvious that $\widetilde{\mathcal{M}}_\delta = \mathcal{M}_\gamma$, 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 $k \geq 1$ in a way such that

$$\delta n^{1-2\kappa-\tau} \rightarrow \infty \quad \text{and} \quad \delta n^{\theta_0} \rightarrow 0 \quad \text{as } n \rightarrow \infty, \quad (40)$$

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

$$\gamma n^{r(1-2\kappa-\tau)} \rightarrow \infty \quad \text{and} \quad \gamma n^{\theta_1} \rightarrow 0 \quad \text{as } n \rightarrow \infty. \quad (41)$$

Before going further, let us make two important observations. First, for any principal submatrix Σ^0 of Σ corresponding to a subset of variables, Condition 4 ensures that

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

Second, by definition, Property C in (15) holds for any $n \times \tilde{p}$ submatrix $\tilde{\mathbf{Z}}$ of \mathbf{Z} with $cn < \tilde{p} \leq p$, where $c > 1$ is some constant. Thus, the probability bound in (17) is uniform over dimension $\tilde{p} \in (cn, p]$. Therefore, for some $C > 0$, by (40) and (17) we have in each step $1 \leq i \leq k$ of the above dimensionality reduction,

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

which along with Bonferroni's inequality gives

$$P(\mathcal{M}_* \subset \mathcal{M}_\gamma) = 1 - O(k \exp(-Cn^{1-2\kappa}/\log n)). \quad (42)$$

It follows from (40) 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 (42) yields

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

Finally, in view of (41), 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

One observes that (8) uses only the order of componentwise magnitudes of $\boldsymbol{\omega}^\lambda$, so it is invariant under scaling. Therefore, in view of (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

$$\lambda\boldsymbol{\omega}^\lambda = \boldsymbol{\omega} - \left[I_p - (I_p + \lambda^{-1}\mathbf{X}^T\mathbf{X})^{-1} \right] \boldsymbol{\omega}.$$

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

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

and thus

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

which along with (38), Conditions 2 and 4, and Bonferroni's inequality shows that

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

Again, by Bonferroni's inequality and (38), any λ satisfying $\lambda^{-1}n^{\frac{1+3\tau}{2}}p^{3/2} = o(n^{1-\kappa})$ can be used. Note that $\kappa + \tau/2 < 1/2$ by assumption. So in particular, we can choose any λ satisfying $\lambda(p^{3/2}n)^{-1} \rightarrow \infty$ as $n \rightarrow \infty$.

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 Sections A.4–A.6 below, we assume that $p > n$ and the distribution of \mathbf{z} is continuous and spherically symmetric, that is, 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 centered 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 $\mathbf{S} = (\mathbf{Z}^T\mathbf{Z})^+ \mathbf{Z}^T\mathbf{Z}$

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,

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

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

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

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,

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

Clearly, this group action is transitive, i.e. for any $g_1, g_2 \in \mathcal{G}_{p,n}$, there exists 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 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

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

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

$$\mathbf{Z} = \mathbf{V} \mathbf{D}_1 \mathbf{U}, \tag{43}$$

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 $\sqrt{\mu_1}, \dots, \sqrt{\mu_n}$, respectively. Thus,

$$\mathbf{Z}^T \mathbf{Z} = \mathbf{U}^T \text{diag}(\mu_1, \dots, \mu_n, 0, \dots, 0) \mathbf{U} \tag{44}$$

and its Moore-Penrose generalized inverse is

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

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

$$\mathbf{S} = (\mathbf{Z}^T \mathbf{Z})^+ \mathbf{Z}^T \mathbf{Z} = \mathbf{U}^T \text{diag}(I_n, 0) \mathbf{U}, \quad \mathbf{U} \in \mathcal{O}(p). \tag{45}$$

From (43), we know that $\mathbf{Z} = \mathbf{V} \text{diag}(\sqrt{\mu_1}, \dots, \sqrt{\mu_n}) (I_n, \mathbf{0})_{n \times p} \mathbf{U}$, and thus

$$(I_n, \mathbf{0})_{n \times p} \mathbf{U} = \text{diag}(1/\sqrt{\mu_1}, \dots, 1/\sqrt{\mu_n}) \mathbf{V}^T \mathbf{Z}.$$

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.,

$$\mathbf{Z} Q \stackrel{(d)}{=} \mathbf{Z} \quad \text{for any } Q \in \mathcal{O}(p).$$

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

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

where $\tilde{\mathbf{U}}$ is uniformly distributed on the orthogonal group $\mathcal{O}(p)$. In particular, we see that $(\mu_1, \dots, \mu_n)^T$ is independent of $(I_n, \mathbf{0})_{n \times p} \mathbf{U}$. Therefore, these facts along with (45) yield the following lemma.

Lemma 1. $\mathcal{L} \left((I_n, \mathbf{0})_{n \times p} \mathbf{U} \right) = \mathcal{L} \left((I_n, \mathbf{0})_{n \times p} \tilde{\mathbf{U}} \right)$ and $(\mu_1, \dots, \mu_n)^T$ 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 μ_1, \dots, μ_n are n eigenvalues of $\mathbf{Z}\mathbf{Z}^T$. 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 the above singular value decomposition (43).

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

Lemma 2. $\mathcal{L}(\langle \mathbf{S}\mathbf{e}_1, \mathbf{e}_1 \rangle) = \frac{\chi_n^2}{\chi_n^2 + \chi_{p-n}^2}$, 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, that is, $\langle \mathbf{S}\mathbf{e}_1, \mathbf{e}_1 \rangle$ has a beta distribution with parameters $n/2$ and $(p - n)/2$.

Proof. Lemma 1 gives $\mathcal{L}(\mathbf{S}) = \mathcal{L}(\mathbf{U}^T \text{diag}(I_n, 0) \mathbf{U})$, 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 $\mathbf{W} = (W_1, \dots, W_p)^T \sim \mathcal{N}(\mathbf{0}, I_p)$. Then, we have $\mathbf{U}\mathbf{e}_1 \stackrel{(d)}{=} \mathbf{W}/\|\mathbf{W}\|$ and

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

This proves Lemma 2. □

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

Lemma 3. (Moderate deviation). Let ξ_1, \dots, ξ_n be i.i.d. χ_1^2 -distributed random variables. Then,

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

$$P(n^{-1}(\xi_1 + \dots + \xi_n) > 1 + \varepsilon) \leq e^{-A_\varepsilon n}, \quad (46)$$

where $A_\varepsilon = [\varepsilon - \log(1 + \varepsilon)]/2 > 0$.

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

$$P(n^{-1}(\xi_1 + \dots + \xi_n) < 1 - \varepsilon) \leq e^{-B_\varepsilon n}, \quad (47)$$

where $B_\varepsilon = [-\varepsilon - \log(1 - \varepsilon)]/2 > 0$.

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

$$M(t) = Ee^{t\xi} = (1 - 2t)^{-1/2}, \quad t \in (-\infty, 1/2). \quad (48)$$

Thus, for any $\varepsilon > 0$ and $0 < t < 1/2$, by Chebyshev's inequality (see, e.g. van der Vaart and Wellner, 1996) we have

$$P\left(\frac{\xi_1 + \cdots + \xi_n}{n} > 1 + \varepsilon\right) \leq \frac{1}{e^{(t+1)n\varepsilon}} E \exp\{t(\xi_1 + \cdots + \xi_n)\} = \exp(-nf_\varepsilon(t)),$$

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

$$P(n^{-1}(\xi_1 + \cdots + \xi_n) > 1 + \varepsilon) \leq e^{-A_\varepsilon n}.$$

This proves (46).

(ii) For any $0 < \varepsilon < 1$ and $t > 0$, by Chebyshev's inequality and (48), we have

$$P(n^{-1}(\xi_1 + \cdots + \xi_n) < 1 - \varepsilon) \leq \frac{1}{e^{tn\varepsilon}} E \exp\{t(1 - \xi_1) + \cdots + t(1 - \xi_n)\} = \exp(-ng_\varepsilon(t)),$$

where $g_\varepsilon(t) = \frac{1}{2} \log(1 + 2t) - (1 - \varepsilon)t$. Taking $t = \varepsilon/(2(1 - \varepsilon))$ yields (47). \square

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

$$P\left(\langle \mathbf{Se}_1, \mathbf{e}_1 \rangle < c_1 \frac{n}{p} \text{ or } > c_2 \frac{n}{p}\right) \leq 4e^{-Cn}. \quad (49)$$

Proof. From Lemma 3, we know that $\langle \mathbf{Se}_1, \mathbf{e}_1 \rangle \stackrel{(d)}{=} \xi/\eta$, where ξ is χ_n^2 -distributed and η is χ_p^2 -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 exist \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

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

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

$$\left\{ \langle \mathbf{Se}_1, \mathbf{e}_1 \rangle < c_1 \frac{n}{p} \text{ or } > c_2 \frac{n}{p} \right\} \subset \mathcal{A} \cup \mathcal{B}. \quad (50)$$

It follows from (46) and (47) and the choice of \tilde{c}_1 and \tilde{c}_2 above that

$$P(\mathcal{A}) \leq 2e^{-Cn} \quad \text{and} \quad P(\mathcal{B}) \leq 2e^{-Cp}. \quad (51)$$

Therefore, by $p \geq n$ and Bonferroni's inequality, the results follow from (50) and (51). \square

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

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

$$P(|V_2| > c\sqrt{np}^{-1}|W|) \leq 3e^{-Cn}, \quad (52)$$

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

Proof. In view of (45), it follows that

$$\|bV\|^2 = \mathbf{e}_1^T \mathbf{S} \mathbf{e}_1 = V_1,$$

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

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

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

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

$$V_2 \stackrel{(d)}{=} \sqrt{V_1 - V_1^2} \frac{W_1}{\sqrt{W_1^2 + \dots + W_{p-1}^2}}. \quad (53)$$

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

$$P(V_1 > c_2 n/p) \leq 2e^{-Cn}. \quad (54)$$

It follows from (47) that there exists some $0 < c_1 < 1$ such that

$$P(W_1^2 + \dots + W_{p-1}^2 < c_1(p-1)) \leq e^{-C(p-1)} \leq e^{-Cn}, \quad (55)$$

since $p > n$. Let $c = \sqrt{c_2/c_1}$. Then, by $V_1 - V_1^2 \leq V_1$ and Bonferroni's inequality, (52) follows immediately from (53)–(55). \square

A.7 Verifying Property C for Gaussian distributions

In this section, we check Property C in (15) for Gaussian distributions. Assume \mathbf{x} has a p -variate Gaussian distribution. Then, the $n \times p$ design matrix $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, I_n \otimes \Sigma)$ 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 i.i.d. $\mathcal{N}(0, 1)$ random variables, where the symbol \otimes denotes the Kronecker product of two matrices. We will 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 C in (15).

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

$$P\left(\left|\sqrt{\lambda} - E(\sqrt{\lambda})\right| > r\right) \leq Ce^{-pr^2/C}.$$

Moreover, for each λ , the same inequality holds for a median of $\sqrt{\lambda}$ 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). \square

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

$$\lim_{n \rightarrow \infty} \text{median} \left(\sqrt{\lambda_{\max}(p^{-1}\mathbf{Z}\mathbf{Z}^T)} \right) = 1 + \gamma^{-1/2}$$

and

$$\liminf_{n \rightarrow \infty} E \left(\sqrt{\lambda_{\min}(p^{-1}\mathbf{Z}\mathbf{Z}^T)} \right) \geq 1 - \gamma^{-1/2}.$$

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

$$\lambda_{\max}(p^{-1}\mathbf{Z}\mathbf{Z}^T) \xrightarrow{\text{a.s.}} \left(1 + \gamma^{-1/2}\right)^2 \quad \text{as } n \rightarrow \infty.$$

For the smallest eigenvalue, it is well known that (see, e.g., Silverstein, 1985 or Bai, 1999)

$$\lambda_{\min}(p^{-1}\mathbf{Z}\mathbf{Z}^T) \xrightarrow{\text{a.s.}} \left(1 - \gamma^{-1/2}\right)^2 \quad \text{as } n \rightarrow \infty.$$

This and Fatou's lemma entails the second result. \square

References

- [1] Antoniadis, A. and Fan, J. (2001). Regularization of wavelets approximations (with discussion). *J. Amer. Statist. Assoc.* **96**, 939–967.
- [2] Bai, Z.D. (1999). Methodologies in spectral analysis of large dimensional random matrices, A review. *Statistica Sinica* **9**, 611–677.
- [3] Bai, Z.D. and Yin, Y.Q. (1993). Limit of smallest eigenvalue of a large dimensional sample covariance matrix. *Ann. Prob.* **21**, 1275–1294.
- [4] Baron, D., Wakin, M. B., Duarte, M. F., Sarvotham, S. and Baraniuk, R. G. (2005). Distributed compressed sensing. *Manuscript*.
- [5] 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.
- [6] Bickel, P. J. and Levina, E. (2007). Regularized estimation of large covariance matrices. *Ann. Statist.*, to appear.

- [7] Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics* **37**, 373–384.
- [8] Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *Ann. Statist.* **24**, 2350–2383.
- [9] Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n . *Ann. Statist.*, to appear.
- [10] Chikuse, Y. (2003). *Statistics on Special Manifolds*. Lecture Notes in Statistics. Springer-Verlag, Berlin.
- [11] Donoho, D. L. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. Aide-Memoire of a Lecture at AMS Conference on Math Challenges of the 21st Century.
- [12] Donoho, D. L. and Elad, M. (2003). Maximal sparsity representation via l_1 minimization. *Proc. Nat. Aca. Sci.* **100**, 2197–2202.
- [13] Donoho, D. L. and Huo, X. (2001). Uncertainty principles and ideal atomic decomposition. *IEEE Trans. Inform. Theory* **47**, 2845–2862.
- [14] Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425–455.
- [15] Eaton, M. L. (1989). *Group Invariance Applications in Statistics*. Institute of Mathematical Statistics, Hayward, California.
- [16] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression (with discussion). *Ann. Statist.* **32**, 407–499.
- [17] Fan, J. (1997). Comments on “Wavelets in statistics: A review,” by A. Antoniadis. *J. Italian Statist. Assoc.* **6**, 131–138.
- [18] Fan, J. and Fan, Y. (2007). High dimensional classification using features annealed independence rules. *Ann. Statist.*, to appear.
- [19] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348–1360.
- [20] Fan, J. and Li, R. (2002). Variable selection for Cox’s proportional hazards model and frailty model. *Ann. Statist.* **30**, 74–99.
- [21] Fan, J. and Li, R. (2006). Statistical challenges with high dimensionality: feature selection in knowledge discovery. *Proceedings of the International Congress of Mathematicians* (M. Sanz-Sole, J. Soria, J.L. Varona, J. Verdera, eds.) Vol. III, 595–622.
- [22] Fan, J and Ren, Y. (2006). Statistical analysis of DNA microarray data. *Clinical Cancer Research* **12**, 4469–4473.

- [23] Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with diverging number of parameters. *Ann. Statist.* **32**, 928–961.
- [24] Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools (with discussion). *Technometrics* **35**, 109–148.
- [25] Geman, S. (1980). A limit theorem for the norm of random matrices. *Ann. Probab.* **8**, 252–261.
- [26] George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica* **7**, 339–373.
- [27] 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.
- [28] Greenshtein, E. (2006). Best subset selection, persistence in high dimensional statistical learning and optimization under l_1 constraint. *Ann. Statist.* **34**, 2367–2386.
- [29] Greenshtein, E. and Ritov, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli* **10**, 971–988.
- [30] Grenander, U. and Szegö, G. (1984). *Toeplitz Forms and Their Applications*. Chelsea, New York.
- [31] Gribonval, R., Mailhe, B., Rauhut, H., Schnass, K. and Vandergheynst, P. (2007). Average case analysis of multichannel thresholding. In *Proc. ICASSP*.
- [32] Hall, P., Marron, J. S. and Neeman, A. (2005). Geometric representation of high dimension, low sample size data. *J. Roy. Statist. Soc. Ser. B* **67**, 427–444.
- [33] Huang, J., Horowitz, J. and Ma, S. (2007). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann. Statist.*, to appear.
- [34] Hunter, D. and Li, R. (2005). Variable selection using MM algorithms. *Ann. Statist.* **33**, 1617–1642.
- [35] Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29**, 295–327.
- [36] Knight, K. and Fu, W. (2000). Asymptotics for Lasso-type estimators. *Ann. Statist.* **28**, 1356–1378.
- [37] Ledoux, M. (2001). *The Concentration of Measure Phenomenon*. Mathematical Surveys and Monographs **89**, AMS.
- [38] Ledoux, M. (2005). *Deviation Inequalities on Largest Eigenvalues*. GAFA Seminar Notes, to appear.
- [39] Meinshausen, N. (2006). Relaxed Lasso. *Computational Statistics and Data Analysis*, to appear.

- [40] Meinshausen, N. and Bühlmann, P. (2006). High dimensional graphs and variable selection with the Lasso. *Ann. Statist.* **34**, 1436–1462.
- [41] Nikolova, M. (2000). Local strong homogeneity of a regularized estimator. *SIAM J. Appl. Math.* **61**, 633–658.
- [42] Paul, D., Bair, E., Hastie, T., Tibshirani, R. (2007). “Pre-conditioning” for feature selection and regression in high-dimensional problems. *Ann. Statist.*, to appear.
- [43] Silverstein, J. W. (1985). The smallest eigenvalue of a large dimensional Wishart matrix. *Ann. Prob.* **13**, 1364–1368.
- [44] Storey, J. D. and Tibshirani R. (2003). Statistical significance for genome-wide studies. *Proc. Natl. Aca. Sci.* **100**, 9440–9445.
- [45] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 267–288.
- [46] Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci.* **99**, 6567–6572.
- [47] van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer-Verlag, New York.
- [48] Zhang, C.-H. (2007). Penalized linear unbiased selection. *Manuscript*.
- [49] Zhang, C.-H. and Huang, J. (2007). The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Ann. Statist.*, to appear.
- [50] Zhao, P. and Yu, B. (2006). On Model Selection Consistency of Lasso. *J. Machine Learning Res.* **7**, 2541–2567.
- [51] Zou, H. (2006). The adaptive Lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 1418–1429.
- [52] Zou, H. and Li, R. (2006). One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.*, to appear.