High-dimensional variable selection via tilting

Haeran Cho and Piotr Fryzlewicz [*]

September 2, 2011

Abstract

This paper considers variable selection in linear regression models where the number of covariates is possibly much larger than the number of observations. High dimensionality of the data brings in many complications, such as (possibly spurious) high correlations among the variables, which result in marginal correlation being unreliable as a measure of association between the variables and the response. We propose a new way of measuring the contribution of each variable to the response which takes into account high correlations among the variables in a data-driven way. The proposed *tilting* procedure provides an adaptive choice between the use of marginal correlation and *tilted correlation* for each variable, where the choice is made depending on the values of the hard-thresholded sample correlation of the design matrix. We study the conditions under which this measure can successfully discriminate between the relevant and the irrelevant variables and thus be used as a tool for variable selection. Finally, an iterative variable screening algorithm is constructed to exploit the theoretical properties of tilted correlation, and its good practical performance is demonstrated in a comparative simulation study.

1 Introduction

Inferring the relationship between the response and the explanatory variables in linear models is an extremely important and widely studied statistical problem, from the point of view of both practical applications and theory. In this work, we consider the following linear model:

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

where $\mathbf{y} = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$ is an n-vector of the response, $\mathbf{X} = (X_1, \ldots, X_p)$ is an $n \times p$ design matrix and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T \in \mathbb{R}^n$ is an n-vector of i.i.d. random errors.

Recent technological advances have led to the explosion of data across many scientific disciplines, where the dimensionality of the data p can be very large; examples can be found in genomics, functional MRI, tomography and finance, to name but a few. In such settings, difficulties arise in estimating the coefficient vector $\boldsymbol{\beta}$. Over the last two decades, substantial progress has been made in tackling this problem under the assumption that only a small number of variables actually contribute to the response, i.e., $\mathcal{S} = \{1 \le j \le p : \ \beta_j \neq 0\}$ is of cardinality $|\mathcal{S}| \ll p$. By identifying $\mathcal{S}$, we can improve both model interpretability and estimation accuracy.

[*] Department of Statistics, Columbia House, London School of Economics, Houghton Street, London, WC2A 2AE, UK. E-mail: {h.cho1, p.fryzlewicz}@lse.ac.uk

There exists a long list of literature devoted to the high-dimensional variable selection problem and an exhaustive survey can be found in Fan and Lv (2010). The Lasso (Tibshirani, 1996) belongs to a class of penalised least squares estimators where the penalty is on the l_1-norm of $\boldsymbol{\beta}$, which leads to a sparse solution by setting certain coefficients to be exactly zero. It has enjoyed considerable attention and substantial efforts in studying the consistency of the methodology and its extension can be found e.g. in Meinshausen and Bühlmann (2008), Zhang and Huang (2008), Zhao and Yu (2006), Zou (2006), Meinshausen and Bühlmann (2010).

Efron et al. (2004) proposed the Least Angle Regression (LARS) algorithm, which can be modified to compute the Lasso solution path for a range of penalty parameters. The main criterion for determining which variables should enter the model in the progression of the LARS algorithm is the screening of the marginal correlations between each variable and the current residual. That is, denoting the current residual by $\mathbf{z}$, the Lasso solution path is computed by taking a step of a suitably chosen size in the equiangular direction between those variables which achieve the maximum $|X_j^T\mathbf{z}|$ at each iteration. The Sure Independence Screening (SIS) proposed in Fan and Lv (2008) is a dimension reduction procedure, which screens the marginal correlations $X_j^T\mathbf{y}$ to choose which variables should remain in the model.

While the aforementioned methods show good theoretical properties as well as performing well in practice, we note that they heavily rely on marginal correlation to measure the strength of association between X_j and $\mathbf{y}$. Fan and Lv (2008) observed that, even when $X_1, \ldots, X_p$ were generated as i.i.d. Gaussian variables, there might exist spurious correlations among the variables with growing dimensionality p. In general, when there are non-negligible correlations among the variables, whether spurious or not, an irrelevant variable ($X_j,\ j \notin \mathcal{S}$) can have large marginal correlation with $\mathbf{y}$ due to its association with the relevant variables ($X_j,\ j \in \mathcal{S}$), which implies that marginal correlation can be misleading, especially if p is large.

There have been some efforts to introduce new measures of association between each variable and the response in order to deal with the issue of high correlations among the variables. Bühlmann et al. (2009) proposed the PC-simple algorithm, which uses partial correlation in order to infer the association between each variable and the response conditional on other variables. Also, we note that "greedy" algorithms such as the traditional forward selection (see e.g. Chapter 8.5 of Weisberg (1980)) or the forward regression (Wang, 2009) have an interpretation in this context due to their greediness (unlike less greedy algorithms generating a solution path, e.g. LARS). At each iteration, both forward selection and forward regression algorithms update the current residual $\mathbf{z}$ by taking the greediest step towards the variables included in the current model, i.e., $\mathbf{z}$ is obtained by projecting $\mathbf{y}$ onto the orthogonal complement of the current model space and this greedy progression can be seen as taking into account the correlations between those variables which are in the current model and those which are not. Radchenko and James (2011) proposed the forward-Lasso adaptive shrinkage (FLASH) which includes the Lasso and forward selection as special cases at two extreme ends. FLASH iteratively adds one variable at a time and adjusts each step size by introducing a new parameter so that their procedure is greedier than the Lasso, yet not as greedy as the forward selection. The regression framework proposed in Witten and Tibshirani (2009) accounts for correlations among the variables using the so-called "scout" procedure, which obtains a shrunken estimate of the inverse covariance matrix of $\mathbf{X}$ by maximising a penalised likelihood and then applies it to the estimation of $\boldsymbol{\beta}$. A more detailed description of the aforementioned methods, in comparison with our proposed methodology, is provided later in Section 3.3.

In this paper, we propose a new way of measuring the contribution of each variable to the response, which also accounts for the correlation structure among variables. It is accomplished by "tilting" each column X_j (so that it becomes X_j^*) such that the impact of other variables $X_k,\ k \neq j$ on the "tilted" correlation between X_j^* and $\mathbf{y}$ is reduced and thus the relationship between the jth covariate and the response can be identified more accurately. One key ingredient of this methodology, which sets it apart from other approaches listed above, is the adaptive choice of the set $\mathcal{C}_j$ of variables X_k whose impact on X_j is to be removed. Informally speaking, we note that $\mathcal{C}_j$ cannot include "too many" variables, as this would distort the association between the jth covariate and the response due to the large dimensionality p. However, we also observe that those X_k's which have low marginal correlations with X_j do not individually cause distortion in measuring this association anyway, so they can safely be omitted from the set $\mathcal{C}_j$. Therefore, it appears natural to include in $\mathcal{C}_j$ only those variables X_k whose correlations with X_j exceed a certain threshold in magnitude, and this hard thresholding step is an important element of our methodology.

Other key steps in our methodology are: projection of each variable onto a subspace chosen in the hard-thresholding step; and rescaling of such projected variables. We show that under certain conditions the tilted correlation can successfully discriminate between relevant and irrelevant variables and thus can be applied as a tool for variable selection. We also propose an iterative algorithm based on tilting and present its unique features in relation to the existing methods discussed above.

The remainder of the paper is organised as follows. In Section 2, we introduce the tilting procedure and study the theoretical properties of tilted correlation in various scenarios. Then in Section 3, we propose the TCS algorithm, which iteratively screens the tilted correlations to identify relevant variables, and compare it in detail to other existing methods. Section 4 reports the outcome of extensive comparative simulation studies and the performance of TCS algorithm is further demonstrated in Section 5 on a real world dataset predicting real estate prices. Section 6 concludes the paper and the proofs of theoretical results are in the Appendix.

2 Tilting: motivation, definition and properties

2.1 Notation and model description

For an n-vector $\mathbf{u} \in \mathbb{R}^n$, we define the l_1 and l_2-norms as $\|\mathbf{u}\|_1 = \sum_j |u_j|$ and $\|\mathbf{u}\|_2 = \sqrt{\sum_j u_j^2}$, and the latter is frequently referred to as the norm. Each column of $\mathbf{X}$ is assumed to have a unit norm, and thus the sample correlation matrix of $\mathbf{X}$ is defined as $\mathbf{C} = \mathbf{X}^T\mathbf{X} = (c_{j,k})_{j,k=1}^p$. We assume that $\epsilon_i,\ i = 1, \ldots, n$ are i.i.d. random noise following a normal distribution $\mathcal{N}(0, \sigma^2/n)$ with $\sigma^2 < \infty$, where the n^{-1} in the noise variance is required due to our normalisation of the columns of $\mathbf{X}$. We denote the ith row of $\mathbf{X}$ as $\mathbf{x}_i = (X_{i,1}, \ldots, X_{i,p})$. Let $\mathcal{D}$ denote a subset of the index set $\mathcal{J} = \{1, \ldots, p\}$. Then $\mathbf{X}_{\mathcal{D}}$ denotes an $n \times |\mathcal{D}|$-submatrix of $\mathbf{X}$ with $X_j,\ j \in \mathcal{D}$ as its columns for any $n \times p$ matrix $\mathbf{X}$. In a similar manner, $\boldsymbol{\beta}_{\mathcal{D}}$ denotes a $|\mathcal{D}|$-subvector of a p-vector

$\boldsymbol{\beta}$ with $\beta_j,\ j \in \mathcal{D}$ as its elements. For a given submatrix $\mathbf{X}_{\mathcal{D}}$, we denote the projection matrix onto the column space of $\mathbf{X}_{\mathcal{D}}$ by $\Pi_{\mathcal{D}}$. Finally, C and C' are used to denote generic positive constants.

2.2 Tilting: motivation and definition

In this section, we introduce the procedure of "tilting" a variable and define the *tilted correlation* between each variable and the response. We first list typical difficulties encountered in high-dimensional problems, which were originally pointed out in Fan and Lv (2008).

(a) Irrelevant variables which are highly correlated with the relevant ones can have high priority to be selected in marginal correlation screening.

(b) A relevant variable can be marginally uncorrelated but jointly correlated with the response.

(c) Collinearity can exist among the variables, i.e., $|c_{j,k}| = |X_j^T X_k|$ for $j \neq k$ can be close to 1.

We note that the marginal correlation between each variable X_j and $\mathbf{y}$ has the following decomposition,

$$X_j^T \mathbf{y} = X_j^T \left(\sum_{k=1}^{p} \beta_k X_k + \boldsymbol{\epsilon} \right) = \beta_j + \underline{\underline{\sum_{k \in \mathcal{S} \setminus \{j\}} \beta_k X_j^T X_k}} + X_j^T \boldsymbol{\epsilon}, \tag{2}$$

which shows that the issues (a) and (b) arise from the underlined summand in (2). The main idea behind tilting is to transform each X_j in such a way that the corresponding underlined summand for the transformed X_j is zero or negligible, while not distorting the contribution of the jth covariate to the response. By examining the form of the underlined summand and viewing it as a "bias" term, it is apparent that its components are particularly large for those k's for which the corresponding term $X_j^T X_k$ is large. If we were to transform X_j by projecting it on the space orthogonal to those X_k's, a corresponding bias term for a thus-transformed X_j would be significantly reduced.

For each X_j, denote the set of such X_k's by $\mathcal{C}_j$. Without prior knowledge of $\mathcal{S}$, one way of selecting $\mathcal{C}_j$ for each X_j is to identify those variables $X_k,\ k \neq j$ which have non-negligible correlations with X_j. A careful choice of $\mathcal{C}_j$ is especially important when the dimensionality p is high; when $\mathcal{C}_j$ is chosen to include too many variables, any vector in $\mathbb{R}^n$ may be well-approximated by $X_k,\ k \in \mathcal{C}_j$, which would result in the association between the transformed X_j and $\mathbf{y}$ failing to reflect the true contribution of the jth covariate to the response. Intuitively, those X_k's having small sample correlations with X_j do not significantly contribute to the underlined bias term, and thus can be safely omitted from the set $\mathcal{C}_j$. Below, we propose a procedure for selecting $\mathcal{C}_j$ adaptively for each j, depending on the sample correlation structure of $\mathbf{X}$.

We first find $\pi_n \in (0, 1)$ which will act as a threshold on each off-diagonal entry $c_{j,k},\ j \neq k$ of the sample correlation matrix $\mathbf{C}$ of $\mathbf{X}$, identifying whether the sample correlation between X_j and X_k is non-negligible. Then, the subset $\mathcal{C}_j$ is identified as $\mathcal{C}_j = \{k \neq j : \ |X_j^T X_k| = |c_{j,k}| > \pi_n\}$

separately for each variable X_j. We note that although the subset $\mathcal{C}_j$ is obviously different for each j, the thresholding procedure for selecting it is always the same. Our procedure for selecting π_n itself is described in Section 3.4. Tilting a variable X_j is defined as the procedure of projecting X_j onto the orthogonal complement of the space spanned by $X_k,\ k \in \mathcal{C}_j$, which reduces to zero the impact of those X_k's on the association between the projected version of X_j and $\mathbf{y}$.

Hard-thresholding was previously adopted for the estimation of a high-dimensional covariance matrix, although we emphasise that this was not in the context of variable selection. In Bickel and Levina (2008), an estimator obtained by hard-thresholding the sample covariance matrix was shown to be consistent with the choice of $C\sqrt{\log p/n}$ as the threshold, provided the covariance matrix was appropriately sparse and the dimensionality p satisfied $\log p/n \to 0$. A similar result was reported in El Karoui (2008) with the threshold of magnitude $Cn^{-\gamma}$ for some $\gamma \in (0, 1/2)$. Our theoretical choice of threshold π_n is described in Section 2.3, where we also briefly compare it to the aforementioned thresholds. In practice, we choose π_n by controlling the false discovery rate, as presented in Section 3.4.

Let $\tilde{\mathbf{X}}_j$ denote a submatrix of $\mathbf{X}$ with $X_k,\ k \in \mathcal{C}_j$ as its columns, and Π_j the projection matrix onto the space spanned by $X_k,\ k \in \mathcal{C}_j$, i.e., $\Pi_j \equiv \tilde{\mathbf{X}}_j(\tilde{\mathbf{X}}_j^T\tilde{\mathbf{X}}_j)^{-1}\tilde{\mathbf{X}}_j^T$. The tilted variable X_j^* for each X_j is defined as $X_j^* \equiv (\mathbf{I}_n - \Pi_j)X_j$. Then the correlation between the tilted variable X_j^* and $X_k,\ k \in \mathcal{C}_j$ is reduced to zero, and therefore such X_k's no longer have any impact on $(X_j^*)^T\mathbf{y}$. However, $(X_j^*)^T\mathbf{y}$ cannot directly be used as a measure of association between X_j and $\mathbf{y}$, since the norm of the tilted variable X_j^*, provided $\mathcal{C}_j$ is non-empty, satisfies $\|X_j^*\|_2 = X_j^T(\mathbf{I}_n - \Pi_j)X_j < X_j^TX_j = 1$. Therefore, we need to rescale $(X_j^*)^T\mathbf{y}$ so as to make it a reliable criterion for gauging the contribution of each X_j to $\mathbf{y}$.

Let a_j and a_{jy} denote the squared proportion of X_j and $\mathbf{y}$ (respectively) represented by $X_k,\ k \in \mathcal{C}_j$, i.e., $a_j \equiv \|\Pi_jX_j\|_2^2/\|X_j\|_2^2$ and $a_{jy} \equiv \|\Pi_j\mathbf{y}\|_2^2/\|\mathbf{y}\|_2^2$. We denote the tilted correlation between X_j and $\mathbf{y}$ with respect to a rescaling factor s_j by $c_j^*(s_j) \equiv s_j^{-1}\cdot(X_j^*)^T\mathbf{y}$, and propose two rescaling rules below.

Rescaling 1. Decompose $(X_j^*)^T\mathbf{y}$ as

$$\begin{aligned}(X_j^*)^T\mathbf{y} &= X_j^T(\mathbf{I}_n - \Pi_j)\mathbf{y} = X_j^T\left\{\sum_{k=1}^p \beta_k(\mathbf{I}_n - \Pi_j)X_k + (\mathbf{I}_n - \Pi_j)\boldsymbol{\epsilon}\right\} \\ &= \beta_jX_j^T(\mathbf{I}_n - \Pi_j)X_j + \sum_{k\in\mathcal{S}\setminus\mathcal{C}_j, k\neq j}\beta_kX_j^T(\mathbf{I}_n - \Pi_j)X_k + X_j^T(\mathbf{I}_n - \Pi_j)\boldsymbol{\epsilon}. \quad (3)\end{aligned}$$

Provided the second and third summands in (3) are negligible in comparison with the first, rescaling the inner product $(X_j^*)^T\mathbf{y}$ by $1 - a_j = X_j^T(\mathbf{I}_n - \Pi_j)X_j$ can "isolate" β_j, which amounts to the contribution of X_j to $\mathbf{y}$, in the sense that $(X_j^*)^T\mathbf{y}/(1 - a_j)$ can be represented as β_j plus a "small" term (our theoretical results later make this statement more precise). Motivated by this, we use the rescaling factor of $\lambda_j \equiv (1 - a_j)$ to define a rescaled version of X_j^* as $X_j^\bullet \equiv (1 - a_j)^{-1}\cdot X_j^*$ and the corresponding tilted correlation as $c_j^*(\lambda_j) = (1 - a_j)^{-1}\cdot(X_j^*)^T\mathbf{y} = (X_j^\bullet)^T\mathbf{y}$.

Rescaling 2. Since $\mathbf{I}_n - \Pi_j$ is also a projection matrix, we note that $(X_j^*)^T\mathbf{y}$ is equal to the inner product between $X_j^* = (\mathbf{I}_n - \Pi_j)X_j$ and $\mathbf{y}_j^* = (\mathbf{I}_n - \Pi_j)\mathbf{y}$, with their norms satisfying $\|X_j^*\|_2 = \sqrt{1-a_j}$ and $\|\mathbf{y}_j^*\|_2 = \sqrt{1-a_{jy}} \cdot \|\mathbf{y}\|_2$. By rescaling X_j^* and $\mathbf{y}_j^*$ by $\sqrt{1-a_j}$ and $\sqrt{1-a_{jy}}$ respectively, we obtain vectors $X_j^\circ \equiv (1-a_j)^{-1/2} \cdot X_j^*$ and $\mathbf{y}_j^\circ \equiv (1-a_{jy})^{-1/2} \cdot \mathbf{y}_j^*$, whose norms satisfy $\|X_j^\circ\|_2 = \|X_j\|_2$ and $\|\mathbf{y}_j^\circ\|_2 = \|\mathbf{y}\|_2$. Therefore, with the rescaling factor set equal to $\Lambda_j \equiv \{(1-a_j)(1-a_{jy})\}^{1/2}$, we define the tilted correlation as $c_j^*(\Lambda_j) = \{(1-a_j)(1-a_{jy})\}^{-1/2} \cdot (X_j^*)^T\mathbf{y} = (X_j^\circ)^T\mathbf{y}_j^\circ$.

We note that, with the rescaling factor λ_j (rescaling 1), the tilted correlation $c_j^*(\lambda_j)$ coincides with the ordinary least squares estimate of β_j when regressing $\mathbf{y}$ onto $X_k,\ k \in \mathcal{C}_j \cup \{j\}$. When rescaled by Λ_j (rescaling 2), the tilted correlation coincides with the sample partial correlation between X_j and $\mathbf{y}$ given $X_k,\ k \in \mathcal{C}_j$ (denoted by $\hat{\rho}_n(j, \mathbf{y}|\mathcal{C}_j)$), up to a constant multiplicative factor $\|\mathbf{y}\|_2$, i.e., $c_j^*(\Lambda_j) = \|\mathbf{y}\|_2 \cdot \hat{\rho}_n(j, \mathbf{y}|\mathcal{C}_j)$. Although partial correlation is also used in the PC-simple algorithm (Bühlmann et al., 2009), we emphasise that a crucial difference between tilting and PC-simple is that tilting makes an adaptive choice of the conditioning subset $\mathcal{C}_j$ for each X_j, as described earlier in this section. For a detailed discussion of this point, see Section 3.3. In what follows, whenever the tilted correlation is denoted by c_j^* without specifying the rescaling factor s_j, the relevant statement is valid for either of the rescaling factors λ_j and Λ_j. Finally, we note that if the set $\mathcal{C}_j$ turns out to be empty for a certain index j, then for such X_j, our tilted correlation with either rescaling factor would reduce to standard marginal correlation, which in this case is expected to work well (in measuring the association between the jth covariate and the response) due to the fact that no other variables X_k are significantly correlated with X_j. In summary, our proposed tilting procedure enables an adaptive choice between the use of marginal correlation and tilted correlation for each variable X_j, depending on the sample correlation structure of $\mathbf{X}$.

In the following section, we study some properties of tilted correlation and show that the corresponding properties do not always hold for marginal correlation. This prepares the ground for the algorithm proposed in Section 3.1 which adopts tilted correlation for variable screening.

2.3 Properties of the tilted correlation

In studying the theoretical properties of tilted correlation, we make the following assumptions on the linear model in (1).

(A1) The number of non-zero coefficients $|\mathcal{S}|$ satisfies $|\mathcal{S}| = O(n^\delta)$ for $\delta \in [0, 1/2)$.

(A2) The number of variables satisfies $\log p = O(n^\theta)$ with $\theta \in [0, 1-2\gamma)$ for $\gamma \in (\delta, 1/2)$.

(A3) With the same γ as in (A2), the threshold is chosen as $\pi_n = C_1 n^{-\gamma}$ for some $C_1 > 0$. We assume that there exists $C > 0$ such that $\mathcal{C}_j = \{k \neq j : \ |c_{j,k}| > \pi_n\}$ is of cardinality $|\mathcal{C}_j| \leq Cn^\xi$ uniformly over all j, where $\xi \in [0, 2(\gamma - \delta))$.

(A4) Non-zero coefficients satisfy $\max_{j \in \mathcal{S}} |\beta_j| < M$ for $M \in (0, \infty)$ and $n^\mu \cdot \min_{j \in \mathcal{S}} |\beta_j| \to \infty$ for $\mu \in [0, \gamma - \delta - \xi/2)$.

(A5) There exists $\alpha \in (0,1)$ satisfying $1 - X_j^T \Pi_j X_j = 1 - a_j > \alpha$ for all j.

(A6) For those j whose corresponding $\mathcal{C}_j$ satisfies $\mathcal{S} \nsubseteq \mathcal{C}_j$, we have

$$n^{\kappa} \cdot \frac{\|(\mathbf{I}_n - \Pi_j)\mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}}\|_2^2}{\|\mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}}\|_2^2} \to \infty,$$

for κ satisfying $\kappa/2 + \mu \in [0, \gamma - \delta - \xi/2)$.

In (A1) and (A2), we let the sparsity $|\mathcal{S}|$ and dimensionality p of the linear model grow with the sample size n. Intuitively, if some non-zero coefficients tend to zero too rapidly, identifying them as relevant variables is difficult. Therefore (A4) imposes a lower bound on the magnitudes of the non-zero coefficients, which still allows the minimum non-zero coefficient to decay to 0 as n grows. It also imposes an upper bound, which is needed to ensure that the ratio between the largest and smallest coefficients in absolute value does not grow too quickly with n.

We now clarify the rest of assumptions which are imposed on the correlation structure of $\mathbf{X}$, and compare them to related conditions in existing literature. It is common practice in high-dimensional variable selection literature to study the performance of proposed methods under some conditions on $\mathbf{X}$. For the Lasso, it was shown that the irrepresentable condition (Zhao and Yu, 2006), also referred to as the neighbourhood stability condition (Meinshausen and Bühlmann, 2008) on $\mathbf{X}$ was sufficient and almost necessary for consistent variable selection. This condition required that

$$\max_{j \notin \mathcal{S}} \left|\text{sign}(\boldsymbol{\beta}_{\mathcal{S}})^T (\mathbf{X}_{\mathcal{S}}^T \mathbf{X}_{\mathcal{S}})^{-1} \mathbf{X}_{\mathcal{S}}^T X_j\right| < 1,$$

which can roughly be interpreted as saying that the portion of the irrelevant variable $X_j,\ j \notin \mathcal{S}$, represented by relevant variables $\mathbf{X}_{\mathcal{S}}$ is bounded from above by 1. Zhang and Huang (2008) showed the variable selection consistency of Lasso under the sparse Riesz condition. It requires the existence of $C > 0$ for which the eigenvalues of $\mathbf{X}_{\mathcal{D}}^T \mathbf{X}_{\mathcal{D}}$ are bounded uniformly over any $\mathcal{D} \subset \mathcal{J}$ with $|\mathcal{D}| \leq C|\mathcal{S}|$. Candès and Tao (2007) showed the consistency of the Dantzig selector under the uniform uncertainty principle (UUP), which also similarly restricts the behaviour of the sparse eigenvalues of $\mathbf{X}_{\mathcal{D}}^T \mathbf{X}_{\mathcal{D}}$.

We note that the assumption (A3) is not directly comparable to the above conditions in the sense that it requires the number of highly correlated variables for each variable not to exceed a certain polynomial rate in n. This bound is needed in order to guarantee the existence of the projection matrix Π_j, as well as to prevent tilted correlations from being distorted by high dimensionality as explained in Section 2.2. We now give an example of when (A3) is satisfied. Suppose for instance that each observation $\mathbf{x}_i,\ i = 1, \ldots, n$ is independently generated from a multivariate normal distribution $\mathcal{N}_p(\mathbf{0}, \Sigma)$ with $\Sigma_{j,k} = \varphi^{|j-k|}$ for some $\varphi \in (-1, 1)$. Then using Lemma 1 in Kalisch and Bühlmann (2007), we have that

$$\mathbb{P}\left(\max_{j \neq k} |c_{j,k} - \Sigma_{j,k}| \leq C_2 n^{-\gamma}\right) \geq 1 - \frac{Cnp(p-1)}{2} \cdot \exp\left(-\frac{C_2(n-4)n^{-2\gamma}}{2}\right), \tag{4}$$

for some $C_2 \in (0, C_1)$ and $C > 0$. The right-hand side of (4) tends to 1, provided $\log p = O(n^{\theta})$ with $\theta \in [0, 1/2 - \gamma)$. Then (A3) holds with probability tending to 1 since $|c_{j,k}| \leq |\varphi|^{|j-k|} + C_2 n^{-\gamma} < \pi_n$ for $|j-k| \gg \log n$ ($|a_n| \gg |b_n|$ means $|a_n b_n^{-1}| \to \infty$). The choice of $\pi_n = C_1 n^{-\gamma}$ is in agreement with Bickel and Levina (2008) and El Karoui (2008) in the sense that their threshold is also greater than $n^{-1/2}$. However, as we describe in Section 3.4, our procedure requires a data-dependent, rather than a fixed threshold, and we propose to choose it by controlling the false discovery rate.

(A5) is required to rule out strong collinearity among the variables. From the fact that $1-a_j = \det\left(\mathbf{X}^T_{\mathcal{C}_j\cup\{j\}}\mathbf{X}_{\mathcal{C}_j\cup\{j\}}\right) / \det\left(\tilde{\mathbf{X}}_j^T\tilde{\mathbf{X}}_j\right)$, we can find a connection between (A5) and the condition requiring strict positive definiteness of the population covariance matrix of $\mathbf{X}$, which is often found in the variable selection literature including Fan and Li (2001), Bühlmann et al. (2009) and Zou (2006).

Further, we show in Appendix D that assumptions (A5) and (A6) are satisfied under a certain mild assumption on $\mathbf{X}$ and $\boldsymbol{\epsilon}$, also used e.g. in Wang (2009).

As far as variable selection is concerned, if the absolute values of tilted correlations for $j \in \mathcal{S}$ are markedly larger than those for $j \notin \mathcal{S}$, we can use the tilted correlations for the purpose of variable screening. Before studying the properties of the tilted correlation in details, we provide a simple example to throw light on the situations where tilted correlation screening is successful while marginal correlation is not. The following set-up is consistent with Condition 3 in Section 2.3.1: $p=3$, $\mathcal{S}=\{1,2\}$, noise is not present, $|c_{1,3}|$ and $|c_{2,3}|$ exceed the threshold. Then, even when $c_{1,2}, c_{1,3}, c_{2,3}$ and the non-zero coefficients β_1, β_2 are chosen so that the marginal correlation screening fails (i.e., $|X_3^T\mathbf{y}| > \max(|X_1^T\mathbf{y}|, |X_2^T\mathbf{y}|)$), it is still the case that $|(X_3^*)^T\mathbf{y}| = 0$ and thus tilted correlation screening can avoid picking up X_3 as relevant.

In the following Sections 2.3.1–2.3.3, we introduce different scenarios under which the tilted correlation screening (with either rescaling factor) achieves separation between relevant and irrelevant variables.

2.3.1 Scenario 1

In the first scenario, we assume the following condition on $\mathbf{X}$.

Condition 1. *There exists $C>0$ such that $\left|(\Pi_j X_j)^T X_k\right| \le Cn^{-\gamma}$ for all $j \in \mathcal{J}$ and $k \in \mathcal{S}\setminus\mathcal{C}_j,\ k \ne j$.*

This condition implies that when X_j is projected onto the space spanned by $X_l,\ l \in \mathcal{C}_j$, any $X_k \in \mathcal{S}$ which are not close to X_j (in the sense that $k \notin \mathcal{C}_j$) remain not "too close" to the projected X_j ($\Pi_j X_j$). In Appendix A.1, it is shown that Condition 1 holds asymptotically when each column X_j is generated independently as a random vector on a sphere of radius 1, which is the surface of the Euclidean ball $B_2^n = \left\{\mathbf{x} \in \mathbb{R}^n : \sum_{i=1}^n x_i^2 \le 1\right\}$. The following theorem states that, under Condition 1, the tilted correlations of the relevant variables dominate those of the irrelevant variables.

Theorem 1. *Under assumptions (A1)–(A6), if Condition 1 holds, then $\mathbb{P}(\mathcal{E}_1) \to 1$ where*

$$\mathcal{E}_1 = \left\{\frac{|c_k^*(s_k)|}{\min_{j\in\mathcal{S}}|c_j^*(s_j)|} \to 0 \text{ for all } k \notin \mathcal{S}\right\}, \tag{5}$$

regardless of the choice of the rescaling factor (that is, with $s_j = \lambda_j$ or $s_j = \Lambda_j$). On the event $\mathcal{E}_1$, the following holds.

- *$n^{\mu} \cdot c_j^* \to 0$ for $j \notin \mathcal{S}$.*
- *$n^{\mu} \cdot |c_j^*| \to \infty$ for $j \in \mathcal{S}$.*
- *With the rescaling 1, $c_j^*(\lambda_j)/\beta_j \to 1$ when $\beta_j \neq 0$.*

As noted in the Introduction, in high-dimensional problems, the maximum sample correlation of the columns of $\mathbf{X}$ can be non-negligible, even if the columns are generated as independent. Therefore marginal correlations $X_j^T\mathbf{y}$ for $j \in \mathcal{S}$ cannot be expected to have the same dominance over those for $j \notin \mathcal{S}$ as in (5).

2.3.2 Scenario 2

Let $\mathcal{K}$ denote a subset of $\mathcal{J}$ such that $X_k,\ k \in \mathcal{K}$ are either relevant ($k \in \mathcal{S}$) or highly correlated with at least one of relevant variables ($k \in \cup_{j\in\mathcal{S}}\mathcal{C}_j$). That is, $\mathcal{K} = \mathcal{S} \cup \{\cup_{j\in\mathcal{S}}\mathcal{C}_j\}$, and we impose the following condition on the sample correlation structure of $\mathbf{X}_{\mathcal{K}}$.

Condition 2. *For each $j \in \mathcal{S}$, if $k \in \mathcal{K} \setminus \{\mathcal{C}_j \cup \{j\}\}$, then $\mathcal{C}_k \cap \mathcal{C}_j = \emptyset$.*

In other words, this condition implies that for each relevant variable X_j, if $X_k,\ k \in \mathcal{K}$ is not highly correlated with X_j, there does not exist an $X_l,\ l \neq j,k$, which achieves sample correlations greater than the threshold π_n with both X_j and X_k simultaneously.
Suppose that the sample correlation matrix of $\mathbf{X}_{\mathcal{K}}$ is "approximately bandable", i.e., $|c_{j,k}| > \pi_n$ for any $j,k \in \mathcal{K}$ satisfying $|j-k| \leq B$ and $|c_{j,k}| < \pi_n$ otherwise, with the band width B satisfying $B|\mathcal{S}|^2/p \to 0$. Then, if $\mathcal{S}$ is selected randomly from $\mathcal{J}$ with each $j \in \mathcal{J}$ having equal probability to be included in $\mathcal{S}$, Condition 2 holds with probability bounded from below by

$$\left(1 - \frac{4B}{p-1}\right) \cdot \left(1 - \frac{8B}{p-2}\right) \cdots \left(1 - \frac{4(|\mathcal{S}|-1)B}{p-|\mathcal{S}|+1}\right) \geq \left(1 - \frac{4|\mathcal{S}|B}{p-|\mathcal{S}|+1}\right)^{|\mathcal{S}|-1} \to 1.$$

Another example satisfying Condition 2 is when each column of $\mathbf{X}_{\mathcal{K}}$ is generated as a linear combination of common factors in such a way that every off-diagonal element of the sample correlation matrix of $\mathbf{X}_{\mathcal{K}}$ exceeds the threshold π_n.
Under this condition, we can derive a similar result as in Scenario 1, with the dominance of the tilted correlations for relevant variables restricted within $\mathcal{K}$.

Theorem 2. *Under (A1)–(A6), if Condition 2 holds, then $\mathbb{P}(\mathcal{E}_2) \to 1$ where*

$$\mathcal{E}_2 = \left\{ \frac{|c_k^*(s_k)|}{\min_{j\in\mathcal{S}} |c_j^*(s_j)|} \to 0 \text{ for all } k \in \mathcal{K} \setminus \mathcal{S} \right\},$$

regardless of the choice of the rescaling factor (that is, with $s_j = \lambda_j$ or $s_j = \Lambda_j$). On the event $\mathcal{E}_2$, the following holds.

- *$n^{\mu} \cdot c_j^* \to 0$ for $j \notin \mathcal{S}$.*
- *$n^{\mu} \cdot |c_j^*| \to \infty$ for $j \in \mathcal{S}$.*
- *With the rescaling 1, $c_j^*(\lambda_j)/\beta_j \to 1$ when $\beta_j \neq 0$.*

2.3.3 Scenario 3

Finally, we consider a case when $\mathbf{X}$ satisfies a condition weaker than Condition 2.

Condition 3. *(C1) For each $j \in \mathcal{S}$, if $k \in \mathcal{K} \setminus \{\mathcal{C}_j \cup \mathcal{S}\}$, then $\mathcal{C}_k \cap \mathcal{C}_j = \emptyset$.*

(C2) The marginal correlation between $X_j^* = (\mathbf{I}_n - \Pi_j)X_j$ for $j \in \mathcal{S}$ and $\mathbb{E}\mathbf{y} = \mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}}$ satisfies $n^{\mu} \cdot \inf_{j \in \mathcal{S}} \left|(X_j^*)^T \mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}}\right| \to \infty$.

It is clear that Condition 2 is stronger than (C1), as the latter does not impose any restriction between $\mathcal{C}_j$ and $\mathcal{C}_k$ if both $j, k \in \mathcal{S}$. Bühlmann et al. (2009) placed a similar lower bound as that in (C2) on the population partial correlation $\rho_n(j, \mathbf{y}|\mathcal{D})$ of relevant variables $X_j,\ j \in \mathcal{S}$ for *any* subset $\mathcal{D} \subset \mathcal{J} \setminus \{j\}$ satisfying $|\mathcal{D}| \le |\mathcal{S}|$. Combined with the assumptions (A4)–(A5), (C2) rules out an ill-posed case where the parameters $\beta_j,\ j \in \mathcal{S}$ take values which cancel out the "tilted covariance" among the relevant variables (this statement is explained more precisely in the proof of Theorem 3). It is shown in Appendix C that Condition 3 is satisfied if Condition 2 holds and thus Condition 3 is indeed weaker than Condition 2. With Condition 3, we can show similar results to those in Theorem 2.

Theorem 3. *Under (A1)–(A6), if Condition 3 holds, then $\mathbb{P}(\mathcal{E}_3) \to 1$ where*

$$\mathcal{E}_3 = \left\{ \frac{|c_k^*(s_k)|}{\min_{j \in \mathcal{S}} |c_j^*(s_j)|} \to 0 \text{ for all } k \in \mathcal{K} \setminus \mathcal{S} \right\},$$

regardless of the choice of the rescaling factor (that is, with $s_j = \lambda_j$ or $s_j = \Lambda_j$). On the event $\mathcal{E}_3$, the following holds.

- *$n^{\mu} \cdot c_j^* \to 0$ for $j \notin \mathcal{S}$.*
- *$n^{\mu} \cdot |c_j^*| \to \infty$ for $j \in \mathcal{S}$.*

In contrast to Scenario 2, tilted correlations $c_j^*(\lambda_j)$ no longer necessarily converge to β_j as $n \to \infty$ in this scenario.
In the next section, we use the theoretical properties of tilted correlations derived in this section to construct a variable screening algorithm.

3 Application of tilting

Recalling issues (a)–(c) listed at the beginning of Section 2 which are typically encountered in high-dimensional problems, it is clear that tilting is specifically designed to tackle the occurrence of (a) and (b). First turning to (a), for an irrelevant variable X_j which attains high marginal correlation with $\mathbf{y}$ due to its high correlations with relevant variables $X_k,\ k \in \mathcal{C}_j \cap \mathcal{S}$, the impact of those high correlations is reduced to 0 in the tilted correlation of X_j and $\mathbf{y}$, and thus tilted correlation provides a more accurate measure of its association with $\mathbf{y}$, as demonstrated in our theoretical results of the previous section. Similar arguments apply to (b), where tilting is capable of fixing *low* marginal correlations between *relevant variables* and $\mathbf{y}$. (As for (c), it is common practice to impose assumptions which rule out strong collinearity among variables, and we have also followed this route.) In what follows, we present an algorithm, specifically constructed to exploit our theoretical study in Section 2.3 by iteratively applying the tilting procedure.

3.1 Tilted correlation screening algorithm

In Scenario 3, under a relatively weaker condition than those in Scenarios 1–2, it is shown that the tilted correlations of relevant variables dominate those of irrelevant variables within $\mathcal{K} = \mathcal{S} \cup (\cup_{j \in \mathcal{S}} \mathcal{C}_j)$. Even though $\mathcal{K}$ is unknown in practice, we can exploit the theoretical results by iteratively screening both marginal correlations and tilted correlations within a specifically chosen subset of variables.

When every off-diagonal entry of the sample correlation matrix is small, marginal correlation screening can be used as a reliable way of measuring the strength of association between each X_j and $\mathbf{y}$, and indeed, c_j^* for the variable X_j with an empty $\mathcal{C}_j$ is equal to the marginal correlation $X_j^T \mathbf{y}$, with either choice of the rescaling factor s_j. Therefore if a variable X_j with $\mathcal{C}_j = \emptyset$ achieves the maximum marginal correlation, such X_j is likely to be relevant. On the other hand, if $\mathcal{C}_j \neq \emptyset$, then high marginal correlation between X_j and $\mathbf{y}$ may have resulted from the high correlations of X_j with $X_k,\ k \in \mathcal{C}_j \cap \mathcal{S}$, even when $j \notin \mathcal{S}$. In this case, by screening the tilted correlations of $X_k,\ k \in \mathcal{C}_j \cup \{j\}$, we can choose the variable attaining the maximum $|c_k^*|$ as a relevant variable. In either case, one variable is selected and added to the *active set* $\mathcal{A}$ which represents the currently chosen model. The next step is to update the linear model by projecting it onto the orthogonal complement of the current model space $\mathbf{X}_\mathcal{A}$, i.e.,

$$(\mathbf{I}_n - \Pi_\mathcal{A})\mathbf{y} = (\mathbf{I}_n - \Pi_\mathcal{A})\mathbf{X}\boldsymbol{\beta} + (\mathbf{I}_n - \Pi_\mathcal{A})\boldsymbol{\epsilon}. \tag{6}$$

With the updated response and design matrix, we iteratively continue the above screening procedure. Below we present the algorithm which is referred to as the tilted correlation screening algorithm (TCS algorithm) throughout the paper.

Step 0 Start with an empty active set $\mathcal{A} = \emptyset$, current residual $\mathbf{z} = \mathbf{y}$, and current design matrix $\mathbf{Z} = \mathbf{X}$.

Step 1 Find the variable which achieves the maximum marginal correlation with $\mathbf{z}$ and let $k = \arg\max_{j \notin \mathcal{A}} |Z_j^T \mathbf{z}|$. Identify $\mathcal{C}_k = \{j \notin \mathcal{A}, j \neq k :\ |Z_k^T Z_j| > \pi_n\}$ and if $\mathcal{C}_k = \emptyset$, let $k^* = k$ and go to Step 3.

Step 2 If $\mathcal{C}_k \neq \emptyset$, screen the tilted correlations c_j^* between Z_j and $\mathbf{z}$ for $j \in \mathcal{C}_k \cup \{k\}$ and find $k^* = \arg\max_{j \in \mathcal{C}_k \cup \{k\}} |c_j^*|$.

Step 3 Add k^* to $\mathcal{A}$ and update the current residual and the current design matrix $\mathbf{z} \leftarrow (\mathbf{I}_n - \Pi_\mathcal{A})\mathbf{y}$ and $\mathbf{Z} \leftarrow (\mathbf{I}_n - \Pi_\mathcal{A})\mathbf{X}$, respectively. Further, rescale each column $j \notin \mathcal{A}$ of $\mathbf{Z}$ to have norm one.

Step 4 Repeat Steps 1–3 until the cardinality of active set $|\mathcal{A}|$ reaches a pre-specified $m < n$.

We note that Theorems 1–3 do not guarantee the selection consistency of the TCS algorithm itself. However, they do demonstrate a certain 'separation' property of the tilted correlation (as a measure of association). Steps 1–2 of the above algorithm exploit this property in the sense that they attempt to "operate" within the set $\mathcal{K}$ (which is unknown without the knowledge of $\mathcal{S}$), since we either directly choose a variable indexed k which is believed to lie in the set $\mathcal{S}$ or screen its corresponding set $\mathcal{C}_k$ (recall that $\mathcal{K} = \mathcal{S} \cup \{\cup_{j \in \mathcal{S}} \mathcal{C}_j\}$).

In Step 4, we need to specify m which acts as a stopping index in the TCS algorithm. The TCS algorithm iteratively builds a solution path of the active set $\mathcal{A}_{(1)} \subset \cdots \subset \mathcal{A}_{(m)} = \mathcal{A}$, and the final model $\hat{\mathcal{S}}$ can be chosen as either one of the sub-models $\mathcal{A}_{(i)}$ or a subset of $\mathcal{A}$. We discuss the selection of the final model $\hat{\mathcal{S}}$ in Section 3.2. In the simulation study, we used $m = \lfloor n/2 \rfloor$, which was an empirical choice made in order to ensure that the projections performed in the algorithm were numerically stable, while a sufficiently large number of variables were selected in the final model, if necessary. In practice however, if the TCS algorithm combined with the chosen model selection criterion returned m variables (i.e. if it reached the maximum permitted number of active variables), we would advise re-running the TCS algorithm with the limit of m slightly raised, until the number of final active variables was less than the current value of m.

During the application of the TCS algorithm, the linear regression model (1) is updated in Step 3 by projecting both $\mathbf{y}$ and $\mathbf{X}$ onto the orthogonal complement of the current model space spanned by $\mathbf{X}_{\mathcal{A}}$. Therefore, with a non-empty active set $\mathcal{A}$, it is interesting to observe that the tilted correlation c_j^* measures the association between X_j and $\mathbf{y}$ conditional on both the current model $X_k,\ k \in \mathcal{A}$ and the following subset of variables adaptively chosen for each $j \notin \mathcal{A}$,

$$\mathcal{C}_{j|\mathcal{A}} = \{k \notin \mathcal{A}, k \neq j :\ \hat{\rho}_n(j, k|\mathcal{A}) > \pi_n\}, \tag{7}$$

where $\hat{\rho}_n(j, k|\mathcal{A})$ denotes the sample partial correlation between X_j and X_k conditional on $\mathbf{X}_{\mathcal{A}}$.

Finally, we discuss the computational cost of the TCS algorithm. When $p \gg n$, the computational complexity of the algorithm is dominated by the computation of the threshold at Step 1, which is $O(np + np^2 + p^2 \log p + p^2) = O(np^2)$. Since the procedure is repeated m times, with m set to satisfy $m = O(n)$, the computational complexity of the entire algorithm is $O(n^2p^2)$, which is n times the cost of computing a $p \times p$ sample covariance matrix.

3.2 Final model selection

Once the size of the active set reaches a pre-specified value m, the final model $\hat{\mathcal{S}}$ needs to be chosen from $\mathcal{A}$. In this section, we present two methods which can be applied in our framework.

One of the most commonly used methods for model selection is cross-validation (CV), in which the observations would be divided into a training set and a test set such that the models returned after each iteration (i.e. $\mathcal{A}_{(1)} \subset \cdots \subset \mathcal{A}_{(m)} = \mathcal{A}$) could be tested using an appropriate error measure. However, we expect that for a CV-based method to work well, it would have to be computationally intensive: for example, a leave-one-out CV or a leave-half-out CV with averaging over different test and training sets.

One less computationally demanding option is to use e.g. an extended version of the Bayesian information criterion (BIC) proposed in Bogdan et al. (2004) and Chen and Chen (2008) as

$$\mathrm{BIC}(\mathcal{A}) = \log\left\{\frac{1}{n}\|(\mathbf{I}_n - \Pi_{\mathcal{A}})\mathbf{y}\|_2^2\right\} + \frac{|\mathcal{A}|}{n}(\log n + 2\log p). \tag{8}$$

This new BIC takes into account high dimensionality of the data by adding a penalty term dependent on p. Since the TCS algorithm generates a solution path which consists of m submodels $\mathcal{A}_{(1)} \subset \cdots \subset \mathcal{A}_{(m)} = \mathcal{A}$, we can choose our final model as $\hat{\mathcal{S}} = \mathcal{A}_{(m^*)}$ where $m^* = \arg\min_{1 \le i \le m} \mathrm{BIC}(\mathcal{A}_{(i)})$.

Chen and Chen (2008) showed the consistency of this new BIC under stronger conditions than those imposed in (A1), (A2) and (A4): the level of sparsity was $|\mathcal{S}| = O(1)$, the dimensionality was $p = O(n^C)$ for $C > 0$, and non-zero coefficients satisfied $\min_{j \in \mathcal{S}} |\beta_j| > C'$ for $C' > 0$. Then, under the asymptotical identifiability condition introduced in Chen and Chen (2008), (see (12) in Appendix D), the modified BIC as defined in (8) was shown to be consistent in the sense that

$$\mathbb{P}\left(\min_{|\mathcal{D}| \le m,\ \mathcal{D} \ne \mathcal{S}} \mathrm{BIC}(\mathcal{D}) > \mathrm{BIC}(\mathcal{S}) \right) \to 1 \text{ for } m \ge |\mathcal{S}|,$$

i.e., the probability of selecting any model other than $\mathcal{S}$ converged to zero. It was also noted that the original BIC was likely to fail when $p > n^{1/2}$. At the price of replacing $\log n/n$ with $n^{-\kappa}$ in (12), the consistency of the new BIC (8) can be shown with the level of sparsity growing with n as in (A1) and the dimensionality increasing exponentially with n as in (A2). The proof of this statement follows the exact line of proof in Chen and Chen (2008) and so we omit the details.

3.3 Relation to existing literature

We first note that our use of the term "tilting" is different from the use of the same term in Hall et al. (2009), where it applies to distance-based classification and denotes an entirely different procedure.

In the Introduction, we briefly discuss a list of existing variable selection techniques in which care is taken of the correlations among the variables in measuring the association between each variable and the response. Having now a complete picture of the TCS algorithm, we provide a more detailed comparison between our methodology and the aforementioned methods.

Bühlmann et al. (2009) proposed the PC-simple algorithm, which iteratively removes variables having small association with the response. Sample partial correlations $\hat{\rho}_n(j, \mathbf{y}|\mathcal{D})$ are used as the measure of association between X_j and $\mathbf{y}$, where $\mathcal{D}$ is *any* subset of the active set $\mathcal{A}$ (those variables still remaining in the model excluding X_j) with its cardinality $|\mathcal{D}|$ equal to the number of iterations taken so far. Behind the use of partial correlations lies the concept of partial faithfulness which implies that, at the population level, if $\rho_n(j, \mathbf{y}|\mathcal{D}) = 0$ for some $\mathcal{D} \subset \mathcal{J} \setminus \{j\}$, then $\rho_n(j, \mathbf{y}|\mathcal{J} \setminus \{j\}) = 0$. Their PC-simple algorithm starts with $\mathcal{A} = \mathcal{J}$ and iteratively repeats the following: (i) screening sample partial correlations $\hat{\rho}_n(j, \mathbf{y}|\mathcal{D})$ for all $j \in \mathcal{A}$ and for all $\mathcal{D}$ satisfying the cardinality condition, (ii) applying Fisher's Z-transform to test the null hypotheses $H_0 : \rho_n(j, \mathbf{y}|\mathcal{D}) = 0$, (iii) removing irrelevant variables from $\mathcal{A}$, until $|\mathcal{A}|$ falls below the number of iterations taken so far. Recalling the definition of the rescaling factor Λ_j, we can see the connection between $c_j^*(\Lambda_j)$ and $\hat{\rho}_n(j, \mathbf{y}|\mathcal{D})$, as both are (up to a multiplicative factor $\|\mathbf{y}\|_2$) partial correlations between X_j and $\mathbf{y}$ conditional on a certain subset of variables. However, a significant difference comes from the fact that the PC-simple algorithm takes all $\mathcal{D} \subset \mathcal{A} \setminus \{j\}$ with fixed $|\mathcal{D}|$ at each iteration, whereas our TCS algorithm adaptively selects $\mathcal{C}_j$

Table 1: Comparison of variable selection methods.

	TCS algorithm	PC-simple	FR	FS
Step 0	$\mathcal{A} = \emptyset$	$\mathcal{A} = \mathcal{J}$	$\mathcal{A} = \emptyset$	$\mathcal{A} = \emptyset$
action	one selected	multiple removed	one selected	one selected
conditioning set $\mathcal{D}$	$\mathcal{A} \cup \mathcal{C}_{j\|\mathcal{A}}$ $= \mathcal{A} \cup \{k \notin \mathcal{A}, k \neq j :$ $\|\hat{\rho}_n(j, k\|\mathcal{A})\| > \pi_n\}$	remaining variables, $\|\mathcal{D}\|$ fixed	current model $\mathcal{A}$	current model $\mathcal{A}$
rescaling	λ_j or Λ_j	Λ_j	λ_j	none

(or $\mathcal{C}_{j|\mathcal{A}}$ when $\mathcal{A} \neq \emptyset$) for each j. Also, while λ_j is also a valid rescaling factor in our tilted correlation methodology, partial correlations are by definition computed using Λ_j only.

As for the forward regression (Wang, 2009, FR) and the forward selection (FS), although the initial stage of the two techniques is simple marginal correlation screening, their progression has a new interpretation given a non-empty active set ($\mathcal{A} \neq \emptyset$). Both algorithms obtain the current residual $\mathbf{z}$ by projecting the response $\mathbf{y}$ onto the orthogonal complement of the current model space, i.e., $\mathbf{z} = (\mathbf{I}_n - \Pi_\mathcal{A})\mathbf{y}$. Therefore they also measure the association between each $X_j,\ j \notin \mathcal{A}$ and $\mathbf{y}$ conditional on the current model space spanned by $\mathbf{X}_\mathcal{A}$ and thus take into account the correlations between $X_j,\ j \notin \mathcal{A}$ and $X_j,\ j \in \mathcal{A}$. The difference between FR and FS comes from the fact that FR updates not only the current residual $\mathbf{z}$ but also the current design matrix as $\mathbf{Z} = (\mathbf{I}_n - \Pi_\mathcal{A})\mathbf{X}$ (as in Step 3 of the TCS algorithm). Therefore FR eventually screens the rescaled version of $X_j^T(\mathbf{I}_n - \Pi_\mathcal{A})\mathbf{y}$ with the rescaling factor defined similarly to λ_j, replacing $\mathcal{C}_j$ with $\mathcal{A}$, i.e., $X_j^T(\mathbf{I}_n - \Pi_\mathcal{A})X_j = 1 - X_j^T\Pi_\mathcal{A}X_j$. On the other hand, there is no rescaling step in FS and it screens the terms $X_j^T(\mathbf{I}_n - \Pi_\mathcal{A})\mathbf{y},\ j \notin \mathcal{A}$, themselves.

By contrast, we note that while both FR and FS apply straight marginal correlation at each stage of their progression, our method, if and as necessary, uses the tilted correlation, which provides an adaptive choice between the marginal correlation and conditional correlation, depending on the correlation structure of the current design matrix. Indeed, in the extreme case where $\pi_n = 1$ is used, we have $\mathcal{C}_j = \emptyset$ and therefore the TCS algorithm becomes identical to FR. Another crucial difference is as already mentioned above in the context of the PC-simple algorithm: the tilting algorithm employs an adaptive choice of the conditioning set, unlike FR and FS.

In conclusion, the TCS algorithm, the PC-simple algorithm, FR and FS share the common ingredient of measuring the contribution of each variable X_j to $\mathbf{y}$ conditional on certain other variables; however, there are also important differences between them, and Table 1 summarises this comparison. We emphasise yet again that the TCS algorithm is distinguished from the rest in its adaptive choice of the conditioning subset via hard-thresholding of the sample correlations among the variables. Also, we note that the theoretical results of Section 2.3 hold for *both* rescaling methods, while the other algorithms use only one of them (FR, PC-simple) or none (FS).

Finally, we note the relationship between the TCS algorithm and the covariance-regularised regression method proposed in Witten and Tibshirani (2009). A key difference between the two is that the TCS algorithm works with the sample marginal correlations among the variables whereas in the scout procedure, it is the conditional correlations among the variables (i.e.,

$\rho_n(j, k|\mathcal{J} \setminus \{j, k\}) \neq 0)$ that are subject to regularisation. Also, the scout procedure achieves such regularisation by maximising a penalised likelihood function rather than hard-thresholding, and the thus-obtained estimate of the covariance structure of $\mathbf{X}$ is applied to estimate $\boldsymbol{\beta}$, again by solving an optimisation problem. By contrast, the tilted correlation method uses the outcome from thresholding the sample correlation structure to compute the tilted correlations and select the variable with maximum tilted correlation in an iterative algorithm, and therefore does not involve any optimisation problems.

3.4 Choice of threshold

In this section, we discuss the practical choice of the unknown threshold π_n from the sample correlation matrix $\mathbf{C}$. Due to the lack of information on the correlation structure of $\mathbf{X}$ in general and the possibility of spurious sample correlation among the variables, a deterministic choice of π_n is not expected to perform well universally and we need a data-driven way of selecting a threshold. Bickel and Levina (2008) proposed a cross-validation method for this purpose, while El Karoui (2008) conjectured the usefulness of a procedure based on controlling the false discovery rate (FDR). Since our aim is different from the accurate estimation of the correlation matrix itself, we propose a threshold selection procedure which is a modified version of the approach taken in the latter paper. In the following, we assume that $\mathbf{X}$ is a realisation of a random matrix with each row generated as $\mathbf{x}_i \sim_{\text{i.i.d.}} (\mathbf{0}, \Sigma)$, where each diagonal element of Σ equals one.

The procedure is a multiple hypothesis testing procedure and thus requires p-values of the $d = p(p-1)/2$ hypotheses $H_0 : \ |\Sigma_{j,k}| = 0$ defined for all $j < k$. We propose to compute the p-values as follows. First, an n-vector with i.i.d. Gaussian entries is repeatedly generated p times, and sample correlations $\{r_{l,m} : \ 1 \leq l < m \leq p\}$ among those vectors are obtained as a reference. Then, the p-value for each null hypothesis $H_0 : \ |\Sigma_{j,k}| = 0$ is defined as $P_{j,k} = d^{-1} \cdot |\{r_{l,m} : \ 1 \leq l < m \leq p, \ |r_{l,m}| \geq |c_{j,k}|\}|$. The next step is to apply the testing technique proposed in Benjamini and Hochberg (1995) to control the false discovery rate. Denoting $P_{(1)} \leq \ldots \leq P_{(d)}$ as the ordered p-values, we find the largest i for which $P_{(i)} \leq i/d \cdot \nu^*$ and reject all $H_{(j)}, \ j = 1, \ldots, i$. Then $\hat{\pi}_{thr}$ is chosen as the absolute value of the correlation corresponding to $P_{(i)}$. FDR is controlled at level ν^* and we use $\nu^* = p^{-1/2}$ as suggested in El Karoui (2008). An extensive simulation study described below confirms good practical performance of the above threshold selection procedure. We also checked the sensitivity of our algorithm to the choice of threshold by applying a grid of thresholds in model (C) below. Apart from the threshold $\hat{\pi}_{thr}$ selected as above, we ran versions of our algorithm where $\hat{\pi}_{thr}$ was multiplied by the constant factors of $0.75, 0.9, 1.1, 1.25$ each time it was used. Performance of our algorithm was similar across the different thresholds, which provides evidence for robustness of our procedure to the choice of threshold within reason.

4 Simulation study

In this section, we compare the performance of the TCS algorithm on simulated data with that of other related methods discussed in the Introduction and Section 3.3, which are the

PC-simple algorithm, FR, FS, iterative SIS (ISIS) and FLASH (for ease of implementation, we adopt the "global" approach for FLASH), as well as Lasso for completeness. Furthermore, some non-convex penalised least squares (PLS) estimation techniques are included in the comparison study, such as the SCAD (Fan and Li, 2001) and the MC+ penalty (Zhang, 2010). Sub-optimality of the Lasso in terms of model selection has been noted in recent literature (see e.g. Zhang and Huang (2008) and Zou and Li (2008)), and non-convex penalties are proposed as a greedier alternative to achieve better variable selection. In the following simulation study, the SCAD estimator is produced using the local linear approximation (Zou and Li, 2008) and the MC+ penalised criterion is optimised using the SparseNet (Mazumder et al., 2009).

The TCS algorithm is applied using both rescaling methods (denoted by TCS1 and TCS2, respectively), with the maximum cardinality of the active set $\mathcal{A}$ (Step 4) set at $m = \lfloor n/2 \rfloor$, a value also used for FR. The extended BIC is adopted (see Section 3.2) to select the final model for the one-at-a-time algorithms, i.e. TCS1, TCS2, FR and FS. For the thus-selected final models, the coefficient values are estimated using least squares. We note that, when the aim is to construct a well-performing predictive model, a shrinkage method can be applied to the least squares estimate. However, since our focus is on the variable selection aspect of the different techniques, we use the plain (i.e. unshrunk) least squares estimates.

As for the rest of the methods, we select the tuning parameters for each method as follows: the data is divided into the training and validation sets such that the training observations are used to compute the solution paths over a range of tuning parameters, and those which give the smallest mean squared error between the response and the predictions on the validation data are selected.

Finally, we note that FS and the Lasso are implemented using the R package `lars`, and the ISIS and the SCAD by the package `SIS`.

4.1 Simulation models

Our simulation models were generated as below. For models (A)–(C) and (F), the procedure for generating the sparse coefficient vectors $\boldsymbol{\beta}$ is outlined below the itemised list which follows.

(A) Factor model with 2 factors: Let ϕ_1 and ϕ_2 be two independent standard normal variables. Each variable X_j, $j = 1, \ldots, p$, is generated as $X_j = f_{j,1}\phi_1 + f_{j,2}\phi_2 + \eta_j$, where $f_{j,1}, f_{j,2}, \eta_j$ are also generated independently from a standard normal distribution. The model is taken from Meinshausen and Bühlmann (2010).

(B) Factor model with 10 factors: Identical to (A) but with 10 instead of 2 factors.

(C) Factor model with 20 factors: Identical to (A) but with 20 instead of 2 factors.

(D) Taken from Fan and Lv (2008) Section 4.2.2:

$$\mathbf{y} = \beta X_1 + \beta X_2 + \beta X_3 - 3\beta\sqrt{\varphi}X_4 + \epsilon,$$

where $\epsilon \sim \mathcal{N}_n(0, \mathbf{I}_n)$ and $(X_{i,1}, \ldots, X_{i,p})^T$ are generated from a multivariate normal distribution $\mathcal{N}_n(\mathbf{0}, \Sigma)$ independently for $i = 1, \ldots, n$. The population covariance matrix

$\Sigma = (\Sigma_{j,k})_{j,k=1}^p$ satisfies $\Sigma_{j,j} = 1$ and $\Sigma_{j,k} = \varphi, j \neq k$, except $\Sigma_{4,k} = \Sigma_{j,4} = \sqrt{\varphi}$, such that X_4 is marginally uncorrelated with $\mathbf{y}$ at the population level. In the original model of Fan and Lv (2008), $\beta = 5$ and $\varphi = 0.5$ were used, but we chose $\beta = 2.5$ and $\varphi = 0.5, 0.95$ to investigate the performance of the variable selection methods in more challenging situations.

(E) Taken from Fan and Lv (2008) Section 4.2.3:

$$\mathbf{y} = \beta X_1 + \beta X_2 + \beta X_3 - 3\beta\sqrt{\varphi} X_4 + 0.25\beta X_5 + \epsilon,$$

with the population covariance matrix of $\mathbf{X}$ as in (D) except $\Sigma_{5,k} = \Sigma_{j,5} = 0$, such that X_5 is uncorrelated with any $X_j,\ j \neq 5$, and relevant. However, it has only a very small contribution to $\mathbf{y}$.

(F) Leukemia data analysis: Golub et al. (1999) analysed the Leukaemia dataset from high-density Affymetrix oligonucloeotide arrays (available on `http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi`), which has 72 observations and 7129 genes (i.e. variables). In Fan and Lv (2008), the dataset was used to investigate the performance of Sure Independence Screening in a feature selection problem. Here, instead of using the actual response from the dataset, we used the design matrix to create simulated models as follows. Each column X_j of the design matrix was normalised to $\|X_j\|_2^2 = n$, and out of 7129 such columns, p were randomly selected to generate an $n \times p$-matrix $\mathbf{X}$. Then we generated a sparse p-vector $\boldsymbol{\beta}$ and the response $\mathbf{y}$ as in (1). In this manner, the knowledge of $\mathcal{S}$ could be used to assess the performance of the competing variable selection techniques. A similar approach was taken in Meinshausen and Bühlmann (2010) to generate simulation models from real datasets.

With the exception of (D)–(E), we generated the simulated data as below. Sparse coefficient vectors $\boldsymbol{\beta}$ were generated by randomly sampling the indices of $\mathcal{S}$ from $1, \ldots, p$, with $|\mathcal{S}| = 10$. The non-zero coefficient vector $\boldsymbol{\beta}_{\mathcal{S}}$ was drawn from a zero-mean normal distribution such that $\mathbf{C}_{\mathcal{S},\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}} \sim \mathcal{N}_{|\mathcal{S}|}(\mathbf{0}, n^{-1}\mathbf{I}_{|\mathcal{S}|})$, where $\mathbf{C}_{\mathcal{S},\mathcal{S}}$ denotes the sample correlation matrix of $\mathbf{X}_{\mathcal{S}}$. In this manner, $\arg\max_{j \in \mathcal{J}} |X_j^T(\mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}_{\mathbf{X}})|$ may not always be attained by $j \in \mathcal{S}$, which makes the correct identification of relevant variables more challenging. The noise level σ was chosen to set $R^2 = \text{var}(\mathbf{x}_i^T\boldsymbol{\beta})/\text{var}(y_i)$ at 0.3, 0.6, or 0.9, adopting a similar approach to that taken in Wang (2009). In models (A)–(E), the number of observations was $n = 100$ while the dimensionality p varied from 500 to 2000 (except (D)–(E) where it was fixed at 1000), and finally, 100 replicates were generated for each set-up.

4.2 Simulation results

For each method and simulation setting, we report the following error measures which are often adopted to evaluate the performance of variable selection: the number of False Positives (FP, the number of irrelevant variables incorrectly identified as relevant), the number of False Negatives (FN, the number of relevant variables incorrectly identified as irrelevant) and the L2 distance $\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_2^2$; all averaged over 100 simulated data sets. The summary of the simulation

Table 2: Simulation results for model (A) with $|\mathcal{S}| = 10$. Results in bold font mean the value of FP+FN is the lowest or within 10% of the lowest; the same for L2. The value of 0 means less than 5×10^{-4}.

p	R^2		TCS1	TCS2	FR	FS	Lasso	ISIS	PCS	MC+	SCAD	FLASH
500	0.3	FP	1.2	0.55	3.8	1.04	44.93	1.06	4.59	5.33	57.28	5.66
		FN	2.47	2.52	1.82	2.2	2.93	9.18	8.45	4.31	1.8	2.9
		FP+FN	3.67	**3.07**	5.62	**3.24**	47.86	10.24	13.04	9.64	59.08	8.56
		L2	**0.012**	**0.012**	**0.012**	**0.013**	0.264	1.006	0.914	0.134	0.096	0.081
	0.6	FP	1.05	0.74	4.49	1.07	47.92	1.09	4.76	3.25	40.76	6.45
		FN	1.07	1.12	0.87	1.16	2.24	9.29	8.45	1.96	1.06	1.94
		FP+FN	2.12	**1.86**	5.36	2.23	50.16	10.38	13.21	5.21	41.82	8.39
		L2	**0.002**	**0.002**	0.003	0.055	0.242	1.021	0.812	0.042	0.04	0.106
	0.9	FP	0.92	0.57	2.64	1.17	47.97	1.06	4.52	1.57	27.48	7.15
		FN	0.43	0.41	0.32	0.62	1.75	9.21	8.37	2.03	0.58	1.49
		FP+FN	1.35	**0.98**	2.96	1.79	49.72	10.27	12.89	3.6	28.06	8.64
		L2	**0**	**0**	0.001	0.074	0.292	1.075	0.982	0.085	0.02	0.205
1000	0.3	FP	1.79	1.38	22.28	1.61	44.56	1.38	5.53	6.6	69.77	10.05
		FN	2.18	2.54	1.41	2.31	4.73	9.48	8.73	5.21	2.34	4.76
		FP+FN	**3.97**	**3.92**	23.69	**3.92**	49.29	10.86	14.26	11.81	72.11	14.81
		L2	**0.01**	0.027	0.035	0.039	0.463	1.073	0.787	0.219	0.159	0.318
	0.6	FP	1.67	1.35	24.91	1.3	46	1.19	5.55	4.39	55.93	7.76
		FN	1.13	1.17	0.78	1.65	4.16	9.33	8.63	2.79	1.51	3.33
		FP+FN	2.8	**2.52**	25.69	2.95	50.16	10.52	14.18	7.18	57.44	11.09
		L2	**0.002**	**0.003**	0.009	0.126	0.498	1.016	0.868	0.13	0.117	0.32
	0.9	FP	1.21	0.8	25.75	1.11	47.38	1.23	5.45	1.84	43.93	7.42
		FN	0.43	0.45	0.3	1.1	3.86	9.38	8.69	2.58	0.94	2.61
		FP+FN	1.64	**1.25**	26.05	2.21	51.24	10.61	14.14	4.42	44.87	10.03
		L2	**0**	**0**	0.002	0.088	0.405	0.916	0.803	0.078	0.063	0.192
2000	0.3	FP	1.77	1.65	41.53	1.64	38.27	1.48	6.71	10.53	80.9	9.07
		FN	2.33	2.36	1.53	3.48	6.48	9.59	8.98	5.79	3.07	6.22
		FP+FN	**4.1**	**4.01**	43.06	5.12	44.75	11.07	15.69	16.32	83.97	15.29
		L2	**0.013**	0.016	0.047	0.116	0.603	0.99	0.804	0.311	0.199	0.467
	0.6	FP	1.89	1.89	40.87	1.39	41.32	1.35	6.37	6.1	66.65	7.82
		FN	1.4	1.46	0.87	2.77	6.18	9.48	8.82	4.06	2.21	5.06
		FP+FN	**3.29**	**3.35**	41.74	4.16	47.5	10.83	15.19	10.16	68.86	12.88
		L2	**0.004**	**0.004**	0.024	0.252	0.752	1.243	0.989	0.338	0.18	0.496
	0.9	FP	1.61	1.32	39.5	1.45	39	1.35	6.87	19.99	59.73	6.96
		FN	0.44	0.56	0.68	2.21	6.32	9.55	8.9	3.88	1.6	5.11
		FP+FN	**2.05**	**1.88**	40.18	3.66	45.32	10.9	15.77	23.87	61.33	12.07
		L2	**0**	0.005	0.314	0.285	0.711	1.126	0.978	0.367	0.147	0.577

results can be found in Tables 2–5. We also present the receiver operating characteristic (ROC) curves, which plot the true positive rate (TPR) against the false positive rate (FPR), in Figures 1–4. Note that the simulation results from model (B) are discussed in the text only and the corresponding figure and table are omitted for brevity. The steep slope of an ROC implies that relevant variables have been selected without including too many irrelevant ones. Vertical lines are plotted as a guideline to indicate when the FPR reaches $2.5|\mathcal{S}|/p$. Since the existing R implementation of ISIS (package `SIS`) returns the final selection of variables only, rather than an entire path, we did not produce the ROC curves for that method.

Overall, compared with other methods, TCS1, TCS2 and FR achieve a high TPR more quickly without including too many irrelevant variables and thus tend to achieve a small L2 distance. While the PC-simple algorithm attains a low FPR, its TPR is also low even when the significant level for the testing procedure is set to be high. For certain set-ups, Lasso or SCAD achieves a high TPR but only at the cost of a high FPR.

Specifically, for factor models (A)–(C), it can be observed that TCS1, TCS2, FR (combined

Table 3: Simulation results for model (C) with $|\mathcal{S}| = 10$. Results in bold font mean the value of FP+FN is the lowest or within 10% of the lowest; the same for L2.

p	R^2		TCS1	TCS2	FR	FS	Lasso	ISIS	PCS	MC+	SCAD	FLASH
500	0.3	FP	4.21	3.57	9.56	8.44	43.82	1.81	5.73	38.84	42.23	19.97
		FN	6.27	5.45	5.81	7.44	3.08	9.81	9.42	4.49	3.69	5.19
		FP+FN	10.48	**9.02**	15.37	15.88	46.9	11.62	15.15	43.33	45.92	25.16
		L2	0.207	**0.172**	0.246	0.427	**0.166**	0.718	0.648	0.322	0.189	0.271
	0.6	FP	6.57	4.44	15.67	15.61	45.36	1.83	5.78	64.69	38.82	19.07
		FN	3.44	2.01	1.57	3.35	1.99	9.83	9.31	5.73	3.4	4.09
		FP+FN	10.01	**6.45**	17.24	18.96	47.35	11.66	15.09	70.42	42.22	23.16
		L2	0.066	0.024	**0.019**	0.114	0.093	0.858	0.782	0.36	0.164	0.207
	0.9	FP	6.89	3.49	16.22	17.58	48.62	1.79	5.9	58.78	39.17	18.66
		FN	1.06	0.86	0.63	1.43	1.01	9.79	9.47	5.7	3.16	3.16
		FP+FN	7.95	**4.35**	16.85	19.01	49.63	11.58	15.37	64.48	42.33	21.82
		L2	0.011	**0.002**	0.025	0.078	0.035	0.82	0.752	0.374	0.157	0.2
1000	0.3	FP	2.29	3.45	8	6.73	45.22	1.92	5.86	109.1	114.8	19.63
		FN	7.9	5.77	7.75	8.67	4.33	9.92	9.58	6.48	3.63	6.92
		FP+FN	10.19	**9.22**	15.75	15.4	49.55	11.84	15.44	115.6	118.4	26.55
		L2	0.558	**0.342**	0.694	0.835	0.414	0.993	0.897	0.588	**0.343**	0.554
	0.6	FP	5.04	4.72	15.21	11.93	48.97	1.92	6.13	90.51	110.8	19.86
		FN	5.79	3.6	4.31	6.41	3.27	9.92	9.6	6.74	2.51	5.97
		FP+FN	10.83	**8.32**	19.52	18.34	52.24	11.84	15.73	97.25	113.3	25.83
		L2	0.286	**0.138**	0.293	0.456	0.287	1.006	0.905	0.537	0.214	0.404
	0.9	FP	9.15	5.44	20.3	15.99	52.41	1.8	6.23	78.06	100.4	20.67
		FN	3.74	1.72	2.18	4.22	2.28	9.8	9.56	6.75	1.75	5.16
		FP+FN	12.89	**7.16**	22.48	20.21	54.69	11.6	15.79	84.81	102.1	25.83
		L2	0.258	**0.058**	0.147	0.52	0.174	1.09	0.985	0.612	0.137	0.43
2000	0.3	FP	1.75	2.25	5.12	4.97	47.13	1.89	6.4	133.6	129.4	19.9
		FN	8.72	7.34	9.13	9.44	5.63	9.89	9.74	7.39	4.81	7.89
		FP+FN	**10.47**	**9.59**	14.25	14.41	52.76	11.78	16.14	141	134.3	27.79
		L2	0.649	**0.446**	0.855	0.894	0.499	0.951	0.87	0.669	**0.438**	0.678
	0.6	FP	3.4	4.76	11.64	6.85	49.4	1.94	6.31	187.3	125.4	20.29
		FN	7.83	4.62	7.27	8.66	4.56	9.94	9.78	6.67	3.68	7.69
		FP+FN	11.23	**9.38**	18.91	15.51	53.96	11.88	16.09	194	129	27.98
		L2	0.512	**0.164**	0.629	0.761	0.418	0.943	0.857	0.566	0.31	0.675
	0.9	FP	7.02	4.93	19.17	10.77	52.8	1.91	6.16	149.3	117.3	20.81
		FN	5.75	2.64	4.3	7.17	3.87	9.91	9.65	7.25	2.85	7.3
		FP+FN	12.77	**7.57**	23.47	17.94	56.67	11.82	15.81	156.6	120.2	28.11
		L2	0.36	**0.104**	0.292	0.516	0.284	0.796	0.708	0.552	0.196	0.56

Table 4: Simulation results for models (D)–(E) with $|\mathcal{S}| = 4$ and 5. Results in bold font mean the value of FP+FN is the lowest or within 10% of the lowest; the same for L2.

p	φ		TCS1	TCS2	FR	FS	Lasso	ISIS	PCS	MC+	SCAD	FLASH
1000	0.5	FP	0.71	2.4	22.41	27.86	58.73	1.21	2.33	27.94	111	26.18
		FN	0	0	0	1	1	3.21	1.65	0.6	1	1
		FP+FN	**0.71**	2.4	22.41	28.86	59.73	4.42	3.98	28.54	112	27.18
		L2	**0.149**	0.351	2.876	33.46	30.92	47.9	38.74	19.12	30.96	31.85
	0.95	FP	0.39	0.76	19.84	7.14	28.37	1.45	1.42	49.58	46.68	12.88
		FN	1.43	3.64	1.89	2.05	1.54	3.71	3.58	1.7	2.07	1.61
		FP+FN	**1.82**	4.4	21.73	9.19	29.91	5.16	5	51.28	48.75	14.49
		L2	**26.71**	71.17	76.23	70.87	65.82	73.73	71.61	67.07	69.23	67.21
1000	0.5	FP	0.85	3.31	30.2	29.06	56.92	1.23	2.31	32.56	112.3	27.04
		FN	0.03	0.11	0.01	1.15	1.05	4.23	2.42	0.79	1.02	1.19
		FP+FN	**0.88**	3.42	30.21	30.21	57.97	5.46	4.73	33.35	113.3	28.23
		L2	**0.177**	0.528	4.102	33.5	31.46	48.83	39.46	22.11	31.46	32.18
	0.95	FP	0.05	0.05	26.08	4.5	28.74	1.03	1.01	35.82	43.73	12.78
		FN	2.76	3.96	1.75	2.32	1.56	4.1	3.77	1.86	2.11	1.83
		FP+FN	**2.81**	4.01	27.83	6.82	30.3	5.13	4.78	37.68	45.84	14.61
		L2	**49.89**	71.56	81.1	69.81	65.9	76.37	71.88	66.78	68.76	67.28

Table 5: Simulation results for model (F) with $|\mathcal{S}| = 10$. Results in bold font mean the value of FP+FN is the lowest or within 10% of the lowest; the same for L2.

p	R^2		TCS1	TCS2	FR	FS	Lasso	ISIS	PCS	MC+	SCAD	FLASH
1000	0.3	FP	2.27	2.08	13.68	1.65	23.69	0.87	6.09	130.6	23.61	7.97
		FN	7.2	6.45	5.12	8.94	8.22	9.92	8.33	7.81	8.42	5.96
		FP+FN	9.47	**8.53**	18.8	10.59	31.91	10.79	14.42	138.4	32.03	13.93
		L2	3.376	**2.579**	3.549	6.487	6.33	7.577	5.144	6.654	6.346	**2.605**
	0.6	FP	3.97	3.87	16.36	1.58	21.89	0.78	5.98	106.4	23.54	8.48
		FN	4.65	4.11	4.07	9.1	8.24	9.89	8.37	7.88	8.46	5.22
		FP+FN	**8.62**	**7.98**	20.43	10.68	30.13	10.67	14.35	114.2	32	13.7
		L2	3.029	**2.515**	6.604	10.53	10.25	11.5	7.181	10.64	10.38	4.229
	0.9	FP	5.97	5.17	14.54	1.77	20.29	0.83	6.1	115.2	20.72	7.73
		FN	1.95	2.42	3.45	9.14	8.7	9.88	8.3	8.03	8.87	4.81
		FP+FN	**7.92**	**7.59**	17.99	10.91	28.99	10.71	14.4	123.2	29.59	12.54
		L2	**0.573**	2.055	5.81	9.555	9.501	10.65	8.428	9.736	9.51	5.428
2000	0.3	FP	1.76	1.53	12.56	1.49	21.06	0.84	6.89	154.2	26.63	8.88
		FN	8.66	8.25	7.73	9.48	8.89	9.9	8.75	8.37	8.86	7.06
		FP+FN	**10.42**	**9.78**	20.29	10.97	29.95	**10.74**	15.64	162.6	35.49	15.94
		L2	4.774	**3.952**	5.626	6.371	6.267	7.756	5.484	6.403	6.286	**4.27**
	0.6	FP	3.18	2.51	16.9	1.62	20.89	0.85	6.45	250.1	29.89	8.46
		FN	6.94	7.04	6.56	9.51	8.83	9.9	8.56	8.05	8.86	6.56
		FP+FN	**10.12**	**9.55**	23.46	11.13	29.72	10.75	15.01	258.2	38.75	15.02
		L2	**2.424**	2.9	5.74	6.891	6.901	8.071	6.072	7.013	6.902	4.79
	0.9	FP	5.4	4.42	18.96	1.83	22.73	0.83	6.73	202.3	29.23	9.04
		FN	4.29	3.98	5.17	9	8.72	9.92	8.64	8.25	8.99	5.86
		FP+FN	9.69	**8.4**	24.13	10.83	31.45	10.75	15.37	210.6	38.22	14.9
		L2	**1.675**	**1.745**	3.64	5.232	5.254	6.67	4.133	5.401	5.275	2.841

with the extended BIC) and SCAD are superior to other methods in terms of achieving small FN, especially when R^2 is sufficiently high. However, the FR and SCAD tend to result in a model with too large an FP in comparison to the TCS algorithm, and therefore the L2 distance obtained from TCS2 is often the smallest. This becomes more obvious as the dimensionality grows and the number of factors increases, and the ROC curves in Figures 1–2 also support this conclusion, as those from the TCS algorithm attain a higher TPR for a similar level of FPR. Note that from our extensive numerical experiments, we observed that increasing number of factors led to an increased chance of marginal correlation screening being misleading at the very first iteration in the sense that $\arg\max_j |X_j^T \mathbf{y}| \notin \mathcal{S}$. In such set-ups, the adaptive choice of $\mathcal{C}_j$ used by the TCS algorithm turned out to be helpful in correctly identifying a relevant variable more often than marginal correlation screening. Between TCS1 and TCS2, while the two perform as well as each other for the two factor models from (A), it is TCS2 which outperforms the other for the models with more factors. As for the rest of the methods, FS performs as well as FR for lower dimensionality, and even better in terms of FP, but its FN is larger than that of FR as p and the number of factors increase. Both PCS algorithm and ISIS return final models which are too small and therefore obtain large FN and small FP; especially ISIS almost always misses the entire set of variables in $\mathcal{S}$. Lasso is not significantly inferior to, and occasionally better than, TCS1, TCS2 and FR in terms of FN, but it tends to select a model with a large FP like SCAD. While the ROC curves of MC+ and FLASH behave better than that of SCAD for certain set-ups (e.g. for two factor models), final selected models for these methods achieve larger FN. Finally, in terms of FP, FLASH tends to be better than SCAD, MC+ and Lasso.

For models (D) and (E), the TCS algorithm and FR outperform the rest when $\varphi = 0.5$, rapidly identifying all the relevant variables before the FPR reaches $2.5|\mathcal{S}|/p$ (left column of Figure 3). However, when correlations among the variables increase with $\varphi = 0.95$, ROC curves show that TCS1 is the only method that can identify all the relevant variables (right column of Figure 3). Other methods, including TCS2 and FR, often neglect to include X_4 due to its high correlations with the other variables, $\sqrt{\varphi}$ being almost 0.975. We note that while the ROC curves indicate that very often all the relevant variables are recovered by TCS1, the models selected by the extended BIC leave out some of them. Since the final models from TCS1 tend to contain the smallest number of noisy variables, we conclude that the extended BIC tends to choose final models which are too small for these particular examples. The rest of methods behave similarly as in the case of factor models; while Lasso, MC+, SCAD and FLASH achieve relatively small FN, the FP of their final models is too large and therefore they end up with a larger L2 distance than that of TCS1.

For the examples generated from the Leukemia dataset (model (F), Figure 4), the TCS algorithm with either of the rescaling methods always performs the best, with its ROC curves always dominating those of others. FR performs the second best and then follows FLASH. The remaining methods are not able to identify as many relevant variables as the TCS algorithm or FR even for a high FPR. The results reported in Table 5 also support this observation, where it is clear that the smallest FP and L2 distance are attained by either TCS1 or TCS2. Sometimes FR outperforms the two in terms of FN but TCS1 or TCS2 still achieves a smaller L2 distance, which implies that TCS algorithm, when combined with the extended BIC, can pick up a smaller model that better mimics the true coefficient vector than that yielded by FR with the same criterion. Interestingly, when it comes to the final model, FLASH achieves similar FN and much smaller FP than FR.

We have observed that the two rescaling methods sometimes select variables in different orders, although it does not necessarily imply that the resulting models are different. Overall, TCS2 performs better than TCS1 except for the examples from (D)–(E). In these two models, the variables $X_1, \ldots, X_p$ have a very special correlation structure in that e.g. X_4, a significant variable, can often appear uncorrelated with $\mathbf{y}$ in marginal correlation screening. Since TCS1 involves the term $\|(\mathbf{I}_n - \Pi_{\mathcal{A}})X_j\|_2^2$ in the denominator of the tilted correlation, as opposed to the term $\|(\mathbf{I}_n - \Pi_{\mathcal{A}})X_j\|_2$ in TCS2, it is better at picking up X_4 than TCS2. In the factor model examples, while the overall correlations among the variables are high, such "masking" does not take place as often among the significant variables. Therefore we conclude that unless the correlations are particularly high, TCS2 usually performs well.

5 Boston housing data analysis

In this section, we apply the TCS algorithm as well as the methods used in the simulation study in Section 4 to the Boston housing data, which was previously used to compare the performance of different regression techniques e.g. in Radchenko and James (2011). Originally, the dataset contains 13 variables which may have influence over the house prices. As in Radchenko and James (2011), we include the interaction terms between the variables in the analysis such that the data has $p = 91$ variables and $n = 506$ observations. Note that, due to the way the variables

Table 6: Boston housing data: test errors and the number of selected variables averaged over 20 test data sets.

	TCS1	TCS2	FR	PC-simple	MC+	SCAD	FLASH
test error	27.03	26.43	33.10	32.47	36.47	34.95	30.14
number of variables	19.5	13.5	16.0	2.0	83.5	36.0	26.0

are produced, there exist large sample correlations across the columns of the design matrix $\mathbf{X}$. We split the data into three with $n_1 = 91(= p)$, $n_2 = 46$ and $n_3 = 369$ observations each, and use the first n_1 observations as the training data (to compute a solution path for each method), next n_2 observations as the validation data (to choose the solution along the path that minimises the sum of the squared residuals for each method), and the last n_3 for computing the test error ($n_3^{-1}\|\mathbf{y}-\hat{\mathbf{y}}\|_2^2$). Random splitting of the data is repeated 20 times and Table 6 reports the average test error and number of selected variables, which shows that TCS2 achieves the minimum test error with the fewest variables in the model (except for the PC-simple algorithm). TCS1 also performs second best with more variables selected during the validation step. FR performs well in terms of both test error and the number of selected variables, and then follows FLASH. We note that the PC-simple algorithm chooses too few variables to describe the data well, while the non-convex penalty algorithms (MC+, SCAD) tend to include many more variables than the rest.

6 Conclusions

In this paper, we proposed a new way of measuring strength of association between the variables and the response in a linear model with a possibly large number of covariates, by adaptively taking into account correlations among the variables. We conclude by listing the new contributions made in this paper.

- Although tilting is not the only procedure which measures the association between a variable and the response conditional on other variables, its selection of the conditioning variables is a step further from simply using the current model itself or its sub-models, as is done in existing iterative algorithms. The hard-thresholding step in the tilting procedure enables an adaptive choice of the conditioning subset $\mathcal{C}_j$ for each variable X_j. Recalling the decomposition of the marginal correlation in (2), this adaptive choice can be seen as a vital step in capturing the contribution of each variable to the response. Also, in the case $\mathcal{C}_j = \emptyset$, tilted correlation is identical to marginal correlation, which can be viewed as "adaptivity" of our procedure.

- We propose two rescaling factors to obtain the tilted correlation c_j^*. Rescaling 1 (λ_j) is also adopted by the forward regression and rescaling 2 (Λ_j) is also adopted by the PC-simple algorithm, yet tilting is the only method to meaningfully use both rescaling factors and our theoretical results in Section 2.3 are valid for either of the two factors. It would be of interest to identify a way of combining the two rescaling methods, which we leave as a topic for future research.

- The separation of relevant and irrelevant variables, achieved by tilted correlation (as in our Theorems 1–3), cannot always be achieved by marginal correlation, and similar results to these theorems have not been reported previously to the best of our knowledge.
- The proposed TCS algorithm is designed to fully exploit the theoretical properties of the tilted correlation, and in particular its asymptotic consistency in separating between the relevant and irrelevant variables. Although we have not yet been able to demonstrate the model selection consistency of the TCS algorithm, numerical experiments confirm its good performance in comparison with other well-performing methods, showing that it can achieve high true positive rate without including many irrelevant variables. The algorithm is simple, easy to implement and does not require the use of advanced computational tools.

Ending on a slightly more general note, since correlation is arguably the most widely used statistical measure of association, we would expect our tilted correlation (which can be viewed as an "adaptive" extension of standard correlation) to be more widely applicable in various statistical contexts beyond the simple linear regression model.

Acknowledgements
We would like to thank the Joint Editor, Associate Editor and two Referees for very helpful comments which led to a substantial improvement of this manuscript.

A Proof of Theorem 1

The proof of Theorem 1 is divided into Steps 1–3. Recalling the decomposition of $(X_j^*)^T\mathbf{y}$ in (3), we first control the inner product between X_j^* and $\boldsymbol{\epsilon}$ uniformly over all j in Step 1. In Steps 2–3, we control the second summand $I \equiv \sum_{k\in\mathcal{S}\setminus\mathcal{C}_j,k\neq j}\beta_k X_j^T(\mathbf{I}_n - \Pi_j)X_k$ for j falling into two different categories, and thus derive the result.

Step 1 For $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, n^{-1}\sigma^2\cdot\mathbf{I}_n)$, we observe that, with probability converging to 1, $\max_{1\le j\le p}|\langle\boldsymbol{\epsilon}, Z_j\rangle| \le \sigma\sqrt{2\log p/n}$ for $Z_1,\ldots,Z_p \in \mathbb{R}^n$ having unit norm as $\|Z_j\|_2 = 1$. From (A2), we have $\sigma\sqrt{2\log p/n} \le Cn^{-\gamma}$ for some $C > 0$, and from (A5), $\|X_j^*\|_2 > \sqrt{\alpha} > 0$. Therefore by defining $\mathcal{E}_0 = \{\max_j|(X_j^*)^T\boldsymbol{\epsilon}| < Cn^{-\gamma}\}$, it follows that $\mathbb{P}(\mathcal{E}_0) \to 1$.

Step 2 In this step, we turn our attention to those j whose corresponding $\mathcal{C}_j$ satisfy $\mathcal{S}\setminus\{j\} \subseteq \mathcal{C}_j$ and thus the corresponding $I = 0$ and $(X_j^*)^T\mathbf{y} = \beta_j(1-a_j) + (X_j^*)^T\boldsymbol{\epsilon}$.

Rescaling 1. With the rescaling factor $\lambda_j = (1-a_j)$ which is bounded away from 0 by (A5), it can be shown that if such j belongs to $\mathcal{S}$, its tilted correlation satisfies $c_j^*(\lambda_j)/\beta_j \to 1$ on $\mathcal{E}_0$, as $|\beta_j| \gg n^{-\mu}$. On the other hand, if $j \notin \mathcal{S}$, we have $\beta_j(1-a_j) = 0$ which leads to $n^\mu \cdot c_j^*(\lambda_j) \le n^\mu \cdot Cn^{-\gamma} \to 0$ on $\mathcal{E}_0$.

Rescaling 2. Note that j whose $\mathcal{C}_j$ include all the members of $\mathcal{S}$ cannot be a member of $\mathcal{S}$ itself, and in this case, $(\mathbf{I}_n - \Pi_j)\mathbf{y}$ is reduced to $(\mathbf{I}_n - \Pi_j)\boldsymbol{\epsilon}$. Since (A3) assumes that each $\mathcal{C}_j$ has its cardinality bounded by Cn^ξ, it can be shown that $\mathbb{P}\left(\max_j\|\Pi_j\boldsymbol{\epsilon}\|_2 \le C'n^{-(\gamma-\xi/2)}\right) \to 1$ for some $C' > 0$, similarly to Step 1. Also, Lemma 3 from Fan and Lv (2008) implies that $\mathbb{P}\left(\sigma^{-2} \cdot \|\boldsymbol{\epsilon}\|_2^2 < 1 - \omega\right) \to 0$ for any $\omega \in (0, 1)$. Combining these observations with (A1) and (A4), we derive that $1 - a_{jy} = \|(\mathbf{I}_n - \Pi_j)\boldsymbol{\epsilon}\|_2^2/\|\mathbf{y}\|_2^2 \geq Cn^{-\delta}$ with probability tending to 1, and eventually we have $\Lambda_j \geq C'n^{-\delta/2}$ from (A5). Therefore, if $\mathcal{S} \subseteq \mathcal{C}_j$ for some $j \notin \mathcal{S}$, its corresponding tilted correlation satisfies $n^\mu \cdot c_j^*(\Lambda_j) \leq n^\mu \cdot Cn^{-(\gamma-\delta/2)} \to 0$ on $\mathcal{E}_0$. In the case of $\mathcal{S} \nsubseteq \mathcal{C}_j$, we can derive from (A6) that for such j, $\|(\mathbf{I}_n - \Pi_j)\mathbf{y}\|_2^2/\|\mathbf{y}\|_2^2 = 1 - a_{jy} \gg n^{-\kappa}$, which, combined with (A5), implies that $\Lambda_j \gg n^{-\kappa/2}$. Then the following holds for such j on $\mathcal{E}_0$: $n^\mu \cdot |c_j^*(\Lambda_j)| \geq n^\mu \cdot C|\beta_j| \to \infty$ if $j \in \mathcal{S}$, while $n^\mu \cdot c_j^*(\Lambda_j) \leq n^\mu \cdot Cn^{-(\gamma-\kappa/2)} \to 0$ if $j \notin \mathcal{S}$.

Step 3 We now consider those $j \in \mathcal{J}$ for which $\mathcal{S} \setminus \{j\} \nsubseteq \mathcal{C}_j$ and consequently the corresponding term $I \neq 0$ in general. From (A3) and Condition 1, we derive that for each j, there exists some $C > 0$ satisfying the following for all $k \in \mathcal{S} \setminus \mathcal{C}_j,\ k \neq j$,

$$|X_j^T(\mathbf{I}_n - \Pi_j)X_k| \leq |X_j^T X_k| + |(\Pi_j X_j)^T X_k| \leq Cn^{-\gamma}. \tag{9}$$

Then from (A1) and (A4), we can bound I as $|I| \leq C'n^{-(\gamma-\delta)}$. Also when $\mathcal{S} \setminus \{j\} \nsubseteq \mathcal{C}_j$, (A5)–(A6) imply that $\Lambda_j \gg n^{-\kappa/2}$. In summary, we can show that the following claims hold on $\mathcal{E}_0$, similarly as in Step 2: if $j \notin \mathcal{S}$, with either of the rescaling factors, $n^\mu \cdot c_j^*(\lambda_j) \leq n^\mu \cdot Cn^{-(\gamma-\delta-\kappa/2)} \to 0$, whereas if $j \in \mathcal{S}$, its coefficient satisfies $|\beta_j| \gg n^{-\mu}$ and therefore $n^\mu \cdot |c_j^*| \geq n^\mu \cdot C|\beta_j| \to \infty$ with $c_j^*(\lambda_j)/\beta_j \to 1$ for $j \in \mathcal{S}$. □

A.1 An example satisfying Condition 1

In this section, we verify the claim made in Section 2.3.1, which states that Condition 1 holds with probability tending to 1 when each column X_j is generated independently as a random vector on a n-dimensional unit sphere. We first introduce a result from modern convex geometry reported in Lecture 2 of Ball (1997), which essentially implies that, as the dimension n grows, it is not likely for any two vectors on a n-dimensional unit sphere to be within a close distance to each other.

Lemma 1. *Let S^{n-1} denote the surface of the Euclidean ball $B_2^n = \{\mathbf{x} \in \mathbb{R}^n : \sum_{i=1}^n x_i^2 \leq 1\}$ and $\mathbf{u} \in \mathbb{R}^n$ be a vector on S^{n-1} such that $\|\mathbf{u}\|_2 = 1$. Then the proportion of spherical cone defined as $\{\mathbf{v} \in S^{n-1} : |\mathbf{u}^T\mathbf{v}| \geq \omega\}$ for any $\mathbf{u}$ is bounded from above by* $\exp(-n\omega^2/2)$.

We first note that any $X_k,\ k \neq j$ can be decomposed as the summation of its projection onto X_j and the remainder, i.e., $X_k = c_{j,k}X_j + (\mathbf{I}_n - X_jX_j^T)X_k$. Then

$$(\Pi_j X_j)^T X_k = c_{j,k}(\Pi_j X_j)^T X_j + \left\{(\mathbf{I}_n - X_jX_j^T)\Pi_j X_j\right\}^T X_k,$$

and for $k \in \mathcal{S} \setminus \mathcal{C}_j,\ k \neq j$, the first summand is bounded from above by $a_j \cdot \pi_n \leq C_1 n^{-\gamma}$. As for the second summand, note that

$$\|(\mathbf{I}_n - X_jX_j^T)\Pi_j X_j\|_2^2 = (\Pi_j X_j)^T(\mathbf{I}_n - X_jX_j^T)\Pi_j X_j = a_j(1 - a_j),$$

and thus $\mathbf{w} = \{a_j(1 - a_j)\}^{-1/2} \cdot (\mathbf{I}_n - X_jX_j^T)\Pi_j X_j$ satisfies $\mathbf{w} \in S^{n-1}$. Then the probability of $|\mathbf{w}^T X_k| > Cn^{-\gamma}$ for any $k \in \mathcal{S} \setminus \mathcal{C}_j,\ k \neq j$ is bounded from above by the proportion of the spherical cone $\left\{X_k \in S^{n-1} :\ |\mathbf{w}^T X_k| > Cn^{-\gamma}\right\}$ in the unit sphere S^{n-1}. Applying Lemma 1, we can show that such proportion is bounded by $\exp\left(-C^2 n^{1-2\gamma}/2\right)$ for each j and k. Therefore, we can find some $C > 0$ satisfying

$$\mathbb{P}\left(\max_{j\in\mathcal{J};\ k\in\mathcal{S}\setminus\mathcal{C}_j,\ k\neq j} |(\Pi_j X_j)^T X_k| > Cn^{-\gamma}\right) \geq 1 - p|\mathcal{S}|\exp\left(-C' n^{1-2\gamma}/2\right),$$

where the right-hand side converges to 1 from assumptions (A1)–(A2).

B Proof of Theorem 2

For those $j \in \mathcal{K} = \mathcal{S} \cup \{\cup_{j\in\mathcal{S}}\mathcal{C}_j\}$, Condition 3 implies that $\mathcal{C}_k \cap \mathcal{C}_j = \emptyset$ if $k \in \mathcal{S} \setminus \mathcal{C}_j$. Then from (A3), we have $\|\Pi_j X_k\|_2 \leq Cn^{-(\gamma-\xi/2)}$ and therefore

$$\left|X_j^T(\mathbf{I}_n - \Pi_j)X_k\right| = \left|X_j^T X_k - (\Pi_j X_j)^T \Pi_j X_k\right| \leq Cn^{-\gamma} + C'n^{-(\gamma-\xi/2)},$$

which leads to

$$\left|\sum_{k\in\mathcal{S}\setminus\mathcal{C}_j, k\neq j} \beta_k X_j^T(\mathbf{I}_n - \Pi_j)X_k\right| = O\left(n^{-(\gamma-\delta-\xi/2)}\right) \tag{10}$$

for all $j \in \mathcal{K}$. Using Step 1 of Appendix A, we derive that

$$\mathcal{E}_{01} = \left\{\max_{j\in\mathcal{K}}\left|\sum_{k\in\mathcal{S}\setminus\mathcal{C}_j, k\neq j} \beta_k X_j^T(\mathbf{I}_n - \Pi_j)X_k + X_j^T(\mathbf{I}_n - \Pi_j)\boldsymbol{\epsilon}\right| \leq Cn^{-(\gamma-\delta-\xi/2)}\right\}$$

satisfies $\mathbb{P}(\mathcal{E}_{01}) = \mathbb{P}(\mathcal{E}_0) \to 1$. Since $\mu + \kappa/2 < \gamma - \delta - \xi/2$, we have $n^\mu \cdot c_j^* \to 0$ for $j \notin \mathcal{S}$ on $\mathcal{E}_{01}$, whereas $n^\mu \cdot |c_j^*| \to \infty$ and $c_j^*(\lambda_j)/\beta_j \to 1$ for those $j \in \mathcal{S}$. Therefore the dominance of tilted correlations for $j \in \mathcal{S}$ over those for $j \in \mathcal{K} \setminus \mathcal{S}$ follows. □

C Proof of Theorem 3

Compared to Condition 2, Condition 3 does not require any restriction on $\mathcal{C}_j \cap \mathcal{C}_k$ when both X_j and X_k are relevant, although it has an additional assumption (C2). Since $n^\mu \cdot |\beta_j|(1-a_j) \to \infty$ for $j \in \mathcal{S}$ from (A4)–(A5), (C2) implies that for any $j \in \mathcal{S}$, non-zero coefficients $\beta_k,\ k \in \mathcal{S} \setminus \mathcal{C}_j$ do not cancel out all the summands in the following to 0,

$$X_j^T(\mathbf{I}_n - \Pi_j)\mathbf{X}_\mathcal{S}\boldsymbol{\beta}_\mathcal{S} = \beta_j(1-a_j) + \sum_{k\in\mathcal{S}\setminus\mathcal{C}_j, k\neq j} \beta_k X_j^T(\mathbf{I}_n - \Pi_j)X_k.$$

If (10) in Appendix B holds, (C2) follows and therefore it can be seen that Condition 2 is stronger than Condition 3.

On the event $\mathcal{E}_0$ (Step 1 of Appendix A), $|X_j^T(\mathbf{I}_n - \Pi_j)\mathbf{y}| \gg n^{-\mu}$ for $j \in \mathcal{S}$ under (C2) and therefore the tilted correlations of relevant variables satisfy $|c_j^*| \gg n^{-\mu}$ with either of the rescaling factors. On the other hand, for $j \in \mathcal{K} \setminus \mathcal{S}$, we can use the arguments in Appendix B to show that $n^{\mu} \cdot c_j^* \to 0$. □

D Study of the assumptions (A5) and (A6)

In this section, we show that the assumptions (A5) and (A6) are satisfied under the following condition from Wang (2009). Let $\lambda_*(\mathbf{A})$ and $\lambda^*(\mathbf{A})$ represent the smallest and the largest eigenvalues of an arbitrary positive definite matrix $\mathbf{A}$, respectively.

- Both $\mathbf{X}$ and $\boldsymbol{\epsilon}$ follow normal distributions.
- There exist two positive constants $0 < \tau_* < \tau^* < \infty$ such that $\tau_* < \lambda_*(\boldsymbol{\Sigma}) \leq \lambda^*(\boldsymbol{\Sigma}) < \tau^*$, where $\text{cov}(\mathbf{x}_i) = \boldsymbol{\Sigma}$ for $i = 1, \ldots, n$.

Then, Wang (2009) showed that there exists $\eta \in (0, 1)$ satisfying

$$\tau_* \leq \min_{\mathcal{D}} \lambda_*(\mathbf{X}_{\mathcal{D}}^T \mathbf{X}_{\mathcal{D}}) \leq \max_{\mathcal{D}} \lambda^*(\mathbf{X}_{\mathcal{D}}^T \mathbf{X}_{\mathcal{D}}) \leq \tau^* \tag{11}$$

with probability tending to 1, for any $\mathcal{D} \subset \{1, \ldots, p\}$ with $|\mathcal{D}| \leq n^{\eta}$. We use the result from (11) in the following arguments.

(A5) Recalling the notations $\tilde{\mathbf{X}}_j = \mathbf{X}_{\mathcal{C}_j}$ and $\Pi_j = \tilde{\mathbf{X}}_j(\tilde{\mathbf{X}}_j^T \tilde{\mathbf{X}}_j)^{-1}\tilde{\mathbf{X}}_j^T$, we have

$$1 - X_j^T \Pi_j X_j = \left\| X_j - \tilde{\mathbf{X}}_j(\tilde{\mathbf{X}}_j^T \tilde{\mathbf{X}}_j)^{-1}\tilde{\mathbf{X}}_j^T X_j \right\|_2^2.$$

We let $\boldsymbol{\theta} = (\tilde{\mathbf{X}}_j^T \tilde{\mathbf{X}}_j)^{-1}\tilde{\mathbf{X}}_j^T X_j$ and assume that ξ from assumption (A3) satisfies $\xi \leq \eta$ such that, by applying (11), we obtain the following;

$$\begin{aligned} 1 - X_j^T \Pi_j X_j &= (1, \boldsymbol{\theta}) \left(X_j, \tilde{\mathbf{X}}_j \right)^T \left(X_j, \tilde{\mathbf{X}}_j \right) (1, \boldsymbol{\theta})^T \\ \geq \quad & (1, \boldsymbol{\theta}) \lambda_* \left((X_j, \tilde{\mathbf{X}}_j)^T (X_j, \tilde{\mathbf{X}}_j) \right) (1, \boldsymbol{\theta})^T \geq (1 + \|\boldsymbol{\theta}\|_2^2)\tau_* \geq \tau_* > 0. \end{aligned}$$

(A6) We note the link between (A6) and the asymptotic identifiability condition for high-dimensional problems first introduced in Chen and Chen (2008). The condition can be re-written as

$$\lim_{n \to \infty} \min_{\mathcal{D} \subset \mathcal{J}, |\mathcal{D}| \leq |\mathcal{S}|, \mathcal{D} \neq \mathcal{S}} n(\log n)^{-1} \cdot \frac{\|(\mathbf{I}_n - \Pi_{\mathcal{D}})\mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}}\|_2^2}{\|\mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}}\|_2^2} \to \infty, \tag{12}$$

after taking into account the column-wise normalisation of $\mathbf{X}$. Although the rate n^{κ} is less favourable than $n(\log n)^{-1}$, following exactly the same arguments as in Section 3 of Chen and Chen (2008), we are able to show that (A6) is implied by the condition in (11).

That is, letting $\boldsymbol{\theta} = (\tilde{\mathbf{X}}_j^T \tilde{\mathbf{X}}_j)^{-1} \tilde{\mathbf{X}}_j^T \mathbf{X}_{\mathcal{S}} \boldsymbol{\beta}_{\mathcal{S}}$, we have

$$\begin{aligned} & n^{\kappa} \cdot \frac{\|(\mathbf{I}_n - \Pi_j)\mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}}\|_2^2}{\|\mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}}\|_2^2} \geq n^{\kappa} \inf_{j \notin \mathcal{S}} \frac{\|\mathbf{X}_{\mathcal{S}\cap\mathcal{C}_j^c}\boldsymbol{\beta}_{\mathcal{S}\cap\mathcal{C}_j^c} - \tilde{\mathbf{X}}_j\boldsymbol{\theta}\|_2^2}{\|\mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}}\|_2^2} \\ \geq \;\; & C n^{\kappa-2\delta} \inf_{j \notin \mathcal{S}} \left(\boldsymbol{\beta}_{\mathcal{S}\cap\mathcal{C}_j^c}^T, -\boldsymbol{\theta}\right)^T \mathbf{X}_{\mathcal{S}\cup\mathcal{C}_j}^T \mathbf{X}_{\mathcal{S}\cup\mathcal{C}_j} \left(\boldsymbol{\beta}_{\mathcal{S}\cap\mathcal{C}_j^c}^T, -\boldsymbol{\theta}\right) \\ \geq \;\; & C n^{\kappa-2\delta} \lambda_*(\mathcal{S}\cup\mathcal{C}_j)\|\boldsymbol{\beta}_{\mathcal{S}\cap\mathcal{C}_j}\|_2^2 \end{aligned} \tag{13}$$

for some positive constant C, where the second inequality is derived under the assumptions (A1) and (A4). Then a constraint can be imposed on the relationship between κ, δ and ξ such that the right-hand side of the above (13) diverges to infinity.

References

Ball, K. (1997), "An elementary introduction to modern convex geometry," *Flavors of Geometry*, 31, 1–58.

Benjamini, Y. and Hochberg, Y. (1995), "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society, Series B*, 289–300.

Bickel, P. J. and Levina, E. (2008), "Covariance regularization by thresholding," *Annals of Statistics*, 36, 2577–2604.

Bogdan, M., Ghosh, J., and Doerge, R. (2004), "Modifying the Schwarz Bayesian information criterion to locate multiple interacting quantitative trait loci," *Genetics*, 167, 989–999.

Bühlmann, P., Kalisch, M., and Maathuis, M. (2009), "Variable selection for high-dimensional models: partially faithful distributions and the PC-simple algorithm," *Biometrika*, 97, 1–19.

Candès, E. and Tao, T. (2007), "The Dantzig selector: statistical estimation when p is much larger than n," *Annals of Statistics*, 6, 2313–2351.

Chen, J. and Chen, Z. (2008), "Extended Bayesian information criteria for model selection with large model spaces," *Biometrika*, 95, 759–771.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least angle regression," *Annals of Statistics*, 32, 407–499.

El Karoui, N. (2008), "Operator norm consistent estimation of large dimensional sparse covariance matrices," *Annals of Statistics*, 36, 2717–2756.

Fan, J. and Li, R. (2001), "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, 96, 1348–1360.

Fan, J. and Lv, J. (2008), "Sure independence screening for ultrahigh dimensional feature space," *Journal of the Royal Statistical Society, Series B*, 70, 849–911.

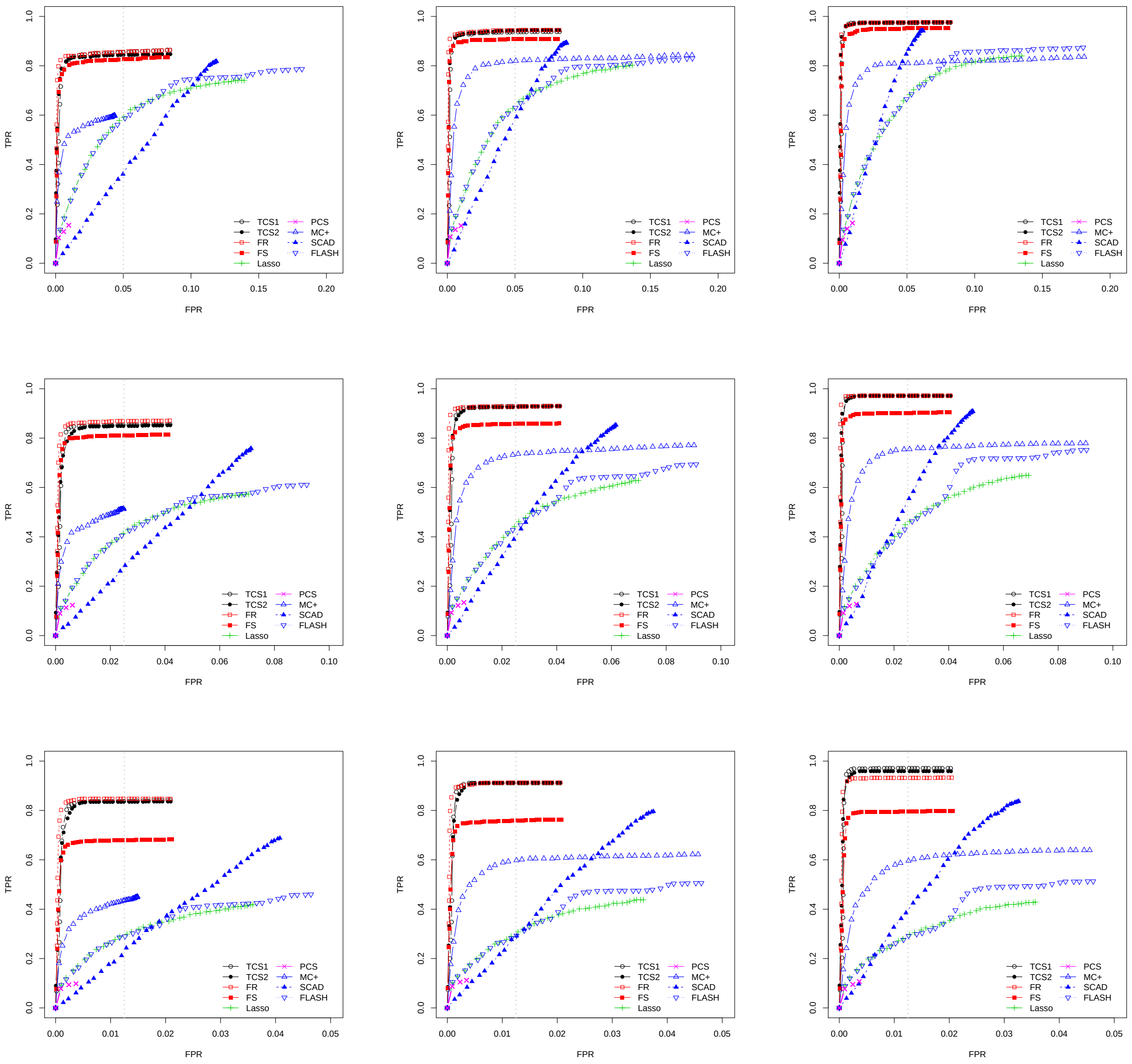

Figure 1: ROC curves for the simulation model (A) with $n = 100$: TCS1 (black empty circle), TCS2 (black filled circle), FR (red empty square), FS (red filled square), Lasso (green crossed circle) PC-simple algorithm (magenta two triangles), MC+ (blue empty triangle), SCAD (blue filled triangle) and FLASH (blue reversed triangle); FPR= $2.5|\mathcal{S}|/p$ (vertical dotted); first row: $p = 500$, second row: $p = 1000$, third row: $p = 2000$; first column: $R^2 = 0.3$, second column: $R^2 = 0.6$, third column: $R^2 = 0.9$.

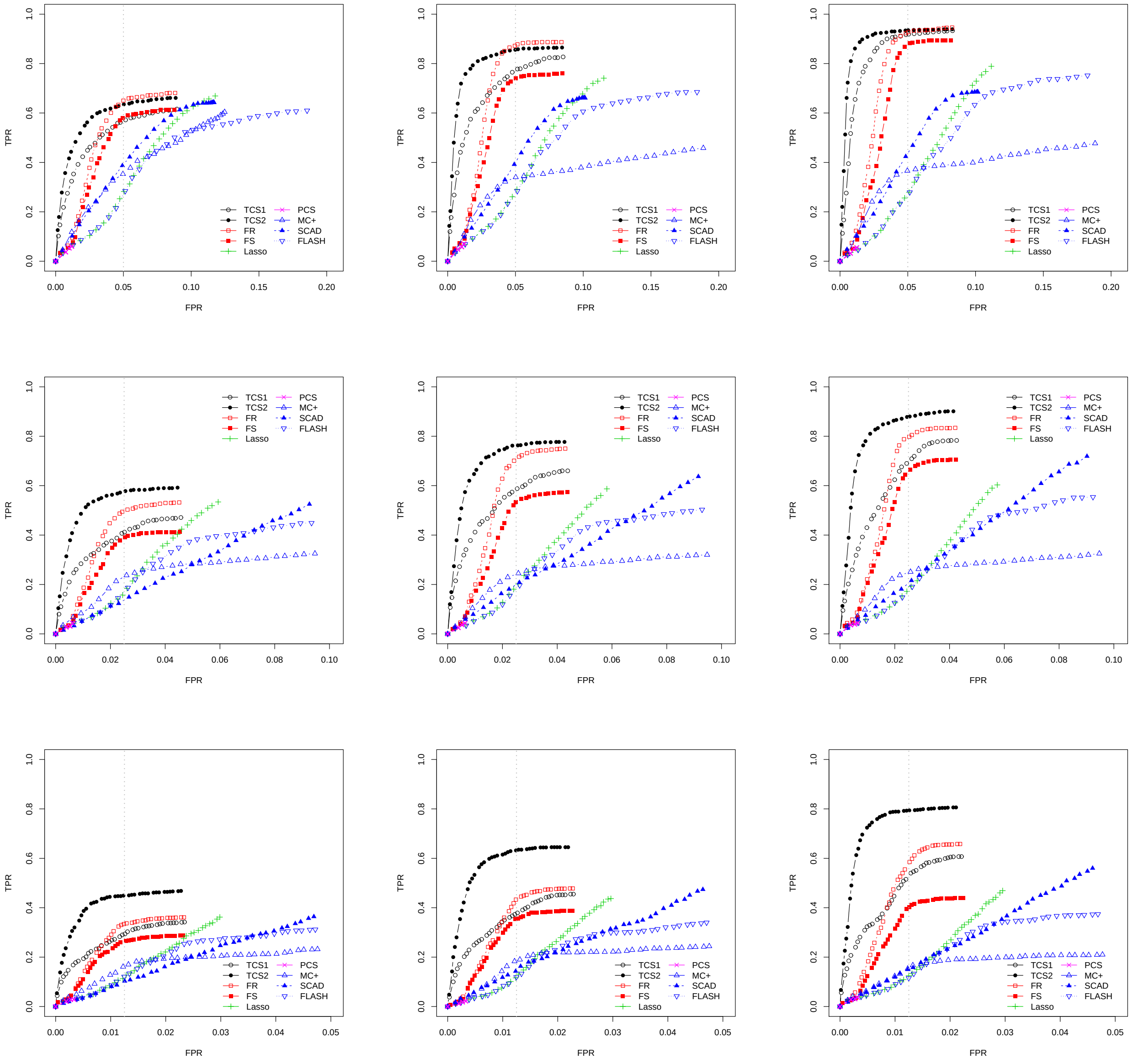

Figure 2: ROC curves for the simulation model (C) with $n = 100$.

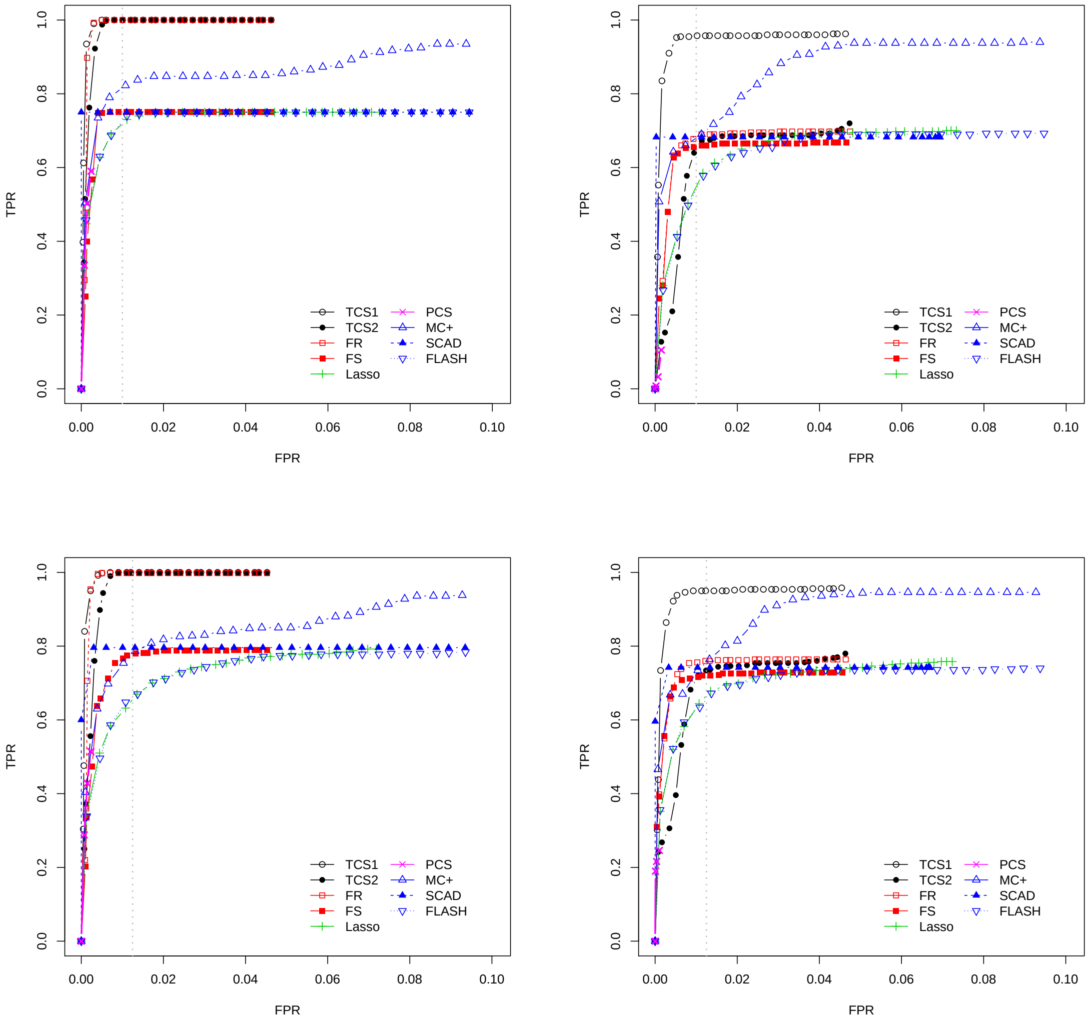

Figure 3: ROC curves for the simulation models (D) (first row) and (E) (second row) with $n = 100$; first column: $\varphi = 0.5$, second column: $\varphi = 0.95$.

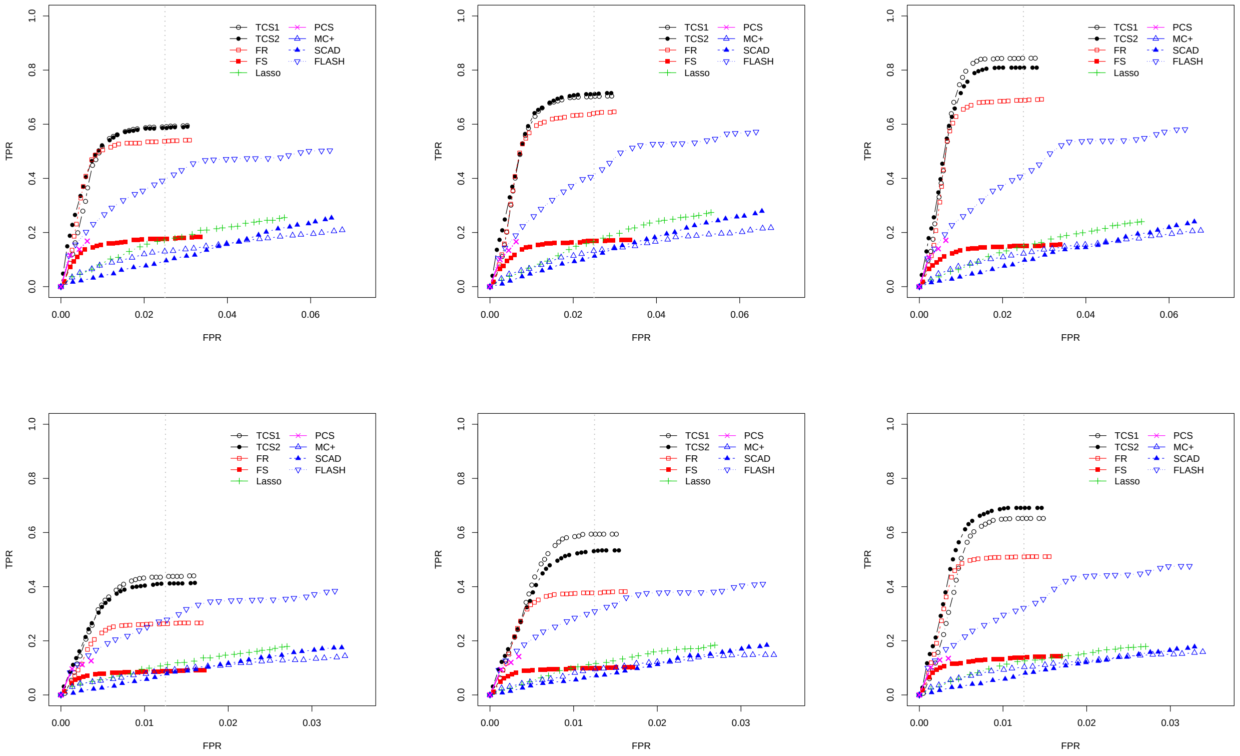

Figure 4: ROC curves for the simulation model (F) with $n = 72$.

— (2010), “A selective overview of variable selection in high dimensional feature space (invited review article),” *Statistica Sinica*, 20, 101–148.

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

Hall, P., Titterington, D. M., and Xue, J. H. (2009), “Tilting methods for assessing the influence of components in a classifier,” *Journal of the Royal Statistical Society, Series B*, 71, 783–803.

Kalisch, M. and Bühlmann, P. (2007), “Estimating high-dimensional directed acyclic graphs with the PC-algorithm,” *Journal of Machine Learning Research*, 8, 613–636.

Mazumder, R., Friedman, J., and Hastie, T. (2009), “SparseNet: Coordinate descent with non-convex penalties,” *Technical report, Stanford University.*

Meinshausen, N. and Bühlmann, P. (2008), “High dimensional graphs and variable selection with the Lasso,” *Annals of Statistics*, 34, 1436–1462.

— (2010), “Stability selection,” *Journal of the Royal Statistical Society, Series B*, 72, 417–473.

Radchenko, P. and James, G. (2011), “Forward-lasso with adaptive shrinkage,” *Annals of Applied Statistics (To appear).*

Tibshirani, R. (1996), “Regression shrinkage and selection via the Lasso,” *Journal of the Royal Statistical Society, Series B*, 58, 267–288.

Wang, H. (2009), “Forward regression for ultra-high dimensional variable screening,” *Journal of the American Statistical Association*, 104, 1512–1524.

Wasserman, L. and Roeder, K. (2009), “High-dimensional variable selection,” *Annals of Statistics*, 37, 2178–2201.

Weisberg, S. (1980), *Applied Linear Regression*, Wiley-Blackwell.

Witten, D. M. and Tibshirani, R. (2009), “Covariance-regularized regression and classification for high-dimensional problems,” *Journal of Royal Statistical Society, Series B*, 71, 615–636.

Zhang, C. (2010), “Nearly unbiased variable selection under minimax concave penalty,” *Annals of Statistics*, 38, 894–942.

Zhang, C. H. and Huang, J. (2008), “The sparsity and bias of the Lasso selection in high-dimensional linear regression,” *Annals of Statistics*, 36, 1567–1594.

Zhao, P. and Yu, B. (2006), “On model selection consistency of Lasso,” *Journal of Machine Learning Research*, 7, 2541–2563.

Zou, H. (2006), “The adaptive Lasso and its oracle properties,” *Journal of the American Statistical Association*, 101, 1418–1429.

Zou, H. and Li, R. (2008), “One-step sparse estimates in nonconcave penalized likelihood models,” *Annals of Statistics*, 36, 1509–1553.