Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: www.tandfonline.com/journals/uasa20

Conditional Sure Independence Screening

Emre Barut, Jianqing Fan & Anneleen Verhasselt

To cite this article: Emre Barut, Jianqing Fan & Anneleen Verhasselt (2016) Conditional Sure Independence Screening, Journal of the American Statistical Association, 111:515, 1266-1277, DOI: 10.1080/01621459.2015.1092974

To link to this article: https://doi.org/10.1080/01621459.2015.1092974

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2016, VOL. 111, NO. 515, 1266–1277, Theory and Methods
http://dx.doi.org/10.1080/01621459.2015.1092974

Taylor & Francis
Taylor & Francis Group

Conditional Sure Independence Screening

Emre Barut[a], Jianqing Fan[b], and Anneleen Verhasselt[c]

[a]Department of Statistics, George Washington University, Washington, DC, USA; [b]Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ, USA, and School of Data Science, Fudan University, Shanghai, China; [c]Interuniversity Institute for Biostatistics and Statistical Bioinformatics, CenStat, Universiteit Hasselt, Hasselt, Belgium

ABSTRACT

Independence screening is powerful for variable selection when the number of variables is massive. Commonly used independence screening methods are based on marginal correlations or its variants. When some prior knowledge on a certain important set of variables is available, a natural assessment on the relative importance of the other predictors is their conditional contributions to the response given the known set of variables. This results in conditional sure independence screening (CSIS). CSIS produces a rich family of alternative screening methods by different choices of the conditioning set and can help reduce the number of false positive and false negative selections when covariates are highly correlated. This article proposes and studies CSIS in generalized linear models. 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 and the properties of CSIS when a data-driven conditioning set is used. 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 datasets. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received August 2012
Revised August 2015

KEYWORDS
False selection rate; Generalized linear models; Sparsity; Sure screening; Variable selection

1. Introduction

Statisticians are nowadays frequently confronted with massive datasets 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. To explore this sparsity, variable selection techniques are needed.

Over the last 10 years, there have been many exciting developments on variable selection techniques for ultrahigh dimensional feature space. They can basically be classified into two classes: penalization and screening. Penalization techniques include 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, Ritov, and Tsybakov 2009), among others. These techniques select variables and estimate parameters simultaneously by solving a high-dimensional optimization problem. Despite various efficient algorithms (Osborne, Presnell, and Turlach 2000; Efron et al. 2004; Fan and Lv 2011), statisticians and machine learners still face huge computational challenges 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) proposed a two-scale approach, called iterative sure independence screening (ISIS), which screens and selects variables iteratively. The approach is further developed by Fan, Samworth, and Wu (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, Titterington, and Xue 2009), generalized correlation screening (Hall and Miller 2009), nonparametric screening (Fan, Feng, and Song 2011), and robust rank correlation-based screening (Li et al. 2012a; Li, Zhong, and Zhu 2012b), 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 small 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 weakly dependent with the response given other variables. Fan and Lv (2008) and Fan, Samworth, and Wu (2009) used 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. It becomes SIS when conditioning set includes no variables. It provides an alternative measure of the contribution of each variable and provides a useful

CONTACT Emre Barut barut@gwu.edu Department of Statistics, George Washington University, Washington, DC 20052.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JASA.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

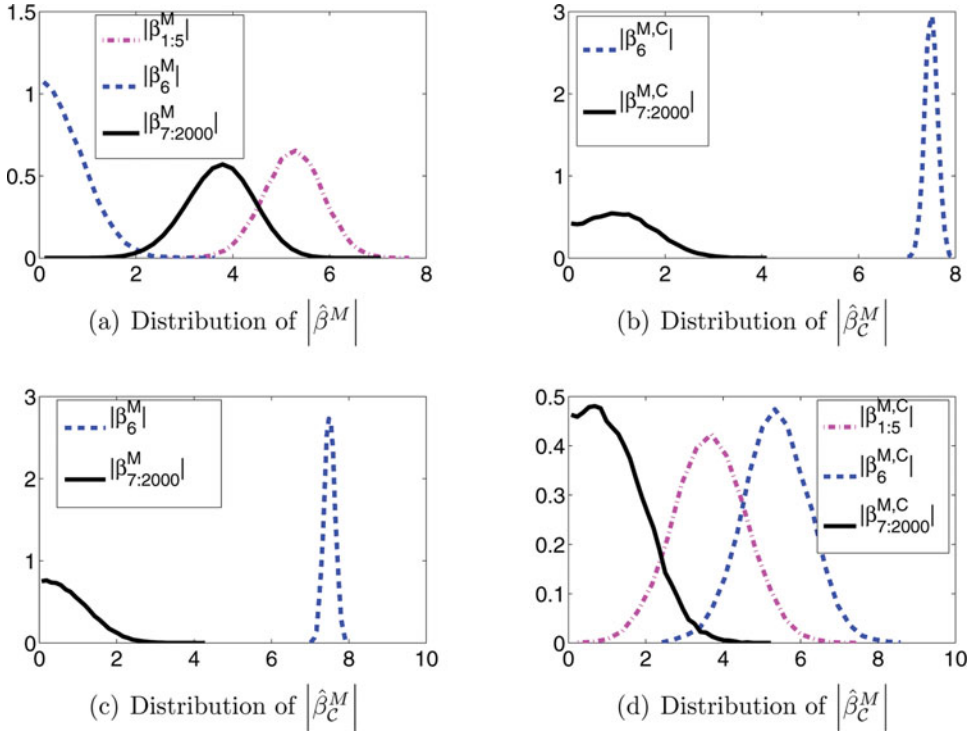


Figure 1. Benefits of conditioning against false negatives. (a) The distributions of the averages of magnitudes $|\hat{\beta}^M_j|$ of marginal regression coefficients over three groups of variables 1:5, 6, 7:2000. The rest plots are similar to (a) except conditioned on (b) the first five active variables, (c) five active variables and five randomly chosen inactive variables, and (d) five randomly chosen inactive variables.

alternative when unconditional ones are not effective. When researchers know from previous investigations that certain variables $\mathbf{X}_{\mathcal{C}}$ are responsible for the outcomes, this knowledge should be taken into account. 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 $\mathbf{X}_{\mathcal{C}}$.

Conditional screening significantly widens the methodology of screening. Since it does not require (though prefers to) $\mathbf{X}_{\mathcal{C}}$ to contain active variables, one can probe the utilities of variables by different choices of $\mathcal{C}$. In the absence of the prior knowledge on the usefulness of variables, one can take those variables that survive the screening and selection as in Fan and Lv (2008). One can also take the first few variables selected from forward regression (Wang 2009) or the least angle regression (LARS) algorithm (Efron et al. 2004). In many situations, we have high statistical evidence that these first few variables are important and can be taken as $\mathbf{X}_{\mathcal{C}}$. This is particularly when they are both selected by the LARS and forward regression algorithm. In particular, forward regression can be regarded as iteratively conditional screening, recruiting one variable at a time. In contrast, conditional screening allows recruiting multiple variables at a time and typically avoids conditioning on a large set $\mathcal{C}$. Hence, the conditional screening is much faster than the stepwise regression.

Conditional screening makes it possible to recover the hidden significant variables. Consider the following linear model:

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

with $\boldsymbol{\beta}^\star = (\beta^\star_1, \ldots, \beta^\star_p)^T$. It can easily be shown that the covariance between X_j and Y is zero if $\beta^\star_j = -\sum_{k\neq j}\beta^\star_k\Sigma_{kj}/\Sigma_{jj}$, where Σ_{kj} is the (k, j) element of $\boldsymbol{\Sigma} = \text{var}(\mathbf{X})$. Yet, $\beta^\star_j$ can be far away from zero. In other words, X_j is a hidden signature variable, and cannot be selected by SIS. To demonstrate that, let us consider the case in which $p = 2000$, $\boldsymbol{\beta}^\star = (3, 3, 3, 3, -7.5, 0, \ldots, 0)^T$, and all predictors 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, we fit marginal regression and obtain the marginal estimates $\{\hat{\beta}^M_j\}^p_{j=1}$. The magnitudes of these estimates are summarized by over three groups: indices 1 to 5 (denoted by $\boldsymbol{\beta}^M_{1:5}$), 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 10,000 simulations. Clearly variable X_6 cannot be selected by marginal screening.

Adapting the conditional screening approach gives a very different result. Conditioning upon the first five variables, conditional correlation between X_6 and Y is large. 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 the presence of the first five variables. Again, the magnitudes $\{|\hat{\beta}^M_{\mathcal{C}j}|\}^{2000}_{j=6}$ are summarized into two values: $|\hat{\beta}^M_{\mathcal{C}6}|$ and $\{|\hat{\beta}^M_{\mathcal{C}j}|\}^{2000}_{j=7}$. The distributions of those over 10,000 simulations are also depicted in Figure 1(b). Clearly, the variable X_6 has higher marginal contributions than inactive variables. That is, conditioning helps recruiting the hidden signature variable.

To see the merits of conditioning, we have repeated the previous experiment with conditioning on five more randomly

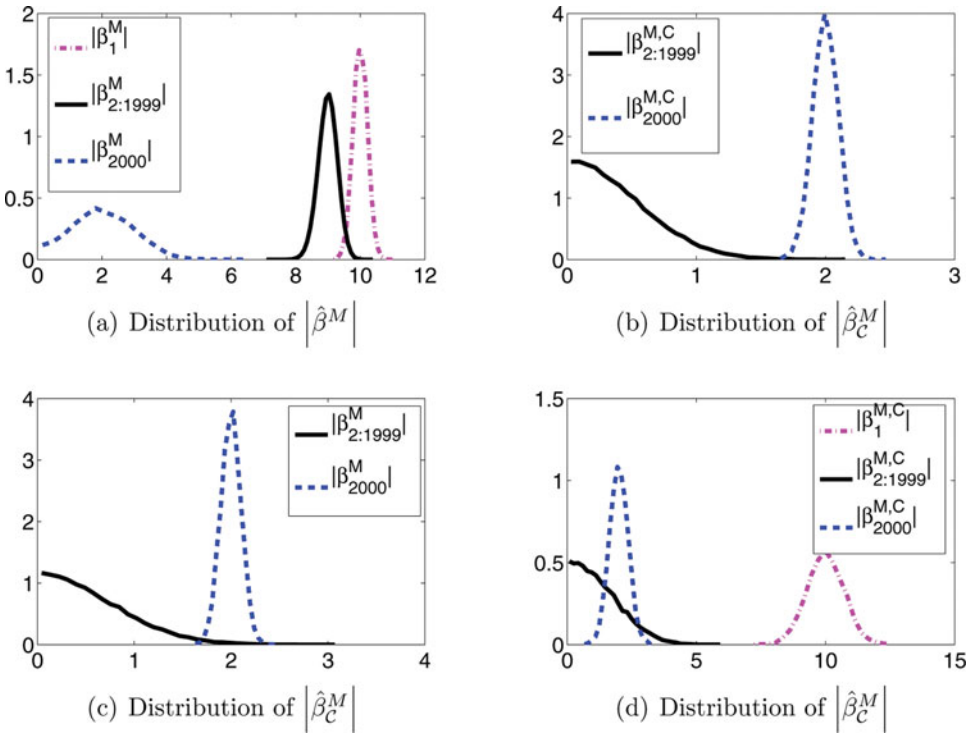


Figure 2. Benefits of conditioning against false positives. Plotted are the distributions on the averages of $|\hat{\beta}^M_{\mathcal{C}j}|$ similar to those in Figure 1, conditioned on (a) null set (unconditional), (b) X_1, (c) X_1 and five randomly selected variables, and (d) five randomly selected variables.

chosen variables (10 in total) and on 5 completely randomly selected variables from the inactive set. The distributions of the magnitudes are depicted in Figure 1(c) and 1(d). Clearly, conditioning provides useful alternative ranks that help recover the hidden signature variables X_6.

Conditional screening can also help reduce the number of false negatives when covariates are highly correlated. To appreciate this, consider the linear model (1) again with $\boldsymbol{\beta}^\star = (10, 0, \ldots, 0, 1)^T$, equi-correlation 0.9 among all covariates except X_{2000}, which is independent of the rest of the covariates. This setting gives

$$\begin{aligned} &\operatorname{cov}(X_1, Y) = 10, \quad \operatorname{cov}(X_{2000}, Y) = 1, \\ \text{and} \quad &\operatorname{cov}(X_j, Y) = 9 \quad \text{for } j \neq 1, 2000. \end{aligned}$$

In this case, marginal utilities for all inactive variables are higher than that for the active variable X_{2000}. A summary similar to Figure 1 is shown in Figure 2. Therefore, based on SIS in Fan and Lv (2008), the active variable X_{2000} has the least priority to be included. Using conditional screening, the magnitude of the hidden active variable X_{2000} is comparatively larger and hence it is more likely to be recruited during screening.

The theoretical basis of the above observation can be understood as follows. As shown by Fan and Lv (2008) and Fan and Song (2010), the size of the selected variables (closely related to false positives) depends on the largest eigenvalue of $\boldsymbol{\Sigma}$: $\lambda_{\max}(\boldsymbol{\Sigma})$. The larger the quantity, the more variables have to be selected to have a sure screening property. By using conditional screening, the relevant quantity now becomes $\lambda_{\max}(\boldsymbol{\Sigma}_{\mathbf{X}_{\mathcal{D}}|\mathbf{X}_{\mathcal{C}}})$, where $\mathbf{X}_{\mathcal{C}}$ refers to the q covariates that we condition upon and $\mathbf{X}_{\mathcal{D}}$ is the rest of the variables. Conditioning helps reduce 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 this, we consider the case where $\mathbf{X}$ is given by equally correlated normal random variables. Simple calculations yield that $\lambda_{\max}(\boldsymbol{\Sigma}_{\mathbf{X}_{\mathcal{D}}}) = (1-r) + rd$, where r is the common correlation and $d = p - q$ and $\lambda_{\max}(\boldsymbol{\Sigma}_{\mathbf{X}_{\mathcal{D}}|\mathbf{X}_{\mathcal{C}}}) = (1-r) + rd\frac{1-r}{1-r+rq}$. It is clear that conditioning helps reduce the correlation among the variables. To quantify this, Figure 3 depicts the ratio $\lambda_{\max}(\boldsymbol{\Sigma}_{\mathbf{X}_{\mathcal{D}}})/\lambda_{\max}(\boldsymbol{\Sigma}_{\mathbf{X}_{\mathcal{D}}|\mathbf{X}_{\mathcal{C}}})$ 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 rest of the article is organized as follows. In Section 2, we introduce the conditional sure independence screening procedure. The sure independence screening 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

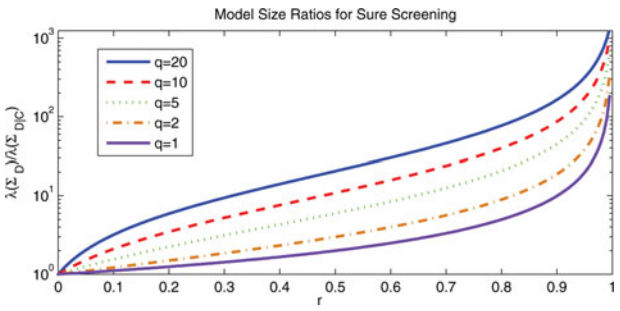


Figure 3. Ratio of maximum eigenvalues of unconditioned and conditioned covariance matrix.

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 $\mathbf{X} = \mathbf{x}$ belongs to an exponential family

$$f(y|\mathbf{x};\theta) = \exp\Big(y\theta(\mathbf{x}) - b(\theta(\mathbf{x})) + c(\mathbf{x};y)\Big), \tag{2}$$

where $b(\cdot)$ and $c(\cdot)$ are known functions. Under model (2), 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. 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 $\boldsymbol{\beta}^\star$.

2.2. Conditional Screening

Given a set of variables $\mathbf{X}_\mathcal{C}$, we wish to recruit additional variables from rest of the variables, $\mathbf{X}_\mathcal{D}$, to better explain the response variable Y. Without loss of generality, assume $\mathcal{C}$ is the set of first q variables and $\mathcal{D}$ is the remaining set of $d = p - q$ variables. We will use the notation

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

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 $\mathbb{E}(X_j) = 0$ and $\mathbb{E}(X_j^2) = 1$, for $j \in \mathcal{D}$. In practice, $\mathbf{X}_\mathcal{C}$ is typically the set of variables that are known to be related to Y from a prior study or those that have high marginal utilities according to SIS and hence are expected to be highly related to Y.

Given a random sample $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ from the generalized linear model (2) 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

$$\hat{\boldsymbol{\beta}}_{\mathcal{C}j}^M = \text{argmin}_{\boldsymbol{\beta}_\mathcal{C}, \beta_j} \mathbb{P}_n\big\{l(\mathbf{X}_\mathcal{C}^T\boldsymbol{\beta}_\mathcal{C} + X_j\beta_j, Y)\big\}, \tag{3}$$

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 the last element of $\hat{\boldsymbol{\beta}}_{\mathcal{C}j}^M$ by $\hat{\beta}_j^M$. 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

$$\hat{\mathcal{M}}_{\mathcal{D},\gamma} = \{j \in \mathcal{D} : |\hat{\beta}_j^M| > \gamma\}, \tag{4}$$

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

$$\hat{R}_{\mathcal{C}j} = \min_{\boldsymbol{\beta}_\mathcal{C}, \beta_j} \mathbb{P}_n\big\{l(\mathbf{X}_\mathcal{C}^T\boldsymbol{\beta}_\mathcal{C} + X_j\beta_j, Y)\big\}, \tag{5}$$

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 the $\hat{R}_{\mathcal{C}j}$, the more the variable X_j contributes in the presence of $\mathbf{X}_\mathcal{C}$. This leads to the method based on the likelihood ratio statistics: recruit additional variables according to

$$\tilde{\mathcal{M}}_{\mathcal{D},\tilde{\gamma}} = \{j \in \mathcal{D} : \hat{R}_{\mathcal{C}j} < \tilde{\gamma}\}, \tag{6}$$

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

The set of variables $\mathbf{X}_\mathcal{C}$ does not necessarily have to contain active variables. Conditional screening uses only the fact that the effects of important variables can be more visible in the presence of $\mathbf{X}_\mathcal{C}$ and the correlations of variables are weakened upon conditioning. This is commonly the case in many applications such as finance and biostatistics, where the variables share some common factors. It gives hidden signature variables a chance to survive. In fact, it was demonstrated in the introduction that conditioning can be beneficial even if the set $\mathbf{X}_\mathcal{C}$ is chosen randomly. Our theoretical study gives a formal theoretical consideration of the iterated method proposed in Fan and Lv (2008) and Fan et al. (2009).

3. Sure Screening Properties

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

$$\boldsymbol{\beta}_{\mathcal{C}j}^M = \text{argmin}_{\boldsymbol{\beta}_\mathcal{C}, \beta_j} \mathbb{E}\, l(\mathbf{X}_\mathcal{C}^T\boldsymbol{\beta}_\mathcal{C} + X_j\beta_j, Y), \tag{7}$$

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 only marginal regressions, we introduce model misspecifications. Hence, in general, the marginal regression coefficient β_j^M differs from 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. Therefore, the marginal conditional regression coefficients provide useful probes for the joint regression.

By (7), the marginal regression coefficients $\boldsymbol{\beta}_{\mathcal{C}j}^M$ satisfy the score equations

$$\mathbb{E}\, b'(\mathbf{X}_{\mathcal{C}j}^T\boldsymbol{\beta}_{\mathcal{C}j}^M)\mathbf{X}_{\mathcal{C}j} = \mathbb{E}\, Y\mathbf{X}_{\mathcal{C}j} = \mathbb{E}\, b'(\mathbf{X}^T\boldsymbol{\beta}^\star)\mathbf{X}_{\mathcal{C}j}, \tag{8}$$

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

$$\boldsymbol{\beta}_{\mathcal{C}}^M = \text{argmin}_{\boldsymbol{\beta}_{\mathcal{C}}}\, \mathbb{E}\, l(\mathbf{X}_{\mathcal{C}}^T\boldsymbol{\beta}_{\mathcal{C}}, Y), \tag{9}$$

and satisfies the equations

$$\mathbb{E}\, b'(\mathbf{X}_{\mathcal{C}}^T\boldsymbol{\beta}_{\mathcal{C}}^M)\mathbf{X}_{\mathcal{C}} = \mathbb{E}\, Y\mathbf{X}_{\mathcal{C}} = \mathbb{E}\, b'(\mathbf{X}^T\boldsymbol{\beta}^\star)\mathbf{X}_{\mathcal{C}}. \tag{10}$$

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

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

$$\text{and} \qquad \mathbb{E}_L(Y|\mathbf{X}_{\mathcal{C}j}) = b'(\mathbf{X}_{\mathcal{C}j}^T\boldsymbol{\beta}_{\mathcal{C}j}^M), \tag{11}$$

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 (10) can be more intuitively expressed as

$$\mathbb{E}(Y - \mathbb{E}_L(Y|\mathbf{X}_{\mathcal{C}}))\mathbf{X}_{\mathcal{C}} = 0. \tag{12}$$

Note that the conditioning in this article is really a conditional linear fit. 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 signals $\{\beta_j^M\}_{j=1}^p$. When $\beta_j^M = 0$, by (8), 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 (10). Then, Equation (8) 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 (12), the above condition can be more comprehensively expressed as

$$\text{cov}_L\left(Y, X_j\middle|\mathbf{X}_{\mathcal{C}}\right) \equiv \mathbb{E}(X_j - \mathbb{E}_L(X_j|\mathbf{X}_{\mathcal{C}}))(Y - \mathbb{E}_L(Y|\mathbf{X}_{\mathcal{C}})) = 0. \tag{13}$$

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 $\text{cov}_L(Y, X_j|\mathbf{X}_{\mathcal{C}}) = 0$.

Proof of the sufficient part is given in Appendix A.1. To have the sure screening property at the population level, 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 nonvanishing. The sure screening property of conditional MLE (CMLE), given by Equation (4), 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 $|\text{cov}_L(Y, X_j|\mathbf{X}_{\mathcal{C}})| \geq c_1 n^{-\kappa}$.

(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, $\mathbb{E}\, m_jX_j^2 \leq c_2$ 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, that is, $b(\theta) = \theta^2/2$, 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 $\mathbb{E}\, X_j^2 = 1$. More generally, if $b'(\theta)$ satisfies Lipschitz's condition, the condition holds automatically. In particular, logistic regression satisfies this condition.

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

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

To gain insights on Condition 1, consider a linear regression problem $Y = \mathbf{X}^T\boldsymbol{\beta} + \varepsilon$ with $\mathbb{E}(\mathbf{X}\varepsilon) = 0$. For a given set $\mathcal{C}$, we can calculate the conditional linear covariance as follows:

$$\text{cov}_L\left(Y, X_j|\mathbf{X}_{\mathcal{C}}\right) = \sum_{i=1}^p \left(\Sigma_{ij} - \boldsymbol{\Sigma}_{\mathcal{C}j}^T\boldsymbol{\Sigma}_{\mathcal{C}}^{-1}\boldsymbol{\Sigma}_{\mathcal{C}i}\right)\beta_i = (\boldsymbol{\Gamma}_j)^T\beta_{\mathcal{D}},$$

where $\Gamma_{ij} = (\Sigma_{\mathbf{X}_{\mathcal{D}}|\mathbf{X}_{\mathcal{C}}})_{i,j} = \Sigma_{ij} - \boldsymbol{\Sigma}_{\mathcal{C}j}^T\boldsymbol{\Sigma}_{\mathcal{C}}^{-1}\boldsymbol{\Sigma}_{\mathcal{C}i}$, which is the conditional covariance between X_i and X_j given $\mathbf{X}_{\mathcal{C}}$ under the normality assumption. Therefore, if a variable is important, that is, $\beta_j \neq 0$, ideally we would want $\mathcal{C}$ to be chosen such that for all other important terms, $i \in \mathcal{M}^\star$, Γ_{ij} is of small order.

The unconditional screening corresponds to $\mathcal{C}$ being the empty set. It is easily seen that CSIS probes into a very different quantity, which depends on the choices of $\mathcal{C}$. Hence, CSIS provides a wide class of screening methods with different choices of $\mathcal{C}$.

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 $\mathcal{B} = \{|\beta_1^M| \leq B, \ldots, |\beta_q^M| \leq B, |\beta_j^M| \leq B\}$ for a sufficiently large 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 are needed.

Condition 2.

(i) The operator norm $\|I_j(\boldsymbol{\beta}_{\mathcal{C}j})\|_{\mathcal{B}}$ of the Fisher information $I_j(\boldsymbol{\beta}_{\mathcal{C}j}) = \mathbb{E}(b''(\mathbf{X}_{\mathcal{C}j}^T\boldsymbol{\beta}_{\mathcal{C}j})\mathbf{X}_{\mathcal{C}j}\mathbf{X}_{\mathcal{C}j}^T)$ is bounded, where $\|I_j(\boldsymbol{\beta}_{\mathcal{C}j})\|_{\mathcal{B}} = \sup_{\boldsymbol{\beta}_{\mathcal{C}j}\in\mathcal{B}, \|\mathbf{x}_{\mathcal{C}j}\|=1} \|I_j(\boldsymbol{\beta}_{\mathcal{C}j})^{1/2}\mathbf{x}_{\mathcal{C}j}\|$.

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

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

for sufficiently large t and that

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

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

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

where 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 $\Lambda_n = \{\mathbf{x}, y : \|\mathbf{x}\|_\infty \le K_n, |y| \le K_n^\star\}$ with $K_n^* = r_0 K_n^\alpha / s_0$.

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

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

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.

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 the supplementary material (Barut, Fan, and Verhasselt 2015).

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 and assume that $n^{1-2\kappa} k_n^{-2} K_n^{-2} \to \infty$.

(i) For any $c_3 > 0$, there exists a positive constant c_4 such that

$$\mathbb{P}\left(\max_{q+1 \le j \le p} |\hat{\beta}^M_j - \beta^M_j| \ge c_3 n^{-\kappa}\right) \le d \exp\left(-c_4 n^{1-2\kappa}(k_n K_n)^{-2}\right) + d n r_2 \exp\left(-r_0 K_n^\alpha\right),$$

where $r_2 = q r_1 + s_1$.

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

$$\mathbb{P}\left(\mathcal{M}_{\star\mathcal{D}} \subset \hat{\mathcal{M}}_{\mathcal{D},\gamma}\right) \ge 1 - s \exp\left(-c_4 n^{1-2\kappa}(k_n K_n)^{-2}\right) - n r_2 s \exp\left(-r_0 K_n^\alpha\right),$$

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 sparse elements but not on d or p. This 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 $\hat{\beta}^M_j$ 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 $\log d = o(n^{-(1-2\kappa)\alpha/A})$, 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. However, a good screening method should also retain a small set of variables after screening to have low false positives. 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 $\mathbb{E}[m_j \mathbf{X}_{\mathcal{C}j}\mathbf{X}^T_{\mathcal{C}j}]$ is larger than a positive constant, uniformly over j, where m_j is defined in Condition 1(ii).

(iii) Letting

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

it holds that $\|\mathbf{Z}\|_2^2 = o\{\lambda_{\max}(\boldsymbol{\Sigma}_{\mathcal{D}|\mathcal{C}})\}$, with $\lambda_{\max}(\boldsymbol{\Sigma}_{\mathcal{D}|\mathcal{C}})$ 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 $\mathbb{E}\,\mathbf{X}_{\mathcal{C}j}\mathbf{X}^T_{\mathcal{C}j}$ 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 (10), $\mathbb{E}\,\mathbf{X}_{\mathcal{C}}\mathbf{X}^T_{\mathcal{C}}\boldsymbol{\beta}^M_{\mathcal{C}} = \mathbb{E}\,\mathbf{X}_{\mathcal{C}}\mathbf{X}^T\boldsymbol{\beta}^\star$ 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.

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

$$\mathbb{P}\left(|\hat{\mathcal{M}}_{\mathcal{D},\gamma}| \le O\left(n^{2\kappa}\lambda_{\max}\left(\boldsymbol{\Sigma}_{\mathcal{D}|\mathcal{C}}\right)\right)\right) \ge 1 - d\left(\exp\left(-c_4 n^{1-2\kappa}(k_n K_n)^{-2}\right) + n r_2 \exp\left(-r_0 K_n^\alpha\right)\right).$$

This theorem is proved in Appendix A.3. From its proof, without Condition 3(iii), Theorem 4 continues to hold with $\boldsymbol{\Sigma}_{\mathcal{D}|\mathcal{C}}$ replaced by $\boldsymbol{\Sigma}_{\mathcal{D}|\mathcal{C}} + \mathbf{Z}\mathbf{Z}^T$.

When there is no prior knowledge on $\mathcal{C}$, data analysts often use a data-driven conditioning set $\widehat{\mathcal{C}}$. For example, $\widehat{\mathcal{C}}$ can be the top L variables from SIS. It can also be top L variables from forward regression (Wang 2009). The latter can be viewed as the

iterative applications of CSIS with $L = 1$. Typically, one applies a high threshold γ to unconditional SIS in (4) or a low threshold $\tilde{\gamma}$ in (6) to obtain a small set $\widehat{\mathcal{C}}$. For example, with the initial $\mathcal{C}$ being the empty set (i.e., unconditional SIS), the top L of $\{|\hat{\beta}_j^M|\}_{j=1}^p$ give a data-driven set $\widehat{\mathcal{C}}$ that consistently estimates the set $\mathcal{C}$, consisting of the L largest values of $\{|\beta_j^M|\}_{j=1}^p$, when they exceed the rest of $|\beta_j^M|$ by an order of magnitude larger than $n^{-\kappa}$, namely, when the gap exceeds the maximum stochastic errors shown in Theorem 3 by an order of magnitude. This follows directly from Theorem 3(i). Therefore, the following theorem focuses on the situation in which $P(\widehat{\mathcal{C}} = \mathcal{C}) \to 1$.

Theorem 5. Suppose that the conditioning set $\widehat{\mathcal{C}}$ is data-driven, satisfying $P(\widehat{\mathcal{C}} = \mathcal{C}) \to 1$.

(i) If the conditions of Theorem 3(ii) holds for $\mathcal{C}$, then

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

(ii) If the conditions of Theorem 4 holds for $\mathcal{C}$, then

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

The proof of this theorem follows directly from the results of Theorems 3 and 4 by considering only the event $\{\widehat{\mathcal{C}} = \mathcal{C}\}$ and the union bound of probability. As we do not impose conditions on $\mathcal{C}$, the condition $P\{\widehat{\mathcal{C}} = \mathcal{C}\} \to 1$ is really a stability condition. This theorem can further be extended to the situation where $\widehat{\mathcal{C}}$ have multiple limit points with positive probability, using a similar argument. We omit the details.

4. Selection of Thresholding Parameter

In practice, the thresholding parameter γ, which relates to the minimum strength of marginal signals in the data, has to be estimated from the data. Underestimating γ will result in a lot variables after screening, which leads to a large number of false positives, and similarly overestimate of γ will prevent sure screening. In this section, we present two procedures that select a thresholding level for CSIS.

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 $I_j(\hat{\beta}_j^M)$ denotes the element that corresponds to β_j in the information matrix $I_j(\boldsymbol{\beta}_{\mathcal{C}j})$. Using this property, we can recruit variables according to high-criticism t-tests. For a small δ, define $\hat{\mathcal{M}}_{\mathcal{D},\delta} = \{j : I_j(\hat{\beta}_j^M)^{1/2}|\hat{\beta}_j^M| \geq \delta\}$, which controls the false discovery rate $\mathbb{E}(|\hat{\mathcal{M}}_{\mathcal{D},\delta} \cap (\mathcal{M}_{\star\mathcal{D}})^c|/|(\mathcal{M}_{\star\mathcal{D}})^c|)$ defined by Zhao and Li (2012). The high-criticism t-tests provide an alternative ranking to the ranking in (4).

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 $\sup_{i\geq 1}\mathbb{E}|e_i|^{2+\chi} < \infty$ for some $\chi > 0$.
2. For $j \in (\mathcal{M}_{\star\mathcal{D}})^c$, we have that $\mathrm{cov}_L(Y, X_j|\mathbf{X}_{\mathcal{C}}) = 0$.

Theorem 6. Under Conditions 2–4, if we choose $\hat{\mathcal{M}}_{\mathcal{D},\delta} = \{j : I_j(\hat{\beta}_j^M)^{1/2}|\hat{\beta}_j^M| \geq \delta\}$, where $\delta = \Phi^{-1}(1 - f/(2d))$ 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 can be used to create a null model, in which the pseudo data $\{(\mathbf{X}_i^*, Y_i)\}_{i=1}^n$ 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. Because of random decoupling, the last d variables are unrelated to Y. Hence, the regression coefficient $\hat{\beta}_j^{M*}$, obtained from the last element of the regression coefficients of Y on $\mathbf{X}_{\mathcal{C}j}^*$ whose rows of the design matrix corresponding to X_j have been randomly permuted, 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 $\hat{\gamma}^* = \max_{q+1\leq j\leq p}|\hat{\beta}_j^{M*}|$. 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

$$\{|\hat{\beta}_{kj}^{M*}|,\ j = q+1,\ldots,p\}_{k=1}^K, \tag{14}$$

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

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 (14), denoted it by γ_τ^*. A useful range for τ is [0.95, 1]. Note that for $\tau = 1$, $\gamma_1^* = \hat{\gamma}_{\max}^*$. The selected variables are then

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

In our numerical implementations, we do coupling five times, that is, $K = 5$, and take $\tau = 0.99$. A similar idea for unconditional SIS appears already in Fan, Feng, and Song (2011) for additive models, and in Zhu et al. (2011) for more general regression functions.

5. Numerical Studies

In this section, we demonstrate the performance of CSIS on simulated data and empirical datasets. We compare CSIS versus sure independence screening and penalized least-square 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), forward regression (FR, Wang 2009) 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 from $p = 2000$ to 40,000. We present results for both linear regression and logistic regression.

We evaluate different screening methods on 200 simulated datasets 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 FDR method, we have chosen the number of tolerated false positives as $n/\log n$. For the experiments with $p = 5000$ 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 the regularization parameter for model selection is not necessarily optimal for variable screening. We have computed the results for both magnitude-based screening (4) and the likelihood ratio-based method (6). For brevity, we omit some of those numerical results.

5.1.1. Normal Model

The first two simulated examples concern linear models 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 (to make fair comparisons, the MMMS for LASSO and forward regression has been subtracted by 5, the size of the conditioning set). Surprisingly, forward regression performs better than LASSO, since the hidden variable is detected more easily once the first five variables are included. Both the FDR and the random decoupling methods return no false negatives under almost all of the simulations. In other words, both of the data-driven thresholding methods ensure 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 $p = 2000$.

	Example 1					
	SIS	MLR	CSIS	CMLR	LASSO	FR
MMMS	1995 (0)	1995 (0)	1 (0)	1 (0)	16 (1)	1 (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	FR
MMMS	1999 (0)	1999 (0)	1 (0)	1 (0)	16 (1)	1 (1)
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, $p = 5000$ and $p = 40{,}000$. Following Fan and Song (2010), we generate the covariates from

$$X_j = (\varepsilon_j + a_j\varepsilon)/\sqrt{1 + a_j^2}, \tag{15}$$

where ε and $\{\varepsilon_j\}_{j=1}^{p/3}$ are iid standard normal random variables, $\{\varepsilon_j\}_{j=p/3+1}^{2p/3}$ are iid double exponential variables with location parameter zero and scale parameter one and $\{\varepsilon_j\}_{j=2p/3+1}^{p}$ are iid 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, $p = 5000$ 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 $a_{101} = \cdots = a_{5,000} = 0$.

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 $a_{51} = \cdots, a_{40{,}000} = 0$. 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 four covariates for Example 3 and on the first two covariates for Example 4. Results are presented in Tables 2 and 3. For forward regression (FR), we also present the frequency of sure screening $\mathbb{P}(\text{SS})$, which is defined as the proportion of simulations in which FR recruits all of the active variables in the first n covariates. As the correlation among covariates increases, $\mathbb{P}(\text{SS})$ decreases, this can be seen in Tables 2 and 3.

As expected, CSIS needs a smaller model size to possess the sure screening property. The effect is more pronounced for larger p and more correlated variables. An unexpected 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 use its advantages. We also see that, both methods for choosing the thresholding parameter are very effective. Both the

Table 2. The MMMS, its RSD (in parentheses), the "false positive," and "false negative" for Example 3 with $p = 5000$ and $s = 4 + 8$.

		SIS					FR	
ρ	n	MMMS	FP_π	FN_π	FP_{FDR}	FN_{FDR}	MMMS	$\mathbb{P}$(SS)
0.00	300	86 (150)	0.21	4.61	20.75	1.23	12 (0)	1.000
0.20	100	43 (19)	34.17	0.82	87.70	0.03	12 (1)	0.920
0.40	100	56 (20)	87.38	0.00	101.75	0.00	13 (1)	0.875
0.60	100	58 (24)	88.20	0.00	101.68	0.00	14 (65)	0.660
0.80	100	63 (19)	88.17	0.00	101.64	0.00	*100* (0)	0.040
		Conditional SIS					Conditional MLR	
ρ	n	MMMS	FP_π	FN_π	FP_{FDR}	FN_{FDR}	MMMS	FP_π	FN_π
0.00	300	57 (92)	0.16	3.74	21.09	0.97	18 (25)	0.72	1.65
0.20	100	31 (38)	2.74	2.97	29.93	0.69	23 (24)	5.71	1.44
0.40	100	29 (21)	17.65	0.99	48.03	0.42	23 (17)	16.45	0.76
0.60	100	32 (18)	44.93	0.23	55.60	0.29	28 (19)	23.81	0.55
0.80	100	42 (20)	67.55	0.06	50.01	0.66	33 (22)	26.09	0.69

Table 3. The MMMS, its RSD (in parentheses), the "false positive," and "false negative" for Example 4 with $n = 200$, $p = 40{,}000$, and $s = 2 + 4$.

	MLR			Conditional MLR			FR	
ρ	MMMS	FP_π	FN_π	MMMS	FP_π	FN_π	MMMS	$\mathbb{P}$(SS)
0.00	1133 (8246)	13.61	0.19	14 (261)	5.42	0.07	6 (0)	0.965
0.20	41 (1503)	31.62	0.11	10 (21)	13.02	0.05	6 (1)	0.980
0.40	37 (12)	39.24	0.06	7 (10)	18.04	0.02	6 (1)	0.960
0.60	37 (11)	42.51	0.05	6 (5)	21.66	0.01	7 (2)	0.825
0.80	36 (12)	44.45	0.00	6 (3)	25.00	0.00	*200* (144)	0.455

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 iid 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 $\mathbb{P}(\mathbf{x}) = \exp(\mathbf{x}^T\boldsymbol{\beta}^\star)(1 + \exp(\mathbf{x}^T\boldsymbol{\beta}^\star))^{-1}$. The first two settings use the same setup of covariates and the same values for $\boldsymbol{\beta}^*$ as that in Example 1. For forward regression (FR), we fit a logistic regression at each step. The results are given in Table 4.

The results are almost the same as in the normal model, only FR does not perform well. FR selected 100 inactive variables in each of the simulations, that is, the maximum allowable when $n = 100$. Conditional screening always lists the active variable as the most important one and LASSO only needs 16 variables.

Table 4. The MMMS, its RSD (in parentheses) for the binomial model with the "false negative" and "false positive" settings for $n = 100$ and $p = 2000$.

	Example 1					
	SIS	MLR	CSIS	CMLR	LASSO	FR
MMMS	1995 (1.5)	1995 (1.5)	1 (0)	1 (0)	16 (1)	100 (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	FR
MMMS	1999 (0)	1999 (0)	1 (0)	1(0)	16 (1)	100 (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 $n = 300$, $p = 5000$, and $s = 4 + 8$.

	SIS					MLR		
ρ	MMMS	FP_π	FN_π	FP_{FDR}	FN_{FDR}	MMMS	FP_π	FN_π
0.00	215 (312)	0.19	5.78	23.06	1.77	210 (312)	20.18	0.08
0.20	27 (14)	73.22	0.02	109.56	0.00	28 (17)	107.08	0.00
0.40	49 (21)	88.19	0.00	110.15	0.00	47 (24)	107.82	0.00
0.60	56 (20)	88.17	0.00	110.00	0.00	60 (22)	107.47	0.00
0.80	68 (19)	88.20	0.00	110.34	0.00	67 (19)	107.30	0.00
	Conditional SIS					Conditional MLR		
ρ	MMMS	FP_π	FN_π	FP_{FDR}	FN_{FDR}	MMMS	FP_π	FN_π
0.00	87 (173)	20.15	1.24	24.03	1.11	83 (173)	20.18	1.21
0.20	19 (13)	49.25	0.14	53.87	0.11	20 (14)	45.27	0.20
0.40	34 (23)	67.82	0.17	61.72	0.31	39 (30)	53.48	0.49
0.60	43 (24)	77.36	0.21	53.83	1.01	71 (87)	49.47	1.15
0.80	66 (55)	78.33	0.51	36.16	3.42	402 (561)	35.42	3.43

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.

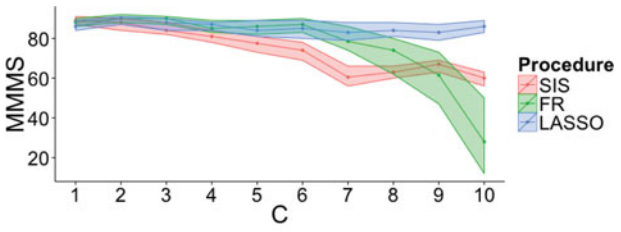
The results are similar to those for the normal model. Due to the nonlinear nature of the problem, the minimum model size is slightly higher and the thresholding methods are less efficient. However, even though the covariates are not too correlated, overall advantage of conditional sure independence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

5.1.3. Comparison of Selection Procedures for $\mathcal{C}$

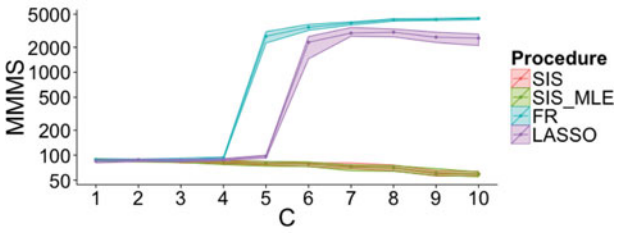
In this section, we compare a variety of variable selection tools for the selection of the conditioning set, $\mathcal{C}$. The purpose of this study is to demonstrate the performance of CSIS if $\mathcal{C}$ is chosen by data-driven methods. We select the conditioned set using four methods, (i) sure independence screening (SIS), (ii) sure independence likelihood screening (SIS-MLE), (iii) forward regression (FR), and (iv) LASSO.

Table 6. The MMMS, its RSD (in parentheses), the "false positive," and "false negative" for Example 4 with the binomial model with $n = 500$, $p = 40{,}000$, and $s = 2 + 4$.

	SIS					MLR		
ρ	MMMS	FP_π	FN_π	FP_{FDR}	FN_{FDR}	MMMS	FP_π	FN_π
0.00	318 (7038)	12.04	1.22	51.32	0.79	309 (7030)	14.06	0.22
0.20	38 (428)	32.47	0.57	68.46	0.38	37 (255)	34.10	0.09
0.40	38 (12)	38.66	0.27	73.42	0.19	35.5 (11)	40.50	0.05
0.60	38 (12)	41.99	0.16	76.11	0.10	35.5 (12)	42.89	0.03
0.80	35 (12)	43.84	0.03	77.38	0.02	33.5 (14)	44.39	0.00
	Conditional SIS					Conditional MLR		
ρ	MMMS	FP_π	FN_π	FP_{FDR}	FN_{FDR}	MMMS	FP_π	FN_π
0.00	13 (354)	5.96	0.66	42.51	0.49	25 (892)	5.96	0.14
0.20	15 (16)	14.51	0.39	49.79	0.27	13 (62)	12.38	0.09
0.40	16 (13)	19.11	0.24	51.68	0.22	13 (22)	14.17	0.08
0.60	19 (10)	22.80	0.21	51.78	0.24	15.5 (17)	13.75	0.11
0.80	19 (10)	26.39	0.14	46.49	0.64	22 (72)	9.30	0.28



(a) MMMS for the Normal Model



(b) MMMS for the Binomial Model, y-axis is given on a log scale

Figure 4. Median minimum model size (MMMS) and its 95% confidence interval with respect to the size of the conditioned set for different data-driven choices for $\mathcal{C}$.

The covariates are generated using the setup in Example 3 and we set the correlation level between the first 100 covariates, ρ, to 0.4, and set the number of parameters, p, to 5000. Sample size, n, is fixed as 100 for linear model and 300 for the binomial model. The number of parameters that are active is set to 12. We vary the size of the conditioned set, $|\mathcal{C}|$ from 1 to 10. We report the median minimum model sizes (MMMS) in the form of the 2.5th percentile and the 97.5th percentile across 200 simulations. The result is depicted in Figure 4.

Performances of different methods vary significantly across problem setups. For the normal model, as the conditioned set size grows, both FR and SIS perform better; whereas no improvements are seen for LASSO. For smaller choices of the conditioned set size, for example, $|\mathcal{C}| \leq 6$, SIS has a better overall performance compared to FR and LASSO, though the difference is substantial, as expected. In fact, the first few variables selected by the forward regression and the LARS algorithm are likely to be the same. As the conditioned set size increases, FR starts to over-perform.

For the binomial model, the results paint a completely different picture. For this setup, SIS and SIS-MLE are significantly more stable than FR and LASSO. In fact, both FR and LASSO recruit a large number of inactive variables. As a result, once the conditioned set size is larger than 5, these two methods completely breakdown. On the other hand, covariates recruited by SIS and SIS-MLE contain more true positives, and MMMS decreases monotonically over $|\mathcal{C}|$. Overall, the results of this study demonstrate that SIS (and SIS-MLE) provide robust and powerful conditioning sets for both linear and logistic regression.

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 that 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.

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 literature 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.

Appendix: Proofs

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 $\text{cov}_L(Y, X_j|\mathbf{X}_\mathcal{C}) = 0$ is 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 (10) imply that $((\boldsymbol{\beta}_\mathcal{C}^M)^T, 0)^T$ is a solution to Equation (8). 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. □

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, that is, Equations (8) and (10), 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 (12), 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 (11) 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. □

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

$$\|\boldsymbol{\beta}_{\mathcal{D}}\|^2 = O\left(\lambda_{\max}\left(\boldsymbol{\Sigma}_{\mathcal{D}|\mathcal{C}}\right)\right). \tag{A.5}$$

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

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

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 $O\left(n^{2\kappa}\lambda_{\max}\left(\boldsymbol{\Sigma}_{\mathcal{D}|\mathcal{C}}\right)\right)$. If we take $\varepsilon = c_5/2$, we obtain that

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

Finally, by Theorem 3, we obtain that

$$\mathbb{P}(\mathcal{B}_n) \geq 1 - d\Big(\exp\left(-c_4 n^{1-2\kappa}(k_n K_n)^{-2}\right) + nr_2\exp\left(-r_0 K_n^{\alpha}\right)\Big)$$

and therefore the statement of the theorem follows.

We now prove (A.5) 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

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

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

By (A.3), (8) and Lipschitz continuity of $b'(\cdot)$, we have

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

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 square, we have $\mathbb{E}[\mathbb{E}_L(\mathbf{X}_{\mathcal{D}}|\mathbf{X}_{\mathcal{C}})\mathbf{X}_{\mathcal{C}}^T] = \mathbb{E}[\mathbf{X}_{\mathcal{D}}\mathbf{X}_{\mathcal{C}}^T]$. Recalling the definition of $\boldsymbol{\Sigma}_{\mathcal{D}|\mathcal{C}}$ in Condition 3 (iii), the above expression can be written as

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

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

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

and the last two terms are $o(\lambda_{\max}([\boldsymbol{\Sigma}_{\mathcal{D}|\mathcal{C}}]))$ 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. □

Supplementary Materials

Due to space constraints, the proofs of Theorems 3 and 6, the results of a simulation study on robustness of CSIS with respect to the conditioning set and the results of a real life financial dataset study are relegated to the supplementary material (Barut, Fan, and Verhasselt 2015).

Acknowledgments

The authors are grateful to the editor, associate editor, and two referees for their valuable comments that lead to improvements in the presentation and the results of the article.

Funding

The article was initiated while Emre Barut was a graduate student and Anneleen Verhasselt was a visiting postdoctoral fellow at Princeton University. This research was partly supported by NSF Grant DMS-1206464, NIH Grants R01-GM072611, and R01-GM100474, FWO Travel Grant V422811N and FWO research grant 1.5.137.13N.

References

Barut, E., Fan, J., and Verhasselt, A. (2015), "Supplementary Material for Conditional Sure Independence Screening," available at *http://amstat.tandfonline.com/doi/suppl/10.1080/01621459.2015.1092974#.Vt3YtJMrK7o*. [1271]

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009), "Simultaneous Analysis of Lasso and Dantzig selector," *The Annals of Statistics*, 37, 1705–1732. [1266]

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. [1266]

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression," *The Annals of Statistics*, 32, 407–499. [1266,1267]

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. [1266,1272]

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. [1266]

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. [1266,1267,1268,1269]

——— (2011), "Nonconcave Penalized Likelihood With NP-Dimensionality," *IEEE Transactions on Information Theory*, 57, 5467–5484. [1266]

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. [1266,1269]

Fan, J., and Song, R. (2010), "Sure Independence Screening in Generalized Linear Models with NP-dimensionality," *The Annals of Statistics*, 38, 3567–3604. [1266,1268,1271,1273,1276]

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. [1272]

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. [1275]

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. [1266]

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. [1266]

Heyde, C. C. (1997), *Quasi-Likelihood and its Application: A General Approach to Optimal Parameter Estimation*, New York: Springer. [1272]

Li, G., Peng, H., Zhang, J., and Zhu, L. (2012a), "Robust Rank Correlation Based Screening," *The Annals of Statistics*, 40, 1846–1877. [1266]

Li, R., Zhong, W., and Zhu, L. (2012b), "Feature Screening via Distance Correlation Learning," *Journal of American Statistical Association*, 107, 1129–1139. [1266]

Osborne, M. R., Presnell, B., and Turlach, B. A. (2000), "On the LASSO and its Dual," *Journal of Computational and Graphical Statistics*, 9, 319–337. [1266]

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 Analysis 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. [1275]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society,* Series B, 58, 267–288. [1266,1273]

Wang, H. (2009), "Forward Regression for Ultra-high Dimensional Variable Screening," *Journal of the American Statistical Association*, 104, 1512–1524. [1267,1271,1273]

Zhang, C., and Zhang, T. (2012), "A General Theory of Concave Regularization for High Dimensional Sparse Estimation Problems," *Statistical Science*, 27, 576–593. [1266]

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. [1272]

Zhu, L., Li, L., Li, R., and Zhu, L. (2011), "Model-Free Feature Screening for Ultrahigh-Dimensional Data," *Journal of the American Statistical Association*, 106, 1464–1475. [1272]