Conditional Sure Independence Screening *

Emre Barut, Jianqing Fan
Princeton University

Anneleen Verhasselt
University of Antwerp

Abstract

Independence screening is a powerful method for variable selection for 'Big Data' when the number of variables is massive. Commonly used independence screening methods are based on marginal correlations or variations of it. In many applications, researchers often have some prior knowledge that a certain set of variables is related to the response. In such a situation, a natural assessment on the relative importance of the other predictors is the conditional contributions of the individual predictors in presence of the known set of variables. This results in conditional sure independence screening (CSIS). Conditioning helps for reducing the false positive and the false negative rates in the variable selection process. In this paper, we propose and study CSIS in the context of generalized linear models. For ultrahigh-dimensional statistical problems, we give conditions under which sure screening is possible and derive an upper bound on the number of selected variables. We also spell out the situation under which CSIS yields model selection consistency. Moreover, we provide two data-driven methods to select the thresholding parameter of conditional screening. The utility of the procedure is illustrated by simulation studies and analysis of two real data sets.

*Emre Barut is graduate student (Email: abarut@princeton.edu), Jianqing Fan is Frederick L. Moore'18 professor (Email: jqfan@princeton.edu), Department of Operations Research & Financial Engineering, Princeton University, Princeton, NJ 08544, USA. Anneleen Verhasselt is assistant professor, Department of Mathematics and Computer Science, University of Antwerp, Belgium. The paper was initiated while Anneleen Verhasselt was a visiting postdoctoral fellow at Princeton University. This research was partly supported by NSF Grants DMS-DMS-1206464, NIH Grant R01-GM072611 and NIH R01-GM100474 and FWO Travel Grant V422811N.

Keywords and phrases: False selection rate; Generalized linear models; Sparsity; Sure screening; Variable selection.

1 INTRODUCTION

Statisticians are nowadays frequently confronted with massive data sets from various frontiers of scientific research. Fields such as genomics, neuroscience, finance and earth sciences have different concerns on their subject matters, but nevertheless share a common theme: They rely heavily on extracting useful information from massive data and the number of covariates p can be huge in comparison with the sample size n. In such a situation, the parameters are identifiable only when the number of the predictors that are relevant to the response is small, namely, the vector of regression coefficients is sparse. This sparsity assumption has a nice interpretation that only a limited number of variables have a prediction power on the response. To explore the sparsity, variable selection techniques are needed.

Over the last ten years, there has been many exciting developments in statistics and machine learning on variable selection techniques for ultrahigh dimensional feature space. They can basically be classified into two classes: penalized likelihood and screening. Penalized likelihood techniques are well known in statistics: Bridge regression (Frank and Friedman 1993), Lasso (Tibshirani 1996), SCAD or other folded concave regularization methods (Fan and Li 2001; Fan and Lv 2011; Zhang and Zhang 2012), and Dantzig selector (Candes and Tao 2007; Bickel et al. 2009), among others. These techniques select variables and estimate parameters simultaneously by solving a high-dimensional optimization problem. See Hastie et al. (2009) and Bühlmann and van de Geer (2011) for an overview of the field. Despite the fact that various efficient algorithms have been proposed (Osborne et al. 2000a,b; Efron et al. 2004; Fan and Lv 2011), statisticians and machine learners still face huge computational challenges when the number of variables is in tens of thousands of dimensions or higher. This is particularly the case as we are entering the era of "Big Data" in which both sample size and dimensionality are large.

With this background, Fan and Lv (2008) propose a two-scale approach, called iterative sure independence screening (ISIS), which screens and selects variables iteratively. The approach is further developed by Fan et al. (2009) in the context of generalized linear models. Theoretical properties of sure independence screening for generalized linear models have been thoroughly studied by Fan and Song (2010). Other marginal screening methods include tilting methods (Hall et al. 2009), generalized correlation screening (Hall and Miller 2009), nonparametric screening (Fan et al. 2011), and robust screening (Li et al. 2012), among others. The merits of screening include expediences in distributed computation and implementation. By ranking marginal utility such as marginal correlation with the response, variables with weak marginal utilities are screened out by a simple thresholding.

The simple marginal screening faces a number of challenges. As pointed out in Fan and Lv (2008), it can screen out those hidden signature variables: those who have a big impact on response but are weakly correlated with the response. It can have large false positives too, namely recruiting those variables who have strong marginal utilities but are conditionally independent with the response given other variables. Fan and Lv (2008) and Fan et al. (2009) use a residual based approach to circumvent the problem but the idea of conditional screening has never been formally developed.

Conditional marginal screening is a natural extension of simple independent screening. In many applications, researchers know from previous investigations that certain variables $\mathbf{X}_{\mathcal{C}}$ are responsible for the outcomes. This knowledge should be taken into account when applying a variable selection technique in order not to remove these predictors from the model and to improve the selection process. Conditional screening recruits additional variables to strengthen the prediction power of $\mathbf{X}_{\mathcal{C}}$, via ranking conditional marginal utility of each variable in presence of $\mathbf{X}_{\mathcal{C}}$. In absence of such a prior knowledge, one can take those variables that survive the screening and selection

as in Fan and Lv (2008).

Conditional screening has several advantages. First of all, it makes it possible to recover the hidden significant variables. This can be seen by considering the following linear regression model

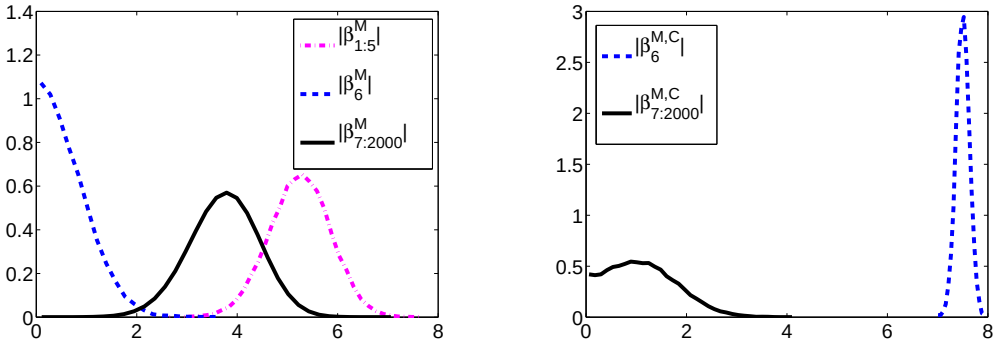
$$Y = \mathbf{X}^T\boldsymbol{\beta}^\star + \varepsilon, \qquad E\mathbf{X}\varepsilon = 0, \tag{1}$$

with $\boldsymbol{\beta}^\star = (\beta_1^\star, \ldots, \beta_p^\star)^T$. The marginal covariance between X_j and Y is given by

$$\mathrm{Cov}(X_j, Y) = \mathrm{Cov}(X_j, \mathbf{X}\boldsymbol{\beta}) = \mathbf{e}_j^T\boldsymbol{\Sigma}\boldsymbol{\beta}^\star,$$

where $\mathbf{e}_j \in I\!R^p$ is equal to 0, except for its jth element which equals to 1. This shows that the marginal covariance between X_j and Y is zero if $\beta_j^\star = -\sum_{k\neq j}\beta_k^\star\sigma_{kj}$, where σ_{kj} is the (k,j) element of $\boldsymbol{\Sigma} = \mathrm{Var}(\mathbf{X})$, with $\mathbf{X} = (X_1, \ldots, X_p)^T$. Yet, $\beta_j^\star$ can be far away from zero. In other words, under the conditions listed above, X_j is a hidden signature variable. To demonstrate that, let us consider the case in which $p = 2000$, with true regression coefficients $\boldsymbol{\beta}^\star = (3, 3, 3, 3, 3, -7.5, 0, \cdots, 0)^T$, and all variables follow the standard normal distribution with equal correlation 0.5, and ε follows the standard normal distribution. By design, X_6 is a hidden signature variable, which is marginally uncorrelated with the response Y. Based on a random sample of size 100 from the model, we fit marginal regression and obtain the marginal estimates $\{\hat{\beta}_j^M\}_{j=1}^p$. The magnitudes of these estimates are summarized by their averages over three groups: indices 1 to 5 (denoted by $\boldsymbol{\beta}_{1:5}^M$), 6 and indices 7 to 2000. Clearly, the magnitude on the first group should be the largest, followed by the third group. Figure 1(a) depicts the distributions of those marginal magnitudes based on 10000 simulations. Clearly variable X_6 can not be selected by marginal screening.

Adapting the conditional screening approach gives a very different result. Condi-

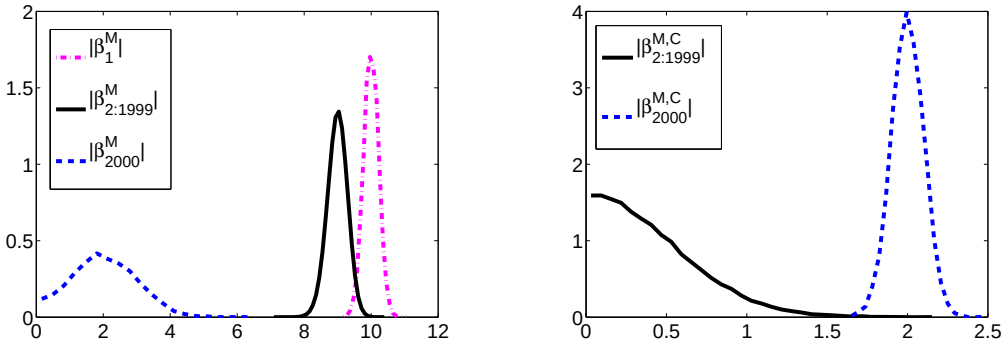


(a) Distribution of $\left|\hat{\beta}^M\right|$ (b) Distribution of $\left|\hat{\beta}^M_{\mathcal{C}}\right|$

Figure 1: Benefits of conditioning against false negatives. Left panel: the distributions of the averages of magnitudes $|\hat{\beta}^M_j|$ of marginal regression coefficients over three groups of variables 1:5, 6, 7:2000. Right panel: the distributions of the averages of the magnitude $|\hat{\beta}^M_{\mathcal{C}j}|$ of conditional marginal regression coefficients over two groups of variables: 6 and 7:2000.

tioning upon the first five variables, conditional correlation between X_6 and Y has a large magnitude. With the same simulated data as in the above example, the regression coefficient $\hat{\beta}^M_{\mathcal{C}j}$ of X_j in the joint model with the first five variables is computed. This measures the conditional contribution of variable X_j in presence of the first five variables. Again, the magnitudes $\{|\hat{\beta}^M_{\mathcal{C}j}|\}_{j=6}^{2000}$ are summarized into two values: $|\hat{\beta}^M_{\mathcal{C}6}|$ and the average of $\{|\hat{\beta}^M_{\mathcal{C}j}|\}_{j=7}^{2000}$. The distributions of those over 10000 simulations are also depicted in Figure 1(b). Clearly, the variable X_6 has higher marginal contributions than others. That is, conditioning helps recruiting the hidden signature variable.

Secondly, conditional screening helps for reducing the number of false negatives. Marginal screening can fail when there are covariates in the non-active set that are highly correlated with active variables. To appreciate this, consider the linear model (1) again with sparse regression coefficients $\boldsymbol{\beta}^\star = (10, 0, \cdots, 0, 1)^T$, equi-correlation 0.9 among all covariates except X_{2000}, which is independent of the rest of the covari-



(a) Distribution of $\left|\hat{\beta}^M\right|$ (b) Distribution of $\left|\hat{\beta}^M_{\mathcal{C}}\right|$

Figure 2: Benefits of conditioning against false positives. Left panel: the distributions of the magnitude $|\hat{\beta}^M_j|$ of marginal regression coefficients over three groups of variables 1, 2:1999 and 2000. Right panel: the distributions of the magnitude $|\hat{\beta}^M_{\mathcal{C}j}|$ of conditional marginal regression coefficients over two groups of variables: 2:1999 and 2000.

ates. This setting gives

$$\mathrm{Cov}(X_1, Y) = 10, \quad \mathrm{Cov}(X_{2000}, Y) = 1, \quad \text{and} \quad \mathrm{Cov}(X_j, Y) = 9, \quad \text{for } j \neq 1, 2000.$$

In this case, marginal utilities for all nonactive variables are higher than that for the active variable X_{2000}. A summary similar to Figure 1 is shown in the left panel of Figure 2. Therefore, based on SIS (sure independence screening) in Fan and Lv (2008), the active variable X_{2000} has the least priority to be included. By using the conditional screening approach in which the covariate X_1 is conditioned upon (used in the joint fit), marginal utilities of the spurious variables are significantly reduced. The distributions of the average of the magnitude of the conditional fitted coefficients $\{|\hat{\beta}^M_{\mathcal{C}j}|\}_{j=2}^{1999}$ and $|\beta^M_{\mathcal{C}2000}|$ are shown in the right panel of Figure 2. Clearly, the nonactive variables are significantly demoted by conditioning.

Finally, as shown by Fan and Lv (2008) and Fan and Song (2010), for a given threshold of marginal utility, the size of the selected variables depends on the correlation among covariates, as measured by the largest eigenvalue of $\mathbf{\Sigma}$: $\lambda_{\max}(\mathbf{\Sigma})$.

The larger the quantity, the more variables have to be selected in order to have a sure screening property. By using conditional screening, the relevant quantity now becomes $\lambda_{\max}\left(\mathbf{\Sigma}_{\mathbf{X}_{\mathcal{D}}|\mathbf{X}_{\mathcal{C}}}\right)$, where $\mathbf{X}_{\mathcal{C}}$ refers to the q covariates that we will condition upon and $\mathbf{X}_{\mathcal{D}}$ is the rest of the variables. Conditioning helps reducing correlation among covariates $\mathbf{X}_{\mathcal{D}}$. This is particularly the case when covariates $\mathbf{X}$ share some common factors, as in many biological (e.g. treatment effects) and financial studies (e.g. market risk factors). To illustrate the benefits we consider the case where $\mathbf{X}$ is given by equally correlated normal random variables. Simple calculations yield that $\lambda_{\max}\left(\mathbf{\Sigma}_{\mathbf{X}_{\mathcal{D}}}\right) = (1-r) + rd$ where r is the common correlation and $d = p - q$. As $\mathbf{X}$ has a normal distribution, the conditional covariance matrix can be calculated easily and it can be shown that

$$\lambda_{\max}\left(\mathbf{\Sigma}_{\mathbf{X}_{\mathcal{D}}|\mathbf{X}_{\mathcal{C}}}\right) = (1-r) + rd\frac{1-r}{1-r+rq}. \tag{2}$$

Note that when $q = 0$, the formula reduces to the unconditional one. It is clear that conditioning helps reducing the correlation among the variables. To quantify the degree of de-correlation, Figure 3 depicts the ratio $\lambda_{\max}\left(\mathbf{\Sigma}_{\mathbf{X}_{\mathcal{D}}}\right)/\lambda_{\max}\left(\mathbf{\Sigma}_{\mathbf{X}_{\mathcal{D}}|\mathbf{X}_{\mathcal{C}}}\right)$ as a function of r for various choices of q when $d = 1000$. The reduction is dramatic, in particular when r is large or q is large. The benefits of conditioning are clearly evidenced.

In this paper, we propose the conditional screening technique and formally establish the conditions under which it has a sure screening property. We also give an upper bound for the number of selected variables for each given threshold value. Two data-driven methods for choosing the thresholding parameter are proposed to facilitate the practical use of the conditional screening technique.

The rest of the paper is organized as follows. In Section 2, we introduce the conditional sure independence screening procedure. The sure independence screening

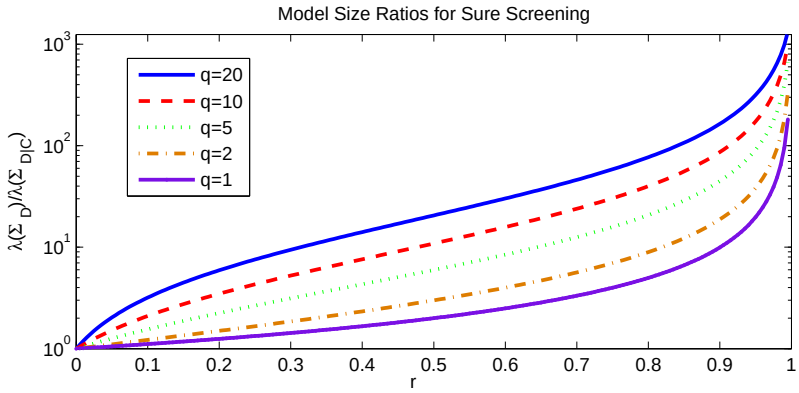


Figure 3: Ratio of maximum eigenvalues of unconditioned and conditioned covariance matrix.

property and the uniform convergence of the conditional marginal maximum likelihood estimator are presented in Section 3. In Section 4, two approaches are proposed to choose the thresholding parameter for CSIS. Finally, we examine the performance of our procedure in Section 5 on simulated and real data. The details of the proofs are deferred to the Appendix.

2 CONDITIONAL INDEPENDENCE SCREENING

2.1 Generalized Linear Models

Generalized linear models assume that the conditional probability density of the random variable Y given $\mathbf{X} = \mathbf{x} = (x_1 \ldots, x_p)^T$ belongs to an exponential family

$$f(y|\mathbf{x};\theta) = \exp\Big(y\theta(\mathbf{x}) - b(\theta(\mathbf{x})) + c(\mathbf{x};y)\Big), \tag{3}$$

where $b(\cdot)$ and $c(\cdot)$ are specific known functions in the canonical parameter $\theta(\mathbf{x})$. Note that we ignore the dispersion parameter ϕ, since the interest only focuses on

estimation of the mean regression function. However, it is easy to include a dispersion parameter ϕ. Under model (3), we have the regression function

$$\mathbb{E}(Y|\mathbf{X} = \mathbf{x}) = b'(\theta(\mathbf{x})).$$

The canonical parameter is further parameterized as

$$\theta(\mathbf{x}) = \mathbf{x}^T\boldsymbol{\beta}^\star,$$

namely the canonical link is used in modeling the mean regression function. Well known distributions in this exponential family include the normal, binomial, Poisson, and Gamma distributions.

In the ultrahigh dimensional sparse linear model, we assume that the true parameter $\boldsymbol{\beta}^\star = (\beta_1^\star, \ldots, \beta_p^\star)^T$ is sparse. Namely, the set

$$\mathcal{M}_\star = \{j = 1, \ldots, p : \beta_j^\star \neq 0\},$$

is small. Our aim is to estimate the set $\mathcal{M}_\star$ and coefficient vector $\boldsymbol{\beta}^\star$, as well as predicting the outcome Y. This is a more challenging task than just predicting Y as in many machine learning problems. When the dimensionality is ultrahigh, one often employs a screening technique first to reduce the model size. It is particularly effective in distributed computation for dealing with "Big Data".

2.2 Conditional Screening

Conditional screening assumes that there is a set of variables $\mathbf{X}_\mathcal{C}$ that are known to be related to the response Y and we wish to recruit additional variables from the rest of variables, given by $\mathbf{X}_\mathcal{D}$, to better explain the response variable Y. For simplicity

of notation, we assume without loss of generality that $\mathcal{C}$ is the set of first q variables and $\mathcal{D}$ is the remaining set of $d = p - q$ variables. We will use the notation

$$\boldsymbol{\beta}_{\mathcal{C}} = (\beta_1, \ldots, \beta_q)^T \in I\!R^q, \quad \text{and} \quad \boldsymbol{\beta}_{\mathcal{D}} = (\beta_{q+1}, \ldots, \beta_p)^T \in I\!R^d,$$

and similar notation for $\mathbf{X}_{\mathcal{C}}$ and $\mathbf{X}_{\mathcal{D}}$. Assume without loss of generality that the covariates have been standardized so that

$$\mathbb{E}(X_j) = 0 \quad \text{and} \quad \mathbb{E}(X_j^2) = 1 \quad \text{for } j \in \mathcal{D}.$$

Given a random sample $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ from the generalized linear model (3) with the canonical link, the conditional maximum marginal likelihood estimator $\hat{\boldsymbol{\beta}}_{\mathcal{C}j}^M$ for $j = q+1, \ldots, p$ is defined as the minimizer of the (negative) marginal log-likelihood

$$\hat{\boldsymbol{\beta}}_{\mathcal{C}j}^M = \text{argmin}_{\boldsymbol{\beta}_{\mathcal{C}}, \beta_j} \mathbb{P}_n \big\{ l(\mathbf{X}_{\mathcal{C}}^T \boldsymbol{\beta}_{\mathcal{C}} + X_j \beta_j, Y) \big\}, \tag{4}$$

where $l(\theta, Y) = b(\theta) - \theta Y$ and $\mathbb{P}_n f(X, Y) = n^{-1} \sum_{i=1}^n f(X_i, Y_i)$ is the empirical measure. Denote from now on by $\hat{\beta}_j^M$ the last element of $\hat{\boldsymbol{\beta}}_{\mathcal{C}j}^M$. It measures the strength of the conditional contribution of X_j given $\mathbf{X}_{\mathcal{C}}$. In the above notation, we assume that the intercept is used and is incorporated in the vector $\mathbf{X}_{\mathcal{C}}$. Conditional marginal screening based on the estimated marginal magnitude is to keep the variables

$$\hat{\mathcal{M}}_{\mathcal{D},\gamma} = \{ j \in \mathcal{D} : |\hat{\beta}_j^M| > \gamma \}, \tag{5}$$

for a given thresholding parameter γ. Namely, we recruit variables with large additional contribution given $\mathbf{X}_{\mathcal{C}}$. This method will be referred to as conditional sure independence screening (CSIS). It depends, however, on the scale of $\mathbb{E}_L(X_j | \mathbf{X}_{\mathcal{C}})$ and $\mathbb{E}_L(Y | \mathbf{X}_{\mathcal{C}})$ to be defined in Section 3.1. A scale-free method is to use the likelihood

reduction of the variable X_j given $\mathbf{X}_{\mathcal{C}}$, which is equivalent to computing

$$\hat{R}_{\mathcal{C}j} = \min_{\boldsymbol{\beta}_{\mathcal{C}},\beta_j} \mathbb{P}_n\big\{l(\mathbf{X}_{\mathcal{C}}^T\boldsymbol{\beta}_{\mathcal{C}} + X_j\beta_j, Y)\big\}, \tag{6}$$

after ignoring the common constant $\min_{\boldsymbol{\beta}_{\mathcal{C}}} \mathbb{P}_n\big\{l(\mathbf{X}_{\mathcal{C}}^T\boldsymbol{\beta}_{\mathcal{C}}, Y)\big\}$. The smaller $\hat{R}_{\mathcal{C}j}$, the more the variable X_j contributes in presence of $\mathbf{X}_{\mathcal{C}}$. This leads to an alternative method based on the likelihood ratio statistics: recruit additional variables according to

$$\tilde{\mathcal{M}}_{\mathcal{D},\tilde{\gamma}} = \{j \in \mathcal{D} : \hat{R}_{\mathcal{C}j} < \tilde{\gamma}\}, \tag{7}$$

where $\tilde{\gamma}$ is a thresholding parameter. This method will be referred to as conditional maximum likelihood ratio screening (CMLR).

3 SURE SCREENING PROPERTIES

In order to prove the sure screening property of our method, we first need some properties on the population level. Let $\boldsymbol{\beta}_{\mathcal{C}j} = (\boldsymbol{\beta}_{\mathcal{C}}^T, \beta_j)^T$, $\mathbf{X}_{\mathcal{C}j} = (\mathbf{X}_{\mathcal{C}}^T, X_j)^T$, and

$$\boldsymbol{\beta}_{\mathcal{C}j}^M = \text{argmin}_{\boldsymbol{\beta}_{\mathcal{C}},\beta_j} \mathbb{E}\, l(\mathbf{X}_{\mathcal{C}}^T\boldsymbol{\beta}_{\mathcal{C}} + X_j\beta_j, Y), \tag{8}$$

with the expectation taken under the true model. Then, $\boldsymbol{\beta}_{\mathcal{C}j}^M$ is the population version of $\hat{\boldsymbol{\beta}}_{\mathcal{C}j}^M$. To establish the sure screening property, we need to show that the marginal regression coefficient β_j^M, the last component of $\boldsymbol{\beta}_{\mathcal{C}j}^M$, provides useful probes for the variables in the joint model $\mathcal{M}_\star$ and its sample version $\hat{\beta}_j^M$ is uniformly close to the population counterpart β_j^M. Therefore, the vector of marginal fitted regression coefficients $\hat{\boldsymbol{\beta}}_{\mathcal{C}j}^M$ is useful for finding the variables in $\mathcal{M}_\star$.

3.1 Properties on Population Level

Since we are fitting d marginal regressions, that is we are using only $q+1$ out of the p original predictors, we need to introduce model misspecifications. Thus, we do not expect that the marginal regression coefficient β_j^M is equal to the joint regression parameter $\beta_j^\star$. However, we hope that when the joint regression coefficient $|\beta_j^\star|$ exceeds a certain threshold, $|\beta_j^M|$ exceeds another threshold in most cases. Therefore, the marginal conditional regression coefficients provide useful probes for the joint regression.

By (8), the marginal regression coefficients $\boldsymbol{\beta}_{\mathcal{C}j}^M$ satisfy the score equation

$$\mathbb{E}\, b'(\mathbf{X}_{\mathcal{C}j}^T \boldsymbol{\beta}_{\mathcal{C}j}^M)\mathbf{X}_{\mathcal{C}j} = \mathbb{E}\, Y\mathbf{X}_{\mathcal{C}j} = \mathbb{E}\, b'(\mathbf{X}^T\boldsymbol{\beta}^\star)\mathbf{X}_{\mathcal{C}j}, \tag{9}$$

where the second equality follows from the fact that $\mathbb{E}(Y|\mathbf{X}) = b'(\mathbf{X}^T\boldsymbol{\beta}^\star)$. Without using the additional variable X_j, the baseline parameter is given by

$$\boldsymbol{\beta}_{\mathcal{C}}^M = \mathrm{argmin}_{\boldsymbol{\beta}_{\mathcal{C}}}\, \mathbb{E}\, l(\mathbf{X}_{\mathcal{C}}^T\boldsymbol{\beta}_{\mathcal{C}}, Y), \tag{10}$$

and satisfies the equation

$$\mathbb{E}\, b'(\mathbf{X}_{\mathcal{C}}^T \boldsymbol{\beta}_{\mathcal{C}}^M)\mathbf{X}_{\mathcal{C}} = \mathbb{E}\, Y\mathbf{X}_{\mathcal{C}} = \mathbb{E}\, b'(\mathbf{X}^T\boldsymbol{\beta}^\star)\mathbf{X}_{\mathcal{C}}. \tag{11}$$

We assume that the problems at marginal level are fully identifiable, namely, the solutions $\boldsymbol{\beta}_{\mathcal{C}}^M$ and $\boldsymbol{\beta}_{\mathcal{C}j}^M$ are unique.

To understand the conditional contribution, we introduce the concept of the conditional linear expectation. We use the notation

$$\mathbb{E}_L(Y|\mathbf{X}_{\mathcal{C}}) = b'(\mathbf{X}_{\mathcal{C}}^T\boldsymbol{\beta}_{\mathcal{C}}^M), \quad \text{and} \quad \mathbb{E}_L(Y|\mathbf{X}_{\mathcal{C}j}) = b'(\mathbf{X}_{\mathcal{C}j}^T\boldsymbol{\beta}_{\mathcal{C}j}^M), \tag{12}$$

which is the best linearly fitted regression within the class of linear functions. Similarly, we use the notation $\mathbb{E}_L(X_j|\mathbf{X}_\mathcal{C})$ to denote the best linear regression fit of X_j by using $\mathbf{X}_\mathcal{C}$. Then, equation (11) can be more intuitively expressed as

$$\mathbb{E}(Y - \mathbb{E}_L(Y|\mathbf{X}_\mathcal{C}))\mathbf{X}_\mathcal{C} = 0. \tag{13}$$

Note that the conditioning in this paper is really a conditioning linear fit and the conditional expectation is really the conditional linear expectation. This facilitates the implementation of the conditional (linear) screening in high-dimensional, but adds some technical challenges in the proof.

Let us examine the implication marginal signal, i.e. β_j^M. When $\beta_j^M = 0$, by (9), the first q components of $\boldsymbol{\beta}_{\mathcal{C}j}^M$, denoted by $\boldsymbol{\beta}_{\mathcal{C}j1}^M$, should be equal to $\boldsymbol{\beta}_\mathcal{C}^M$ by uniqueness of equation (11). Then, equation (9) on the component X_j entails

$$\mathbb{E}\, b'(\mathbf{X}_\mathcal{C}^T\boldsymbol{\beta}_\mathcal{C}^M)X_j = \mathbb{E}\, YX_j, \quad \text{or} \quad \mathbb{E}\, X_j(Y - \mathbb{E}_L(Y|\mathbf{X}_\mathcal{C})) = 0.$$

Using (13), the above condition can be more comprehensively expressed as

$$\mathrm{Cov}_L\left(Y, X_j\middle|\mathbf{X}_\mathcal{C}\right) \equiv \mathbb{E}(X_j - \mathbb{E}_L(X_j|\mathbf{X}_\mathcal{C}))(Y - \mathbb{E}_L(Y|\mathbf{X}_\mathcal{C})) = 0. \tag{14}$$

This proves the necessary condition of the following theorem.

Theorem 1. *For $j \in \mathcal{D}$, the marginal regression parameters $\beta_j^M = 0$ if and only if $\mathrm{Cov}_L\left(Y, X_j\middle|\mathbf{X}_\mathcal{C}\right) = 0$.*

Proof of the sufficient part is given in Appendix A.1. In order to have the sure screening property at the population level of equation (8), the important variables $\{X_j, j \in \mathcal{M}_{\star\mathcal{D}}\}$ should be conditionally correlated with the response, where $\mathcal{M}_{\star\mathcal{D}} = \mathcal{M}_\star \cap \mathcal{D}$. Moreover, if X_j (with $j \in \mathcal{M}_{\star\mathcal{D}}$) is conditionally correlated with

the response, the regression coefficient β_j^M is non-vanishing. The sure screening property of conditional MLE (CMLE), given by equation (5), will be guaranteed if the minimum marginal signal strength is stronger than the estimation error. This will be shown in Theorem 2 and requires Condition 1. The details of the proof are relegated to Appendix A.2.

Condition 1.

(i) For $j \in \mathcal{M}_{\star\mathcal{D}}$, there exists a positive constant $c_1 > 0$ and $\kappa < 1/2$ such that $|\mathrm{Cov}_L(Y, X_j|\mathbf{X}_\mathcal{C})| \geq c_1 n^{-\kappa}$.

(ii) Let m_j be the random variable defined by

$$m_j = \frac{b'\left(\mathbf{X}_{\mathcal{C}j}^T \boldsymbol{\beta}_{\mathcal{C}j}^M\right) - b'\left(\mathbf{X}_\mathcal{C}^T \boldsymbol{\beta}_\mathcal{C}^M\right)}{\mathbf{X}_{\mathcal{C}j}^T \boldsymbol{\beta}_{\mathcal{C}j}^M - \mathbf{X}_\mathcal{C}^T \boldsymbol{\beta}_\mathcal{C}^M}.$$

Then, $\mathbb{E}\, m_j X_j^2 \leq c_2$ uniformly in $j = q+1, \ldots, p$.

Note that, by strict convexity of $b(\theta)$, $m_j > 0$ almost surely. When we are dealing with linear models, i.e. $b(\theta) = \theta^2/2$, then $m_j = 1$ and Condition 1(ii) requires that $\mathbb{E}X_j^2$ is bounded uniformly, which is automatically satisfied by the normalization condition $\mathbb{E}\, X_j^2 = 1$.

Theorem 2. *If Condition 1 holds, then there exists a $c_3 > 0$ such that*

$$\min_{j \in \mathcal{M}_{\mathcal{D}\star}} \left|\beta_j^M\right| \geq c_3 n^{-\kappa}.$$

3.2 Properties on Sample Level

In this section, we prove the uniform convergence of the conditional marginal maximum likelihood estimator and the sure screening property of the conditional sure

independence screening method. In addition we provide an upper bound on the size of the set of selected variables $\hat{\mathcal{M}}_{\mathcal{D},\gamma}$.

Since the log-likelihood of a generalized linear model with the canonical link is concave, $\mathbb{E}(l(Y, \mathbf{X}_{\mathcal{C}j}^T \boldsymbol{\beta}_{\mathcal{C}j}))$ has a unique minimizer over $\boldsymbol{\beta}_{\mathcal{C}j} \in \mathcal{B}$ at an interior point $\boldsymbol{\beta}_{\mathcal{C}j}^M$, where $\mathcal{B} = \{|\beta_1^M| \leq B, \ldots, |\beta_q^M| \leq B, |\beta_j^M| \leq B\}$ is the set over which the marginal likelihood is maximized. To obtain the uniform convergence result at the sample level, a few more conditions on the conditional marginal likelihood are needed.

Condition 2.

(i) For the Fisher information $I_j(\boldsymbol{\beta}_{\mathcal{C}j}) = \mathbb{E}(b''(\mathbf{X}_{\mathcal{C}j}^T \boldsymbol{\beta}_{\mathcal{C}j}) \mathbf{X}_{\mathcal{C}j} \mathbf{X}_{\mathcal{C}j}^T)$, its operator norm, $\|I_j(\boldsymbol{\beta}_{\mathcal{C}j})\|_{\mathcal{B}}$ is bounded, where

$$\|I_j(\boldsymbol{\beta}_{\mathcal{C}j})\|_{\mathcal{B}} = \sup_{\boldsymbol{\beta}_{\mathcal{C}j} \in \mathcal{B}, \|\mathbf{x}_{\mathcal{C}j}\|=1} \|I_j(\boldsymbol{\beta}_{\mathcal{C}j})^{1/2} \mathbf{x}_{\mathcal{C}j}\|,$$

and $\|\cdot\|$ is the Euclidian norm.

(ii) There exists some positive constants r_0, r_1, s_0, s_1 and α such that for sufficiently large t

$$P(|X_j| > t) \leq r_1 \exp(-r_0 t^\alpha) \quad \text{for } j = 1, \ldots, p$$

and that

$$\mathbb{E}\left(b(\mathbf{X}^T \boldsymbol{\beta}^\star + s_0) - b(\mathbf{X}^T \boldsymbol{\beta}^\star)\right) + \mathbb{E}\left(b(\mathbf{X}^T \boldsymbol{\beta}^\star - s_0) - b(\mathbf{X}^T \boldsymbol{\beta}^\star)\right) \leq s_1.$$

(iii) The second derivative of $b(\theta)$ is continuous and positive. There exists an $\varepsilon_1 > 0$ such that for all $j = q+1, \ldots, p$:

$$\sup_{\boldsymbol{\beta}_{\mathcal{C}j} \in \mathcal{B}, \|\boldsymbol{\beta}_{\mathcal{C}j} - \boldsymbol{\beta}_{\mathcal{C}j}^M\| \leq \varepsilon_1} |\mathbb{E}\, b(\mathbf{X}_{\mathcal{C}j}^T \boldsymbol{\beta}_{\mathcal{C}j}) I(|X_j| > K_n)| \leq o(n^{-1}),$$

where $I(\cdot)$ is the indicator function and K_n is an arbitrarily large constant such that for a given $\boldsymbol{\beta}$ in $\mathcal{B}$, the function $l(\mathbf{x}^T\boldsymbol{\beta}, y)$ is Lipschitz for all $(\mathbf{x}, y)$ in $\Lambda_n = \{\mathbf{x}, y : \|\mathbf{x}\|_\infty \leq K_n, |y| \leq K_n^\star\}$ with $K_n^* = r_0 K_n^\alpha / s_0$.

(iv) For all $\boldsymbol{\beta}_{\mathcal{C}j} \in \mathcal{B}$, we have

$$\mathbb{E}\left(l(\mathbf{X}_{\mathcal{C}j}^T\boldsymbol{\beta}_{\mathcal{C}j}, Y) - l(\mathbf{X}_{\mathcal{C}j}^T\boldsymbol{\beta}_{\mathcal{C}j}^M, Y)\right) \geq V\|\boldsymbol{\beta}_{\mathcal{C}j} - \boldsymbol{\beta}_{\mathcal{C}j}^M\|^2,$$

for some positive V, bounded from below uniformly over $j = q+1, \ldots, p$.

The first three conditions given in Condition 2 are satisfied for almost all of the commonly used generalized linear models. Examples include linear regression, logistic regression, and Poisson regression. The first part of Condition 2(ii) puts an exponential bound on the tails of X_j.

In the following theorem, the uniform convergence of our conditional marginal maximum likelihood estimator is stated as well as the sure screening property of the procedure. The proof of this theorem is deferred to Appendix A.3.

Theorem 3. *Suppose that Condition 2 holds. Let* $k_n = b'(K_n B(q+1)) + r_0 K_n^\alpha / s_0$, *with* K_n *given in Condition 2.*

(i) If $n^{1-2\kappa} k_n^{-2} K_n^{-2} \to \infty$, *then for any* $c_3 > 0$, *there exists a positive constant* c_4 *such that*

$$\begin{aligned} &\mathbb{P}\left(\max_{q+1\leq j\leq p} |\hat{\beta}_j^M - \beta_j^M| \geq c_3 n^{-\kappa}\right) \\ \leq\ & d\exp\left(-c_4 n^{1-2\kappa}(k_n K_n)^{-2}\right) + dnr_2 \exp\left(-r_0 K_n^\alpha\right), \end{aligned}$$

where $r_2 = qr_1 + s_1$.

(ii) If in addition, Condition 1 holds, then by taking $\gamma = c_5 n^{-\kappa}$ *with* $c_5 \leq c_3/2$, *we*

have

$$\mathbb{P}\left(\mathcal{M}_{\star\mathcal{D}} \subset \hat{\mathcal{M}}_{\mathcal{D},\gamma}\right) \geq 1 - s\exp\left(-c_4 n^{1-2\kappa}(k_n K_n)^{-2}\right) - nr_2 s\exp\left(-r_0 K_n^{\alpha}\right),$$

for some constant c_5, where $s = |\mathcal{M}_{\star\mathcal{D}}|$ the size of the set of nonsparse elements.

Note that the sure screening property, stated in the second conclusion of Theorem 3, depends only on the size s of the set of nonsparse elements and not on the dimensionality d or p. This can be seen in the second conclusion above. This result is understandable since we only need the elements in $\mathcal{M}_{\star\mathcal{D}}$ to pass the threshold, and this only requires the uniform convergence of $\hat{\beta}_j^M$ over $j \in \mathcal{M}_{\star\mathcal{D}}$.

The truncation parameter K_n appears on both terms of the upper bound of the probability. There is a trade-off on this choice. For the Bernoulli model with logistic link, $b'(\cdot)$ is bounded and the optimal order for K_n is $n^{(1-2\kappa)/(\alpha+2)}$. In this case, the conditional sure independence screening method can handle the dimensionality

$$\log d = o\left(n^{(1-2\kappa)\alpha/(\alpha+2)}\right),$$

which guarantees that the upper bound in Theorem 3 converges to zero. A similar result for unconditional screening is shown in Fan and Song (2010). In particular, when the covariates are bounded, we can take $\alpha = \infty$, and when covariates are normal, we have that $\alpha = 2$. For the normal linear model, following the same argument as in Fan and Song (2010), the optimal choice is $K_n = n^{(1-2\kappa)/A}$ where $A = \max\{\alpha + 4, 3\alpha + 2\}$. Then, conditional sure independence screening can handle dimensionality

$$\log d = o\left(n^{-(1-2\kappa)\alpha/A}\right),$$

which is of order $o(n^{-(1-2\kappa)/4})$ when $\alpha = 2$.

We have just stated the sure screening property of our CSIS method, that is $\hat{\mathcal{M}}_{\mathcal{D},\gamma} \supset \hat{\mathcal{M}}_{\star\mathcal{D}}$. However, a good screening method does not only possess sure screening, but also retains a small set of variables after thresholding. Below, we give a bound on the size of the selected set of variables, under the following additional conditions.

Condition 3.

(i) The variance $\mathrm{Var}(\mathbf{X}^T\boldsymbol{\beta}^\star) = \boldsymbol{\beta}^{\star T}\boldsymbol{\Sigma}\boldsymbol{\beta}^\star$ and $b''(\cdot)$ are bounded.

(ii) The minimum eigenvalue of the matrix $\mathbb{E}[m_j\mathbf{X}_{\mathcal{C}j}\mathbf{X}_{\mathcal{C}j}^T]$ is larger than a positive constant, uniformly over j, where m_j is defined in Condition 1(ii).

(iii) Letting

$$\mathbf{Z} = \mathbb{E}\Big\{\mathbb{E}\left[\mathbf{X}_{\mathcal{D}}|\mathbf{X}_{\mathcal{C}}\right]\left[\mathbf{X}^T\boldsymbol{\beta}^\star - \mathbf{X}_{\mathcal{C}}^T\boldsymbol{\beta}_{\mathcal{C}}^M\right]\Big\},$$

it holds that $\|\mathbf{Z}\|_2^2 = o\Big\{\lambda_{\max}\big(\boldsymbol{\Sigma}_{\mathcal{D}|\mathcal{C}}\big)\Big\}$, with $\lambda_{\max}\big(\boldsymbol{\Sigma}_{\mathcal{D}|\mathcal{C}}\big)$ the largest eigenvalue of $\boldsymbol{\Sigma}_{\mathcal{D}|\mathcal{C}} = \mathbb{E}[\mathbf{X}_{\mathcal{D}} - \mathbb{E}_L(\mathbf{X}_{\mathcal{D}}|\mathbf{X}_{\mathcal{C}})][\mathbf{X}_{\mathcal{D}} - \mathbb{E}_L(\mathbf{X}_{\mathcal{D}}|\mathbf{X}_{\mathcal{C}})]^T$.

As noted above, for the normal linear model, $b(\theta) = \theta^2/2$. Condition 3 (ii) requires that the minimum eigenvalue of $\mathbb{E}\,\mathbf{X}_{\mathcal{C}j}\mathbf{X}_{\mathcal{C}j}^T$ be bounded away from zero. In general, by strict convexity of $b(\theta)$, $m_j > 0$ almost surely. Thus, Condition 3(ii) is mild.

For the linear model with $b'(\theta) = \theta$, by (11),

$$\mathbb{E}\,\mathbf{X}_{\mathcal{C}}\mathbf{X}_{\mathcal{C}}^T\boldsymbol{\beta}_{\mathcal{C}}^M = \mathbb{E}\,\mathbf{X}_{\mathcal{C}}\mathbf{X}^T\boldsymbol{\beta}^\star$$

and hence $\mathbf{Z} = 0$ since $\mathbb{E}_L\left[\mathbf{X}_{\mathcal{D}}|\mathbf{X}_{\mathcal{C}}\right]$ is linear in $\mathbf{X}_{\mathcal{C}}$ by definition. Thus, Condition 3(ii) holds automatically.

From the proof of Theorem 4, without Condition 3(iii), Theorem 4 below continues to hold with $\boldsymbol{\Sigma}_{\mathcal{D}|\mathcal{C}}$ replaced by $\boldsymbol{\Sigma}_{\mathcal{D}|\mathcal{C}} + \mathbf{Z}\mathbf{Z}^T$.

Theorem 4. *Under Conditions 2 and 3, we have for* $\gamma = c_6 n^{-2\kappa}$*, there exists a* $c_4 > 0$ *such that*

$$\begin{aligned} & \mathbb{P}\big(|\hat{\mathcal{M}}_{\mathcal{D},\gamma}| \leq O\big(n^{2\kappa}\lambda_{\max}\big(\mathbf{\Sigma}_{\mathcal{D}|\mathcal{C}}\big)\big)\big) \\ \geq \; & 1 - d\Big(\exp\big(-c_4 n^{1-2\kappa}(k_n K_n)^{-2}\big) + nr_2 \exp\big(-r_0 K_n^{\alpha}\big)\Big). \end{aligned}$$

This theorem is proved in Appendix A.4.

4 SELECTION OF THE THRESHOLDING PARAMETER

In the previous section, we have shown that CSIS has the sure screening property when the thresholding level γ is chosen such that $\gamma \propto n^{-\kappa}$. Unfortunately, in practice γ, which relates to the minimum strength of marginal signals in the data, is always unknown. Therefore, γ has to be estimated from the data itself. Underestimating γ will result in a lot variables after screening, which leads to a large number of false positives, and similarly overestimation of γ will prevent sure screening.

In this section, we present two procedures that select a thresholding level for CSIS. The first approach is based on controlling the number of false positives by bounding the false discovery rate (FDR). This method uses the fact that quasi-likelihood estimates for GLMs enjoy asymptotic normality. The second approach, that we call random decoupling, uses a resampling technique to create the null model and to measure the maximum strength of noise. In random decoupling, we use marginal regression on the null model to obtain the marginal regression coefficients that are known to be zero. We use the maximum of these marginal coefficients of the null

model as a thresholding level.

4.1 Controlling FDR

It is well known that quasi-maximum likelihood estimates have an asymptotically normal distribution under general conditions (Heyde 1997; Gao et al. 2008). Then, for covariates j such that, $\beta_j^M = 0$, asymptotically it follows that

$$\left[I_j\left(\hat{\beta}_j^M\right)\right]^{1/2} \hat{\beta}_j^M \sim \mathcal{N}(0,1),$$

where $I_j\left(\hat{\beta}_j^M\right)$ denotes the element that corresponds to β_j in the information matrix $I_j(\boldsymbol{\beta}_{\mathcal{C}j})$.

Using this property, we can build a thresholding technique that bounds the proportion of elements j such that, $\beta_j^M = 0$. For the case, when $\beta_j^M = 0$ for all $j \in (\mathcal{M}_{\star\mathcal{D}})^c$, this rate is also called the false discovery rate in Zhao and Li (2012) and is given by $\mathbb{E}\left(\left|\hat{\mathcal{M}}_{\mathcal{D},\delta} \cap (\mathcal{M}_{\star\mathcal{D}})^c\right| / \left|(\mathcal{M}_{\star\mathcal{D}})^c\right|\right)$.

By choosing $\hat{\mathcal{M}}_{\mathcal{D},\delta} = \left\{j : I_j\left(\hat{\beta}_j^M\right)^{1/2} \left|\hat{\beta}_j^M\right| \geq \delta\right\}$, the expected false discovery rate is bounded above by $2\left(1 - \Phi\left(\delta\right)\right)$, where $\Phi(\cdot)$ is the distribution function of a standard normal random variable. This approach can also be seen as a modification of the method introduced by Zhao and Li (2012) for the Cox model. By setting δ to $\Phi^{-1}\left(1 - f/(2d)\right)$ where f is the maximum number of false positives we can tolerate, we obtain an expected false positive rate that is less than $f/(d - |\mathcal{M}_{\star\mathcal{D}}|)$ as the following theorem shows. The proof of this theorem is given in Appendix A.5.

Condition 4.

1. For any j, let $e_i = Y_i - b'(\mathbf{X}_{i,\mathcal{C}j}^T \boldsymbol{\beta}_{\mathcal{C}j})$ for $i = 1, \ldots, n$. For a given j, $\mathrm{Var}(e_i) \geq c_6$ for some positive c_6 and $i = 1, \ldots, n$ and $\sup_{i \geq 1} \mathbb{E}\left|e_i\right|^{2+\chi} < \infty$ for some $\chi > 0$.

2. For $j \in (\mathcal{M}_{\star\mathcal{D}})^c$, we have that $\mathrm{Cov}_L\left(Y, X_j \middle| \mathbf{X}_{\mathcal{C}}\right) = 0$.

Theorem 5. *Under Conditions 2, 3 and 4, if we choose*

$$\hat{\mathcal{M}}_{\mathcal{D},\delta} = \left\{ j : I_j \left(\hat{\beta}_j^M\right)^{1/2} \left|\hat{\beta}_j^M\right| \geq \delta \right\},$$

where $\delta = \Phi^{-1}\left(1 - f/(2d)\right)$ *and* f *is the number of false positives that can be tolerated, then, for some constant* $c_7 > 0$ *it holds that*

$$\mathbb{E}\left(\frac{\left|\hat{\mathcal{M}}_{\mathcal{D},\delta} \cap (\mathcal{M}_{\star\mathcal{D}})^c\right|}{|(\mathcal{M}_{\star\mathcal{D}})^c|}\right) \leq \frac{f}{d} + \frac{c_7}{\sqrt{n}}.$$

4.2 Random Decoupling

Random decoupling is an another procedure to select the thresholding parameter γ. It is used to create a null model, in which the data is formed by randomly permuting the rows of the last d columns of the design matrix, while keeping the first q columns of the design matrix intact. It is easy to see that by regressing Y on $\mathbf{X}^*_{\mathcal{C}j}$ where the rows of the design matrix corresponding to X_j ($j \notin \mathcal{C}$) have been randomly permuted, the obtained marginal values of $\hat{\beta}_j^{M*}$ is a statistical estimate of zero. These marginal estimates based on decoupled data measure the noise level of the estimates under the null model. Let $\hat{\gamma}^* = \max_{q+1 \leq j \leq p} |\hat{\beta}_j^{M*}|$. If $\hat{\gamma}^*$ is used as the thresholding value, all variables will be screened out based on the permuted data, which leads to no false positives in this case. In other words, it is the minimum thresholding parameter that makes no false positives. However, this $\hat{\gamma}^*$ depends on the realization of the permutation. To stabilize the thresholding value, one can repeat this exercise K times (e.g. 5 or 10 times), resulting in the values

$$\{|\hat{\beta}_{kj}^{M*}|, j = q+1, \cdots, p\}_{k=1}^K, \tag{15}$$

$\{\gamma_k^*\}_{k=1}^K$, where $\gamma_k^* = \max_{q+1\le j\le p} |\hat{\beta}_{kj}^{M*}|$.

Now, one can choose the maximum of $\{\gamma_k^*\}_{k=1}^K$, denoted by $\hat{\gamma}_{\max}^*$, as a thresholding value. A more stable choice is the τ-quantile of the values in (15), denoted it by γ_τ^*. A useful range for τ is $[.95, 1]$. Note that for $\tau = 1$, $\gamma_1^* = \hat{\gamma}_{\max}^*$. The selected variables are then

$$\hat{\mathcal{M}}_{\mathcal{D},\tau} = \{j : |\hat{\beta}_j^M| \ge \gamma_\tau^*\}.$$

In our numerical implementations, we do coupling five times, i.e. $K = 5$, and take $\tau = 0.99$. A similar idea for unconditional SIS appears already in Fan et al. (2011) for additive models.

5 NUMERICAL STUDIES

In this section, we demonstrate the performance of CSIS on simulated data and two empirical datasets. We compare CSIS versus sure independence screening and penalized least squares methods in a variety of settings.

5.1 Simulation Study

In the simulation study, we compare the performance of the proposed CSIS with Lasso (Tibshirani 1996) and unconditional SIS (Fan and Song 2010), in terms of variable screening. We vary the sample size from 100 to 500 for different scenarios and the number of predictors range from $p = 2,000$ to $40,000$. We present results with both the linear regression and the logistic regression.

We evaluate different screening methods on 200 simulated data sets based on the following criteria:

1. MMMS: median minimum model size of the selected models that are required to have a sure screening. The sampling variability of minimum model size (MMS) is measured by the robust standard deviation (RSD), which is defined as the associated interquartile range of MMS divided by 1.34 across 200 simulations.

2. FP: average number of false positives across the 200 simulations,

3. FN: average number of false negatives across 200 simulations.

We consider two different methods for selecting thresholding parameters: controlling FDR and random decoupling as outlined in the previous section, and we present false negatives and false positives for each method. Number of average false positives and false negatives are denoted by FP_π and FN_π for the random decoupling method and FP_{FDR} and FN_{FDR} for the FDR method. For the experiments with $p = 5,000$ and $p = 40,000$, we do not report the corresponding results for Lasso, since it is not proposed for variable screening, and the data-driven choice of regularization parameter for model selection is not necessarily optimal for variable screening.

5.1.1 Normal model

The first two simulated examples concern linear models introduced in the introduction, regarding the false positives and false negatives of unconditional SIS. We report the simulation results in Table 1 in which the column labeled "**Example 1**" refers to the first setting and column labeled "**Example 2**" referred to the second setting. These examples are designed to fail the unconditional SIS. Not surprisingly, SIS performs poorly in sure screening the variables, and conditional SIS easily resolves the problem. Also, we note that CSIS needs only one additional variable to have sure screening, whereas Lasso needs 15 additional variables. Both the FDR and the random decoupling methods return no false negatives under almost all of the simu-

lations. In other words, both of the data-driven thresholding methods ensured the sure screening property. However, they tend to be conservative, as the numbers of the false positives are high. The FDR approach has a relatively small number of false positives when used for conditional sure independent screening. For these settings, FDR method was found to be less conservative than the random decoupling method.

Table 1: The MMMS, its RSD (in parentheses), the "false negative" and "false positive" for the linear model with $n = 100$ and $p = 2,000$.

			Example 1		
	SIS	MLR	CSIS	CMLR	Lasso
MMMS	1995 (0)	1995 (0)	1 (0)	1 (0)	16 (0)
FP_π, FN_π	1531, 0.07	1859, 1.00	175, 0	112, 0	-
FP_{FDR}, FN_{FDR}	1934, 0.07	-	164, 0	-	-
			Example 2		
	SIS	MLR	CSIS	CMLR	Lasso
MMMS	1999 (0)	1999 (0)	1 (0)	1 (0)	16 (0)
FP_π, FN_π	1998, 0.01	1998, 0.04	543.1, 0	174, 0	-
FP_{FDR}, FN_{FDR}	1998, 0.01	-	15.66, 0	-	-

In the next two settings, we work with higher dimensions, $p = 5,000$ and $p = 40,000$. Following Fan and Song (2010), we generate the covariates from

$$X_j = \frac{\varepsilon_j + a_j \varepsilon}{\sqrt{1 + a_j^2}}, \tag{16}$$

where ε and $\{\varepsilon_j\}_{j=1}^{p/3}$ are i.i.d. standard normal random variables, $\{\varepsilon_j\}_{j=p/3+1}^{2p/3}$ are i.i.d. double exponential variables with location parameter zero and scale parameter one and $\{\varepsilon_j\}_{j=2p/3+1}^{p}$ are i.i.d. and follow a mixture normal distribution with two components $N(-1, 1)$, $N(1, 0.5)$ and equal mixture proportion. The covariates are standardized to have mean zero and variance one. Specifically, we consider the following two settings.

Example 3. In this setting, $p = 5,000$ and $s = 12$. The constants $a_1, \ldots, a_{100}$ are the same and chosen such that the correlation $\rho = \text{Corr}(X_i, X_j) = 0, 0.2, 0.4, 0.6$ and 0.8 among the first 100 variables and $a_{101} = \ldots = a_{5,000} = 0$.

Example 4. In this setting, $p = 40,000$ and $s = 6$. The constants $a_1, \ldots, a_{50}$ are generated from the normal random distribution with mean a and variance 1 and $a_{51} = \ldots, a_{40,000} = 0$. The constant a is taken such that $\mathbb{E}(\text{Corr}(X_i, X_j)) = 0, 0.2, 0.4, 0.6$ and 0.8 among the first r variables.

In both of the settings $\boldsymbol{\beta}^\star$ is generated from an alternating sequence of 1 and 1.3. For conditional sure independence screening, we condition on the first 2 covariates if $s = 6$ and we condition on the first 4 covariates if $s = 12$. Results are presented in Tables 2 and 3.

As expected, CSIS needs a smaller model size to have all the relevant variables, i.e. to possess the sure screening property. The effect is more pronounced for higher p and when more of the variables are correlated. A surprising result is that the advantage of conditioning is less when the correlation levels are higher. This is probably because of the fact that only 50 or 100 of the covariates are correlated, hence conditioning cannot fully utilize its advantages. We also see that, both methods for choosing the thresholding parameter are very effective. Both the FDR and empirical decoupling methods tend to have the sure screening property (no false negatives) and low number of false positives.

5.1.2 Binomial model

In this section data are given by i.i.d. copies of $(\mathbf{X}^T, Y)$, where the conditional distribution of Y given $\mathbf{X} = \mathbf{x}$ is a binomial distribution with probability of success

Table 2: The MMMS, its RSD (in parentheses), the "false positive" and "false negative" for Example 3 with $p = 5,000$ and $s = 4 + 8$.

Sure Independence Screening						
ρ	n	MMMS	FP_π	FN_π	$\mathrm{FP}_{\mathrm{FDR}}$	$\mathrm{FN}_{\mathrm{FDR}}$
0.00	300	86 (150)	0.21	4.61	20.75	1.23
0.20	100	43 (19)	34.17	0.82	87.70	0.03
0.40	100	56 (20)	87.38	0.00	101.75	0.00
0.60	100	58 (24)	88.20	0.00	101.68	0.00
0.80	100	63 (19)	88.17	0.00	101.64	0.00
Conditional Sure Independence Screening						
ρ	n	MMMS	FP_π	FN_π	$\mathrm{FP}_{\mathrm{FDR}}$	$\mathrm{FN}_{\mathrm{FDR}}$
0.00	300	57 (92)	0.16	3.74	21.09	0.97
0.20	100	31 (38)	2.74	2.97	29.93	0.69
0.40	100	29 (21)	17.65	0.99	48.03	0.42
0.60	100	32 (18)	44.93	0.23	55.60	0.29
0.80	100	42 (20)	67.55	0.06	50.01	0.66

Maximum Likelihood Ratio				
ρ	n	MMMS	FP_π	FN_π
0.00	300	86 (141)	0.77	0.23
0.20	100	43 (20)	47.88	0.03
0.40	100	52 (19)	88.48	0.00
0.60	100	58 (18)	88.78	0.00
0.80	100	60 (19)	88.75	0.00
Conditional Maximum Likelihood Ratio				
ρ	n	MMMS	FP_π	FN_π
0.00	300	18 (25)	0.72	1.65
0.20	100	23 (24)	5.71	1.44
0.40	100	23 (17)	16.45	0.76
0.60	100	28 (19)	23.81	0.55
0.80	100	33 (22)	26.09	0.69

Table 3: The MMMS, its RSD (in parentheses), the "false positive" and "false negative" for Example 4 with $p = 40,000$ and $s = 2 + 4$.

Sure Independence Screening						
ρ	n	MMMS	FP_π	FN_π	$\mathrm{FP}_{\mathrm{FDR}}$	$\mathrm{FN}_{\mathrm{FDR}}$
0.00	200	1133 (8246)	11.46	1.35	40.70	0.89
0.20	200	37 (1079)	30.37	0.61	57.83	0.46
0.40	200	37 (12)	37.92	0.32	62.71	0.24
0.60	200	37 (11)	41.35	0.17	65.61	0.13
0.80	200	36 (12)	43.73	0.02	66.89	0.02
Conditional Sure Independence Screening						
ρ	n	MMMS	FP_π	FN_π	$\mathrm{FP}_{\mathrm{FDR}}$	$\mathrm{FN}_{\mathrm{FDR}}$
0.00	200	13 (84)	5.83	0.57	31.04	0.43
0.20	200	16 (18)	16.62	0.31	41.07	0.23
0.40	200	16 (12)	23.89	0.11	45.61	0.08
0.60	200	17 (10)	29.83	0.03	50.05	0.01
0.80	200	17 (10)	37.41	0.00	54.34	0.02

Maximum Likelihood Ratio				
ρ	n	MMMS	FP_π	FN_π
0.00	200	1133 (8246)	13.61	0.19
0.20	200	41 (1503)	31.62	0.11
0.40	200	37 (12)	39.24	0.06
0.60	200	37 (11)	42.51	0.05
0.80	200	36 (12)	44.45	0.00
Conditional Maximum Likelihood Ratio				
ρ	n	MMMS	FP_π	FN_π
0.00	200	14 (261)	5.42	0.07
0.20	200	10 (21)	13.02	0.05
0.40	200	7 (10)	18.04	0.02
0.60	200	6 (5)	21.66	0.01
0.80	200	6 (3)	25.00	0.00

$\mathbb{P}(\mathbf{x}) = \exp\left(\mathbf{x}^T\boldsymbol{\beta}^\star\right)\left(1+\exp\left(\mathbf{x}^T\boldsymbol{\beta}^\star\right)\right)^{-1}$. The first two settings use the same setup of covariates and the same values for $\boldsymbol{\beta}^*$ as that in Example 1. The results are given in Table 4.

The results are almost the same as in the normal model. Conditional screening always lists the active variable as the most important one and Lasso only needs 16 variables. We also see that FDR and random decoupling methods are still successful, even though the setting is nonlinear.

The final settings for the binomial model use the same construction for the covariates as those in Examples 3 and 4. We again work with $s = 6$ and $s = 12$. For settings 2 and 3, $\boldsymbol{\beta}^\star$ is again given by a sequence of 1s and 1.3s. Results are given in Tables 5 and 6.

The results are the same as for the normal model. Due to the nonlinear nature of the problem, the minimum model size is slightly higher and the thresholding methods are less efficient. However, even though the covariates are not too correlated, overall advantage of conditional sure independence screening can easily be observed.

Table 4: The MMMS, its RSD (in parentheses) for the binomial model with the "false negative" and "false positive" settings for $n = 100$ and $p = 2,000$.

			Example 1		
	SIS	MLR	CSIS	CMLR	Lasso
MMMS	1995 (1.5)	1995 (1.5)	1 (0)	1 (0)	16 (0)
FP_π, FN_π	726, 0.07	1282, 1.00	35.72, 0	31.11, 0.01	-
FP_{FDR}, FN_{FDR}	1344, 0.07	-	34.05, 0	-	-
			Example 2		
	SIS	MLR	CSIS	CMLR	Lasso
MMMS	1999 (0)	1999 (0)	1 (0)	1(0)	16 (0)
FP_π, FN_π	1998, 0.03	1998, 0.14	462, 0	157, 0.01	-
FP_{FDR}, FN_{FDR}	1998, 0.04	-	5.65, 0	-	-

Table 5: The MMMS, its RSD (in parentheses), the "false positive" and "false negative" for Example 3 with the binomial model with $p = 5,000$ and $s = 4 + 8$.

Sure Independence Screening						
ρ	n	MMMS	FP_π	FN_π	$\mathrm{FP}_{\mathrm{FDR}}$	$\mathrm{FN}_{\mathrm{FDR}}$
0.00	300	215 (312)	0.19	5.78	23.06	1.77
0.20	300	27 (14)	73.22	0.02	109.56	0.00
0.40	300	49 (21)	88.19	0.00	110.15	0.00
0.60	300	56 (20)	88.17	0.00	110.00	0.00
0.80	300	68 (19)	88.20	0.00	110.34	0.00
Conditional Sure Independence Screening						
ρ	n	MMMS	FP_π	FN_π	$\mathrm{FP}_{\mathrm{FDR}}$	$\mathrm{FN}_{\mathrm{FDR}}$
0.00	300	87 (173)	20.15	1.24	24.03	1.11
0.20	300	19 (13)	49.25	0.14	53.87	0.11
0.40	300	34 (23)	67.82	0.17	61.72	0.31
0.60	300	43 (24)	77.36	0.21	53.83	1.01
0.80	300	66 (55)	78.33	0.51	36.16	3.42

Maximum Likelihood Ratio				
ρ	n	MMMS	FP_π	FN_π
0.00	300	210 (312)	20.18	0.08
0.20	300	28 (17)	107.08	0.00
0.40	300	47 (24)	107.82	0.00
0.60	300	60 (22)	107.47	0.00
0.80	300	67 (19)	107.30	0.00
Conditional Maximum Likelihood Ratio				
ρ	n	MMMS	FP_π	FN_π
0.00	300	83 (173)	20.18	1.21
0.20	300	20 (14)	45.27	0.20
0.40	300	39 (30)	53.48	0.49
0.60	300	71 (87)	49.47	1.15
0.80	300	402 (561)	35.42	3.43

Table 6: The MMMS, its RSD (in parentheses), the "false positive" and "false negative" for Example 4 with the binomial model with $p = 40,000$ and $s = 2 + 4$.

Sure Independence Screening						
ρ	n	MMMS	FP_π	FN_π	$\mathrm{FP}_{\mathrm{FDR}}$	$\mathrm{FN}_{\mathrm{FDR}}$
0.00	500	318 (7038)	12.04	1.22	51.32	0.79
0.20	500	38 (428)	32.47	0.57	68.46	0.38
0.40	500	38 (12)	38.66	0.27	73.42	0.19
0.60	500	38 (12)	41.99	0.16	76.11	0.10
0.80	500	35 (12)	43.84	0.03	77.38	0.02
Conditional Sure Independence Screening						
ρ	n	MMMS	FP_π	FN_π	$\mathrm{FP}_{\mathrm{FDR}}$	$\mathrm{FN}_{\mathrm{FDR}}$
0.00	500	13 (354)	5.96	0.66	42.51	0.49
0.20	500	15 (16)	14.51	0.39	49.79	0.27
0.40	500	16 (13)	19.11	0.24	51.68	0.22
0.60	500	19 (10)	22.80	0.21	51.78	0.24
0.80	500	19 (10)	26.39	0.14	46.49	0.64

Maximum Likelihood Ratio				
ρ	n	MMMS	FP_π	FN_π
0.00	500	309 (7030)	14.06	0.22
0.20	500	37 (255)	34.10	0.09
0.40	500	35.5 (11)	40.50	0.05
0.60	500	35.5 (12)	42.89	0.03
0.80	500	33.5 (14)	44.39	0.00
Conditional Maximum Likelihood Ratio				
ρ	n	MMMS	FP_π	FN_π
0.00	500	25 (892)	5.96	0.14
0.20	500	13 (62)	12.38	0.09
0.40	500	13 (22)	14.17	0.08
0.60	500	15.5 (17)	13.75	0.11
0.80	500	22 (72)	9.30	0.28

5.2 Leukemia Data

In this section, we demonstrate how CSIS can be used to do variable selection with an empirical dataset. We consider the leukemia dataset which was first studied by Golub et al. (1999) and is available at http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi. The data come from a study of gene expression in two types of acute leukemias, acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). Gene expression levels were measured using Affymetrix oligonucleotide arrays containing 7129 genes and 72 samples coming from two classes, namely 47 in class ALL and 25 in class AML. Among these 72 samples, 38 (27 ALL and 11 AML) are set to be training samples and 34 (20 ALL and 14 AML) are set as test samples. For this dataset we want to select the relevant genes, and based on the selected genes estimate whether the patient has ALL or AML. AML progresses very fast and has a poor prognosis. Therefore, a consistent classification method that relies on gene expression levels would be very beneficial for the diagnosis.

In order to choose the conditioning genes, we take a pair of genes described in Golub et al. (1999) that result in low test errors. First is Zyxin and the second one is Transcriptional activator hSNF2b. Both genes have empirically high correlations for the difference between people with AML and ALL.

After conditioning on the aforementioned genes, we implement our conditional selection procedure using logistic regression. Using the random decoupling method, we select a single gene, TCRD (T-cell receptor delta locus). Although this gene has not been discovered by the ALL/AML studies so far, it is known to have a relation with T-Cell ALL, a subgroup of ALL (Szczepaski et al. 2003). By using only these three genes, we are able to obtain a training error of 0 out of 38, and a test error of 1 out of 34. Similar studies in the past using sparse linear discriminant analysis or nearest shrunken centroids methods have obtained test errors of 1 by using more

than 10 variables. We conjecture that this is due to the high correlation between the Zyxin gene and others, and that this correlation masks the information contained in the TCRD gene.

5.3 Financial Data

In this section we illustrate the advantages of conditional sure independence screening on a factor model with financial data. From the website http://mba.tuck.dartmouth.edu/pages /faculty/ken.french/ we obtain 30 portfolios formed with respect to their industries. The returns for each portfolio are denoted by y^j (for $j = 1, \ldots 30$). The Fama-French three-factor model suggests that these returns follow the following equation

$$y_i^j = b_1^j f_i^1 + b_2^j f_i^2 + b_3^j f_i^3 + \varepsilon_i, \tag{17}$$

where f^1 is the excess return of the proxy market portfolio (given by the difference of the one-month T-Bill yield and the value weighted return of all stocks on NYSE, AMEX and NASDAQ), f^2 is the difference between the return of small and big companies (measured by the difference of returns of two portfolios, one with companies that have small market cap and one with companies with large market cap) and finally f^3 is the difference of return from value companies and growth companies. This model was first proposed by Fama and French (1993) and has been extensively analyzed since then. Since this seminal work, many other factors have been considered. In our numerical example, we used screening with the permutation test to detect if other factors are necessary. Besides the three factors mentioned above, we consider the momentum factor as an additional factor. This gives us 4 factors that are conditioned upon in CSIS. For each given industrial portfolio, we also consider the returns from the other 29 portfolios as potential prediction factors.

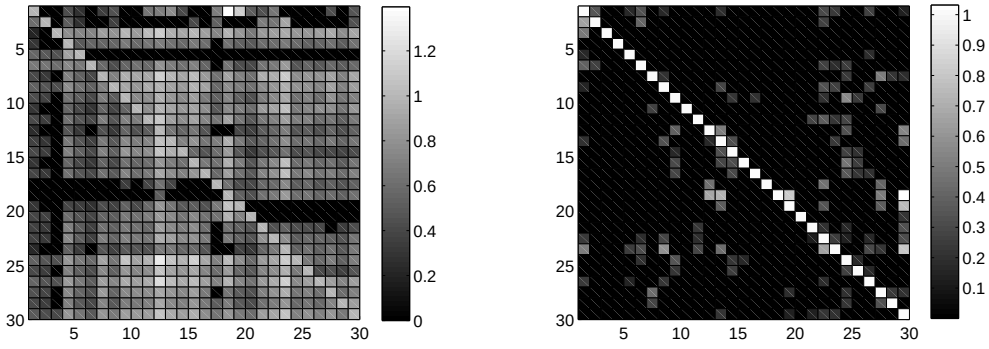
We use daily returns data from 1/3/2002 to 12/31/2007. For each portfolio (30 in total), we first consider the marginal screening without conditioning. On average, for each portfolio, marginal screening picks 25.3 among 29 other industrial portfolios as predictors. This is mainly due to correlations between the returns of different portfolios. We next consider conditional marginal screening, in which the three Fama-French factors and the momentum factor are conditioned upon. As expected, the number of the selected variables decreases significantly to an average of 4.8. That is, about 4.8 portfolios on average can still have some potential prediction power in presence of the aforementioned four major factors. The marginal and conditional fits of the values are given in Figure 4. The black parts indicate the variables which are not included.

It is seen from these results that, conditional screening is more advantageous compared to marginal screening if few of the factors are known to be important. Furthermore, when there is significant correlation between some of the factors, as shown in the introduction, marginal screening considers most of the factors as relevant. In almost all financial models, stock returns are correlated with the return of the market portfolio. Therefore, in variable selection for financial factor models with many variables, one should always consider the returns conditional on the main driving forces of the market.

APPENDIX

A.1 Proof of Theorem 1

Proof of Theorem 1. The necessary part has already been proven in Section 3.1. To prove the sufficient condition, we first note that condition $\mathrm{Cov}_L\left(Y, X_j \middle| \mathbf{X}_{\mathcal{C}}\right) = 0$ is



(a) $\left|\hat{\beta}^M\right|$ using marginal screening (b) $\left|\hat{\beta}^M\right|$ using conditional screening

Figure 4: Chosen factors with marginal (left) and conditional screening (right).

equivalent to

$$\mathbb{E}\, b'(\mathbf{X}_{\mathcal{C}}^T \boldsymbol{\beta}_{\mathcal{C}}^M) X_j = \mathbb{E}\, Y X_j,$$

as shown in Section 3.1. This and (11) imply that $((\boldsymbol{\beta}_{\mathcal{C}}^M)^T, 0)^T$ is a solution to equation (9). By the uniqueness, it follows that $\boldsymbol{\beta}_{\mathcal{C}j}^M = ((\boldsymbol{\beta}_{\mathcal{C}}^M)^T, 0)^T$, namely $\beta_j^M = 0$. This completes the proof.

□

A.2 Proof of Theorem 2

Proof of Theorem 2. We denote the matrix $\mathbb{E}\, m_j \mathbf{X}_{\mathcal{C}j} \mathbf{X}_{\mathcal{C}j}^T$ as Ω_j and partition it as

$$\Omega_j = \begin{bmatrix} \mathbb{E}\, m_j \mathbf{X}_{\mathcal{C}} \mathbf{X}_{\mathcal{C}}^T & \mathbb{E}\, m_j \mathbf{X}_{\mathcal{C}} \mathbf{X}_j \\ \mathbb{E}\, m_j \mathbf{X}_j \mathbf{X}_{\mathcal{C}}^T & \mathbb{E}\, m_j \mathbf{X}_j^2 \end{bmatrix} = \begin{bmatrix} \Omega_{\mathcal{C},\mathcal{C}} & \Omega_{\mathcal{C},j} \\ \Omega_{\mathcal{C},j}^T & \Omega_{j,j} \end{bmatrix}.$$

From the score equations, i.e. equations (9) and (11), we have that

$$\mathbb{E}\, b'\left(\mathbf{X}_{\mathcal{C}}^T \boldsymbol{\beta}_{\mathcal{C}}^M\right) \mathbf{X}_{\mathcal{C}} = \mathbb{E}\, b'\left(\mathbf{X}_{\mathcal{C}j}^T \boldsymbol{\beta}_{\mathcal{C}j}^M\right) \mathbf{X}_{\mathcal{C}}.$$

Using the definition of m_j, the above equation can be written as

$$\mathbb{E}\, m_j(\mathbf{X}_{\mathcal{C}j}^T\boldsymbol{\beta}_{\mathcal{C}j}^M - \mathbf{X}_{\mathcal{C}}^T\boldsymbol{\beta}_{\mathcal{C}}^M)\mathbf{X}_{\mathcal{C}} = 0.$$

By letting $\boldsymbol{\beta}_{\Delta,j} = \boldsymbol{\beta}_{\mathcal{C}j1}^M - \boldsymbol{\beta}_{\mathcal{C}}^M$, we have that

$$\mathbb{E}\, m_j(\mathbf{X}_{\mathcal{C}}^T\boldsymbol{\beta}_{\Delta,j}^M + X_j^T\boldsymbol{\beta}_j^M)\mathbf{X}_{\mathcal{C}} = 0.$$

or equivalently

$$\boldsymbol{\beta}_{\Delta,j} = -\,\Omega_{\mathcal{C},\mathcal{C}}^{-1}\Omega_{\mathcal{C},j}\beta_j^M. \tag{A.1}$$

Furthermore, by (13), we can express $\mathrm{Cov}_L(Y, X_j|\mathbf{X}_{\mathcal{C}})$ as

$$\mathrm{Cov}_L(Y, X_j|\mathbf{X}_{\mathcal{C}}) = \mathbb{E}\, X_j\{Y - \mathbb{E}_L(Y|\mathbf{X}_{\mathcal{C}}^T)\}. \tag{A.2}$$

It follows from (12) that

$$\mathrm{Cov}_L(Y, X_j|\mathbf{X}_{\mathcal{C}}) = \mathbb{E}\, X_j\left\{b'\left(\mathbf{X}_{\mathcal{C}j}^T\boldsymbol{\beta}_{\mathcal{C}j}^M\right) - b'\left(\mathbf{X}_{\mathcal{C}}^T\boldsymbol{\beta}_{\mathcal{C}}^M\right)\right\}. \tag{A.3}$$

Using the definition of m_j again, we have

$$\begin{aligned}\mathrm{Cov}_L(Y, X_j|\mathbf{X}_{\mathcal{C}}) &= \mathbb{E}\, m_j X_j(\mathbf{X}_{\mathcal{C}j}^T\boldsymbol{\beta}_{\mathcal{C}j}^M - \mathbf{X}_{\mathcal{C}}^T\boldsymbol{\beta}_{\mathcal{C}}^M)\\ &= \mathbb{E}\, m_j X_j(\mathbf{X}_{\mathcal{C}}^T\boldsymbol{\beta}_{\Delta,j}^M + X_j^T\boldsymbol{\beta}_j^M)\\ &= \Omega_{\mathcal{C},j}^T\boldsymbol{\beta}_{\Delta,j} + \Omega_{j,j}\beta_j^M.\end{aligned}$$

By (A.1), we conclude that

$$\mathrm{Cov}_L(Y, X_j|\mathbf{X}_{\mathcal{C}}) = (\Omega_{j,j} - \Omega_{\mathcal{C},j}^T\Omega_{\mathcal{C},\mathcal{C}}^{-1}\Omega_{\mathcal{C},j})\beta_j^M. \tag{A.4}$$

Now it is easy to see by Condition 1 that

$$|\beta_j^M| \geq c_2^{-1}|\mathrm{Cov}_L(Y, X_j|\mathbf{X}_{\mathcal{C}})| \geq c_3 n^{-\kappa},$$

where $c_3 = c_1/c_2$. Taking the minimum over all $j \in \mathcal{M}_{\mathcal{D}_\star}$ gives the result. □

A.3 Proof of Theorem 3

The proof of Theorem 3 uses an exponential bound for a quasi maximum likelihood estimator. This bound is shown in Fan and Song (2010) and we repeat their theorem here to facilitate the reading.

Let $\boldsymbol{\beta}_0 = \arg\min_{\boldsymbol{\beta}} \mathbb{E}l(\mathbf{X}^T\boldsymbol{\beta}, Y)$ the population parameter, which is an interior point of a large compact and convex set $\mathbf{B} \subset I\!R^p$.

Condition 5.

1. The Fisher information

$$I(\boldsymbol{\beta}) = \mathbb{E}\left\{\left[\frac{\partial}{\partial\boldsymbol{\beta}} l\left(\mathbf{X}^T\boldsymbol{\beta}, Y\right)\right]\left[\frac{\partial}{\partial\boldsymbol{\beta}} l\left(\mathbf{X}^T\boldsymbol{\beta}, Y\right)\right]^T\right\},$$

   is finite and positive definite at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$. Furthermore, $\sup_{\boldsymbol{\beta}\in\mathbf{B},\mathbf{x}} \left\|I(\boldsymbol{\beta})^{1/2}\mathbf{x}\right\|/\|\mathbf{x}\|$ exists.

2. The function $l(\mathbf{x}^T\boldsymbol{\beta}, y)$ is Lipschitz with a positive constant k_n for any $\boldsymbol{\beta}$ in $\mathbf{B}$, and $(\mathbf{x}, y)$ in $\Lambda_n = \{\mathbf{x}, y : \|\mathbf{x}\|_\infty \leq K_n, |y| \leq K_n^\star\}$ with K_n and $K_n^\star$ arbitrarily large constants. Furthermore, there exists a constant C such that

$$\sup_{\boldsymbol{\beta}\in\mathbf{B}, \left\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\right\| \leq Ck_nV_n^{-1}(p/n)^{1/2}} \left|\mathbb{E}\left[l\left(\mathbf{X}^T\boldsymbol{\beta}, Y\right) - l\left(\mathbf{X}^T\boldsymbol{\beta}_0, Y\right)\right](1 - I_n(\mathbf{X}, Y))\right| \leq o(p/n), \tag{A.5}$$

where $I_n(\mathbf{x}, y) = I((\mathbf{x}, y) \in \Lambda_n)$ with constant V_n defined below.

3. The function $l\left(\mathbf{X}^T\boldsymbol{\beta}, Y\right)$ is convex in $\boldsymbol{\beta}$ and

$$\left|\mathbb{E}\left[l\left(\mathbf{X}^T\boldsymbol{\beta}, Y\right) - l\left(\mathbf{X}^T\boldsymbol{\beta}_0, Y\right)\right]\right| \geq V_n \left\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\right\|^2,$$

for some positive constants V_n, and all $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \leq C k_n V_n^{-1} (p/n)^{1/2}$.

Theorem 6. *(Fan and Song 2010) Under Condition 5, for any $t > 0$ it holds that*

$$\mathbb{P}\left(\sqrt{n}\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right\| \geq 16 k_n (1+t)/V_n\right) \leq \exp\left(-2t^2/K_n^2\right) + n\mathbb{P}\left(\Lambda_n^c\right).$$

The proof of Theorem 3 is based on Theorem 6.

Proof of Theorem 3. By Lemma 1 of Fan and Song (2010), Condition 2(ii) gives the bound

$$P(|Y| \geq u) \leq s_1 \exp(-s_0 u).$$

Hence, we have

$$\mathbb{P}(\Lambda_n^c) \leq P(\|\mathbf{X}\|_\infty > K_n) + P(|Y| \geq K_n^\star) \leq r_2 \exp(-r_0 K_n^\alpha).$$

Using this and Theorem 6, letting $1 + t = c_3 V_n n^{1/2-\kappa}/(16 k_n)$, we have

$$\begin{aligned}\mathbb{P}\left(\left|\hat{\beta}_j^M - \beta_j^M\right| \geq c_3 n^{-\kappa}\right) &\leq \mathbb{P}\left(\left\|\hat{\boldsymbol{\beta}}_{\mathcal{C}j}^M - \boldsymbol{\beta}_{\mathcal{C}j}^M\right\| \geq c_3 n^{-\kappa}\right) \\ &\leq \exp\left(-c_4 n^{1-2\kappa}/(k_n K_n)^2\right) + n r_2 \exp\left(-r_0 K_n^\alpha\right),\end{aligned}$$

for some positive constant c_4. Then, by Bonferroni's inequality, we obtain

$$\mathbb{P}\left(\max_{q+1 \leq j \leq p}\left|\hat{\beta}_j^M - \beta_j^M\right| \geq c_3 n^{-\kappa}\right) \leq d\Big(\exp\left(-c_4 n^{1-2\kappa}(k_n K_n)^{-2}\right) + n r_2 \exp\left(-r_0 K_n^\alpha\right)\Big).$$

This proves the first conclusion.

The second statement can be shown by considering the event

$$\mathcal{A}_n = \left\{ \max_{j \in \mathcal{M}_{\star\mathcal{D}}} \left| \hat{\beta}_j^M - \beta_j^M \right| \leq c_3 n^{-\kappa}/2 \right\}.$$

On the event $\mathcal{A}_n$, by Theorem 2, it holds that for all $j \in \mathcal{M}_{\star\mathcal{D}}$

$$\left| \hat{\beta}_j^M \right| \geq c_3 n^{-\kappa}/2.$$

By letting $\gamma = c_5 n^{-\kappa} \leq c_3 n^{-\kappa}/2$, on the event $\mathcal{A}_n$ we have the sure screening property, that is $\mathcal{M}_{\star\mathcal{D}} \subset \hat{\mathcal{M}}_{\mathcal{D},\gamma}$. The probability bound can be shown by using the first result along with Bonferroni's inequality over all chosen j, which gives

$$\mathbb{P}\left(\mathcal{A}_n^c\right) \leq s \left[\exp\left(-c_4 n^{1-2\kappa} (k_n K_n)^{-2} \right) + n r_2 \exp\left(-r_0 K_n^{\alpha} \right) \right].$$

This completes the proof.

□

A.4 Proof of Theorem 4

Proof of Theorem 4. The first part of the proof is similar to that of Theorem 5 of Fan and Song (2010). The idea of this proof is to show that

$$\|\boldsymbol{\beta}_{\mathcal{D}}\|^2 = O\left(\lambda_{\max}\left(\boldsymbol{\Sigma}_{\mathcal{D}|\mathcal{C}} \right) \right). \tag{A.6}$$

If this holds, the size of the set $\{j = q+1, \ldots, p : |\beta_j^M| > \varepsilon n^{-\kappa}\}$ can not exceed $O\left(n^{2\kappa}\lambda_{\max}\left(\mathbf{\Sigma}_{\mathcal{D}|\mathcal{C}}\right)\right)$ for any $\varepsilon > 0$. Thus on the event

$$\mathcal{B}_n = \left\{ \max_{q+1 \le j \le p} |\hat{\beta}_j^M - \beta_j^M| \le \varepsilon n^{-\kappa} \right\},$$

the set $\{j = q+1, \ldots, p : |\hat{\beta}_j^M| > 2\varepsilon n^{-\kappa}\}$ is a subset of the set $\{j = q+1, \ldots, p : |\beta_j^M| > \varepsilon n^{-\kappa}\}$, whose size is bounded by $O\left(n^{2\kappa}\lambda_{\max}\left(\mathbf{\Sigma}_{\mathcal{D}|\mathcal{C}}\right)\right)$. If we take $\varepsilon = c_5/2$, we obtain that

$$\mathbb{P}\left(|\hat{\mathcal{M}}_{\mathcal{D},\gamma}| \le O\left(n^{2\kappa}\lambda_{\max}\left(\mathbf{\Sigma}_{\mathcal{D}|\mathcal{C}}\right)\right)\right) \ge \mathbb{P}(\mathcal{B}_n).$$

Finally, by Theorem 3, we obtain that

$$\mathbb{P}(\mathcal{B}_n) \ge 1 - d\Big(\exp\big(-c_4 n^{1-2\kappa}(k_n K_n)^{-2}\big) + n r_2 \exp\big(-r_0 K_n^{\alpha}\big)\Big)$$

and therefore the statement of the theorem follows.

We now prove (A.6) by using $\mathrm{Var}(\mathbf{X}^T\boldsymbol{\beta}^\star) = O(1)$ and (A.4). By Condition 3(ii), the Schur's complement $(\Omega_{j,j} - \Omega_{\mathcal{C},j}^T \Omega_{\mathcal{C},\mathcal{C}}^{-1} \Omega_{\mathcal{C},j})$ is uniformly bounded from below. Therefore, by (A.4), we have

$$|\beta_j^M| \le D_1 |\mathrm{Cov}_L(Y, X_j | \mathbf{X}_{\mathcal{C}})|,$$

for a positive constant D_1. Hence, we need only to bound the conditional covariance.

By (A.3), (9) and Lipschitz continuity of $b'(\cdot)$, we have

$$\begin{aligned}
|\mathrm{Cov}_L(Y, X_j|\mathbf{X}_{\mathcal{C}})| &= \mathbb{E}\big| X_j \left\{ b'\left(\mathbf{X}^T\boldsymbol{\beta}^*\right) - b'\left(\mathbf{X}_{\mathcal{C}}^T\boldsymbol{\beta}_{\mathcal{C}}^M\right)\right\}\big| \\
&\le D_2\, \mathbb{E}\big| X_j (\mathbf{X}^T\boldsymbol{\beta}^\star - \mathbf{X}_{\mathcal{C}}^T\boldsymbol{\beta}_{\mathcal{C}}^M)\big| \\
&= D_2\, \mathbb{E}\big| X_j [\mathbf{X}_{\mathcal{C}}^T\boldsymbol{\beta}_{\mathcal{C}}^\Delta + \mathbf{X}_{\mathcal{D}}^T\boldsymbol{\beta}_{\mathcal{D}}^\star]\big|.
\end{aligned}$$

where $\boldsymbol{\beta}_{\mathcal{C}}^{\Delta} = (\boldsymbol{\beta}_{\mathcal{C}}^{\star} - \boldsymbol{\beta}_{\mathcal{C}}^{M})$. Writing the last term in the vector form, we need to bound

$$\| \mathbb{E}\, \mathbf{X}_{\mathcal{D}} \mathbf{X}_{\mathcal{D}}^T \boldsymbol{\beta}_{\mathcal{D}}^{\star} + \mathbf{X}_{\mathcal{D}} \mathbf{X}_{\mathcal{C}}^T \boldsymbol{\beta}_{\mathcal{C}}^{\Delta} \|^2.$$

From the property of the least-squares, we have $\mathbb{E}[\mathbb{E}_L(\mathbf{X}_{\mathcal{D}}|\mathbf{X}_{\mathcal{C}})\mathbf{X}_{\mathcal{C}}^T] = \mathbb{E}[\mathbf{X}_{\mathcal{D}}\mathbf{X}_{\mathcal{C}}^T]$. Thus the above expression can be written as

$$\| \left[\boldsymbol{\Sigma}_{\mathcal{D}|\mathcal{C}}\right] \boldsymbol{\beta}_{\mathcal{D}}^{\star} + \mathbb{E}\, \mathbb{E}_L(\mathbf{X}_{\mathcal{D}}|\mathbf{X}_{\mathcal{C}})[\mathbf{X}_{\mathcal{C}}^T \boldsymbol{\beta}_{\mathcal{C}}^{\Delta} + \mathbb{E}_L(\mathbf{X}_{\mathcal{D}}^T|\mathbf{X}_{\mathcal{C}})\boldsymbol{\beta}_{\mathcal{D}}^{*})] \| \;\; = \;\; \left\| \left[\boldsymbol{\Sigma}_{\mathcal{D}|\mathcal{C}}\right] \boldsymbol{\beta}_{\mathcal{D}}^{\star} + \mathbf{Z} \right\|^2,$$

recalling the definition of $\mathbf{Z} = \mathbb{E}\, \mathbb{E}_L(\mathbf{X}_{\mathcal{D}}|\mathbf{X}_{\mathcal{C}}) \left(\mathbf{X}^T \boldsymbol{\beta}^{\star} - \mathbf{X}_{\mathcal{C}}^T \boldsymbol{\beta}_{\mathcal{C}}^{M}\right)$ in Condition 3.

Using the law of total variance, we have that

$$\begin{aligned}
\left\| \left[\boldsymbol{\Sigma}_{\mathcal{D}|\mathcal{C}}\right] \boldsymbol{\beta}_{\mathcal{D}}^{\star} + \mathbf{Z} \right\|^2 &= \boldsymbol{\beta}_{\mathcal{D}}^{\star\, T} \left[\boldsymbol{\Sigma}_{\mathcal{D}|\mathcal{C}}\right]^2 \boldsymbol{\beta}_{\mathcal{D}}^{\star} + 2\mathbf{Z}^T \left[\boldsymbol{\Sigma}_{\mathcal{D}|\mathcal{C}}\right] + \mathbf{Z}^T \mathbf{Z} \\
&\leq \lambda_{\max}\left(\left[\boldsymbol{\Sigma}_{\mathcal{D}|\mathcal{C}}\right]\right) \left(\boldsymbol{\beta}_{\mathcal{D}}^{\star\, T} \left[\boldsymbol{\Sigma}_{\mathcal{D}|\mathcal{C}}\right] \boldsymbol{\beta}_{\mathcal{D}}^{\star}\right) + 2\mathbf{Z}^T \left[\boldsymbol{\Sigma}_{\mathcal{D}|\mathcal{C}}\right] + \mathbf{Z}^T \mathbf{Z} \\
&\leq \lambda_{\max}\left(\left[\boldsymbol{\Sigma}_{\mathcal{D}|\mathcal{C}}\right]\right) \mathrm{Var}(\mathbf{X}^T \boldsymbol{\beta}^{\star}) + 2\mathbf{Z}^T \left[\boldsymbol{\Sigma}_{\mathcal{D}|\mathcal{C}}\right] + \mathbf{Z}^T \mathbf{Z},
\end{aligned}$$

and the last two terms are $o\left(\lambda_{\max}\left(\left[\boldsymbol{\Sigma}_{\mathcal{D}|\mathcal{C}}\right]\right)\right)$ due to Condition 3. Therefore, we have that

$$\|\boldsymbol{\beta}_{\mathcal{D}}\|^2 = O\left(\lambda_{\max}\left(\left[\boldsymbol{\Sigma}_{\mathcal{D}|\mathcal{C}}\right]\right)\right),$$

and that gives us the desired result. □

A.5 Proof of Theorem 5

Proof of Theorem 5. Note that the false discovery proportion can be rewritten as

$$\mathbb{E}\left(\frac{\left|\hat{\mathcal{M}}_{\mathcal{D},\delta} \cap (\mathcal{M}_{\star\mathcal{D}})^c\right|}{|(\mathcal{M}_{\star\mathcal{D}})^c|}\right) = \frac{1}{d - |\mathcal{M}_{\star\mathcal{D}}|} \sum_{j \in (\mathcal{M}_{\star\mathcal{D}})^c} \mathbb{P}\left(I_j\left(\hat{\beta}_j^M\right)^{1/2} \left|\hat{\beta}_j^M\right| \geq \delta\right).$$

With the given conditions, by Theorem 1, we have $\beta_j^M = 0$. Since $\mathbf{X}_\mathcal{C}$ includes the intercept term, $\mathbb{E}\, e_i = 0$. It is known that $I_j \left(\hat{\beta}_j^M\right)^{1/2} \left|\hat{\beta}_j^M\right|$ (for $j \in (\mathcal{M}_{\star\mathcal{D}})^c$) has an asymptotically standard normal distribution (Gao et al., 2008, Heyde, 1997). Then, it follows that for a $c_7 > 0$

$$\sup_z \left| \mathbb{P}\left(I_j \left(\hat{\beta}_j^M\right)^{1/2} \left|\hat{\beta}_j^M\right| \geq z \right) - \Phi(z) \right| \leq c_7 n^{-1/2}.$$

Combining both equations, we obtain

$$\mathbb{E}\left(\frac{\left|\hat{\mathcal{M}}_{\mathcal{D},\delta} \cap (\mathcal{M}_{\star\mathcal{D}})^c\right|}{|(\mathcal{M}_{\star\mathcal{D}})^c|} \right) \leq \frac{1}{d - |\mathcal{M}_{\star\mathcal{D}}|} \sum_{j \in (\mathcal{M}_{\star\mathcal{D}})^c} \left(2\left(1 - \Phi\left(\delta\right)\right) + c_7 n^{-1/2}\right).$$

Setting $\delta = \Phi^{-1}\left(1 - \frac{f}{2d}\right)$ gives the result. □

References

Bickel, P.J., Ritov, Y., and Tsybakov, A.B. (2009), "Simultaneous Analysis of Lasso and Dantzig selector," *The Annals of Statistics,* 37 1705–1732.

Bühlmann, P., and van de Geer, S. (2011), *Statistics for High-Dimensional Data: Methods, Theory and Applications,* New York: Springer.

Candes, E., and Tao, T. (2007), "The Dantzig Selector: Statistical Estimation When p Is Much Larger Than n" (with discussion), *The Annals of Statistics,* 35, 2313–2351.

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

Fama, E.F., and French, K.R. (1993), "Common Risk Factors in the Returns on Stocks and Bonds," *Journal of Financial Economics,* 33, 3–56.

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

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 (Statistical Methodology),* 70, 849–911.

Fan, J., and Lv, J. (2011), "Nonconcave Penalized Likelihood With NP-Dimensionality," *Information Theory, IEEE Transactions,* 57, 5467–5484.

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

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

Frank, I.E., and Friedman, J. (1993), "A Statistical View of Some Chemometrics Regression Tools," *Technometrics,* 35, 109–135.

Gao, Q., Wu, Y., Zhu, C. and Wang, Z. (2008), "Asymptotic Normality of Maximum Quasi-Likelihood Estimators in Generalized Linear Models with Fixed Design," *Journal of Systems Science and Complexity,* 21, 463–473.

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C., and Lander, E. (1999), "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," *Science,* 286, 531–537.

Hall, P., and Miller, H. (2009), "Using Generalized Correlation to Effect Variable Selection in Very High Dimensional Problems," *Journal of Computational and Graphical Statistics,* 18, 533–550.

Hastie, T.J., Tibshirani, R., and Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference and Prediction,* New York: Springer.

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 (Statistical Methodology),* 71, 783–803.

Heyde, C.C. (1997), *Quasi-likelihood and its Application: a General Approach to Optimal Parameter Estimation,* New York: Springer.

Li, G., Peng, H., Zhang, J., and Zhu, L. (2012), "Robust Sure Independence Screening Based on Rank Correlation for the Ultrahigh Dimensional Models," manuscript, Beijing University of Technology.

Osborne, M.R., Presnell, B. and Turlach, B.A. (2000a), "On the LASSO and its Dual," *Journal of Computational and Graphical Statistics,* 9, 319–337.

Osborne, M.R., Presnell, B. and Turlach, B.A. (2000b), "A New Approach to Variable Selection in Least Squares Problems," *IMA Journal of Numerical Analysis,* 20, 389–403.

Szczepaski, T., van der Velden, V.H., Raff, T., Jacobs, D.C., van Wering, E.R., Brggemann, M., Kneba, M., and van Dongen, J.J. (2003), "Comparative Anal-

ysis of T-cell Receptor Gene Rearrangements at Diagnosis and Relapse of T-cell Acute Lymphoblastic Leukemia (T-ALL) Shows High Stability of Clonal Markers for Monitoring of Minimal Residual Disease and Reveals the Occurrence of Second T-ALL," *Leukemia,* 17, 2149–2156.

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

Wasserman, L., and Roeder, K. (2009), "High-dimensional Variable Selection," *The Annals of Statistics,* 37, 2178–2201.

Zhang, C., and Zhang, T. (2012), "A General Theory of Concave Regularization for High Dimensional Sparse Estimation Problems," manuscript, Rutgers University.

Zhao, S.D., and Li, Y. (2012), "Principled Sure Independence Screening for Cox Models with Ultra-high Dimensional Covariates," *Journal of Multivariate Analysis,* 105, 397–411.