Journal of the
Royal Statistical Society

SERIES B
Statistical
Methodology

J. R. Statist. Soc. B (2012)
74, *Part* 1, *pp.* 37–65

Variance estimation using refitted cross-validation in ultrahigh dimensional regression

Jianqing Fan,

Princeton University, USA

Shaojun Guo

Chinese Academy of Sciences, Beijing, People's Republic of China

and Ning Hao

University of Arizona, Tucson, USA

[Received April 2010. Revised April 2011]

Summary. Variance estimation is a fundamental problem in statistical modelling. In ultrahigh dimensional linear regression where the dimensionality is much larger than the sample size, traditional variance estimation techniques are not applicable. Recent advances in variable selection in ultrahigh dimensional linear regression make this problem accessible. One of the major problems in ultrahigh dimensional regression is the high spurious correlation between the unobserved realized noise and some of the predictors. As a result, the realized noises are actually predicted when extra irrelevant variables are selected, leading to a serious underestimate of the level of noise. We propose a two-stage refitted procedure via a data splitting technique, called refitted cross-validation, to attenuate the influence of irrelevant variables with high spurious correlations. Our asymptotic results show that the resulting procedure performs as well as the oracle estimator, which knows in advance the mean regression function. The simulation studies lend further support to our theoretical claims. The naive two-stage estimator and the plug-in one-stage estimators using the lasso and smoothly clipped absolute deviation are also studied and compared. Their performances can be improved by the refitted cross-validation method proposed.

Keywords: Data splitting; Dimension reduction; High dimensionality; Refitted cross-validation; Sure screening; Variable selection; Variance estimation

1. Introduction

Variance estimation is a fundamental problem in statistical modelling. It is prominently featured in the statistical inference on regression coefficients. It is also important for variable selection criteria such as Akaike's information criterion AIC and the Bayesian information criterion BIC. It provides also a benchmark of forecasting error when an oracle actually knows the regression function and such a benchmark is very important for forecasters to gauge their forecasting performance relative to the oracle. For conventional linear models, the residual variance estimator usually performs well and plays an important role in the inferences after model selection and estimation. However, the ordinary least squares methods do not work for many contemporary data sets which have a greater number of covariates than the sample size. For example, in

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

1369–7412/12/74037

disease classification using microarray data, the number of arrays is usually in tens, yet tens of thousands of gene expressions are potential predictors. When interactions are considered, the dimensionality grows even more quickly; for example considering possible interactions among thousands of genes or single-nucleotide polymorphisms yields a number of parameters in the order of millions. In this paper, we propose and compare several methods for variance estimation in the setting of an ultrahigh dimensional linear model. A key assumption which makes the high dimensional problems solvable is the sparsity condition: the number of non-zero components is small compared with the sample size. With sparsity, variable selection can identify the subset of important predictors and improve the model interpretability and predictability.

Recently, there have been several important advances in model selection and estimation for ultrahigh dimensional problems. The properties of penalized likelihood methods such as the lasso and smoothly clipped absolute deviation (SCAD) have been extensively studied in high and ultrahigh dimensional regression. Various useful results have been obtained. See, for example, Fan and Peng (2004), Zhao and Yu (2006), Bunea *et al.* (2007), Zhang and Huang (2008), Meinshausen and Yu (2009), Kim *et al.* (2008), Meier *et al.* (2008), Lv and Fan (2009) and Fan and Lv (2011). Another important model selection tool is the Dantzig selector that was proposed by Candes and Tao (2007), which can be easily recast as a linear program. It is closely related to the lasso, as demonstrated by Bickel *et al.* (2009). Fan and Lv (2008) showed that correlation ranking has a sure screening property in the Gaussian linear model with Gaussian covariates and proposed the sure independent screening (SIS) and iteratively sure independent screening (ISIS) methods. Fan *et al.* (2009) extended ISIS to a general pseudolikelihood framework, which includes generalized linear models as a special case. Fan and Song (2010) have developed general conditions under which the marginal regression has a sure screening property in the context of generalized linear models. For an overview, see Fan and Lv (2010).

In all the work mentioned above, the primary focus is the consistency of model selection and parameter estimation. The problem of variance estimation in ultrahigh dimensional settings has hardly been touched. A natural approach to estimate the variance is the following two-stage procedure. In the first stage, a model selection tool is applied to select a model which, if is not exactly the true model, includes all important variables with moderate model size (smaller than the sample size). In the terminology of Fan and Lv (2008), the model selected has a sure screening property. In the second stage, the variance is estimated by an ordinary least squares method based on the variables selected in the first stage. Obviously, this method works well if we can recover exactly the true model in the first stage. This is usually difficult to achieve in ultrahigh dimensional problems. Yet, sure screening properties are much easier to obtain. Unfortunately, this naive two-step approach can seriously underestimate the level of noise even with the sure screening property in the first stage owing to spurious correlation that is inherent in ultrahigh dimensional problems. When the number of irrelevant variables is huge, some of these variables have large sample correlations with the realized noises. Hence, almost all variable selection procedures will, with high probability, select those spurious variables in the model when the model is overfitted, and the realized noises are actually predicted by several spurious variables, leading to a serious underestimate of the residual variance.

The above phenomenon can be easily illustrated in the simplest model, in which the true coefficient $\beta = 0$. Suppose that one extra variable is selected by a method such as the lasso or SIS in the first stage. Then, the ordinary least squares estimator $\hat{\sigma}_n^2$ is

$$\hat{\sigma}_n^2 = (1-\gamma_n^2)\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2, \tag{1}$$

where γ_n is the sample correlation of the spurious variable and the response, which is really the realized noise in this null model. Most variable selection procedures such as stepwise addition, SIS and the lasso will first select the covariate that has the highest sample correlation with the response, namely $\gamma_n = \max_{j \leqslant p} |\widehat{\mathrm{corr}}_n(X_j, Y)|$. In other words, this extra variable is selected to predict the realized noise vector best. However, as Fan and Lv (2008) stated, the maximum absolute sample correlation γ_n can be very large, which makes $\hat{\sigma}_n^2$ seriously biased. To illustrate the point, we simulated 500 data sets with sample size $n=50$ and the number of covariates $p=10, 100, 1000, 5000$, with $\{X_j\}_{j=1}^p$ and noise independent and identically distributed (IID) from the standard normal distribution. Fig. 1(a) presents the densities of γ_n across the 500 simulations and Fig. 1(b) depicts the densities of the estimator $\hat{\sigma}_n^2$ defined in equation (1). Clearly, the biases of $\hat{\sigma}_n^2$ become larger as p increases.

The bias becomes larger when more spurious variables are recruited to the model. To illustrate the point, let us use stepwise addition to recruit s variables to the model. Clearly, the realized noises are now better predicted, leading to an even more severe underestimate of the level of noise. Fig. 2 depicts the distributions of spurious multiple correlation with the response (realized noise) and the corresponding naive two-stage estimator of variance for $s=1, 2, 5, 10$, keeping $p=1000$ fixed. Clearly, the biases become much larger with s. For comparison, we also depict

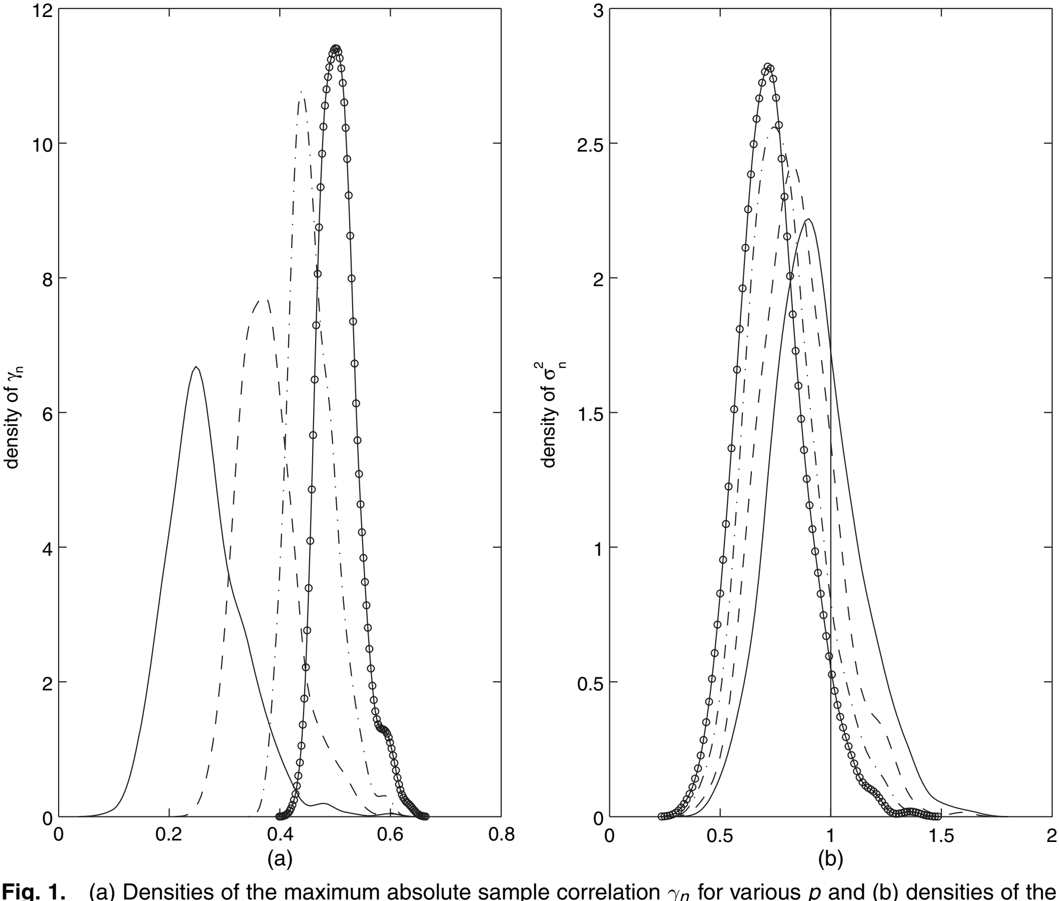


Fig. 1. (a) Densities of the maximum absolute sample correlation γ_n for various p and (b) densities of the corresponding estimates $\hat{\sigma}_n^2$ given by equation (1) (all calculations are based on 500 simulations and the sample size n is 50): |, true variance 1; ———, $p=10$; – – –, $p=100$; · - · - ·, $p=1000$; —o—, $p=5000$

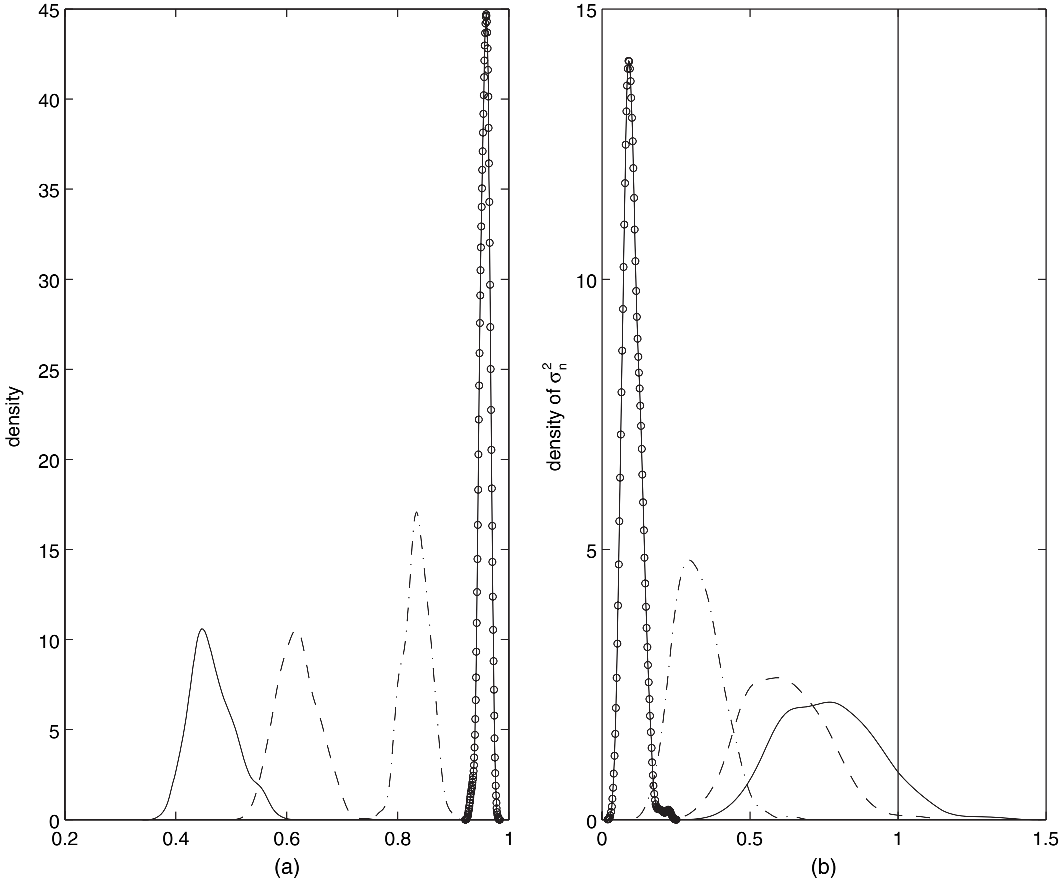


Fig. 2. (a) Densities of spurious multiple correlation with the response for various numbers of spurious variables s and (b) densities of the naive two-stage estimators of variance (all calculations are based on the stepwise addition algorithm with 500 simulations, $n=50$ and $p=1000$): |, true variance 1; ———, $s=1$; – – –, $s=2$; · - · - ·, $s=5$; —o—, $s=10$

similar distributions based on SIS, which selects s variables that are marginally most correlated with the response variable. The results are depicted in Fig. 3(a). Although the biases based on the SIS method are still large, they are smaller than those based on the stepwise addition method, as the latter chose the co-ordinated spurious variables to optimize the prediction of the realized noise.

A similar phenomenon was also observed in classical model selection by Ye (1998). To correct the effects of model selection, Ye (1998) developed the concept of a generalized degree of freedom but it is computationally intensive and can only be applied to some special cases.

To attenuate the influence of spurious variables that are entered into the selected model and to improve the accuracy of estimation, we introduce a refitted cross-validation (RCV) technique. Roughly speaking, we split the data randomly into two halves, do model selection by using the first half of the data set and refit the model on the basis of the variables selected in the first stage, using the second half of the data to estimate the variance, and vice versa. The estimator proposed is just the average of these two estimators. The results of the RCV variance estimators with $s=1,2,5,10$ are presented in Fig. 3(b). The corrections of biases due to spurious correlation are dramatic. The essential difference between this approach and the naive two-stage approach

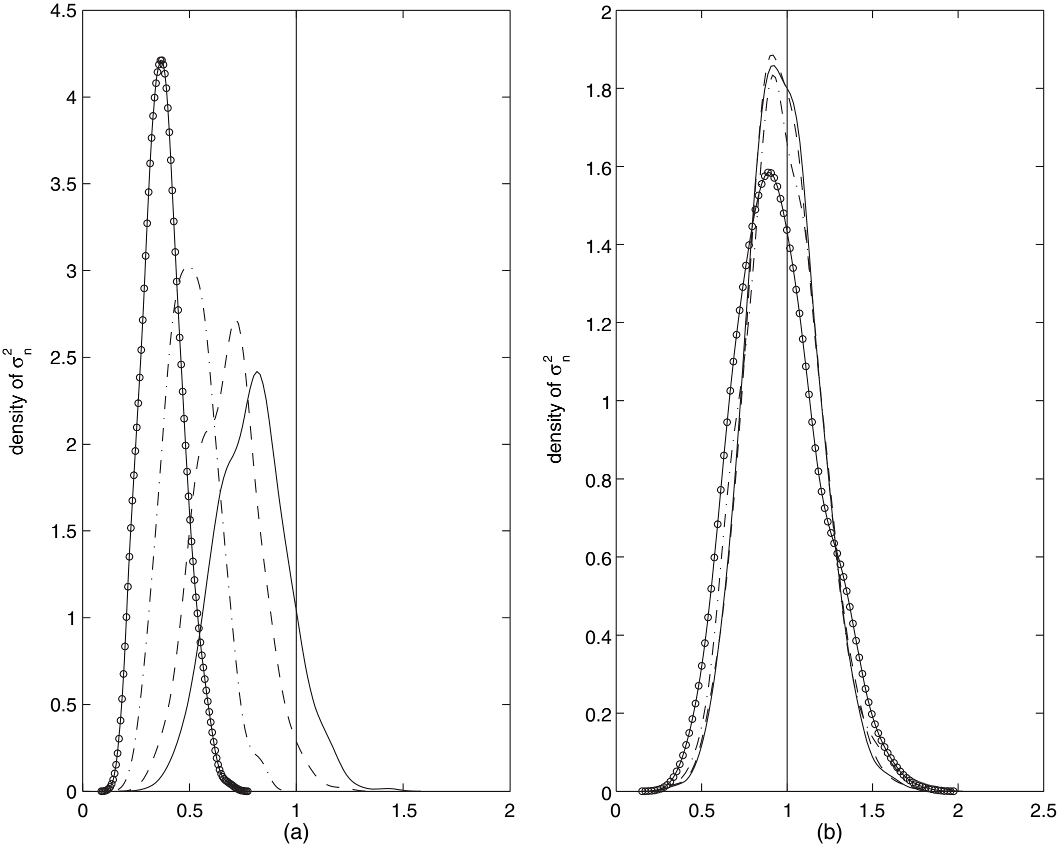


Fig. 3. (a) Densities of the variance estimators based on the naive two-stage approach for various numbers of spurious variables and (b) densities of RCV estimators of variance (all calculations are based on 500 simulations using SIS as a model selector and the sample size n is 50; they show that the biases of the naive two-stage estimator are correctable): |, true variance 1; ———, $s=1$; – – –, $s=2$; · - · - ·, $s=5$; —o—, $s=10$

is that the regression coefficients in the first stage are discarded and refitted by using the second half of the data and hence the spurious correlations in the first stage are significantly reduced at the second stage. The variance estimation is unbiased as long as the models selected in the first stage contain all relevant variables, namely they have a sure screening property. It turns out that this simple RCV method improves dramatically the performance of the naive two-stage procedure. Clearly, the RCV can also be used to do model selection itself, reducing the influence of spurious variables.

To appreciate why, suppose that a predictor has a big sample correlation with the response (realized noise in the null model) over the first half of the data set and is selected into the model by a model selection procedure. Since the two halves of the data set are independent and the chance that a given predictor is highly correlated with realized noise is small, it is very unlikely that this predictor has a large sample correlation with the realized noise over the second half of the data set. Hence, its influence on the variance estimation is very small when refitted and estimating the variance over the second half will not cause any bias. This argument is also true for the non-null models provided that the model selected includes all important variables.

To gain better understanding of the RCV approach, we compare our method with the direct plug-in method, which computes the residual variance based on a regularized fit. This was inspired by Greenshtein and Ritov (2004) on the persistence of the lasso estimator. An interpretation of their results is that such an estimator is consistent. However, a bias term of order $O\{s\log(p)/n\}$ is inherent in the lasso-based estimator, when the regularization parameter is optimally tuned. When the bias is negligible, the lasso-based plug-in estimator is consistent. The plug-in variance estimation based on the general folded concave penalized least squares estimators such as SCAD is also discussed. In some cases, this method is comparable with the RCV approach.

The paper is organized as follows. Section 2 gives some additional insights into the challenges of high dimensionality in variance estimation. In Section 3, the RCV variance estimator is proposed and its sampling properties are established. Section 4 studies the variance-estimation-based penalized likelihood methods. Extensive simulation studies are conducted in Section 5 to illustrate the advantage of the methodology proposed. Section 6 is devoted to a discussion and the detailed proofs are provided in Appendix A.

2. Insights into challenges of high dimensionality in variance estimation

Consider the usual linear model

$$Y_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_i, \qquad \text{or } \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2}$$

where $\mathbf{y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ is an n-vector of responses, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^{\mathrm{T}}$ is an $n \times p$ matrix of IID variables $\mathbf{x}_1, \ldots, \mathbf{x}_n$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is a p-vector of parameters and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{T}}$ is an n-vector of IID random noises with mean 0 and variance σ^2. We always assume that the noise is independent of predictors. For any index set $M \subset \{1, 2, \ldots, p\}$, $\boldsymbol{\beta}_M$ denotes the subvector containing the components of the vector $\boldsymbol{\beta}$ that are indexed by M, $\mathbf{X}_M$ denotes the submatrix containing the columns of $\mathbf{X}$ that are indexed by M and $\mathbf{P}_M = \mathbf{X}_M(\mathbf{X}_M^{\mathrm{T}}\mathbf{X}_M)^{-1}\mathbf{X}_M^{\mathrm{T}}$ is the projection operator onto the linear space that is generated by the column vectors of $\mathbf{X}_M$.

When $p > n$ or $p \gg n$, it is often assumed that the true model $M_0 = \{j : \beta_j \neq 0\}$ is sparse, i.e. the number of non-zero coefficients $s = |M_0|$ is small. It is usually assumed that s is fixed or diverging at a mild rate. Under various sparsity assumptions and regularity conditions, the most popular variable selection tools such as the lasso, SCAD, adaptive lasso, SIS and Dantzig selector have various good properties regarding model selection consistency. Among these properties are the sure screening property, model consistency, sign consistency, the weak oracle property and the oracle property, from weak to strong. Theoretically, under some regularity conditions, all the aforementioned model selection tools can achieve model consistency. In other words, they can exactly pick out the true sparse model with probability tending to 1. However, in practice, these conditions are impossible to check and difficult to meet. Hence, it is often very difficult to extract the exact subset of significant variables among a huge set of covariates. One of the reasons is the spurious correlation, as we now illustrate.

Suppose that unknown to us the true data-generating process in model (2) is

$$\mathbf{Y} = 2\mathbf{X}_1 + 0.3\mathbf{X}_2 + \boldsymbol{\varepsilon}$$

where $\mathbf{X}_j$ is the n-dimensional vector of the realizations of the covariate X_j. Furthermore, let us assume that $\{X_j\}_{j=1}^{p}$ and ε follow independently the standard normal distribution. As illustrated in Fig. 1(a), where p is large, there are realizations of variables that have high correlations with ε. Let us say $\widehat{\mathrm{corr}}(\mathbf{X}_9, \boldsymbol{\varepsilon}) = 0.5$. Then, X_9 can even have a better chance of being selected than X_2. Here and hereafter, we refer the spurious variables to those variables that are selected to predict the realized noise $\boldsymbol{\varepsilon}$ and their associated sample correlations are called spurious correlations.

Continuing with the above example, the naive two-stage estimator will work well when the model selection is consistent. Since we may not obtain model consistency in practice and have no way to check even if we obtain it by chance, it is natural to ask whether the naive two-stage strategy works if only sure screening can be achieved in the first stage. In the aforementioned example, let us say that a model selector chooses the set $\{X_1, X_2, X_9\}$, which contains all true variables. However, in the naive two-stage fitting, $\mathbf{X}_9$ is used to predict ε, resulting in a substantial underestimate of $\sigma^2 = \text{var}(\varepsilon)$. If both variables X_1 and X_2 are selected, all spurious variables are recruited to predict ε. The more spurious variables are selected, the better ε is predicted, and the more serious underestimation of σ^2 by the naive two-stage estimation.

We say that a model selection procedure satisfies the sure screening property if the selected model $\hat{M}$ with model size $\hat{s}$ includes the true model M_0 with probability tending to 1. Explicitly,

$$P(\hat{M} \supset M_0) \to 1 \qquad \text{as } n \to \infty.$$

The sure screening property is a crucial criterion when evaluating a model selection procedure for high or ultrahigh dimensional problems. Among all model consistent properties, the sure screening property is the weakest and the easiest to achieve in practice.

We demonstrate the naive two-stage procedure in detail. Assume that the selected model $\hat{M}$ in the first stage includes the true model M_0. The ordinary least squares estimator $\hat{\sigma}^2_{\hat{M}}$ at the second stage, using only the selected variables in $\hat{M}$, is

$$\hat{\sigma}^2_{\hat{M}} = \frac{\mathbf{y}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{P}_{\hat{M}})\mathbf{y}}{n - \hat{s}} = \frac{\varepsilon^{\mathrm{T}}(\mathbf{I}_n - \mathbf{P}_{\hat{M}})\varepsilon}{n - \hat{s}}, \tag{3}$$

where $\mathbf{I}_n$ is the $n \times n$ identity matrix. How does this estimator perform? To facilitate the notation, denote the naive estimator by $\hat{\sigma}^2_n$. Then, estimator (3) can be written as

$$\hat{\sigma}^2_n = \frac{1}{n - \hat{s}}(1 - \hat{\gamma}^2_n)\varepsilon^{\mathrm{T}}\varepsilon,$$

where $\hat{\gamma}^2_n = \varepsilon^{\mathrm{T}}\mathbf{P}_{\hat{M}}\varepsilon / \varepsilon^{\mathrm{T}}\varepsilon$. Let us analyse the asymptotic behaviour of this naive two-stage estimator.

Theorem 1. Under assumptions 1 and 2 together with 3 and 4 or 5 and 6 in Appendix A, we have the following results.

(a) If a procedure satisfies the sure screening property with $\hat{s} \leqslant b_n$ where $b_n = o(n)$ is given in assumption 2, then $\sigma^2_n/(1 - \hat{\gamma}^2_n)$ converges to σ^2 in probability as $n \to \infty$. Furthermore,

$$\{\hat{\sigma}^2_n/(1 - \hat{\gamma}^2_n) - \sigma^2\}\sqrt{n} \xrightarrow{\mathcal{D}} N(0, E[\varepsilon_1^4] - \sigma^4),$$

where '$\to^{\mathcal{D}}$' stands for 'convergence in distribution'.

(b) If, in addition, $\log(p)/n = O(1)$, then $\hat{\gamma}_n = O_P[\sqrt{\{\hat{s} \log(p)/n\}}]$.

It is perhaps worthwhile to make a remark about theorem 1. $\hat{\gamma}^2_n$ plays an important role in the performance of $\hat{\sigma}^2_n$. It represents the fraction of bias in $\hat{\sigma}^2_n$. The slower $\hat{\gamma}_n$ converges to 0, the worse $\hat{\sigma}^2_n$ performs. Moreover, if $\hat{\gamma}^2_n$ converges to a positive constant with a non-negligible probability, it will lead to an inconsistent estimator. The estimator cannot be root n consistent if $\hat{s} \log(p)/\sqrt{n} \to \infty$. This explains the poor performance of $\hat{\sigma}^2_n$, as demonstrated in Figs 2 and 3. Although theorem 1 gives an upper bound of γ_n, it is often sharp. For instance, if $\{X_j\}_{j=1}^p$ and ε are IID standard normal distributions and $\hat{s} = 1$, then $\hat{\gamma}_n$ is just the maximum absolute sample correlation between ε and $\{X_j\}_{j=1}^p$. Denote the jth sample correlation by $\hat{\gamma}_{nj} = \widehat{\text{corr}}_n(X_j, \varepsilon)$, $j = 1, \ldots, p$. Applying the transformation $T(r) = r/\sqrt{(1 - r^2)}$, we obtain a sequence $\{\xi_{nj}=\surd(n-2)\ T(\hat{\gamma}_{nj})\}_{j=1}^{p}$ with IID Student t-distribution with $n-2$ degrees of freedom. Simple analysis on the extreme statistics of the sequences $\{\xi_{nj}\}$ and $\{\hat{\gamma}_{nj}\}$ shows that, for any $c>0$ such that $\log(p/c)\leqslant n+2$, we have

$$P\left[\hat{\gamma}_n>\sqrt{\left\{\frac{\log(p/c)}{2n}\right\}}\right]>1-\exp(-c), \tag{4}$$

which implies the sharpness of theorem 1 in this specific case. Furthermore, when $\log(p)=o(n^{1/2})$,

$$\hat{\gamma}_n=\surd\{2\log(p)/n\}\{1+o_p(1)\}$$

with the limiting distribution is given by

$$P[\surd\{2\log(2p)\}(\hat{\gamma}_n\surd n-d_{2p})<x]\rightarrow\exp\{-\exp(-x)\}. \tag{5}$$

where

$$d_p=\frac{\surd\{2\log(p)\}-\log\surd\{4\pi\log(p)\}}{\surd\{2\log(p)\}}.$$

See Appendix A.5 for details.

3. Variance estimation based on refitted cross-validation

3.1. Refitted cross-validation

In this section, we introduce the RCV method to remove the influence of spurious variables in the second stage. The method requires only that the model selection procedure in stage 1 has a sure screening property. The idea is as follows. We assume that the sample size n is even for simplicity and split randomly the sample into two groups. In the first stage, an ultrahigh dimensional variable selection method like SIS is applied to these two data sets separately, which yields two small sets of selected variables. In the second stage, the ordinary least squares method is used to re-estimate the coefficient $\boldsymbol{\beta}$ and variance σ^2. Differently from the naive two-stage method, we apply ordinary least squares again to the *first subset* of the data with the variables selected by the *second subset* of the data and vice versa. Taking the average of these two estimators, we obtain our estimator of σ^2. The refitting in the second stage is fundamental to reduce the influence of the spurious variables in the first stage of variable selection.

To implement this idea of RCV, consider a data set with sample size n, which is randomly split into two even data sets $(\mathbf{y}^{(1)},\mathbf{X}^{(1)})$ and $(\mathbf{y}^{(2)},\mathbf{X}^{(2)})$. First, a variable selection tool is performed on $(\mathbf{y}^{(1)},\mathbf{X}^{(1)})$ and let $\hat{M}_1$ denote the set of variables selected. The variance σ^2 is then estimated on the second data set $(\mathbf{y}^{(2)},\mathbf{X}^{(2)}_{\hat{M}_1})$, namely

$$\hat{\sigma}_1^2=\frac{\mathbf{y}^{(2)\mathrm{T}}(\mathbf{I}_{n/2}-\mathbf{P}^{(2)}_{\hat{M}_1})\mathbf{y}^{(2)}}{n/2-|\hat{M}_1|},$$

where $\mathbf{P}^{(2)}_{\hat{M}_1}=\mathbf{X}^{(2)}_{\hat{M}_1}(\mathbf{X}^{(2)\mathrm{T}}_{\hat{M}_1}\mathbf{X}^{(2)}_{\hat{M}_1})^{-1}\mathbf{X}^{(2)\mathrm{T}}_{\hat{M}_1}$. Similarly, we use the second data set $(\mathbf{y}^{(2)},\mathbf{X}^{(2)})$ to select the set of important variables $\hat{M}_2$ and the first data set $(\mathbf{y}^{(1)},\mathbf{X}^{(1)}_{\hat{M}_2})$ for estimation of σ^2, resulting in

$$\hat{\sigma}_2^2=\frac{\mathbf{y}^{(1)\mathrm{T}}(\mathbf{I}_{n/2}-\mathbf{P}^{(1)}_{\hat{M}_2})\mathbf{y}^{(1)}}{n/2-|\hat{M}_2|}.$$

We define the final estimator as

$$\hat{\sigma}^2_{\text{RCV}} = (\hat{\sigma}_1^2 + \hat{\sigma}_2^2)/2. \tag{6}$$

An alternative is the weighted average defined by

$$\hat{\sigma}^2_{\text{WRCV}} = \frac{\mathbf{y}^{(2)\text{T}}(\mathbf{I}_{n/2} - \mathbf{P}^{(2)}_{\hat{M}_1})\mathbf{y}^{(2)} + \mathbf{y}^{(1)\text{T}}(\mathbf{I}_{n/2} - \mathbf{P}^{(1)}_{\hat{M}_2})\mathbf{y}^{(1)}}{n - |\hat{M}_1| - |\hat{M}_2|}. \tag{7}$$

When $|\hat{M}_1| = |\hat{M}_2|$, we have $\hat{\sigma}^2_{\text{RCV}} = \hat{\sigma}^2_{\text{WRCV}}$.

In this procedure, although $\hat{M}_1$ includes some extra unimportant variables besides the important variables, these extra variables will play minor roles when we estimate σ^2 by using the second data set along with refitting since they are just some random unrelated variables over the second data set. Furthermore, even when some important variables are missed in the first stage of model selection, they have a good chance of being well approximated by the other variables selected in the first stage to reduce modelling biases. Thanks to the refitting in the second stage, the best linear approximation of those selected variables is used to reduce the biases. Therefore, a larger selected model size gives us, not only a better chance of sure screening, but also a way to reduce modelling biases in the second stage when some important variables are missing. This explains why the RCV method is relatively insensitive to the model size selected, demonstrated in Fig. 3 and in Fig. 6 in Section 5.1. With a larger model being selected in stage 1, we may lose some degrees of freedom and hence obtain an estimator with slightly larger variance than the oracle estimator at finite sample. Nevertheless, the RCV estimator performs well in practice and is asymptotically optimal when $\hat{s} = o(n)$. The following theorem gives the property of the RCV estimator. It requires only a sure screening property, which was studied by Fan and Lv (2008) for normal multiple regression, Fan and Song (2010) for generalized linear models and Zhao and Li (2010) for the Cox regression model.

Theorem 2. Assume that regularity conditions 1 and 2 in Appendix A hold and $E[\varepsilon^4] < \infty$. If a procedure satisfies the sure screening property with $\hat{s}_1 \leqslant b_n$ and $\hat{s}_2 \leqslant b_n$, then

$$(\hat{\sigma}^2_{\text{RCV}} - \sigma^2)\sqrt{n} \xrightarrow{\mathcal{D}} N(0, E[\varepsilon^4] - \sigma^4). \tag{8}$$

Theorem 2 reveals that the RCV estimator of variance has an oracle property. If the regression coefficient $\boldsymbol{\beta}^*$ is known by oracle, then we can compute the realized noise $\varepsilon_i = Y_i - \mathbf{x}_i^{\text{T}}\boldsymbol{\beta}^*$ and obtain the oracle estimator

$$\hat{\sigma}^2_{\text{O}} = n^{-1} \sum_{i=1}^{n} (Y_i - \mathbf{x}_i^{\text{T}}\boldsymbol{\beta}^*)^2. \tag{9}$$

This oracle estimator has the same asymptotic variance as $\hat{\sigma}^2_{\text{RCV}}$.

There are two natural extensions of the aforementioned RCV techniques.

(a) *K-fold data splitting*: the first natural extension is to use a K-fold data splitting technique rather than twofold splitting. We can divide the data into K groups and select the model with all groups except one, which is used to estimate the variance with refitting. We may improve the sure screening probability with this K-fold method since there are now more data in the first stage. However, there are only n/K data points in the second stage for refitting. This means that the number of variables that are selected in the first stage should be much less than n/K. This makes the ability of sure screening difficult in the first stage. For this reason, we work only on the twofold RCV.

(b) *Repeated data splitting*: there are many ways to split the data randomly. Hence, many RCV variance estimators can be obtained. We may take the average of the resulting estimators. This reduces the influence of the randomness in the data splitting.

Remark 1. The RCV procedure provides an efficient method for variance estimation. The technical conditions in theorem 2 may not be the weakest possible. They are imposed to facilitate the proofs. In particular, we assume that $P\{\phi_{\min}(b_n) \geqslant \lambda_0\} = 1$ for all n, which implies that the variables selected in stage 1 are not highly correlated. Other methods beyond least squares can be applied in the refitted stage when those assumptions are possibly violated in practice. For instance, if some selected variables in stage 1 are highly correlated or the selected model size is relatively large, ridge regression or penalization methods can be applied in the refitted stage. Moreover, if the density of the error ε seems heavy tailed, some classical robust methods can also be employed.

Remark 2. The paper focuses on variance estimation under the exact sparsity assumption and sure screening property. It is possible to extend our results to nearly sparse cases. For example, the parameter $\boldsymbol{\beta}$ is not sparse but satisfies some decay condition such as $\Sigma_k |\boldsymbol{\beta}_i| \leqslant C$ for some positive constant C. In this case, we do not have to worry too much whether a model selection procedure can recover small parameters. In this case, so long as a model selection method can pick up a majority of all variables with large coefficients in the first stage, we would expect that the RCV estimator performs well.

3.2. Applications

Many statistical problems require knowledge of the residual variance, especially for high or ultrahigh dimensional linear regression. Here we briefly outline a couple of applications.

(a) *Constructing confidence intervals for coefficients*: a natural application is to use estimated $\hat{\sigma}_{\mathrm{RCV}}$ to construct confidence intervals for non-vanishing estimated coefficients. For example, it is well known that the SCAD estimator has an oracle property (Fan and Li, 2001; Fan and Lv, 2011). Let $\hat{\boldsymbol{\beta}}_{\hat{M}}$ be the SCAD estimator, with corresponding design matrix $\mathbf{X}_{\hat{M}}$. Then, for each $j \in \hat{M}$, the $1-\alpha$ confidence interval for β_j is

$$\hat{\beta}_j \pm z_{1-\alpha/2} c_j \hat{\sigma}_{\mathrm{RCV}}, \tag{10}$$

in which c_j is the diagonal element of the matrix $(\mathbf{X}_{\hat{M}}^{\mathrm{T}} \mathbf{X}_{\hat{M}})^{-1}$ that corresponds to the jth variable. Our simulation studies show that such a confidence interval is accurate and has a similar performance to the case where σ is known.

The confidence intervals can also be constructed on the basis of the raw materials in the RCV. For example, for each element in $\hat{M} \equiv \hat{M}_1 \cap \hat{M}_2$, we can take the average of the refitted coefficients as the estimate of the regression coefficients in the set $\hat{M}$, and $(\mathbf{S}_1 + \mathbf{S}_2)\hat{\sigma}^2_{\mathrm{RCV}}/4$ as the corresponding estimated covariance matrix, where $\mathbf{S}_1 = (\mathbf{X}_{\hat{M}}^{(1)\mathrm{T}} \mathbf{X}_{\hat{M}}^{(1)})^{-1}$ is computed on the basis of the first half of the data at the refitting stage and $\mathbf{S}_2 = (\mathbf{X}_{\hat{M}}^{(2)\mathrm{T}} \mathbf{X}_{\hat{M}}^{(2)})^{-1}$ is computed on the basis of the second half of the data. In addition, some 'cleaning' techniques through p-values can be also applied here. In particular, Wasserman and Roeder (2009) and Meinshausen *et al.* (2009) studied these techniques to reduce the number of falsely selected variables substantially.

(b) *Genomewide association studies*: let X_j be the coding of the jth single-nucleotide polymorphism and Y be the observed phenotype (e.g. height or blood pressure) or the expression of a gene of interest. In such a quantitative trait loci study, one frequently fits the marginal linear regression

$$E[Y|X_j] = \alpha_j + \beta_j X_j \tag{11}$$

on the basis of a sample of size n individuals, resulting in the marginal least squares estimate $\hat{\beta}_j$. The interest is to test simultaneously the hypotheses $H_{0,j}:\beta_j=0$ $(j=1,\ldots,p)$. If the conditional distribution of Y given $X_1,\ldots,X_p$ is $N\{\mu(X_1,\ldots,X_p),\sigma^2\}$, then it can easily be shown (Han *et al.*, 2011) that $(\hat{\beta}_1,\ldots,\hat{\beta}_p)^{\mathrm{T}}\sim N\{(\beta_1,\ldots,\beta_p)^{\mathrm{T}},\sigma^2\mathbf{S}/n\}$, where the (i,j) element of $\mathbf{S}$ is the sample covariance matrix of X_i and X_j divided by their sample variances. With σ^2 estimated by the RCV, the P-value for testing individual hypothesis $H_{0,j}$ can be computed. In addition, the dependence of the least squares estimates is now known and hence the false discovery proportion or rate can be estimated and controlled (Han *et al.*, 2011).

(c) *Model selection*: popular penalized approaches for variable selection such as the lasso, SCAD, adaptive lasso and elastic net often involve the choice of a tuning or regularization parameter. A proper tuning parameter can improve the efficiency and accuracy for variable selection. Several criteria, such as Mallows's C_p, AIC and BIC, are constructed to choose tuning parameters. All these criteria rely heavily on a common parameter: the error variance. As an illustration, consider estimating the tuning parameter of the lasso (see also Zou *et al.* (2007)). Let λ be the tuning parameter with the fitted value $\hat{\boldsymbol{\mu}}_\lambda=\mathbf{X}\hat{\boldsymbol{\beta}}_\lambda$. Then AIC and BIC for the lasso are written as

$$\mathrm{AIC}(\hat{\boldsymbol{\mu}}_\lambda,\sigma^2)=\frac{\|\mathbf{y}-\hat{\boldsymbol{\mu}}_\lambda\|^2}{n\sigma^2}+\frac{2}{n}\widehat{\mathrm{df}}(\hat{\boldsymbol{\mu}}_\lambda)$$

and

$$\mathrm{BIC}(\hat{\boldsymbol{\mu}}_\lambda,\sigma^2)=\frac{\|\mathbf{y}-\hat{\boldsymbol{\mu}}_\lambda\|^2}{n\sigma^2}+\frac{\log(n)}{n}\widehat{\mathrm{df}}(\hat{\boldsymbol{\mu}}_\lambda).$$

It is easily seen that the variance σ^2 has an important impact on both AIC and BIC.

4. Folded concave penalized least squares

In this section, we discuss some related methods on variance estimation and their corresponding asymptotic properties. The oracle estimator of σ^2 is

$$\hat{R}(\boldsymbol{\beta}^*)=n^{-1}\sum_{i=1}^{n}(Y_i-\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*)^2.$$

A natural candidate to estimate the variance is $\hat{R}(\hat{\boldsymbol{\beta}})$, where $\hat{\boldsymbol{\beta}}$ is the lasso or SCAD estimator of $\boldsymbol{\beta}^*$. Greenshtein and Ritov (2004) showed the persistent property for the lasso estimator $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$. Their result, interpreted in the linear regression setting, implies that $R(\hat{\boldsymbol{\beta}}_{\mathrm{L}})\rightarrow R(\boldsymbol{\beta}^*)=\sigma^2$ in probability, where $R(\boldsymbol{\beta})=E[(Y-\mathbf{X}\boldsymbol{\beta})^2]$. In fact, it is easy to see that their result implies that

$$\hat{R}(\hat{\boldsymbol{\beta}}_{\mathrm{L}})\rightarrow\sigma^2=R(\boldsymbol{\beta}^*).$$

In other words, $\hat{R}(\hat{\boldsymbol{\beta}}_{\mathrm{L}})$ is a consistent estimator for the variance.

Recall that the lasso estimator is defined as

$$\hat{\boldsymbol{\beta}}_{\mathrm{L}}=\arg\min_{\boldsymbol{\beta}}\left\{\frac{1}{n}\sum_{i=1}^{n}(Y_i-\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})^2+\lambda_n\|\boldsymbol{\beta}\|_1\right\}. \qquad (12)$$

To make $\hat{R}(\hat{\boldsymbol{\beta}}_{\mathrm{L}})$ consistent, Greenshtein and Ritov (2004) suggested $\lambda_n=o[\{n/\log(p)\}^{1/2}]$ asymptotically. Wasserman and Roeder (2009) showed that the consistency still holds when λ_n is chosen by cross-validation. Therefore, we define the lasso variance estimator $\hat{\sigma}_{\mathrm{L}}^2$ by

$$\hat{\sigma}_{\mathrm{L}}^2=\frac{1}{n-\hat{s}_{\mathrm{L}}}\sum_{i=1}^{n}(Y_i-\mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{L}}s)^2, \qquad (13)$$

where $\hat{s}_{\mathrm{L}} = \#\{j : (\hat{\beta}_{\mathrm{L}})_j \neq 0\}$.

We shall see that $\hat{\sigma}^2_{\mathrm{L}}$ usually underestimates the variance owing to spurious correlation, as the lasso shares a similar spirit to that of the stepwise addition (see the algorithm LARS by Efron *et al.* (2004)). Thus, we also consider the leave-one-out lasso variance estimator

$$\hat{\sigma}^2_{\mathrm{LL}} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \mathbf{x}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}}_{\mathrm{L}}^{(-i)})^2 \tag{14}$$

where $\hat{\boldsymbol{\beta}}_{\mathrm{L}}^{(-i)}$ is the lasso estimator using all samples except the ith. In practice, a K-fold (K equals 5 or 10) cross-validated lasso estimator is often used and shares the same spirit as that of equation (14). We divide the data set into K parts, say $\mathcal{D}_1, \ldots, \mathcal{D}_K$, and define

$$\hat{\sigma}^2_{\mathrm{CVL}} = \min_{\lambda} \left\{ \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in \mathcal{D}_k} (Y_i - \mathbf{x}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}}_{\lambda}^{(-k)})^2 \right\} \tag{15}$$

where $\hat{\boldsymbol{\beta}}_{\lambda}^{(-k)}$ is the lasso estimator using all data except those in $\mathcal{D}_k$ with tuning parameter λ. This estimator differs from the plug-in method (13) in that multiple estimates from training samples are used to compute residuals from the testing samples. We shall see that the estimator $\hat{\sigma}^2_{\mathrm{CVL}}$ is typically closer to $R(\hat{\boldsymbol{\beta}}_{\mathrm{L}})$ than to $\hat{R}(\hat{\boldsymbol{\beta}}_{\mathrm{L}})$, but it usually somewhat overestimates the true variance from our simulation experience. The following theorem shows the rate of convergence for the lasso estimator.

Theorem 3. Suppose that assumptions 1–4 and 7 in Appendix A hold. If the true model size $s = o(n^{\alpha_0})$ for some $\alpha_0 < 1$, then we have

$$\hat{\sigma}^2_{\mathrm{L}} - \sigma^2 = O_P[\max\{n^{-1/2}, s \log(p)/n\}].$$

If $s \log(p)/\sqrt{n} \to 0$, we have

$$(\hat{\sigma}^2_{\mathrm{L}} - \sigma^2)\sqrt{n} \to N(0, E[\varepsilon^4] - \sigma^4).$$

The factor $s \log(p)/n$ reflects the bias of the penalized L_1-estimator. It can be non-negligible. When it is negligible, the plug-in lasso estimator also has the oracle property. In general, it is difficult to study the asymptotic distribution of the lasso estimator when the bias is not negligible. In particular, we cannot obtain the standard error for the estimator. Even for finite p, Knight and Fu (2000) investigated the asymptotic distribution of lasso-type estimators but it is too complicated to be applied for inference. To tackle this difficulty, Park and Casella (2008) and Kyung *et al.* (2010) used a hierarchical Bayesian formulation to produce a valid standard error for the lasso estimator, and Chatterjee and Lahiri (2011) proposed a modified bootstrap method to approximate the distribution of the lasso estimator. But it is unclear yet whether or not their methods can be applied to a high or ultrahigh dimensional setting.

Recently, Fan and Lv (2011) studied the oracle properties of the non-concave penalized likelihood method in the ultrahigh dimensional setting. Inspired by their results, the variance σ^2 can be consistently and efficiently estimated. The SCAD penalty $\rho_\lambda(t)$ (Fan and Li, 2001) is the function whose derivative is given by

$$\rho'_\lambda(t) = \lambda \left\{ I(t \leqslant \lambda) + \frac{(a\lambda - t)_+}{(a-1)\lambda} I(t > \lambda) \right\}, \qquad t \geqslant 0, \ a > 2,$$

where $a = 3.7$ is often used. Denote by

$$\mathbf{Q}_{n,\lambda_n}(\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + 2n \sum_{j=1}^{p} \rho_{\lambda_n}(|\beta_j|), \tag{16}$$

and let $\hat{\boldsymbol{\beta}}_{\mathrm{SCAD}}$ be a local minimizer of $\mathbf{Q}_{n,\lambda_n}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$. Thus, the variance σ^2 can be estimated by

$$\hat{\sigma}^2_{\mathrm{SCAD}} = \frac{1}{n-\hat{s}} \sum_{i=1}^{n} (Y_i - \mathbf{x}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}}_{\mathrm{SCAD}})^2,$$

where $\hat{s} = \#\{j : (\hat{\beta}_{\mathrm{SCAD}})_j \neq 0\}$.

The following theorem shows the oracle property and rate of convergence for the SCAD estimator.

Theorem 4. Assume that $\log(p) = O(n^{\alpha_0})$ and the true model size $s = O(n^{\alpha_0})$, where $\alpha_0 \in [0, 1)$. Suppose that assumptions 1, 3 and 4 (or 5 and 6) and 8 and 9 in Appendix A are satisfied. Then,

(a) (model consistency) there is a strictly local minimizer $\hat{\boldsymbol{\beta}}_n = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^{\mathrm{T}}$ of $\mathbf{Q}_{n,\lambda_n}(\boldsymbol{\beta})$ such that

$$\{j : \hat{\beta}_j \neq 0\} = M_0$$

with probability tending to 1 and

(b) (asymptotic normality) with this estimator $\hat{\boldsymbol{\beta}}_n$, we have

$$(\hat{\sigma}^2_{\mathrm{SCAD}} - \sigma^2)\sqrt{n} \xrightarrow{\mathcal{D}} N(0, E[\varepsilon^4] - \sigma^4).$$

Theorem 4 reveals that, if λ_n is chosen reasonably, $\hat{\sigma}^2_{\mathrm{SCAD}}$ works as well as the RCV estimator $\hat{\sigma}^2_{\mathrm{RCV}}$ and better than $\hat{\sigma}^2_{\mathrm{L}}$. However, it is difficult to achieve this oracle property sometimes.

Table 1. Simulation results for example 1: bias BIAS, standard error SE and average model size AMS for the oracle, naive and RCV two-stage procedures

Method	*Results for the following values of n:*								
	n = 50			*n = 100*			*n = 200*		
	BIAS	*SE*	*AMS*	*BIAS*	*SE*	*AMS*	*BIAS*	*SE*	*AMS*
p = 10									
Oracle	0.006	0.220	0	−0.023	0.144	0	−0.015	0.109	0
N-SIS	−0.072	0.209	5	−0.064	0.142	5	−0.030	0.109	5
RCV-SIS	0.017	0.234	5	−0.029	0.150	5	−0.013	0.114	5
N-LASSO	−0.052	0.211	1.08	−0.051	0.148	1.01	−0.028	0.108	0.94
RCV-LASSO	−0.003	0.219	1.41	−0.026	0.149	1.24	−0.015	0.110	1.02
p = 100									
Oracle	−0.011	0.205	0	0.023	0.154	0	−0.010	0.154	0
N-SIS	−0.325	0.151	5	−0.164	0.135	5	−0.112	0.135	5
RCV-SIS	−0.004	0.216	5	0.018	0.165	5	−0.009	0.165	5
N-LASSO	−0.272	0.319	5.90	−0.153	0.279	13.56	−0.073	0.279	3.16
RCV-LASSO	0.032	0.359	4.67	0.022	0.171	5.89	−0.010	0.171	12.41
p = 1000									
Oracle	−0.011	0.176	0	−0.015	0.130	0	−0.015	0.095	0
N-SIS	−0.488	0.118	5	−0.314	0.098	5	−0.192	0.079	5
RCV-SIS	−0.017	0.211	5	−0.018	0.144	5	−0.012	0.098	5
N-LASSO	−0.351	0.399	7.47	−0.256	0.330	9.37	−0.196	0.251	9.90
RCV-LASSO	−0.029	0.266	5.03	−0.022	0.186	8.27	−0.014	0.103	8.79

5. Numerical Results

5.1. Simulation study

In this section, we illustrate and compare the finite sample performance of the methods that were described in the last three sections. We applied these methods to three examples: the null model and two sparse models. The null model (example 1) is given by

$$Y = \mathbf{x}^{\mathrm{T}}\mathbf{0} + \varepsilon, \qquad \varepsilon \sim N(0,1) \tag{17}$$

where $X_1, X_2, \ldots, X_p$ are IID random variables, following the standard Gaussian distribution. This is the sparsest possible model. The second sparse model (example 2) is given by

$$Y = b(X_1 + X_2 + X_3) + \varepsilon, \qquad \varepsilon \sim N(0,1), \tag{18}$$

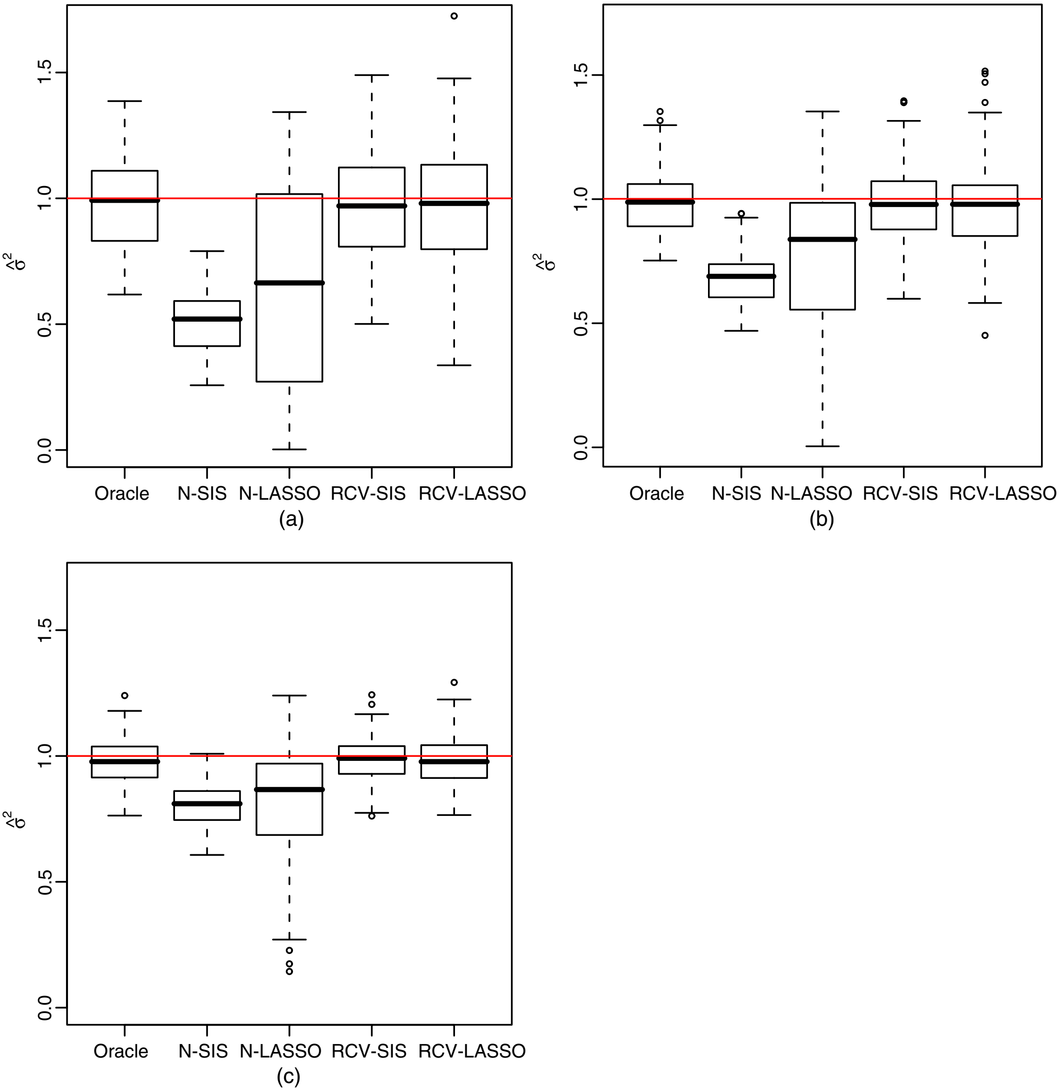


Fig. 4. Boxplots of $\hat{\sigma}_n^2$ when data are generated from the null model (17) with $p = 1000$ and (a) $n = 50$, (b) $n = 100$ and (c) $n = 200$ (the number of simulations is 100): ——, true variance 1

with different b representing different levels of signal-to-noise ratio. The covariates that are associated with model (18) are jointly normal with equal correlation ρ, and marginally $N(0,1)$.

The third sparse model (example 3) is more challenging, with 10 non-trivial coefficients, $\{\beta_j | j = 1,2,3,5,7,11,13,17,19,23\}$. The covariates are jointly normal with $\text{cov}(X_i, X_j) = 0.5^{|i-j|}$. The non-zero coefficients vector is

$$b(1.01,-0.06,0.72,1.55,2.32,-0.36,3.75,-2.04,-0.13,0.61)$$

where b varies to fit different signal-to-noise ratio levels. The random error follows the standard normal distribution.

In each of these settings, we test the following four methods to estimate the variance:

(a) oracle estimator (9), which is not a feasible estimator whose performance provides a benchmark (*method 1*);
(b) a naive two-stage method, denoted by *N-SIS*, if SIS is employed in the model selection step (*method 2*);
(c) RCV variance estimator (6) (*method 3*);
(d) a one-step method via penalized least squares estimators (*method 4*). We introduced this method in Section 4 and recommended two formulae to estimate the variance: a direct plug-in, P, method like formula (13) and a cross-validation, CV, method like formula (15).

In methods 2–4, we employed (I)SIS, SCAD or the lasso as our model selection tools. For SCAD and the lasso, the tuning parameters were chosen by fivefold or 10-fold cross-validation. For (I)SIS, the predetermined model size is always taken to be 5 in the null model and $n/4$ in the sparse model, unless specified explicitly. The principled method of Zhao and Li (2010) can be employed to choose the model size automatically.

5.1.1. *Example 1*

Assume that the response Y is independent of all predictors X_is, which follow an IID standard Gaussian distribution. We consider the cases when the numbers of covariates vary from 10, 100 to 1000 and the sample sizes equal 50, 100 and 200. The simulation results are based on 100 replications and are summarized in Table 1. In Fig. 4, three boxplots are illustrated to compare the performance of the various methods for the case $n = 50, 100, 200$ and $p = 1000$. From the simulation results, we can see that the improved two-stage estimators RCV-SIS and RCV-LASSO

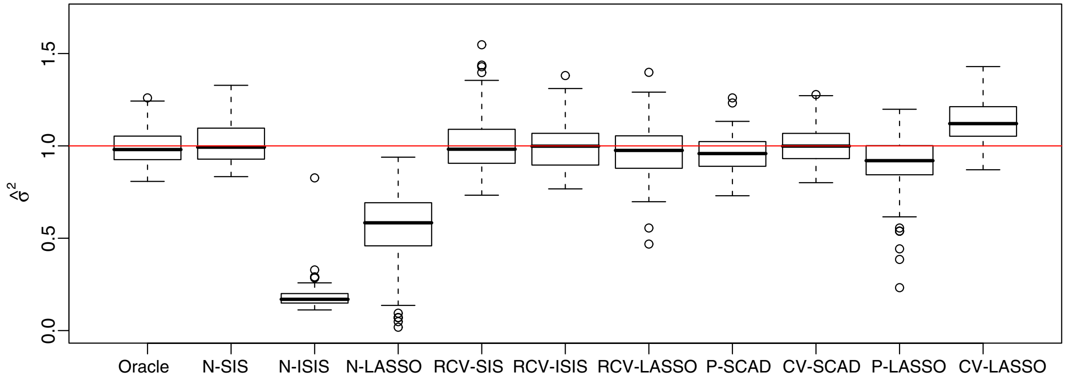


Fig. 5. Comparison of various methods for variance estimation in model (18) with $n = 200$ and $p = 2000$ ($\rho = 0.5$ and $b = 1$): presented are boxplots of $\hat{\sigma}_n^2$ based on 100 replications

Table 2. Simulation results for example 2 with $n=200$ and $p=2000$: bias BIAS, standard error SE, average model size AMS and sure screening probability SSP

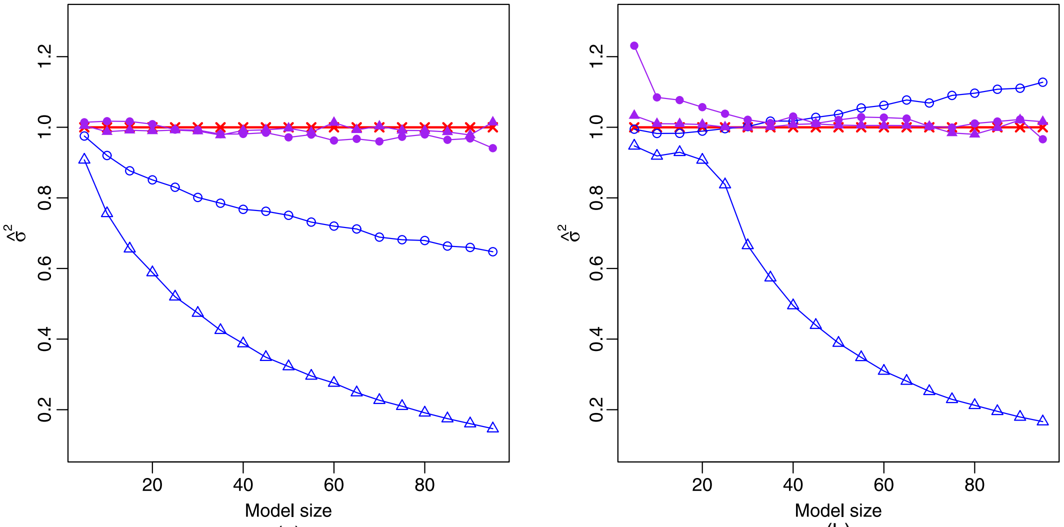
Method	*Results for the following value of ρ:*							
	$\rho=0$				$\rho=0.5$			
	BIAS	*SE*	*AMS*	*SSP*	*BIAS*	*SE*	*AMS*	*SSP*
$b=2$								
Oracle	−0.014	0.089	3.000	1.000	−0.014	0.090	3.000	1.000
N-SIS	−0.111	0.096	50.000	1.000	−0.011	0.102	50.000	1.000
N-ISIS	−0.791	0.073	49.130	1.000	−0.821	0.036	46.870	1.000
N-LASSO	−0.581	0.163	41.460	1.000	−0.526	0.172	43.310	1.000
RCV-SIS	−0.030	0.132	50.000	1.000	0.025	0.279	50.000	0.960
RCV-ISIS	−0.017	0.113	25.770	1.000	−0.020	0.106	22.185	1.000
RCV-LASSO	−0.004	0.130	34.230	1.000	−0.026	0.147	34.990	1.000
P-SCAD	−0.048	0.109	7.810	1.000	−0.036	0.097	6.080	1.000
CV-SCAD	0.000	0.095	7.810	1.000	0.001	0.096	6.080	1.000
P-LASSO	−0.102	0.195	41.460	1.000	−0.113	0.164	43.310	1.000
CV-LASSO	0.141	0.111	41.460	1.000	0.127	0.116	43.310	1.000
$b=1/\surd 3$								
Oracle	−0.014	0.090	3.000	1.000	−0.014	0.090	3.000	1.000
N-SIS	0.010	0.105	50.000	1.000	0.046	0.107	50.000	0.980
N-ISIS	−0.817	0.077	46.400	1.000	−0.809	0.099	46.250	1.000
N-LASSO	−0.445	0.202	39.290	1.000	−0.381	0.239	37.140	1.000
RCV-SIS	0.017	0.164	50.000	0.880	0.057	0.158	50.000	0.430
RCV-ISIS	−0.002	0.122	22.225	0.970	0.113	0.161	22.445	0.150
RCV-LASSO	−0.029	0.147	33.470	0.990	0.046	0.161	31.890	0.450
P-SCAD	−0.036	0.096	6.110	1.000	−0.066	0.102	14.520	1.000
CV-SCAD	0.003	0.096	6.110	1.000	0.079	0.124	14.520	1.000
P-LASSO	−0.097	0.171	39.290	1.000	−0.089	0.171	37.140	1.000
CV-LASSO	0.126	0.116	39.290	1.000	0.125	0.116	37.140	1.000

are comparable with the oracle estimator and much better than the naive estimators, especially in the case when $p \gg n$. This coincides with our theoretical result. RCV improves dramatically the naive (natural) method, no matter whether SIS or the lasso is used.

5.1.2. Example 2

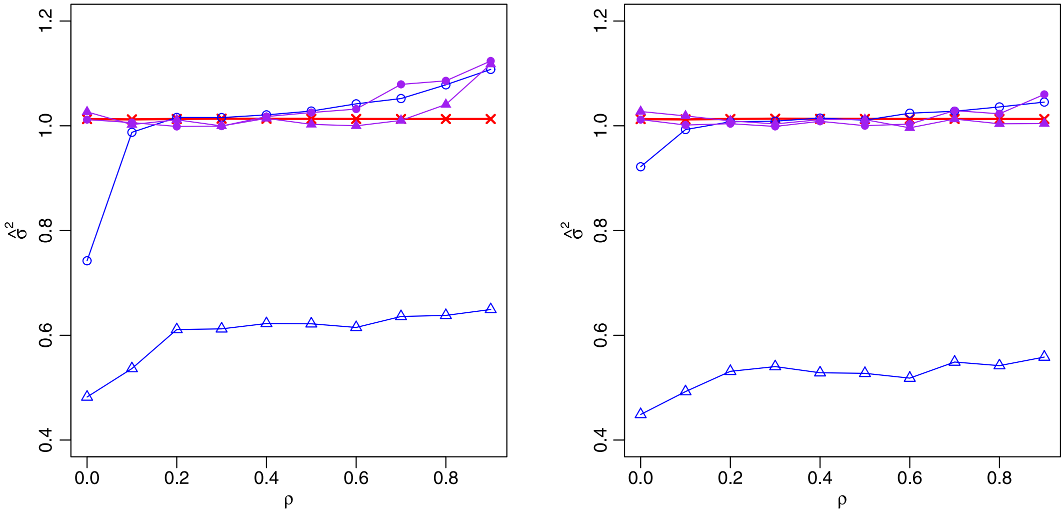
We now consider model (18) with $(n, p)=(200, 2000)$ and $\rho=0$ and $\rho=0.5$. Moreover, we consider three values of coefficients $b=2$, $b=1$ and $b=1/\surd 3$, corresponding to different levels of signal-to-noise ratio $\surd 12$, $\surd 3$ and 1 for each case when $\rho=0$. The results that are depicted in Table 2 are based on 100 replications (the results for $b=1$ are presented in Fig. 5 and have been omitted from Table 2). The boxplots of all estimators for the case $\rho=0.5$ and $b=1$ are shown in Fig. 5. They indicate that the RCV methods behave as well as the oracle, and much better than the naive two-stage methods. Furthermore, the performance of the naive two-stage method depends highly on the model selection technique. The one-step methods perform well also, especially P-SCAD and CV-SCAD. P-LASSO and CV-LASSO behave slightly worse than SCAD methods. These simulation results lend further support to our theoretical conclusions in earlier sections.

To test the sensitivity of the RCV procedure to the model size $\hat{s}$ and covariance structure among predictors, additional simulations have been conducted and their results are summarized



(a) (b)

Fig. 6. Sensitivity of model size $\hat{s}$ on variance estimation for (a) $\rho = 0$ and $b = 1$ and (b) $\rho = 0.5$ and $b = 1$ (presented are the medians of naive and RCV two-stage estimators when $n = 200$ and $p = 2000$ among 100 replications): ×, oracle; ○, N-SIS; △, N-LASSO; ●, RCV-SIS; ▲, RCV-LASSO



(a) (b)

Fig. 7. Effect of covariance structure on variance estimation for (a) $b = 1$ and (b) $b = 2$ (presented are the medians of naive and RCV two-stage estimators when $n = 200$ and $p = 2000$ among 100 replications for various ρ): ×, oracle; ○, N-SIS; △, N-LASSO; ●, RCV-SIS; ▲, RCV-LASSO

in Figs 6 and 7. From Fig. 6, it is clear that the RCV method is insensitive to model size $\hat{s}$, as explained before theorem 2. Fig. 7 shows that the RCV methods are also robust with respect to the covariance structure. In contrast, N-LASSO always underestimates the variance.

To show the effectiveness of $\hat{\sigma}_{\mathrm{RCV}}$ in the construction of confidence intervals, we calculate the coverage probability of the confidence interval (10) based on 10000 simulations. This was conducted for β_1, β_2 and β_3 with $b = 1/\sqrt{3}$, 1, 2 and $\rho = 0$ and $\rho = 0.5$. For brevity we present only one specific case for β_1 with $b = 1$ in Table 3.

Table 3. Simulation results for example 2 with $n=200$, $p=2000$ and $b=1$: coverage probability of confidence intervals of different levels for β_1, based on 10000 replications

Method	*Coverage probabilities for the following values of ρ and confidence intervals:*							
	$\rho=0$				$\rho=0.5$			
	80%	*90%*	*95%*	*99%*	*80%*	*90%*	*95%*	*99%*
Oracle	0.7967	0.8974	0.9476	0.9874	0.7931	0.9006	0.9483	0.9865
RCV	0.7919	0.8928	0.9435	0.9847	0.8042	0.9022	0.9518	0.9871

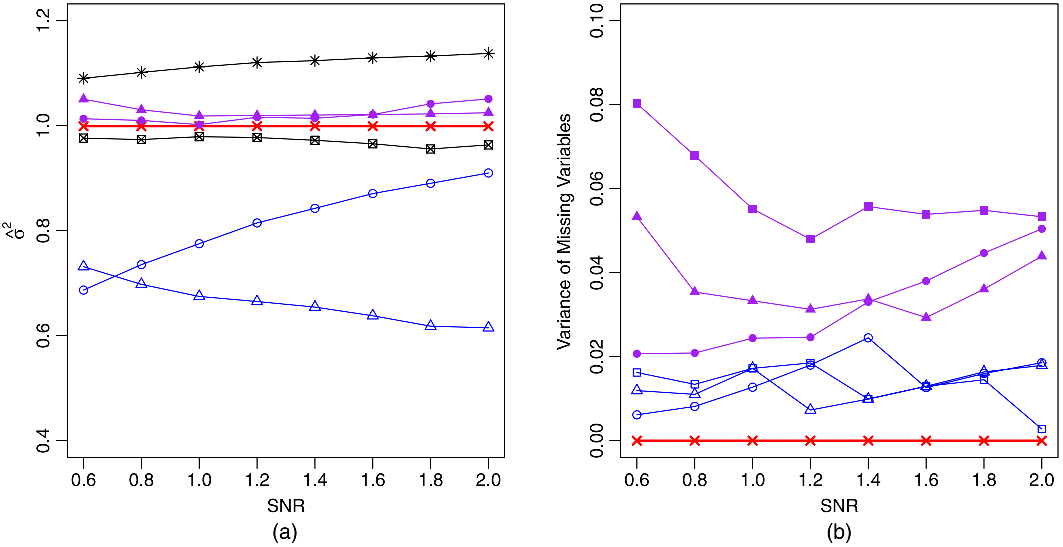


Fig. 8. (a) Medians of various variance estimators when $n=400$ and $p=1000$ among 100 replications for example 3 (×, oracle; ○, N-SIS; △, N-LASSO; ●, RCV-SIS; ▲, RCV-LASSO; □, P-LASSO; $*$, CV-LASSO) and (b) medians of variance of missing variables of various model selection methods (×, oracle; ○, SIS; □, ISIS; △, LASSO; ●, RCV-SIS; ■, RCV-ISIS; ▲, RCV-LASSO)

5.1.3. Example 3

We consider a more realistic model with 10 important predictors, detailed at the beginning of this section. Since some non-vanishing coefficients are very small, no method can guarantee that all relevant variables are chosen in the model selected, i.e. have a sure screening property. To quantify the severity of missing relevant variables, we use the quantity variance of missing variables, $\mathrm{var}(\mathbf{x}_S^{\mathrm{T}}\beta_S)/\sigma^2$, to measure, where S is the set of important variables that are not included in the model selected and β_S are their regression coefficients in the simulated model. For RCV methods, the variance of missing variables is the average of the variances of missing variables for two halves of the data. Fig. 8 summarizes the simulation results for $(n,p)=(400,1000)$, whereas Fig. 9 depicts the results for $(n,p)=(400,10000)$ when the penalization methods are not easily accessible. The naive methods seriously underestimate the variance and are sensitive to the model selection tools, dimensionality and signal-to-noise ratio among others. In contrast, the RCV methods are much more stable and only slightly overestimate the variance when the sure screening condition is not satisfied. The one-step methods, especially plug-in methods, also perform well.

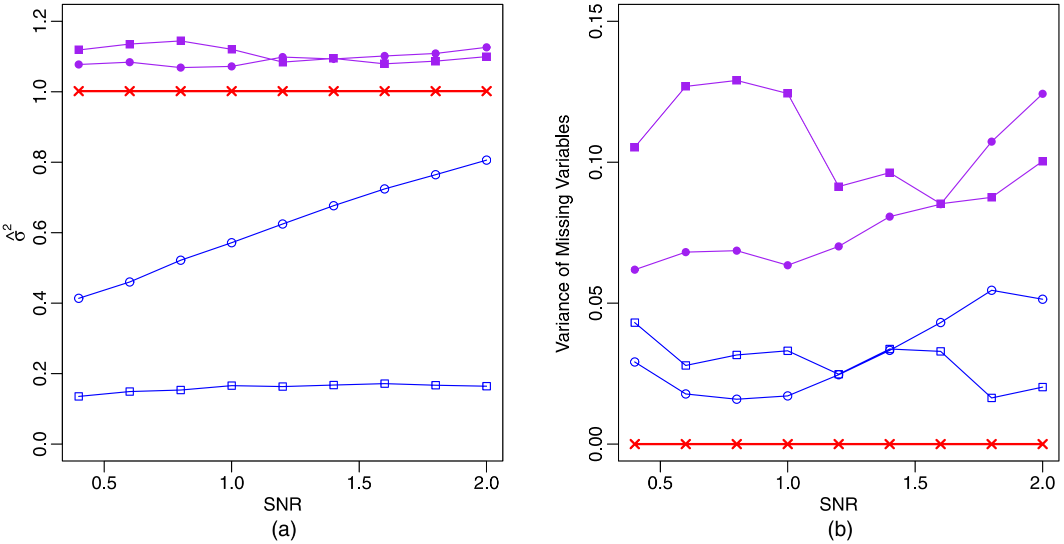


Fig. 9. (a) Medians of various variance estimators when $n=400$ and $p=10000$ among 100 replications (×, oracle; ○, N-SIS; □, N-ISIS; ●, RCV-SIS; ■, RCV-ISIS) and (b) medians of variance of missing variables of various model selection tools (×, oracle; ○, SIS; □, ISIS; ●, RCV–SIS; ■, RCV-ISIS)

Table 4. Estimated residual standard deviation and variance explained by regression for naive two-stage and RCV methods for forecasting HPA in San Francisco and Los Angeles

Model size	*Results for San Francisco*			*Results for Los Angeles*		
	Naive	*RCV*	*Variance explained (%)*	*Naive*	*RCV*	*Variance explained (%)*
2	0.5577	0.5563	76.92	0.5236	0.5255	88.68
3	0.5236	0.5536	79.83	0.4887	0.5214	90.23
5	0.5072	0.5179	81.40	0.4583	0.5210	91.56
10	0.4555	0.5057	85.67	0.4401	0.4995	92.56
15	0.3938	0.4730	89.79	0.3747	0.4794	94.86
20	0.3862	0.4749	90.66	0.3137	0.4596	96.57
30	0.3635	0.4735	92.58	0.2503	0.4621	98.05

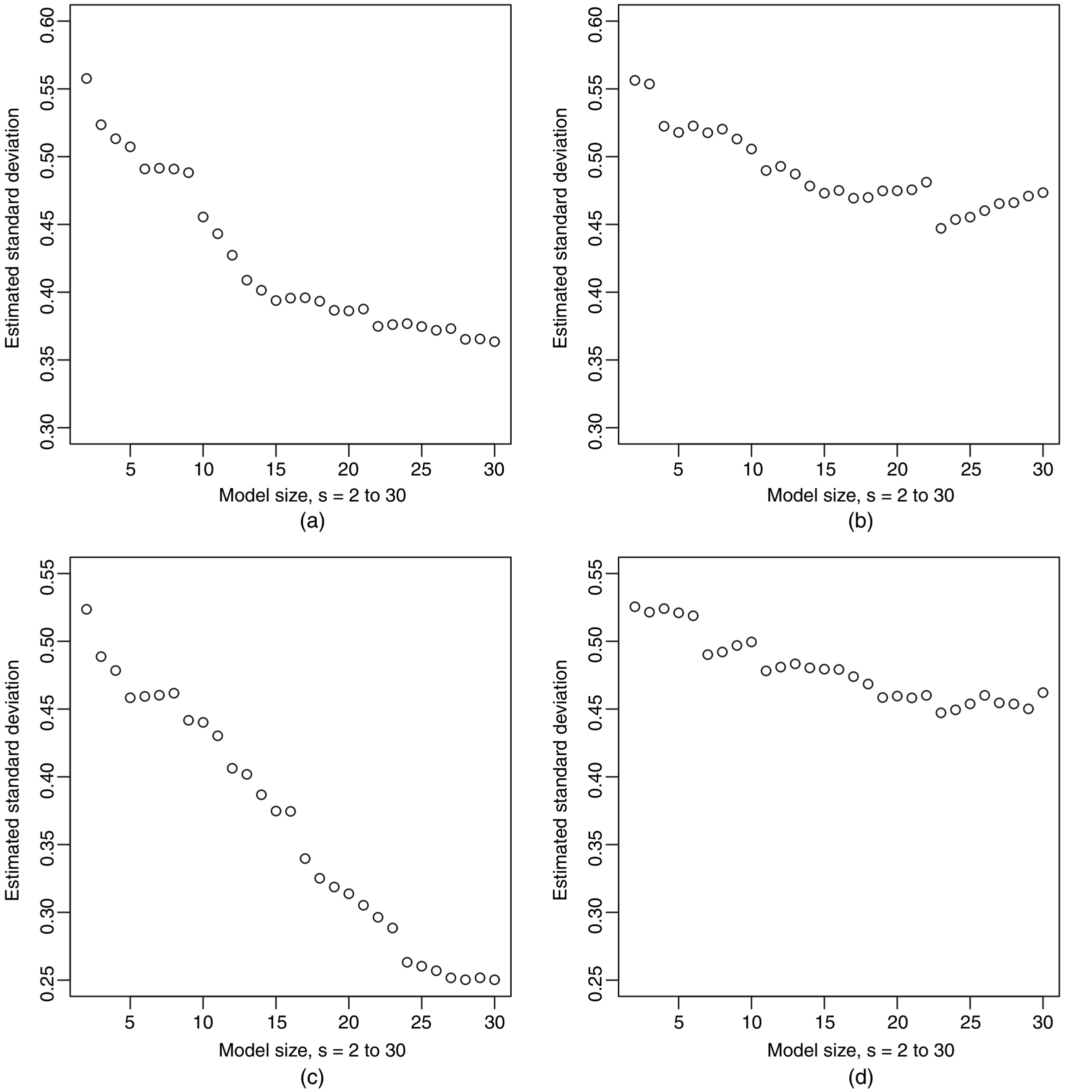
5.2. *Real data analysis*

We now apply our proposed procedure to analyse recent house price data from 1996–2005. The data set consists of 119 months of appreciation of the national house price index HPI, which is defined as the percentage of monthly log-HPI changes in 381 core-based statistical areas (CBSAs) in the USA. The goal is to forecast housing price appreciation (HPA) over those 381 CBSAs over the next several years. Housing prices are geographically dependent. They depend also on macroeconomic variables. Their dependence on macroeconomic variables can be summarized by the national HPA. Therefore, a reasonable model for predicting the next period HPA in a given CBSA is

$$Y_t = \beta_0 + \beta_{\mathrm{N}} X_{t-1,\mathrm{N}} + \sum_{i=1}^{381} \beta_i X_{t-1,i} + \varepsilon_t, \tag{19}$$

where X_{N} stands for the national HPA, $\{X_i\}_{i=1}^{381}$ are the HPAs in those 381 CBSAs and ε is a random error independent of X. This is clearly a problem with the number of predictors more than the number of covariates. However, conditional on the national HPA X_{N}, it is reasonable to expect that only the local neighbourhoods have non-negligible influence, but it is difficult to predetermine those neighbourhoods. In other words, it is reasonable to expect that the coefficients $\{\beta_i\}_{i=1}^{381}$ are sparse.

Our primary interest is to estimate the residual variance σ^2, which is the prediction error



(a)

(b)

(c)

(d)

Fig. 10. Estimated standard deviation of benchmark one-step forecasts of HPA in San Francisco and Los Angeles for various model sizes: (a) San Francisco, naive method; (b) San Francisco, RCV method; (c) Los Angeles, naive method; (d) Los Angeles, RCV method

of the benchmark model. We always keep the variables X_{N} and X_1, which is the lag 1 HPA of the region to be predicted. We applied SCAD using the local linear approximation (Zou and Li, 2008), which is the iteratively reweighted lasso, to estimate coefficients in model (19). We summarize the result, $\hat{\sigma}$, as a function of the selected model size s, to examine the sensitivity to the selected model size. Reported also is the percentage of variance explained, which is defined as

$$R^2 = 1 - \mathrm{RSS} \Bigg/ \sum_{t=1}^{119} (Y_t - \bar{Y})^2,$$

where $\bar{Y}$ is the sample average of the time series. For illustration, we focus only on one CBSA in San Francisco and one in Los Angeles. The results are summarized in Table 4 and Fig. 10, in which the naive two-stage method is also included for comparison.

First, as shown in Fig. 10, the influence of the naive method on the selected model size is much larger than that of the RCV method. This is due to the spurious correlation as we discussed before. The RCV estimate is reasonably stable, but it is also influenced by the selected model size when it is large. This is understandable given the sample size of 119.

In the case of San Francisco, from Fig. 10(b), the RCV method suggests that the standard deviation should be around 0.52%, which is reasonably stable for s in the range of 4–8. By inspection of the solution path of the naive two-stage method, we see that, besides X_{N} and X_1, first selected is the variable X_{306}, which corresponds to CBSA San Jose–Sunnyvale–Santa Clara (San Benito County and Santa Clara County). The variable X_{306} also enters both models when $s \geqslant 3$ in the RCV method. Therefore, we suggest that the model selected consists of at least variables X_1, X_2 and X_{306}. As expected, in the RCV method, the fourth selected variables are not the same for the two split subsamples. The variance explained by regression takes 79.83% of total variance.

Similar analysis can be applied to the Los Angeles case. Fig. 10(d) suggests that the standard deviation should be around 0.50% (when s is between 7 and 10). From the solution path, we suggest that the model selected consists of at least variables X_{N}, X_1 and X_{252}, which corresponds to CBSA Oxnard–Thousand Oaks–Ventura (Ventura County). The variance explained by regression takes 90.23% of the total variance.

6. Discussion

Variance estimation is important and challenging for ultrahigh dimensional sparse regression. One of the challenges is the spurious correlation: covariates can have high correlations with the realized noise and hence are recruited to predict the noise. As a result, the naive (natural) two-stage estimator seriously underestimates the variance. Its performance is very unstable and depends largely on the model selection tool that is employed. The RCV method is proposed to attenuate the influence of the effect of spurious variables. Both the asymptotic theory and the empirical result show that the RCV estimator is the best among all estimators. It is accurate and stable, and insensitive to the model selection tool and the size of the model selected. Therefore, we may employ fast model selection tools like SIS for computational efficiency for the RCV variance estimation. We also compare the RCV method with the direct plug-in method. When choosing tuning parameters of a penalized likelihood method like the lasso, we suggest using a more conservative cross-validation rather than aggressive BIC. However, the lasso method can still yield a non-negligible bias for variance estimation in ultrahigh dimensional regression. The SCAD method is almost as good as the RCV method, but it is computationally more expensive than RCV-SIS.

Acknowledgements

The paper was supported by the National Institutes of Health grant R01-GM072611 and National Science Foundation grants DMS-0704337 and was completed while Shaojun Guo and Ning Hao were postdoctoral fellows at Princeton University. The paper is dedicated to Peter J. Bickel on his occasion of the 70th birthday.

Appendix A: Notation and conditions

We first state the following assumptions, which are standard in the literatures of high dimensional statistical learning. For convenience, define

$$\phi_{\min}(m) = \min_{M:|M| \leqslant m} \left\{ \lambda_{\min}\left(\frac{1}{n} \mathbf{X}_M^{\mathrm{T}} \mathbf{X}_M \right) \right\}$$

and

$$\phi_{\max}(m) = \max_{M:|M| \leqslant m} \left\{ \lambda_{\max}\left(\frac{1}{n} \mathbf{X}_M^{\mathrm{T}} \mathbf{X}_M \right) \right\},$$

where $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ denote the smallest and largest eigenvalues of a matrix $\mathbf{A}$ respectively.

For a vector $\mathbf{v}$, we use the standard notation $\|\mathbf{v}\|_p = (\Sigma_i |v_i|^p)^{1/p}$ and $\|\mathbf{v}\|_\infty = \max_i\{|v_i|\}$. For a matrix $\mathbf{B}$, we use three different norms. $\|\mathbf{B}\|_{2,\infty}$ is defined in assumption 8 below; $\|\mathbf{B}\|_2$ denotes the usual operator norm, i.e. $\|\mathbf{B}\|_2 = \max_{\|\mathbf{v}\|_2 \leqslant 1} \|\mathbf{B}\mathbf{v}\|_2$; $\|\mathbf{B}\|_\infty = \max_{i,j}\{|B_{ij}|\}$ is the usual sup-norm.

Assumption 1. The errors $\varepsilon_1, \ldots, \varepsilon_n$ are IID with zero mean and finite variance σ^2 and independent of the design matrix $\mathbf{X}$.

Assumption 2. There is a constant $\lambda_0 > 0$ and b_n such that $b_n/n \to 0$ such that $P\{\phi_{\min}(b_n) \geqslant \lambda_0\} = 1$ for all n.

Assumption 3. There is a constant L such that $\max_{i,j} |X_{ij}| \leqslant L$, where X_{ij} is the (i, j) element of the design matrix $\mathbf{X}$.

Assumption 4. $E[\exp(|\varepsilon_1|/a)] \leqslant b$ for some finite constants $a, b > 0$.

We have no intention to make the assumptions the weakest possible. For example, assumption 3 can be relaxed to $\max_{i,j} |X_{ij}| \leqslant L\{\log(n)\}^\xi$ for any $\xi > 0$ or further relaxation. The aim of assumptions 3 and 4 is to guarantee that $\hat{\gamma}_n$ in theorem 1 is of the order $\surd\{\hat{s} \log(p)/n\}$.

Theorem 1 still holds under the random design with the assumptions below.

Assumption 5. The random vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are IID and there is a constant α such that $E[\exp\{(|X_{ij}|/\rho)^\alpha\}] \leqslant L$ for all i and j and some constants $\alpha > 1$, and $\rho, L > 0$, where X_{ij} is the (i, j)th element of $\mathbf{X}$.

Assumption 6. ε_1 satisfies the condition that $E[\exp\{(|\varepsilon_1|/a)^\theta\}] \leqslant b$ for some finite positive constants $a, b, \theta > 0$ and $1/\alpha + 1/\theta \leqslant 1$, where α is defined by assumption 5.

For instance, when X_{ij} and ε_i are sub-Gaussian ($\alpha = \theta = 2$) for each i and j, assumptions 5 and 6 are satisfied.

The following assumption 7 is imposed for proving theorem 3. For fixed design matrix $\mathbf{X}$, the corresponding condition was also imposed in Meinshausen and Yu (2009) and some discussions of weaker conditions were shown in Bickel *et al.* (2009).

Assumption 7. There are constants $0 < k_{\min} \leqslant k_{\max} < \infty$ such that

$$P(\liminf_{n\to\infty} [\phi_{\min}\{s \log(n)\}] \geqslant k_{\min}) = 1,$$

and

$$P(\limsup_{n\to\infty} [\phi_{\max}(s + \min\{n, p\}) \leqslant k_{\max}]) = 1.$$

The following two additional assumptions are stated for proving theorem 4. These conditions correspond to conditions 4 and 5 in Fan and Lv (2011). Without loss of generality, assume that the true value

$\boldsymbol{\beta}_0=(\boldsymbol{\beta}_{01}^{\mathrm{T}},\boldsymbol{\beta}_{02}^{\mathrm{T}})^{\mathrm{T}}$ with each component of $\boldsymbol{\beta}_{01}$ non-zero and $\boldsymbol{\beta}_{02}=\mathbf{0}$. Let $\mathbf{X}_1$ and $\mathbf{X}_2$ be the submatrices of $n\times p$ design matrix $\mathbf{X}$ with columns corresponding to $\boldsymbol{\beta}_{01}$ and $\boldsymbol{\beta}_{02}$ respectively.

Assumption 8. There are constants $0<c_1,c_2<\infty$ such that

$$P\left\{\lambda_{\min}\left(\frac{1}{n}\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1\right)\geqslant c_1\right\}\to 1,$$

and

$$P\left(\left\|\frac{1}{n}\mathbf{X}_2^{\mathrm{T}}\mathbf{X}_1\right\|_{2,\infty}\leqslant c_2\right)\to 1,$$

as $n\to\infty$, where $\|\mathbf{B}\|_{2,\infty}=\max_{\|\mathbf{v}\|_2\leqslant 1}\|\mathbf{B}\mathbf{v}\|_{\infty}$.

Assumption 9. Denote $d_n=\frac{1}{2}\min_{j=1,\ldots,s}|\beta_{0j}|$. Assume that $d_n\geqslant n^{-\gamma}\log(n)$ with $\gamma\in(0,\frac{1}{2}]$. Take $\lambda_n\propto n^{-(1-\alpha_0)/2}\log(n)$ and $\lambda_n\ll d_n$, where α_0 is defined in theorem 4.

Remark 3. The norm $\|\mathbf{B}\|_{2,\infty}$ is somewhat abstract. It can easily be shown that

$$\|\mathbf{B}\|_{2,\infty}\leqslant s\|\mathbf{B}\|_{\infty},$$

where s is the number of columns of $\mathbf{B}$, which is a crude upper bound. Using this and the argument in the proof of theorem 4, if

$$P\left(\left\|\frac{1}{n}\mathbf{X}_2^{\mathrm{T}}\mathbf{X}_1\right\|_{\infty}\leqslant c_3\right)\to 1$$

and $\lambda_n\geqslant n^{-(1-3\alpha_0)/2}\log(n)$ and $\lambda_n\ll d_n$, then the conclusion of theorem 4 holds.

A.1. Proof of theorem 1

Part (a) of theorem 1 follows the standard law of large numbers and central limit theorem. Now we prove the second part under assumptions 1–4.

By assumption 2,

$$\varepsilon^{\mathrm{T}}\mathbf{P}_{\hat{M}}\varepsilon=\varepsilon^{\mathrm{T}}\mathbf{X}_{\hat{M}}(\mathbf{X}_{\hat{M}}^{\mathrm{T}}\mathbf{X}_{\hat{M}})^{-1}\mathbf{X}_{\hat{M}}^{\mathrm{T}}\varepsilon\leqslant\frac{1}{\lambda_0 n}\|\mathbf{X}_{\hat{M}}^{\mathrm{T}}\varepsilon\|^2. \tag{20}$$

Let $\mathbf{X}_j$ denote the jth column vector of the design matrix $\mathbf{X}$. For a large constant c, consider the event $\mathcal{E}_n=\{\max_{1\leqslant j\leqslant p}|\mathbf{X}_j^{\mathrm{T}}\varepsilon|\leqslant c\surd\{n\log(p)\}\}$. Under the event $\mathcal{E}_n$, it follows from equation (20) that

$$\varepsilon^{\mathrm{T}}\mathbf{P}_{\hat{M}}\varepsilon\leqslant\frac{1}{\lambda_0}\hat{s}c^2\log(p).$$

Together with the fact $n^{-1}\|\varepsilon\|^2\to\sigma^2$, we obtain

$$\hat{\gamma}_n^2=\varepsilon^{\mathrm{T}}\mathbf{P}_{\hat{M}}\varepsilon/\varepsilon^{\mathrm{T}}\varepsilon=O_P\{\hat{s}\log(p)/n\}.$$

Hence it suffices to show that $P(\mathcal{E}_n)\to 1$ as $n\to\infty$ for some constant c. Observe that, by assumptions 3 and 4, for each j,

$$E|X_{ij}\varepsilon_i|^m\leqslant m!(La)^m E[\exp(|\varepsilon_1|/a)]\leqslant\tfrac{1}{2}m!(2ba^2L^2)(aL)^{m-2}.$$

Using Bernstein's inequality (e.g. lemma 2.2.11 of van der Vaart and Wellner (1996)), we have

$$\begin{aligned}P(\mathcal{E}_n^C)&\leqslant P\left[\max_{1\leqslant j\leqslant p}|\mathbf{X}_j^{\mathrm{T}}\varepsilon|\geqslant c\surd\{n\log(p)\}\right]\\&\leqslant\sum_{j=1}^{p}P[|\mathbf{X}_j^{\mathrm{T}}\varepsilon|\geqslant c\surd\{n\log(p)\}]\\&\leqslant 2p\exp\left(-\frac{c^2 n\log(p)}{2[2ba^2L^2+aLc\surd\{n\log(p)\}]}\right)\\&=2\exp\left(\log(p)\left[1-\frac{1}{4ba^2L^2c^{-2}n^{-1}+2aLc^{-1}\surd\{\log(p)/n\}}\right]\right).\end{aligned} \tag{21}$$

For sufficiently large c, we have $4ba^2L^2c^{-2}n^{-1}+2aLc^{-1}\sqrt{}\log(p)/n<1$ since $\log(p)/n$ is bounded. Therefore, the power in equation (21) goes to $-\infty$ as $p\to\infty$. It follows that $P(\mathcal{E}_n)=1-P(\mathcal{E}_n^C)\to 1$.

Next we show that the second part of the theorem still holds under assumptions 5 and 6 instead of assumptions 3 and 4. It is sufficient to verify that $P(\mathcal{E}_n)\to 1$ as $n\to\infty$ for some constant c. The key step is to establish the inequality

$$E[|X_{ij}\varepsilon_i|^m]\leqslant \tfrac{1}{2}m!\{8(2+L+b)\rho^2a^2\}(2\rho a)^{m-2}, \tag{22}$$

for each $j=1,\ldots,p$.

Note that

$$P(|XY|>t)\leqslant P(|X|>t^{1/\alpha})+P(|Y|>t^{1-1/\alpha})$$

for $\alpha>1$ and random variables X and Y. Thus, for any $t\geqslant 1$ and each i and j,

$$\begin{aligned}P\left(\left|\frac{X_{ij}}{\rho}\right|\left|\frac{\varepsilon_i}{a}\right|>t\right)&\leqslant P\left(\left|\frac{X_{ij}}{\rho}\right|>t^{1/\alpha}\right)+P\left(\left|\frac{\varepsilon_i}{a}\right|>t^{1-1/\alpha}\right)\\&\leqslant L\exp(-t)+b\exp(-t^{\theta(1-1/\alpha)})\\&\leqslant (L+b)\exp(-t).\end{aligned}$$

If X is a non-negative random variable with its distribution $F(t)$ and tail probability $P(X>t)\leqslant C\exp(-t)$ for some constant $C>0$ and each $t\geqslant 1$, then by integration by parts

$$\begin{aligned}E\left[\exp\left(\frac{1}{2}X\right)\right]&=-\int_0^\infty \exp\left(\frac{x}{2}\right)\mathrm{d}\{1-F(x)\}\\&=1+\frac{1}{2}\int_0^\infty\{1-F(x)\}\exp\left(\frac{x}{2}\right)\mathrm{d}x\\&\leqslant 1+\frac{1}{2}\int_0^1\exp\left(\frac{x}{2}\right)\mathrm{d}x+\frac{1}{2}\int_1^\infty C\exp\left(-\frac{x}{2}\right)\mathrm{d}x\\&\leqslant 2+C.\end{aligned}$$

As a result, it follows that, for each i and j,

$$E\exp\left(\frac{1}{2}\left|\frac{X_{ij}}{\rho}\right|\left|\frac{\varepsilon_i}{a}\right|\right)\leqslant 2+L+b.$$

Thus, for each positive integer j and $m\geqslant 2$,

$$\begin{aligned}E[|X_{ij}\varepsilon_i|^m]&\leqslant (2\rho a)^m m!\,E\left[\exp\left(\frac{1}{2}\left|\frac{X_{ij}}{\rho}\right|\left|\frac{\varepsilon_i}{a}\right|\right)\right]\\&\leqslant(2\rho a)^m m!(2+L+b)\\&=\frac{1}{2}m!\{8(2+L+b)\rho^2a^2\}(2\rho a)^{m-2}.\end{aligned}$$

Theorem 1 is proved.

A.2. Proof of theorem 2

Define sequences of events $\mathcal{A}_{n1}=\{M_0\subset\hat{M}_1\}$, $\mathcal{A}_{n2}=\{M_0\subset\hat{M}_2\}$ and $\mathcal{A}_n=\mathcal{A}_{n1}\cap\mathcal{A}_{n2}$. On the event $\mathcal{A}_n$, we have

$$\hat{\sigma}_1^2=\frac{\boldsymbol{\varepsilon}^{(2)\mathrm{T}}(\mathbf{I}_{n/2}-\mathbf{P}^{(2)}_{\hat{M}_1})\boldsymbol{\varepsilon}^{(2)}}{n/2-\hat{s}_1}$$

and

$$\hat{\sigma}_2^2=\frac{\boldsymbol{\varepsilon}^{(1)\mathrm{T}}(\mathbf{I}_{n/2}-\mathbf{P}^{(1)}_{\hat{M}_2})\boldsymbol{\varepsilon}^{(1)}}{n/2-\hat{s}_2},$$

where $\varepsilon^{(1)}$ and $\varepsilon^{(2)}$ correspond to $\mathbf{y}^{(1)}$ and $\mathbf{y}^{(2)}$ respectively. Decompose now $(n/2-\hat{s}_1)(\hat{\sigma}_1^2-\sigma^2)$ on the event $\mathcal{A}_n$ as

$$(\tfrac{1}{2}n-\hat{s}_1)(\hat{\sigma}_1^2-\sigma^2)=\varepsilon^{(2)\mathrm{T}}\varepsilon^{(2)}-\tfrac{1}{2}n\sigma^2-(\varepsilon^{(2)\mathrm{T}}\mathbf{P}^{(2)}_{\hat{M}_1}\varepsilon^{(2)}-\hat{s}_1\sigma^2).$$

We now prove that $\varepsilon^{(2)\mathrm{T}}\mathbf{P}^{(2)}_{\hat{M}_1}\varepsilon^{(2)}-\hat{s}_1\sigma^2=O_P(\surd\hat{s}_1)$.

First, consider the quadratic form $S=\boldsymbol{\xi}^{\mathrm{T}}\mathbf{P}\boldsymbol{\xi}$ where $\mathbf{P}$ is a symmetric $m\times m$ matrix, $\boldsymbol{\xi}=(\xi_1,\ldots,\xi_m)^{\mathrm{T}}$ and ξ_i $(i=1,\ldots,m)$ are IID. Assume that $E[\xi_1]=0$, $E[\xi_1^2]=\sigma^2$ and the fourth moment $E[\xi_1^4]<\infty$. Let P_{ij} be the (i,j)th element of the matrix $\mathbf{P}$. Then,

$$E[S]=E\left[\sum_{i=1}^{m}\xi_i^2P_{ii}\right]=\sigma^2\,\mathrm{tr}(\mathbf{P}),$$

and

$$\begin{aligned}\mathrm{var}(S)&=E\left[\sum_{i,j,l,k}^{m}\xi_i\xi_j\xi_l\xi_kP_{ij}P_{lk}\right]-\sigma^4\left(\sum_{i=1}^{m}P_{ii}\right)^2\\&=E\left[\sum_{i=1}^{m}\xi_i^4P_{ii}^2\right]+E\left[\sum_{i=l\neq j=k}^{m}\xi_i^2\xi_j^2P_{ij}P_{lk}\right]+E\left[\sum_{i=k\neq j=l}^{m}\xi_i^2\xi_j^2P_{ij}P_{lk}\right]\\&\quad+E\left[\sum_{i=j\neq l=k}^{m}\xi_i^2\xi_l^2P_{ij}P_{lk}\right]-\sigma^4\left(\sum_{i=1}^{m}P_{ii}\right)^2\\&=E[\xi_1^4]\left(\sum_{i=1}^{m}P_{ii}^2\right)+2\sigma^4\left(\sum_{i\neq j}^{m}P_{ij}^2\right)+\sigma^4\left(\sum_{i\neq l}^{m}P_{ii}P_{ll}\right)-\sigma^4\left(\sum_{i=1}^{m}P_{ii}\right)^2\\&=(E[\xi_1^4]-\sigma^4)\left(\sum_{i=1}^{m}P_{ii}^2\right)+2\sigma^4\left(\sum_{i\neq j}^{m}P_{ij}^2\right)\\&\leqslant(E[\xi_1^4]+\sigma^4)\,\mathrm{tr}(\mathbf{P}^2),\end{aligned}$$

where the last inequality holds since $\mathrm{tr}(\mathbf{P}^2)=\Sigma_{i,j}^{m}P_{ij}^2$.

Observe that $\mathrm{tr}(\mathbf{P}^{(2)}_{\hat{M}_1})=\mathrm{tr}\{(\mathbf{P}^{(2)}_{\hat{M}_1})^2\}=\hat{s}_1$. Hence, on the event $\mathcal{A}_{n1}$, we have

$$E[\varepsilon^{(2)\mathrm{T}}\mathbf{P}^{(2)}_{\hat{M}_1}\varepsilon^{(2)}|\mathbf{X}^{(2)}_{\hat{M}_1}]=\hat{s}_1\sigma^2,$$

and

$$\mathrm{var}(\varepsilon^{(2)\mathrm{T}}\mathbf{P}^{(2)}_{\hat{M}_1}\varepsilon^{(2)}|\mathbf{X}^{(2)}_{\hat{M}_1})\leqslant(E[\varepsilon^4]+\sigma^4)\hat{s}_1.$$

Using the Markov inequality, it follows that, under the event $\mathcal{A}_{n1}$,

$$\varepsilon^{(2)\mathrm{T}}\mathbf{P}^{(2)}_{\hat{M}_1}\varepsilon^{(2)}-\hat{s}_1\sigma^2=O_P(\surd\hat{s}_1).$$

Combining with the assumptions $\hat{s}_1/n\to^{\mathrm{P}}0$ and $P(\mathcal{A}_{n1})\to^{\mathrm{P}}1$, we obtain that

$$\varepsilon^{(2)\mathrm{T}}\mathbf{P}^{(2)}_{\hat{M}_1}\varepsilon^{(2)}-\hat{s}_1\sigma^2=o_P(\surd n).$$

As a result,

$$(\tfrac{1}{2}n-\hat{s}_1)(\hat{\sigma}_1^2-\sigma^2)=\varepsilon^{(2)\mathrm{T}}\varepsilon^{(2)}-\tfrac{1}{2}n\sigma^2+o_P(\surd n).$$

Similarly, we conclude that

$$(\tfrac{1}{2}n-\hat{s}_2)(\hat{\sigma}_2^2-\sigma^2)=\varepsilon^{(1)\mathrm{T}}\varepsilon^{(1)}-\tfrac{1}{2}n\sigma^2+o_P(\surd n).$$

Therefore, using the last two results, we have

$$\begin{aligned}(\hat{\sigma}^2_{\mathrm{RCV}}-\sigma^2)\surd n&=\frac{\surd n}{n-2\hat{s}_1}\left(\varepsilon^{(2)\mathrm{T}}\varepsilon^{(2)}-\frac{1}{2}n\sigma^2\right)+\frac{\surd n}{n-2\hat{s}_2}\left(\varepsilon^{(1)\mathrm{T}}\varepsilon^{(1)}-\frac{1}{2}n\sigma^2\right)+o_P(1)\\&=\frac{1}{\surd n}\sum_{i=1}^{n}(\varepsilon_i^2-\sigma^2)+o_P(1),\end{aligned}$$

which implies that

$$(\hat{\sigma}^2_{\mathrm{RCV}} - \sigma^2)\sqrt{n} \xrightarrow{\mathcal{D}} N(0, E[\varepsilon^4] - \sigma^4).$$

The proof of theorem 2 is completed.

To prove theorem 3, we shall use the following lemma. The results were stated and proved in Meinshausen and Yu (2009) and Bickel *et al.* (2009).

Lemma 1. Consider the lasso selector $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ defined by equation (12) with λ_n. Under assumptions 1–4 and 7, for $\lambda_n \propto \sigma\sqrt{\log(p)/n}$, there is a constant $M > 0$ such that, with probability tending to 1 for $n \to \infty$,

$$\hat{s}_{\mathrm{L}} \leqslant Ms,$$
$$\|\hat{\boldsymbol{\beta}}_{\mathrm{L}} - \boldsymbol{\beta}_0\|_1 \leqslant M\sigma s\sqrt{\left\{\frac{\log(p)}{n}\right\}},$$

and

$$\|\mathbf{X}(\hat{\boldsymbol{\beta}}_{\mathrm{L}} - \boldsymbol{\beta}_0)\|_2^2 \leqslant M\sigma^2 s \log(p).$$

A.3. Proof of theorem 3

$(n - \hat{s}_{\mathrm{L}})(\hat{\sigma}^2_{\mathrm{L}} - \sigma^2)$ can be decomposed as

$$\begin{aligned}(n - \hat{s}_{\mathrm{L}})(\hat{\sigma}^2_{\mathrm{L}} - \sigma^2) &= (\varepsilon^{\mathrm{T}}\varepsilon - n\sigma^2) - 2\varepsilon^{\mathrm{T}}\mathbf{X}(\hat{\boldsymbol{\beta}}_{\mathrm{L}} - \beta_0) + \|\mathbf{X}(\hat{\boldsymbol{\beta}}_{\mathrm{L}} - \beta_0)\|_2^2 \\ &= R_1 + R_2 + R_3.\end{aligned}$$

The classical central limit theorem yields $R_1 = O_P(n^{1/2})$. Note that

$$|R_2| \leqslant 2\|\mathbf{X}^{\mathrm{T}}\varepsilon\|_\infty \|\hat{\boldsymbol{\beta}}_{\mathrm{L}} - \beta_0\|_1.$$

By equation (21) and lemma 1, it follows that

$$|R_2| = O_P[\sqrt{\{n\log(p)\}}]\; O_P\{s\sqrt{\log(p)/n}\} = O_P\{s\log(p)\}.$$

In addition, by the third conclusion in lemma 1, $|R_3| = O_P\{s\log(p)\}$. Therefore, the conclusion holds.

A.4. Proof of theorem 4

Let $\hat{\boldsymbol{\beta}}^{\mathrm{o}} = (\hat{\boldsymbol{\beta}}_1^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}$ with $\hat{\boldsymbol{\beta}}_1 = (\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1)^{-1}\mathbf{X}_1^{\mathrm{T}}\mathbf{y}$ be the oracle estimator. The key step is to show that, with probability tending to 1, the oracle estimator $\hat{\boldsymbol{\beta}}^{\mathrm{o}}$ is a strictly local minimizer of $\mathbf{Q}_{n,\lambda_n}(\boldsymbol{\beta})$ defined by equation (16). To prove it, by theorem 1 of Fan and Lv (2011), it suffices to show that, with probability tending to 1, $\hat{\boldsymbol{\beta}}^{\mathrm{o}}$ satisfies

$$\mathbf{X}_1^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathrm{o}}) - n\,\tilde{\boldsymbol{\rho}}_{\lambda_n}(\hat{\boldsymbol{\beta}}_1) = 0, \tag{23}$$

$$\|\mathbf{X}_2^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathrm{o}})\|_\infty < n\,\rho'_{\lambda_n}(0+), \tag{24}$$

$$\lambda_{\min}\left(\frac{1}{n}\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1\right) > \kappa_{\lambda_n}(\hat{\boldsymbol{\beta}}_1), \tag{25}$$

where $\tilde{\boldsymbol{\rho}}_{\lambda_n}(\hat{\boldsymbol{\beta}}_1) = (\mathrm{sgn}(\hat{\beta}_1)\rho'_{\lambda_n}(|\hat{\beta}_1|), \ldots, \mathrm{sgn}(\hat{\beta}_s)\rho'_{\lambda_n}(|\hat{\beta}_s|))^{\mathrm{T}}$ and $\kappa_{\lambda_n}(\hat{\boldsymbol{\beta}}_1) = \max_{j=1,\ldots,s}\{-\rho''_{\lambda_n}(|\hat{\beta}_j|)\}$.

Let $\boldsymbol{\xi}_1 = \mathbf{X}_1^{\mathrm{T}}\varepsilon$ and $\boldsymbol{\xi}_2 = \mathbf{X}_2^{\mathrm{T}}\varepsilon$. Consider the events

$$\mathcal{A}_{n1} = \{\|\boldsymbol{\xi}_1\|_\infty \leqslant \sqrt{[n\log(n)\log\{\log(n)\}]}\} \cap \left\{\lambda_{\min}\left(\frac{1}{n}\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1\right) \geqslant c_1\right\}$$

and

$$\mathcal{A}_{n2} = \{\|\boldsymbol{\xi}_2\|_\infty \leqslant \sqrt{[n^{\alpha_0+1}\log\{\log(n)\}]}\} \cap \left\{\left\|\frac{1}{n}\mathbf{X}_2^{\mathrm{T}}\mathbf{X}_1\right\|_{2,\infty} \leqslant c_2\right\}.$$

Observe that $\hat{\boldsymbol{\beta}}_1=(\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1)^{-1}\mathbf{X}_1^{\mathrm{T}}\mathbf{y}$. Then, we obtain $\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_{01}=(\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1)^{-1}\mathbf{X}_1^{\mathrm{T}}\boldsymbol{\varepsilon}$ and hence, under the event $\mathcal{A}_{n1}$,

$$\begin{aligned}\|\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_{01}\|_\infty &\leqslant \|\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_{01}\|_2\\ &\leqslant \left\|\left(\frac{1}{n}\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1\right)^{-1}\right\|_2 \left\|\frac{1}{n}\mathbf{X}_1^{\mathrm{T}}\boldsymbol{\varepsilon}\right\|_2\\ &\leqslant \lambda_{\min}\left(\frac{1}{n}\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1\right)^{-1}\Big/\sqrt{s}\,\left\|\frac{1}{n}\mathbf{X}_1^{\mathrm{T}}\boldsymbol{\varepsilon}\right\|_\infty\\ &\leqslant c\sqrt{[\log(n)\log\{\log(n)\}/n^{1-\alpha_0}]}\ll\lambda_n,\end{aligned}$$

for some constant c not depending on n. Note that, in the above inequalities, we use the facts $s=O(n^{\alpha_0})$ and $\lambda_n\propto n^{-(1-\alpha_0)/2}\log(n)$.

Since $d_n=\frac{1}{2}\min_{j=1,\dots,s}|\beta_{0j}|\geqslant n^{-\gamma}\log(n)$ with $\gamma\in(0,\frac{1}{2}]$ and $d_n\gg\lambda_n$, as addressed in assumption 9, we have, under the event $\mathcal{A}_{n1}$,

$$\begin{aligned}\min_{j=1,\dots,s}|\hat{\beta}_j| &\geqslant \min_{j=1,\dots,s}|\beta_{0j}|-\|\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_{01}\|_\infty\\ &\geqslant 2d_n-c\sqrt{[\log(n)\log\{\log(n)\}/n^{1-\alpha_0}]}\\ &\geqslant d_n\gg\lambda_n\end{aligned}$$

for sufficiently large n. As a result, this leads to $\tilde{\boldsymbol{\rho}}_{\lambda_n}(\hat{\boldsymbol{\beta}}_1)=\mathbf{0}$ and $\kappa_{\lambda_n}(\hat{\boldsymbol{\beta}}_1)=0$ and hence implies that conditions (23) and (25) hold under the event $\mathcal{A}_{n1}$.

Now turn to prove the inequality (24). Under the event $\mathcal{A}_{n1}\cap\mathcal{A}_{n2}$, we have

$$\begin{aligned}\left\|\frac{1}{n}\mathbf{X}_2^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}^{\mathrm{o}})\right\|_\infty &\leqslant \frac{1}{n}\|\boldsymbol{\xi}_2\|_\infty+\frac{1}{n}\|\mathbf{X}_2^{\mathrm{T}}\mathbf{X}_1\|_{2,\infty}\|\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_{01}\|_2 \qquad (26)\\ &\leqslant \sqrt{[n^{\alpha_0-1}\log\{\log(n)\}]}+c_2c\sqrt{[\log(n)\log\{\log(n)/n^{1-\alpha_0}\}]}\\ &\propto \lambda_n(\sqrt{\log\{\log(n)\}}/\log(n)+c_2c\sqrt{[\log\{\log(n)\}/\log(n)]})\\ &\leqslant \frac{1}{2}\lambda_n<\rho'_{\lambda_n}(0+)\end{aligned}$$

for sufficiently large n. This shows that inequality (24) holds for sufficiently large n under the event $\mathcal{A}_{n1}\cap\mathcal{A}_{n2}$. By taking $c=\sqrt{\log\{\log(n)\}}$, similar arguments to those for theorem 1 lead to

$$P(\mathcal{A}_{n1}\cap\mathcal{A}_{n2})\rightarrow 1$$

as $n\rightarrow\infty$. Thus, we have proven that $\hat{\boldsymbol{\beta}}^{\mathrm{o}}$ is a strictly local minimizer of $\mathbf{Q}_{n,\lambda_n}(\boldsymbol{\beta})$ with large probability tending to 1. Consequently, $\hat{\boldsymbol{\beta}}_{\mathrm{SCAD}}=\hat{\boldsymbol{\beta}}^{\mathrm{o}}$.

Now consider the asymptotic distribution of $\hat{\sigma}^2_{\mathrm{SCAD}}-\sigma^2$. Observe that $\hat{\boldsymbol{\beta}}_1=(\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1)^{-1}\mathbf{X}_1^{\mathrm{T}}\mathbf{y}$. Under the event $\mathcal{A}_{n1}\cap\mathcal{A}_{n2}$,

$$\hat{\sigma}^2_{\mathrm{SCAD}}-\sigma^2=\frac{1}{n-s}\boldsymbol{\varepsilon}^{\mathrm{T}}(\mathbf{I}_n-\mathbf{P}_{M_0})\boldsymbol{\varepsilon}-\sigma^2.$$

Hence, we have that

$$(\hat{\sigma}^2_{\mathrm{SCAD}}-\sigma^2)\sqrt{n}\overset{\mathcal{D}}{\rightarrow}N(0,E[\varepsilon^4]-\sigma^4),$$

which also implies that $\hat{\sigma}^2_{\mathrm{SCAD}}-\sigma^2=O_P(n^{-1/2})$. The proof is complete.

A.5. Proof of results (4) and (5)

Let $\Phi(\cdot)$ and $F_{n-2}(\cdot)$ be the cumulative distribution functions of the standard Gaussian and Student's t-distribution with $n-2$ degrees of freedom. For large u,

$$1-F_{n-2}(u)>1-\Phi(u)>\exp(-u^2).$$

Therefore, $u = \surd \log(p/c)$ satisfies $F_{n-2}(u) < \Phi(u) < 1 - c/p$. The classical result that $\{\xi_{nj}\}_{j=1}^{p}$ are IID t_{n-2}-distributions entails that

$$P\{\sup_{1 \leqslant j \leqslant p} (\xi_{nj}) > u\} = P[\sup_{1 \leqslant j \leqslant p} \{F_{n-2}(\xi_{nj})\} > F_{n-2}(u)]$$
$$= 1 - \{1 - F_{n-2}(u)\}^{p},$$

which, by the choice of u, is further bounded from below by

$$1 - (1 - c/p)^{p} \geqslant 1 - \exp(-c).$$

Note that $\gamma_{nj} = \xi_{nj}/(n - 2 + \xi_{nj}^{2})^{1/2}$ is strictly increasing. It follows that

$$P\left\{\sup_{1 \leqslant j \leqslant p} (\gamma_{nj}) > \frac{u}{(n - 2 + u^{2})^{1/2}}\right\} = P\left\{\sup_{1 \leqslant j \leqslant p} (\xi_{nj}) > u\right\} \geqslant 1 - \exp(-c).$$

Result (4) follows from the fact that, when $u^{2} \leqslant n + 2$,

$$\frac{u}{(n - 2 + u^{2})^{1/2}} < \frac{u}{\surd(2n)}.$$

We now derive the limiting distribution (5). For each $x > 0$,

$$P[\surd\{2 \log(p)\}\{\sup_{1 \leqslant j \leqslant p}(\xi_{nj}) - d_{p}\} < x] = P\left[\sup_{1 \leqslant j \leqslant p}(\xi_{nj}) < d_{p} + \frac{x}{\surd\{2 \log(p)\}}\right]$$
$$= \left\{1 - \int_{d_{p}+x/\surd\{2 \log(p)\}}^{\infty} f_{n-2}(t)\,\mathrm{d}t\right\}^{p}.$$

Therefore, it suffices to show

$$p \int_{d_{p}+x/\surd\{2 \log(p)\}}^{\infty} f_{n-2}(t)\,\mathrm{d}t \rightarrow \exp(-x). \tag{27}$$

Let $\nu = n - 2$. The following inequalities are helpful to verify the limit (27)

$$\frac{1}{\surd(2\pi)}\left(\frac{1}{t} - \frac{1}{t^{3}}\right)\exp\left(-\frac{t^{2}}{2}\right) \leqslant \int_{t}^{\infty} \phi(s)\,\mathrm{d}s \leqslant \int_{t}^{\infty} f_{\nu}(s)\,\mathrm{d}s \leqslant C(\nu)\frac{1}{t}\,\frac{\nu}{\nu - 1}\left(1 + \frac{t^{2}}{\nu}\right)^{-(\nu-1)/2}, \tag{28}$$

where

$$C(\nu) = \Gamma\left(\frac{\nu + 1}{2}\right) \bigg/ \surd(\nu\pi)\,\Gamma\left(\frac{\nu}{2}\right).$$

Substituting $t = d_{p} + x/\surd\{2 \log(p)\}$ into the inequalities (28), it is easy to verify that, under the condition $\log(p) = o(n^{1/2})$,

$$\exp(-x) + o(1) < p \int_{d_{p}+x/\surd\{2 \log(p)\}}^{\infty} f_{\nu}(t)\,\mathrm{d}t < \exp(-x) + o(1).$$

This proves limit (27) and hence result (5).

References

Bickel, P. J., Ritov, Y. and Tsybakov, A. (2009) Simultaneous analysis of lasso and dantzig selector. *Ann. Statist.*, **37**, 1705–1732.

Bunea, F., Tsybakov, A. and Wegkamp, M. (2007) Sparsity oracle inequalities for the lasso. *Electron. J. Statist.*, **1**, 169–194.

Candes, E. and Tao, T. (2007) The dantzig selector: statistical estimation when p is much larger than n (with discussion). *Ann. Statist.*, **35**, 2313–2351.

Chatterjee, A. and Lahiri, S. N. (2011) Bootstrapping lasso estimators. *J. Am. Statist. Ass.*, **106**, 608–625.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004) Least angle regression (with discussions). *Ann. Statist.*, **32**, 409–499.

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

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

Fan, J. and Lv, J. (2010) A selective overview of variable selection in high dimensional feature space. *Statist. Sin.*, **20**, 101–148.

Fan, J. and Lv, J. (2011) Non-concave penalized likelihood with NP-dimensionality. *IEEE Trans. Inform. Theor.*, to be published.

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

Fan, J., Samworth, R. and Wu, Y. (2009) Ultrahigh dimensional feature selection: beyond the linear model. *J. Mach. Learn. Res.*, **10**, 2013–2038.

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

Greenshtein, E. and Ritov, Y. (2004) Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, **10**, 971–988.

Han, X., Gu, W. and Fan, J. (2011) Control of the false discovery rate under arbitrary covariance dependence. *J. Am. Statist. Ass.*, to be published.

Kim, Y., Choi, H. and Oh, H.-S. (2008) Smoothly clipped absolute deviation on high dimensions. *J. Am. Statist. Ass.*, **103**, 1665–1673.

Knight, K. and Fu, W. (2000) Asymptotics for lasso-type estimators. *Ann. Statist.*, **28**, 1356–1378.

Kyung, M., Gill, J., Ghosh, M. and Casella, G. (2010) Penalized regression, standard errors and Bayesian lassos. *Baysn Anal.*, **5**, 369–412.

Lv, J. and Fan, Y. (2009) A unified approach to model selection and sparse recovery using regularized least squares. *Ann. Statist.*, **37**, 3498–3528.

Meier, L., van de Geer, S. and Bühlmann, P. (2008) The group lasso for logistic regression. *J. R. Statist. Soc.* B, **70**, 53–71.

Meinshausen, N., Meier, L. and Bühlmann, P. (2009) p-values for high-dimensional regression. *J. Am. Statist. Ass.*, **104**, 1671–1681.

Meinshausen, N. and Yu, B. (2009) LASSO-type recovery of sparse representations for high-dimensional data. *Ann. Statist.*, **37**, 246–270.

Park, T. and Casella, G. (2008) The Bayesian lasso. *J. Am. Statist. Ass.*, **103**, 681–686.

Wasserman, L. and Roeder, K. (2009) High dimensional variable selection. *Ann. Statist.*, **37**, 2178–2201.

van der Vaart, A. W. and Wellner, J. A. (1996) *Weak Convergence and Empirical Processes*. New York: Springer.

Ye, J. (1998) On measuring and correcting the effects of data mining and model selection. *J. Am. Statist. Ass.*, **93**, 120–131.

Zhang, C. H. and Huang, J. (2008) The sparsity and bias of the lasso selection in high-dimensional linear regression. *Ann. Statist.*, **36**, 1567–1594.

Zhao, S. and Li, Y. (2010) Principled sure independence screening for Cox models with ultra-high-dimensional covariates. *Preprint*.

Zhao, P. and Yu, B. (2006) On model selection consistency of lasso. *J. Mach. Learn. Res.*, **7**, 2541–2563.

Zou, H., Hastie, T. and Tibshirani, R. (2007) On the "degrees of freedom" of the lasso. *Ann. Statist.*, **35**, 2173–2192.

Zou, H. and Li, R. (2008) One-step sparse estimates in nonconcave penalized likelihood models (with discussion). *Ann. Statist.*, **36**, 1509–1533.