Published in final edited form as:
J R Stat Soc Series B Stat Methodol. 2012 November 1; 74(5): 849–870. doi:10.1111/j.1467-9868.2011.01026.x.

CORRELATION PURSUIT: FORWARD STEPWISE VARIABLE SELECTION FOR INDEX MODELS

Wenxuan Zhong[*] **[Assistant Professor]**,
Department of Statistics, University of Illinois at Urbana Champaign, Champaign, IL 61820

Tingting Zhang [Assistant Professor],
Department of Statistics, University of Virginia, Charlottesville, VA 22904

Yu Zhu [Associate Professor], and
Department of Statistics, Purdue University, West Lafayette, IN 47907

Jun S. Liu[*] **[Professor]**
Department of Statistics, Harvard University, Cambridge, MA 02138

Abstract

In this article, a stepwise procedure, correlation pursuit (COP), is developed for variable selection under the sufficient dimension reduction framework, in which the response variable Y is influenced by the predictors $X_1, X_2, \dots, X_p$ through an unknown function of a few linear combinations of them. Unlike linear stepwise regression, COP does not impose a special form of relationship (such as linear) between the response variable and the predictor variables. The COP procedure selects variables that attain the maximum correlation between the transformed response and the linear combination of the variables. Various asymptotic properties of the COP procedure are established, and in particular, its variable selection performance under diverging number of predictors and sample size has been investigated. The excellent empirical performance of the COP procedure in comparison with existing methods are demonstrated by both extensive simulation studies and a real example in functional genomics.

Keywords

variable selection; projection pursuit regression; sliced inverse regression; stepwise regression; dimension reduction

1. Introduction

Advances in science and technology in the past few decades have led to an explosive growth of high dimensional data across a variety of areas such as genetics, molecular biology, cognitive sciences, environmental sciences, astrophysics, finance, internet commerce, etc. Compared to their dimensionalities, a large amount of data sets generated from these areas have relatively small sample sizes. Variable (or feature) selection and dimension reduction are more than often key steps in analyzing these data. Much progress has been made in the past few decades on variable selection for linear models (see Shao 1998 and Fan and Lv 2010 for a review). In recent years, shrinkage based procedures for simultaneously estimating regression coefficients and selecting predictors have been particularly attractive to researchers, and many promising algorithms such as LASSO (Tibshirani 1996, Zou 2006,

[*] To whom correspondence should be addressed.

Friedman 2007), least angle regression (LARS) (Efron et al. 2004), SCAD (Fan and Li 2001), etc., have been invented.

Let $Y \in \mathbb{R}$ be a univariate response variable and $X = (X_1, X_2, \cdots, X_p)' \in \mathbb{R}^p$ a vector of p continuous predictor variables. Throughout this article, we consider the following sufficient dimension reduction (SDR) model framework as pioneered by Li (1991) and Cook (1994). Let $\beta_1, \beta_2, \cdots, \beta_K$ be p-dimensional vectors with $\beta_i = (\beta_{1i}, \beta_{2i}, \cdots, \beta_{pi})'$ for $1 \le i \le K$. The SDR model assumes that Y and X are mutually independent conditional on $\beta_1'X, \beta_2'X, \cdots, \beta_K'X$, i.e.,

$$Y \perp X | B'X, \quad (1)$$

where "$\perp$" means "independent of" and $B = (\beta_1, \beta_2, \cdots, \beta_K)$. Expression (1) implies that all the information X contains about Y is contained in the K projections $\beta_1'X, \cdots, \beta_K'X$. A predictor variable X_j $(1 \le j \le p)$ is said to be relevant if there exists at least one i $(1 \le i \le K)$ such that $\beta_{ji} \ne 0$. Let L be the number of relevant predictor variables. When there are a large number of predictors (i.e. p is large), it is usually safe to impose the *sparsity assumption*, which states that only a small subset of the predictors influence Y and the others are irrelevant. In the SDR model, this assumption means that both K and L are small relative to p.

In his seminal paper on dimension reduction, Li (1991) proposed a seemingly different model of the form:

$$Y = f(\beta_1'X, \beta_2'X, \cdots, \beta_K'X, \varepsilon), \quad (2)$$

where f is an unknown $(K + 1)$-variate link function and ε is a stochastic error independent of X. It has been shown that the two models (1) and (2) are in fact equivalent (Zeng and Zhu 2010). We henceforth always refer to $\beta_1, \beta_2, \cdots, \beta_K$ as the SDR directions and the space spanned by these directions as a SDR subspace. In general, SDR subspaces are not unique. To resolve this ambiguity, Cook (1994) introduced the concept of *central subspace*, which is the intersection of all possible SDR subspaces and is a SDR subspace itself, and showed that the central space is well-defined and unique under some general conditions. We denote the central subspace by $\mathscr{S}B$) and assume its existence throughout this article.

Various methods have been developed for estimating $\beta_1, \cdots, \beta_K$ in the literature on SDR. One particular family of methods utilizes inverse regression, which is to regress X against Y. The sliced inversion regression (SIR) method proposed by Li (1991) is the forerunner of this family of methods. Recognizing that estimation of the SDR directions does not automatically lead to variable selection, Cook (2004) derived various χ^2 tests for assessing the contribution of predictor variables to the SDR directions. Based on these tests, Li et al. (2005) proposed a backward subset selection method for selecting significant predictors. Following the recent trend of using the L_1 or L_2 penalty for variable selection, Zhong et al. (2005) proposed to regularize the sample covariance matrix of the predictor variables in SIR and developed a procedure called RSIR for variable selection. Li (2007) proposed Sparse SIR (SSIR) to obtain shrinkage estimates of the SDR directions. Bondell and Li (2009) further adopted the nonnegative garrote method for estimating the SDR directions and showed that the resulting method is consistent in variable selection.

The majority of the aforementioned methods take a two-step approach to variable selection under the SDR model. The first step is to perform dimension reduction, that is, to estimate the SDR directions; and the second step is to select the relevant variables using statistical testing or shrinkage methods. Because these methods need to estimate the covariance and conditional covariance matrices of X, both of which are of dimensions $p \times p$, the effectiveness and robustness of the two-step approach are questionable when p is large relative to n. Zhu et al. (2006) have shown that the estimation accuracy of SDR directions deteriorates as p increases. In other words, the more irrelevant variables there are, the more likely a method fails to estimate the SDR directions accurately, and the less likely the method identifies the true relevant predictor variables.

In this article, we propose correlation pursuit (COP), a stepwise procedure for simultaneous dimension reduction and variable selection under the SDR model. Similar to projection pursuit (Friedman and Tukey 1974, Huber 1985), COP defines a projection function to measure the correlation between the transformed response and the projections of X and pursues a subset of explanatory variables that maximize the projection function. It starts with an randomly selected subset and iterates between finding an explanatory variable (predictor) that significantly improves the current projection function to add to the subset and finding an insignificant predictor to remove from the subset. During each iteration step, COP needs only to consider the predictors currently in the subset and one more predictor outside the subset. Therefore, COP can avoid the estimation and inversion of $p \times p$ covariance and conditional covariance matrices of X and mitigate the curse of dimensionality. Furthermore, COP performs dimension reduction and variable selection simultaneously. Therefore, dimension reduction and variable selection can be mutually enhanced. Our theoretical investigations as well as simulation studies show that COP is a promising tool for dimension reduction and variable selection in high dimensional data analysis.

The rest of the article is organized as follows. In Section 2, we give a brief introduction to SIR, following a correlation interpretation of SIR provided by Chen and Li (1998). This interpretation was also used in Fung et al. (2002) and Zhou and He (2008) for dimension reduction via canonical correlation. In the same section, we describe the COP procedure and derive various test statistics used by the procedure. The asymptotic behavior of the COP procedure is discussed in Section 3. Several implementation issues of the procedure are discussed in Section 4. Simulation and real data examples are reported in Sections 5 and 6, respectively. Additional remarks in Section 7 conclude the article. An abbreviated version of the proofs of the theorems is provided in the Appendix..

2. Correlation Pursuit for Variable Selection

2.1. Profile Correlation and SIR

Let η be an arbitrary direction in $\mathbb{R}^p$. We define the *profile correlation* between Y and $\eta' X$, denoted by $P(\eta)$, as:

$$P(\eta)=\max_{T}\ corr(T(Y),\eta' X), \quad (3)$$

where the maximization is taken over all possible transformations of Y including non-monotone ones. The profile correlation $P(\eta)$ reflects the largest possible correlation between a transformed response $T(Y)$ and the projection $\eta' X$. Let η_1 be the direction that maximizes $P(\eta)$ subject to $\eta'\Sigma\eta = 1$, that is, $\eta_1 = \mathrm{argmax}_{\eta'\Sigma\eta=1} P(\eta)$. We refer to η_1 as the first *principal direction* for the profile correlation between Y and X and call $P(\eta_1)$ the first profile correlation. Direction η_1, or its projection $\eta_1' X$, may not entirely characterize the dependency between Y and X. Using $P(\eta)$ as the projection function again, we can seek for a second direction, denoted by η_2, which is uncorrelated with $\eta_1' X$, i.e. $\eta_2'\Sigma\eta_1=0$, and maximizes $P(\eta)$, that is, $\eta_2 = \mathrm{argmax}_{\eta:\eta'\Sigma\eta_1=0} P(\eta)$. We refer to η_2 as the second principal direction and $P(\eta_2)$ as the second profile correlation. This procedure can be continued until no more directions can be found that are orthogonal to the obtained directions and have a nonzero profile correlation with Y. Suppose $\tilde{K}$ principal directions exist between Y and X, which are η_1, η_2, ⋯, and $\eta_{\tilde{K}}$, with the corresponding profile correlations $P(\eta_1) \geq P(\eta_2) \geq \cdots P(\eta_{\tilde{K}}) > 0$. We need to impose the following condition to establish the connection between the principal directions and the SDR directions under the SDR model.

Condition 1 (Linearity Condition): For any η in $\mathbb{R}^p$, $E(\eta' X|B' X)$ is linear in $B' X$, where B is as defined in equation (1).

Proposition 2.1. *Under the SDR model and the Linearity Condition, the principal directions* η_1, η_2, …, $\eta_{\tilde{K}}$ *are in the central space* $\mathscr{S}$B).

To make this article self-sufficient, we have included the proof of Proposition 2.1 in the Appendix. Based on the proposition, the principal directions are indeed SDR directions. In general, $\tilde{K} < K$. When the link function f is symmetric along a direction, using correlation alone may fail to recover this direction. For example, if $Y=X_1^2+\varepsilon$, the profile correlation between Y and X_1 will always be zero. To exclude this possibility, we follow the convention in the SDR literature to impose the following condition.

Condition 2 (Coverage Condition): The number of principal directions of profile correlation is equal to the dimensionality of the central subspace, that is, $\tilde{K} = K$.

Under both the linearity and coverage conditions, the principal directions η_1, η_2, ⋯, η_K form a special basis of the central subspace $\mathscr{S}$B), that is, $\mathscr{S}$B) = span(η_1, η_2, …., η_K). This basis is uniquely defined and is the estimation target of SIR. In the rest of the article, for ease of discussion, we use β_1, β_2, …., β_k and η_1, η_2, …, η_K, interchangeably.

Chen and Li (1998) showed that, at the population level, there exits an explicit solution for the principal directions. In the proof of their Theorem 3.1, Chen and Li (1998) has derived that

$$P^2(\eta)=\frac{\eta'\mathrm{var}[\mathrm{E}(X|Y)]\eta}{\eta'\Sigma\eta} \equiv \frac{\eta' M\eta}{\eta'\Sigma\eta}, \quad (4)$$

where $M \triangleq \mathrm{var}[\mathrm{E}(X|Y)]$ is the covariance matrix of the expectation of X given Y. Furthermore, the principal directions of profile correlation are the solutions of the following eigenvalue decomposition problem:

$$M\upsilon_i=\lambda_i\Sigma\upsilon_i, \upsilon_i'\Sigma\upsilon_i=1,\ \ for\ i=1,2,\ldots,K; \quad (5)$$

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_K > 0. \quad (6)$$

The principal directions η_1, η_2, …, and η_K are the first K eigenvectors of $\Sigma^{-1} M$, and their corresponding eigenvalues are exactly the squared profile correlations, that is, $P^2(\eta_i) = \lambda_i$ for $i = 1, 2, \ldots, K$.

Given independent observations $\{(\mathbf{x}_i, y_i)\}_{i=1,\cdots,n}$ of (X, Y), where $\mathbf{x}_i = (x_{i1}, \cdots, x_{ip})'$, Σ can be estimated by the sample covariance matrix,

$$\hat{\Sigma}=\frac{\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'}{n-1}, \quad (7)$$

where $\bar{\mathbf{x}}$ is the sample mean of $\{\mathbf{x}_i\}$. Li (1991) proposed the following SIR procedure to estimate M. First, the range of $\{y_i\}_{i=1}^{n}$ is divided into H disjoint intervals, denoted as $S_1, \ldots, S_H$. For $h = 1, \ldots, H$, the mean vector $\bar{\mathbf{x}}_h=n_h^{-1}\sum_{y_i\in S_h}\mathbf{x}_i$ is calculated, where n_h is the number of y_i's in S_h. Then, M is estimated by

$$\hat{\mathrm{M}}=\frac{\sum_{h=1}^{H} n_h(\bar{\mathbf{x}}_h-\bar{\mathbf{x}})(\bar{\mathbf{x}}_h-\bar{\mathbf{x}})'}{n}, \quad (8)$$

and the matrix $\Sigma^{-1} M$ is estimated by $\hat{\Sigma}^{-1} \hat{M}$. The first K eigenvectors of $\hat{\Sigma}^{-1} \hat{M}$, denoted by $\hat{\eta}_1, \hat{\eta}_2, \ldots, \hat{\eta}_K$, are used to estimate the first K eigenvectors of $\Sigma^{-1} M$ or, equivalently, the principal directions $\eta_1, \eta_2, \ldots, \eta_K$, respectively. The first K eigenvalues of $\hat{\Sigma}^{-1} \hat{M}$, denoted by $\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_K$, are used to estimate the eigenvalues of $\Sigma^{-1} M$ or, equivalently, the squared profile correlations $\lambda_1, \lambda_2, \ldots, \lambda_K$, respectively.

2.2. Correlation Pursuit

The SIR method needs to estimate the two $p \times p$ covariance matrices Σ and M, and to obtain the eigenvalue decomposition of $\hat{\Sigma}^{-1} \hat{M}$. When a large number of irrelevant variables are present and the sample size n is relatively small, $\hat{\Sigma}$ and $\hat{M}$ become unstable, which leads to very inaccurate estimates of principal directions $\hat{\eta}_1, \hat{\eta}_2, \ldots, \hat{\eta}_K$ (Zhu et al. 2006). As a consequence, those shrinkage-based variable selection methods that rely on $\hat{\eta}_1, \hat{\eta}_2, \ldots, \hat{\eta}_K$ often perform poorly for the SDR model when p is large. We here propose a stepwise SIR-based procedure for simultaneous dimension reduction (i.e., estimating the principal directions) and variable selection (i.e., identifying true predictors). Our procedure starts with a collection of randomly selected predictors and iterates between an addition step, which selects and adds a predictor to the collection, and a deletion step, which selects and deletes a predictor from the collection. The procedure terminates when no new addition or deletion occurs.

The Addition Step. Let $\mathscr{A}$ denote the collection of the indices of the selected predictors and $X_{\mathscr{A}}$ the collection of the selected variables. Applying SIR to the data involving only the predictors in $X_{\mathscr{A}}$, we obtain the estimated squared profile correlations $\hat{\lambda}_1^{\mathscr{A}}, \hat{\lambda}_2^{\mathscr{A}}, \ldots, \hat{\lambda}_K^{\mathscr{A}}$. Superscript $\mathscr{A}$ indicates that the estimated squared profile correlations depend on the current subset of selected predictors. Let X_t be an arbitrary predictor outside $\mathscr{A}$ and $\mathscr{A} + t = \mathscr{A}\{t\}$. Applying SIR to the data involving the predictors in $\mathscr{A} + t$, we obtain the estimated squared profile correlations $\hat{\lambda}_1^{\mathscr{A}+t}, \hat{\lambda}_2^{\mathscr{A}+t}, \ldots, \hat{\lambda}_K^{\mathscr{A}+t}$. Because $\mathscr{A} \subset \mathscr{A} + t$, it is easy to see that $\hat{\lambda}_1^{\mathscr{A}} \le \hat{\lambda}_1^{\mathscr{A}+t}$. The difference $\hat{\lambda}_1^{\mathscr{A}+t} - \hat{\lambda}_1^{\mathscr{A}}$ reflects the amount of improvement in the first profile correlation due to the incorporation of X_t. We standardize this difference and use the the resulting test statistic

$$\mathrm{COP}_1^{\mathscr{A}+t}=\frac{n(\hat{\lambda}_1^{\mathscr{A}+t}-\hat{\lambda}_1^{\mathscr{A}})}{1-\hat{\lambda}_1^{\mathscr{A}+t}}, \quad (9)$$

to assess the significance of adding X_t to $\mathscr{A}$ in improving the first profile correlation. Similarly, the contributions of adding X_t to the other profile correlations can be assessed by

$$\mathrm{COP}_i^{\mathscr{A}+t}=\frac{n(\hat{\lambda}_i^{\mathscr{A}+t}-\hat{\lambda}_i^{\mathscr{A}})}{1-\hat{\lambda}_i^{\mathscr{A}+t}}, \quad (10)$$

for $2 \le i \le K$. The overall contribution of adding X_t to the improvement in all the K profile correlations can be assessed by combining the statistics $\text{COP}_i^{\mathscr{A}+t}$ into one single test statistic

$$\text{COP}_{1:K}^{\mathscr{A}+t}=\sum_{i=1}^{K}\text{COP}_i^{\mathscr{A}+t}. \quad (11)$$

We further define that

$$\overline{\text{COP}}_{1:K}^{\mathscr{A}}=\max_{t\in\mathscr{A}^c}\text{COP}_{1:K}^{\mathscr{A}+t}. \quad (12)$$

Let $X_{\bar{t}}$ be a predictor that attains $\overline{\text{COP}}_{1:K}^{\mathscr{A}}$, that is, $\overline{\text{COP}}_{1:K}^{\mathscr{A}}=\text{COP}_{1:K}^{\mathscr{A}+\bar{t}}$, and let c_e be a pre-specified threshold (details about its choice are deferred to the next two sections). Then, if $\overline{\text{COP}}_{1:K}^{\mathscr{A}}>c_e$, we add $\bar{t}$ to $\mathscr{A}$ otherwise, we do not add any variable.

The deletion step. Let X_t be an arbitrary predictor in $\mathscr{A}$ and define $\mathscr{A}-t=\mathscr{A}-\{t\}$. Let $\hat{\lambda}_1^{\mathscr{A}-t}, \hat{\lambda}_2^{\mathscr{A}-t}, \ldots, \hat{\lambda}_K^{\mathscr{A}-t}$ be the estimated squared profile correlations based on the data involving the predictors in $\mathscr{A}-t$ only. The impact of deleting X_t from $\mathscr{A}$ on the ith squared profile correlation can be measured by

$$\text{COP}_i^{\mathscr{A}-t}=\frac{n(\hat{\lambda}_i^{\mathscr{A}}-\hat{\lambda}_i^{\mathscr{A}-t})}{1-\hat{\lambda}_i^{\mathscr{A}}}, \quad (13)$$

for $1 \le i \le K$. The overall impact of deleting X_t is measured by

$$\text{COP}_{1:K}^{\mathscr{A}-t}=\sum_{i=1}^{K}\text{COP}_i^{\mathscr{A}-t}, \quad (14)$$

and the least impact from deleting one predictor from $\mathscr{A}$ is then defined to be

$$\underline{\text{COP}}_{1:K}^{\mathscr{A}}=\min_{t\in\mathscr{A}}\text{COP}_{1:K}^{\mathscr{A}-t}. \quad (15)$$

Let $X_{\underline{t}}$ be a predictor that achieves $\underline{\text{COP}}_{1:K}^{\mathscr{A}}$, and let c_d be a pre-specified threshold for deletion. If $\underline{\text{COP}}_{1:K}^{\mathscr{A}}<c_d$, we delete $X_{\underline{t}}$ from $\mathscr{A}$ otherwise, no deletion happens.

The asymptotic distributions of the proposed statistics and the selection of the thresholds will be discussed in the next two sections. Because the described procedure aims to find predictors that can most significantly improve the profile correlations between Y and X, we call it the Correlation Pursuit procedure (COP). Below we summarize COP into a pseudo-code.

The COP Algorithm

```
1. Set the number of principal directions $K$ and the threshold values $c_e$ and $c_d$.
2. Randomly select $K + 1$ variables as the initial collection of selected variables $\mathscr{A}$
3. Iterate until no more addition or deletion of predictors can be performed:
```

The addition step:

- Find $\bar{t}$ such that $\mathrm{COP}_{1:K}^{\mathscr{A}+\bar{t}}=\overline{\mathrm{COP}}_{1:K}^{\mathscr{A}}$;
- If $\overline{\mathrm{COP}}_{1:K}^{\mathscr{A}}>c_e$, add $\bar{t}$ to $\mathscr{A}$ that is, let $\mathscr{A}=\mathscr{A}+\bar{t}$;

The deletion step:

- Find $\underline{t}$ such that $\mathrm{COP}_{1:K}^{\mathscr{A}-\underline{t}}=\underline{\mathrm{COP}}_{1:K}^{\mathscr{A}}$;
- If $\underline{\mathrm{COP}}_{1:K}^{\mathscr{A}}<c_d$, delete $\underline{t}$ from $\mathscr{A}$ that is, let $\mathscr{A}=\mathscr{A}-\underline{t}$;

4. Output $\mathscr{A}$

3. Theoretical Properties

3.1. Asymptotic distributions of test statistics in COP

Let us first consider an addition step. We assume that SIR uses a fixed slicing scheme relative to the number of observations n, that is, the slices $S_1, S_2, \cdots, S_H$ are fixed (defined by the range of the response variable) but the number of observations in each slice goes to infinity. Let X_t be an arbitrary predictor in $\mathscr{A}^c$. Under the null hypothesis H_0 that all the predictors in $\mathscr{A}^c$ are irrelevant, we have $\eta_{t1} = \eta_{t2} = \cdots = \eta_{tK} = 0$. Recall that the statistics we propose to measure the contributions of X_t to the K profile correlations are $(\mathrm{COP}_1^{\mathscr{A}+t}, \mathrm{COP}_2^{\mathscr{A}+t}, \ldots, \mathrm{COP}_K^{\mathscr{A}+t})'$, and to measure the overall contribution of X_t by $\mathrm{COP}_{1:K}^{\mathscr{A}+t}$. To establish the asymptotic distributions of these statistics, we need to impose a condition on the conditional expectation of X_t given $X_{\mathscr{A}}$.

Condition 3 (Regression Condition): $\mathrm{E}(X_t \mid X_{\mathscr{A}})$ is linear in $X_{\mathscr{A}}$.

Theorem 3.1. *Assume that Conditions 1 and 2 hold, Condition 3 holds for* $(X_{\mathscr{A}}, X_t)$ *for any* $X_t \in X_{\mathscr{A}^C}$, *and the squared profile correlations* $\lambda_1, \lambda_2, \ldots, \lambda_K$ *are positive and different from each other. Then, for any given fixed slicing scheme, under the null hypothesis* H_0 *that all the predictors in* $\mathscr{A}^c$ *are irrelevant, we have that*

$$(COP_1^{\mathscr{A}+t}, COP_2^{\mathscr{A}+t}, \cdots, COP_K^{\mathscr{A}+t}) \to (Z_{1t}^2, Z_{2t}^2, \cdots, Z_{Kt}^2) \quad (16)$$

in distribution and

$$COP_{1:K}^{\mathscr{A}+t} \to \sum_{l=1}^{K} Z_{lt}^2 \quad (17)$$

in distribution as $n \to \infty$. *Here,* $(Z_{1t}, Z_{2t}, \cdots, Z_{Kt})$ *follows the multivariate normal distribution with mean zero and covariance matrix* W_{Kt}. *The explicit expression of* W_{Kt} *is given in the* Appendix.

The asymptotic distributions in Theorem 3.1 can be much simplified if we impose the following condition on the variance of the conditional expectation of X_t given $X_{\mathscr{A}}$.

Condition 4 (Constant Variance Condition): $\mathrm{E}[(X_t - \mathrm{E}(X_t|X_{\mathscr{A}}))^2|X_{\mathscr{A}}]$ is a constant.

Corollary 3.1. *Assume that Conditions 1 and 2 hold, Conditions 3 and 4 hold for* $(X_{\mathscr{A}}, X_t)$ *for* $X_t \in X_{\mathscr{A}^c}$, *and the squared profile correlations* $\lambda_1, \lambda_2, \ldots, \lambda_K$ *are positive and different from each other. Then, for any given fixed slicing scheme, under the null hypothesis* H_0 *that* *all the predictors in* $\mathscr{A}^c$ *are irrelevant, we have that* $COP_1^{\mathscr{A}+t}, COP_2^{\mathscr{A}+t}, \cdots, COP_K^{\mathscr{A}+t}$ *are asymptotically independent and identically distributed as* $\chi^2(1)$, *and* $COP_{1:K}^{\mathscr{A}+t}$ *is asymptotically* $\chi^2(K)$.

Theorems 3.1 and Corollary 3.1 characterize the asymptotic behaviors of the test statistics for an arbitrary X_t in $\mathscr{A}^c$. In the COP procedure, however, the predictor that attains the maximum value of $\mathrm{COP}_{1:K}^{\mathscr{A}+t}$ among $t \in \mathscr{A}^c$, which is $\overline{\mathrm{COP}}_{1:K}^{\mathscr{A}}$, is considered a candidate predictor to enter $\mathscr{A}$ Our next theorem characterizes the joint asymptotic behavior of $\{\mathrm{COP}_{1:K}^{\mathscr{A}+t}\}_{t \in \mathscr{A}^c}$ as well as that of $\overline{\mathrm{COP}}_{1:K}^{\mathscr{A}}$.

Note that the linearity, regression, and constant variance conditions together are more general than the normality assumption on X because they only need to hold for the basis of the central subspace (e.g., B or $\eta_1, \cdots, \eta_K$) and a given subset of predictors (e.g., $\mathscr{A}$. If we require that the conditions hold for any projection and any given subset of the predictors, however, then it is equivalent to requiring that X follows a multivariate normal distribution. In order to understand the joint behavior of all the COP statistics, in what follows we impose the normality assumption on X.

Let $\mathscr{A}=\{t_j\}_{j=1}^d$ *and* $\mathscr{A}^c=\{t_j\}_{j=d+1}^p$ denote the collection of currently selected predictors and its complement respectively. Let $\Sigma_{\mathscr{A}} = \mathrm{Cov}(X_{\mathscr{A}})$, $\Sigma_{\mathscr{A}^c} = \mathrm{Cov}(X_{\mathscr{A}^c})$, $\Sigma_{\mathscr{A}^c} = \mathrm{Cov}(X_{\mathscr{A}}, X_{\mathscr{A}^c})$, and $\tilde{\Sigma}_{\mathscr{A}^c}=\Sigma_{\mathscr{A}^c} - \Sigma_{\mathscr{A}^c\mathscr{A}}\Sigma_{\mathscr{A}}^{-1}\Sigma_{\mathscr{A}\mathscr{A}^c}$. Note that $\Sigma_{\mathscr{A}^c\mathscr{A}}=\Sigma'_{\mathscr{A}\mathscr{A}^c}$. Let $\tilde{a} = (\tilde{a}_1, \tilde{a}_2, \ldots, \tilde{a}_{p-d})'$ be the vector of the diagonal elements of $\tilde{\Sigma}_{\mathscr{A}^c}$. Define $D_{\mathscr{A}^c} = \mathrm{diag}(\tilde{a}_1, \tilde{a}_2, \ldots, \tilde{a}_{p-d})$, and define $U_{\mathscr{A}^c}=D_{\mathscr{A}^c}^{-1/2}\tilde{\Sigma}_{\mathscr{A}^c}D_{\mathscr{A}^c}^{-1/2}$.

Theorem 3.2. *Assume that (a) X follows a multivariate normal distribution; (b) the coverage condition holds; and (c) the squared profile correlations* $\lambda_1, \lambda_2, \cdots, \lambda_K$ *are nonzero and different from each other. Then, for any fixed slicing scheme, under the null hypothesis* H_0 *that all the predictors in* $\mathscr{A}^c$ *are irrelevant, we have*

$$(COP_{1:K}^{\mathscr{A}+t_{d+1}}, COP_{1:K}^{\mathscr{A}+t_{d+2}}, \cdots, COP_{1:K}^{\mathscr{A}+t_p}) \xrightarrow{D} (\sum_{k=1}^K z_{k,d+1}^2, \cdots, \sum_{k=1}^K z_{k,p}^2), \quad (18)$$

and

$$\overline{COP}_{1:K}^{\mathscr{A}} \xrightarrow{D} \max_{t \in \mathscr{A}^c} \sum_{k=1}^K z_{k,t}^2 \quad (19)$$

as n goes to ∞. *Here* $z_k = (z_{k,d+1}, \cdots, z_{k,p})$ *for* $k = 1, \cdots, K$ *are mutually independent and each* z_k *follows a multivariate normal with mean zero and covariance matrix* $U_{\mathscr{A}^c}$.

We now consider deletion steps of the COP procedure. We let $\mathscr{A}$ denote the current collection of selected predictors before a deletion step, and let X_t be an arbitrary predictor in $\mathscr{A}$ Note that $\mathrm{COP}_k^{\mathscr{A}-t}=\mathrm{COP}_k^{\tilde{\mathscr{A}}+t}$, where $\tilde{\mathscr{A}} = \mathscr{A} - t$ for $1 \le k \le K$. Therefore, results similar to those stated in Theorem 3.1 and Corollary 3.1 can be obtained for $(\mathrm{COP}_1^{\mathscr{A}-t}, \mathrm{COP}_2^{\mathscr{A}-t}, \cdots, \mathrm{COP}_K^{\mathscr{A}-t})$ *and* $\mathrm{COP}_{1:K}^{\mathscr{A}-t}$ after some modifications described below. First, our current "null hypothesis", denoted as H_{0b}, is that X_t and the predictors in $\mathscr{A}^c$ are irrelevant. Second, the regression and constant variance conditions need to be imposed on the conditional expectation of X_t given $X_{\tilde{\mathscr{A}}}$ instead. The asymptotic distribution of $\underline{\mathrm{COP}}_{1:K}^{\mathscr{A}}$, however, turns out to be fairly complicated if not entirely elusive, due to the fact that there does not exist a common null hypothesis for all $X_t \in \mathscr{A}$ In what follows, we will establish two strong results that have implications for properly selecting the thresholds c_e and c_d, as well as for the consistency of the COP procedure in selecting true predictors.

3.2. Selection consistency of COP

Let $\mathscr{T}$ be the collection of the true predictors under the SDR model. The principal profile correlation directions are $\eta_1, \eta_2, \cdots, \eta_K$, which form a basis of the central subspace. Assume that $S_1, \cdots, S_H$ is a fixed slicing scheme used by SIR.

Let $p_h = P(y \in S_h)$, $\mathbf{v}_K = (\eta_1' X - E(\eta_1' X), \cdots, \eta_K' X - E(\eta_K' X))'$, and

$$M_{H,K} = \sum_{h=1}^{H} p_h \mathbf{L}_{h,K} \mathbf{L}_{h,K}', \quad (20)$$

where $\mathbf{L}_{h,K} = E(\mathbf{v}_K | Y \in S_h)$. A few more conditions are needed for the results we state in the next two theorems.

Condition 5. X follows a multivariate normal distribution with covariance matrix Σ such that $\tau_{\min} \le \lambda_{\min}(\Sigma_p) \le \lambda_{\max}(\Sigma_p) \le \tau_{\max}$, where $\tau_{\min}$ and $\tau_{\max}$ are two positive constants, and $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ are the minimum and maximum eigenvalues of a matrix, respectively.

Condition 6. There exists a constant $\omega_H > 0$ such that $\lambda_{\min}(M_{H,K}) > \omega_H$.

Condition 7. There exist constants σ_0^2 and $\upsilon > 0$ such that for any slice S_h and any two predictors X_i and X_j, $\mathrm{Var}(X_j | Y \in S_h) \le \sigma_0^2$ *and* $Var\ (X_i X_j | y \in S_h) \le \sigma_0^2$ for all $i, j = 1, \cdots, p$, and $h = 1, \cdots, H$. In addition,

$$E(|X_j|^l | Y \in S_h) \le \frac{l!}{2} \mathrm{Var}(X_j | Y \in S_h) \upsilon^{l-2} \ and$$

$$E(|X_i X_j|^l | Y \in S_h) \le \frac{l!}{2} \mathrm{Var}(X_i X_j | Y \in S_h) \upsilon^{l-2}, \ for\ l \ge 2.$$

Condition 8. Let $\eta^j = (\eta_{j1}, \eta_{j2}, \cdots, \eta_{jK})'$ in which η_{jk} is the coefficient of X_j in the kth principal correlation direction η_k. There exist a positive constant $\varpi > 0$ and a nonnegative constant ξ_0; such that $\| \eta^j \|^2 > \varpi \cdot n^{-\xi_0}$ for $j \in \mathscr{T}$ where $\| \cdot \|$ denotes the standard L_2 − norm.

Condition 9. $\lim_{n \to \infty} p = \infty$ and $p = o(n^{\varrho_0})$ with $\varrho_0 \ge 0$ and $2\varrho_0 + 2\xi_0 < 1$.

Condition 5 ensures that the variances of the predictors are on a comparable scale and that they are not strongly correlated. Condition 6 assumes a lower bound for the eigenvalues of $M_{H,K}$, which is slightly stronger than the coverage condition that ensures SIR to recover all the SDR directions. Condition 7 imposes conditions on the moments of the conditional expectations of X given $Y \in S_h$ so that the Bernstein inequalities hold for the conditional sample means. Condition 8 assumes that the coefficients of any true predictors do not decrease to zero too fast as both n and p increase; otherwise, such predictors will not be identifiable asymptotically. Condition 9 allows p to increase as n increases, but their rates are constrained. Similar conditions have been used by others for establishing variable selection results for stepwise procedures in linear regression (Wang 2009, Fan and Lv 2008).

Theorem 3.3. *Let $\mathscr{A}$ be the set of currently selected predictors and let $\mathscr{T}$ be the set of true predictors. Let* $\vartheta = \varpi \cdot \omega_H \cdot \frac{\tau_{\min}^2}{2\tau_{\max}}$. *Assume that Conditions 5–9 hold. Then, we have*

$$P\left(\min_{\mathscr{A}:\mathscr{A}^c\cap\mathscr{T}\neq\varnothing}\ \max_{t\in\mathscr{A}^c\cap\mathscr{T}} COP_{1:K}^{\mathscr{A}+t} \geq \vartheta n^{1-\xi_0}\right) \to 1, \quad (21)$$

for any fixed slicing scheme as n goes to ∞.

The probability statement in (21) is not just about one given collection of predictors. It considers all the possible collections that do not include all the true predictors yet, that is, $\{\mathscr{A}: \mathscr{A}^c \cap \mathscr{T} \neq \varnothing\}$. In other words, it considers all the possible scenarios where the null hypothesis H_0 is not true. Further note that $\max_{t\in\mathscr{A}^c\cap\mathscr{T}} COP_{1:K}^{\mathscr{A}+t} \neq \overline{\text{COP}}_{1:K}^{\mathscr{A}}$. Because $\max_{t\in\mathscr{A}^c} COP_{1:K}^{\mathscr{A}+t} \geq \max_{t\in\mathscr{A}^c\cap\mathscr{T}} COP_{1:K}^{\mathscr{A}+t}$, from equation (21), we have

$$P\left(\min_{\mathscr{A}:\mathscr{A}^c\cap\mathscr{T}\neq\varnothing} \overline{\text{COP}}_{1:K}^{\mathscr{A}} \geq \vartheta n^{1-\xi_0}\right) \to 1. \quad (22)$$

This result implies that by setting c_e to $\vartheta n^{1-\xi_0}$ or smaller, if the COP procedure has not collected all the true predictors yet, then with probability going to 1 (as n goes to ∞), it will continue to select a predictor to the current collection. Thus, the addition step of COP will not stop until all the true predictors are selected. Another way to interpret (22) is that the selection power of the COP procedure converges to 1 asymptotically.

Theorem 3.4. *Assume that Conditions 5–9 hold. Then we have*

$$P\left(\max_{\mathscr{A}:\mathscr{A}^c\cap\mathscr{T}\neq\varnothing}\ \max_{t\in\mathscr{A}^c} COP_{1:K}^{\mathscr{A}+t} < Cn^{\varrho}\right) \to 1, \quad (23)$$

for $\varrho > 1/2 + \varrho_0$, *and any positive constant C, under any fixed slicing scheme with n going to* ∞.

Theorem 3.4 has two implications. The first one regards the addition step of COP. Once all the true predictors are selected, that is, $\mathscr{A}^c \cap \mathscr{T} = \varnothing$, the probability that it will select a false predictor from $\mathscr{A}^c$ converges to zero. The second implication concerns the deletion step. Consider one collection of selected predictors $\tilde{\mathscr{A}}$ and assume that $\tilde{\mathscr{A}}$ contains all the true predictors and also some irrelevant ones, that is, $\tilde{\mathscr{A}} \supset \mathscr{T}$ Clearly,

$$\underline{\text{COP}}_{1:K}^{\tilde{\mathscr{A}}} \leq \min_{t\in\tilde{\mathscr{A}}-\mathscr{T}} COP_{1:K}^{\tilde{\mathscr{A}}-t} \leq \max_{\mathscr{A}:\mathscr{A}^c\cap\mathscr{T}=\varnothing}\ \max_{t\in\mathscr{A}^c} \sum_{k=1}^{K} \text{COP}_k^{\mathscr{A}+t}. \quad (24)$$

Therefore,

$$P\left(\underline{\text{COP}}_{1:K}^{\tilde{\mathscr{A}}} < C_{n^{\varrho}}\right) \to 1. \quad (25)$$

In other words, with probability going to 1, the COP procedure will delete an irrelevant predictor from the current collection.

One possible choice of the thresholds is $\chi_e^2 = \vartheta n^{1-\xi_0}$ *and* $\chi_d^2 = \vartheta \cdot n^{1-\xi_0}/2$. From Theorem 3.3, asymptotically, the COP algorithm will not stop selecting variables until all the true predictors are included. Moreover, once all the true predictors are included, according to Theorem 3.4, all the redundant variables will be removed from the selected variables.

4. Implementation Issues

When implementing the COP algorithm, one needs to specify the number of profile correlation directions K, the thresholds for the addition and deletion steps c_e and c_d, and the slicing scheme particularly the number of slices H. A proper specification of these tuning parameters is critical for the success of the COP algorithm.

4.1. Slicing Schemes and the Choice of *H*

Li (1991) suggested that in terms of estimation, the performance of SIR is robust to the number of slices in general. The COP algorithm uses SIR to derive test statistics for selecting variables. It is of interest to understand the impact of a slicing scheme on the involved testing procedures. Again, we consider an addition step in the COP procedure. Let $\mathscr{A}$ be the current collection of selected predictors. Let X_t be an arbitrary predictor in $\mathscr{A}^c$.

Theorem 4.1. *Assume that X follows a multivariate normal distribution. Then, for any given fixed slicing scheme, we have*

$$P\left(\frac{COP_{1:K}^{\mathscr{A}}}{n} \geq C_{H,\mathscr{A}+t}\right) \to 1, \quad as\ n \to \infty, \quad (26)$$

where

$$C_{H,\mathscr{A}+t}=(\tilde{\eta}_{t,\mathscr{A}})' M_{H,K} \tilde{\eta}_{t,\mathscr{A}} / \sigma_{t,\mathscr{A}}^2, \quad (27)$$

$\sigma_{t,\mathscr{A}}^2 = Var(X_t|X_{\mathscr{A}})$, $\tilde{\eta}_{t,\mathscr{A}} = Cov(X_t, \mathbf{v}_K|X_{\mathscr{A}})$ *and* $M_{H,K}$ *is defined in* equation (20).

The difference between Theorem 3.1 and Theorem 4.1 is that the latter does not assume that X_t is an irrelevant predictor. When X_t is indeed a true predictor, then η^t is not a zero vector and $\max_{t \in \mathscr{A}^c \cap \mathscr{T}} C_{H,\mathscr{A}\,t}$ is greater than zero. The larger $C_{H,\mathscr{A}\,t}$ is, the more likely X_t will be added to $\mathscr{A}$ The next result shows that a finer slicing scheme leads to higher power for the addition step by COP. For any two different slicing schemes $S = (S_1, \cdots, S_{H_1})$ and $S'=(S_1^{'}, \cdots, S_{H_2}^{'})$, we say that S' is a refinement of S, denoted by $S' \precsim S$, if for any $S_{h'}^{'} \in S'$, there exists a $S_h \in S$ such that $S_{h'}^{'} \subseteq S_h$.

Proposition 4.1. *Suppose S and* S' *are two slicing schemes such that* $S' \precsim S$*. Then, for any* $\eta \in \mathbb{R}^K$, *we have*

$$\eta' M_{H_2,K} \eta \geq \eta' M_{H_1,K} \eta, \quad (28)$$

where $M_{H_2,K}$ *and* $M_{H_1,K}$ *are defined as in* (20) *under the slicing schemes* S' *and S, respectively.*

Proposition 4.1 implies that the constant $C_{H,\mathscr{A}}$ in Theorem 4.1 becomes larger when a finer slicing scheme is used. This further suggests that the power of the COP procedure in selecting true predictors tends to increase if a slicing scheme uses a larger number of slices. On the other hand, when a slicing scheme uses a larger number of slices, the number of observations in each slice will decrease, which makes the estimate of $E(X|y \in S_h)$ less accurate and further makes the estimates of $M = \text{Cov}\{E(X|Y)\}$ and its eigenvalues $\lambda_1, \cdots, \lambda_K$ less stable. The success of the COP procedure hinges on a good balance between the number of slices and the number of observations in each slice. We observed from intensive simulation studies that, with a reasonable number of observations in each slice (say ≥ 20), a large number of slices is preferred.

4.2. Choice of c_e and c_d

Section 3 has characterized the asymptotic distributions or behaviors of the test statistics involved in the COP procedure. In theory, these results (Theorems 3.4 and 3.5) can be used for choosing the thresholds c_e and c_d. In practice, however, these thresholds should be used with much caution due to the following concerns. First, the distributions obtained in Section 3 are for a single addition or deletion step and under various assumptions. Second, the distributions are valid only in an asymptotic sense. In what follows, we propose to use a cross-validation procedure for selecting c_e and c_d.

Let $\{\alpha_i\}_{1\le i\le d}$ be a pre-specified grid on a sub-interval in (0, 1) and $\{\chi^2_{\alpha_i,K}\}_{1\le i\le d}$ be the collection of the 100α_ith percentile of χ^2_K. For convenience, we only consider the m pairs of $c_e=\chi^2_{\alpha_i,K}$ *and* $c_d=\chi^2_{\alpha_i-0.05,K}$ for $1 \le i \le m$. Note that $c_d < c_e$ and that there is only one tuning parameter we need to determine. We follow the general 5-fold cross validation scheme to select the best pair of c_e and c_d. We randomly divide the original data into five equal-sized subsets, and then apply the COP procedure to any four subsets to generate the estimation and variable selection results. The remaining subset of the data is used to test the model and generate a performance measurement. The performance measurements are averaged and the result is used as the CV score. We choose the pair of c_e and c_d that mazimizes the CV score.

We define the performance measure used in the CV procedure as follows. Suppose $\mathscr{A}$ is the collection of selected predictors and $\eta_{1,\mathscr{A}}, \cdots, \eta_{K,\mathscr{A}}$ are the estimates of the principal profile correlation directions produced by applying the COP procedure to the training data set. We consider the first principal profile correlation direction first. Recall $\eta_{1,\mathscr{A}}$ is the direction that achieves the maximum correlation of a linear projection of X and the transformed response Y, and the optimal transformation is $T_1(Y)=E(\eta'_{1,\mathscr{A}}X|Y)$ (Theorem 3.1 in Chen and Li (1998)). With $\eta_{1,\mathscr{A}}$ estimated by $\hat{\eta}_{1,\mathscr{A}}$ using the training data, we apply LOESS proposed by Cleveland (1979) to fit $T_1(Y)$ using the training data and denote the fitted transformation as $\hat{T}_1(\cdot)$. Let $\tilde{X}$ and $\tilde{Y}$ be the data matrix and the response vector of the testing data set. Then, the squared profile correlation between $\tilde{X}$ and $\tilde{Y}$ based on the direction $\hat{\eta}_{1,\mathscr{A}}$ and transformation $\hat{T}_1(\cdot)$ is computed as $\mathrm{corr}^2(\hat{\mathrm{T}}_1(\tilde{Y}), \hat{\eta}'_{1,\mathscr{A}}\tilde{\mathrm{X}})$. Similarly, the squared profile correlations between $\tilde{X}$ and $\tilde{Y}$ along $\hat{\eta}_{2,\mathscr{A}}, \cdots, \hat{\eta}_{K,\mathscr{A}}$ can be calculated. The overall performance measure is defined to be

$$\mathrm{PC}=\sum_{k=1}^{K}\mathrm{corr}^2(\hat{\mathrm{T}}_k(\tilde{Y}), \hat{\eta}'_{k,\mathscr{A}}\tilde{\mathrm{X}}). \quad (29)$$

The CV score for any pair (c_e, c_d) is defined to be the average PC over the five possible partitions of the training-test data sets.

4.3. Selection of the number of directions K

To determine K, the number of principal profile correlation directions, we adopt a BIC-type criterion proposed by Zhu et al. (2006). For any given k between 1 and J, where $J \le \max(n, p)$ is a reasonable upper bound chosen by the user, we apply the COP procedure with $K = k$. Suppose the resulting collection of the selected predictors is $\mathscr{A}_k$ and the cardinality of $\mathscr{A}_k$ is p_k. Using the data involving only the selected predictors, we can estimate $M = \mathrm{Cov}(E(X_{\mathscr{A}_k}|$

Y)) as before and denote the result as $\hat{M}$. Let $\hat{\theta}_1 \geq \hat{\theta}_2 \cdots \hat{\theta}_p$ be the eigenvalues of $\hat{M} + I_{p_k}$, where I_{p_k} is the p_k by p_k identity matrix, and let τ be the number of $\hat{\theta}_i$'s that are greater than 1. Define

$$G(k) = -\,log\ L(k) + \frac{\log(n)}{2} k(2p_k - k + 1), \quad (30)$$

where $log\ L(k) = \sum_{i=\min(\tau,k)+1}^{p} (log\ \hat{\theta}_i + 1 - \hat{\theta}_i)$. We choose $K = \text{argmin}_{1 \leq k \leq J}\, G(k)$. In the original criterion proposed by Zhu et al. (2006), the authors show that the criterion produces a consistent estimate of K for fixed p_k. Our simulation study shows that the modified criterion leads to the correct specification of K for the COP procedure and can be generally used in practice.

5. Simulation Study

We have performed extensive simulation studies to compare the COP algorithm with a few existing variable selection methods and will present three examples in this section. When implementing the COP algorithm in these examples, we use the CV procedure and the GIC criterion discussed in the previous section to select the thresholds c_e and c_d and the dimensionality K, respectively. The grids used for selecting c_e is $\{\chi^2_{0.90,K}, \chi^2_{0.95,K}, \chi^2_{0.99,K}, \chi^2_{0.999,K}, \chi^2_{0.9999,K}\}$, and the associated grid for selecting c_d is $\{\chi^2_{0.85,K}, \chi^2_{0.90,K}, \chi^2_{0.94,K}, \chi^2_{0.949,K}, \chi^2_{0.9499,K}\}$. The range used for selecting K is from 1 to 4 (i.e., $J = 4$). For SSIR, we use the grid {0, 0.1, …, 0.9, 1} × {0, 0.1, …, 0.9, 1} to select the pair of tuning parameters that leads to its best performance. Both COP and SSIR involve slicing the range of the response variable, for which we use the same scheme to facilitate fair comparison.

5.1. Linear models

In this example, we consider the linear model

$$Y = X\beta + \sigma\varepsilon, \quad (31)$$

where $X = (X_1, X_2, \cdots, X_p)'$ follows a p-variate normal distribution with mean zero and covariances $\text{Cov}(X_i, X_j) = \rho^{|i-j|}$ for $1 \leq i, j \leq p$, and ε is independent of X and follows $N(0, 1)$. The variable selection methods we compare the COP procedure to include LASSO, SCAD (Fan and Li 2001), MARS, and SSIR (Li 2007). The R packages SIS, lars and mda are used to run SCAD, LASSO and MARS, respectively. The tuning parameters involved in SCAD and LASSO are selected by cross validation. We use the codes provided by the original authors to run SSIR. In this example, we consider two specifications of the linear model given below.

Scenario 1.1 : $p = 8$, $\beta = (3, 1.5, 2, 0, 0, 0, 0, 0)'$, $\sigma = 3$, $\rho = 0.5$;

Scenario 1.2 : $p = 1000$, $\beta = (3, 1.5, 1, 1, 2, 1, 0.9, 1, 1, 1, 0, \cdots, 0)'$, $\sigma = 1$, $\rho = 0.5$.

Under Scenario 1.1, Model (31) involves 3 true predictors and 5 irrelevant variables, and was originally used in Tibshirani (1996) and Fan and Li (2001) to demonstrate the empirical performances of LASSO and SCAD. We randomly generated 100 data sets from Scenario 1.1, each with 40 data points (i.e. n=40), and applied the aforementioned methods to the data sets. Two quantities were used to measure the variable selection performance of each method, which are the average number of irrelevant predictors falsely selected as true predictors (denoted by FP) and the average number of true predictors falsely excluded as irrelevant predictors (denoted by FN). Note that under Scenario 1.1, FPs and FNs range from

0 to 5 and 0 to 3, respectively, with small values indicating good performances in variable selection. The FP and FN values of the tested methods are reported in Table 1.

Under Scenario 1.2, Model (31) involves ten true predictors and 990 irrelevant predictors and is clearly more challenging than Scenario 1.1. We randomly generated 100 data sets each with 200 data points (i.e. n=200) from Scenario 1.2. Note that in each data set, $n < p$. Similar to Scenario 1.1, we applied the methods mentioned above to the data sets and report the FP and FN values of these methods in Table 1. The tuning parameters in all these methods are determined by cross validation.

From the left panel of Table 1, under Scenario 1.1, SSIR has the lowest FP value (FP=0.19), that is, the average number of irrelevant variables selected by SSIR is 0.19; and COP has the third lowest FP values (0.71). The other methods tend to have more false positives than SSIR and COP. In terms of FNs, the order of the methods ranked from the lowest to the highest is MARS, SCAD, LASSO, COP, and SSIR. The relative sub-par performance of COP and SSIR is due to the fact that these two methods are developed for variable selection under models more general than the linear model.

From the right panel of Table 1, under Scenario 1.2, COP has the lowest FP value (FP=2.28). In terms of FN, LASSO and MARS have the lowest value with COP following modestly behind. Compared with MARS, COP has a much lower FP value and a slightly higher FN value. SSIR breaks down under Scenario 1.2 because the variance-covariance matrix of X is no longer invertible. In terms of both FP and FN, COP outperformed SCAD under the scenario. One explanation for this comparison result is that SCAD involves non-convex optimization, and can be unstable in implementation.

5.2. Nonlinear multiple index models

In this example, we consider the following multiple index model,

$$Y=\frac{X_1+X_2+\cdots+X_d}{0.5+(1.5+X_2+X_3+X_4)^2}+\sigma\varepsilon, \quad (32)$$

where $X_1, \cdots,, X_p$ are i.i.d. $N(0, 1)$ random variables, ε is $N(0, 1)$ and independent of X, and d and σ are parameters that need to be further specified. This model was originally used in Li (1991) for demonstrating the performance of SIR. It is not difficult to see that given the two projections $X_1 + X_2 + \cdots + X_d$ and $X_2 + X_3 + X_4$, Y and X are independent with each other. The dimensionality of the central subspace of Model (32) is two, and the collection of true predictors is $\{X_1, \cdots, X_d\} \cup \{X_2, X_3, X_4\}$. Because Model (32) is nonlinear, methods that were designed specifically for linear models such as LASSO and SCAD are clearly at disadvantage. Therefore, in this example, we only compare the performances of MARS, SSIR and COP.

By specifying p, d and σ at different values, we have the following three scenarios,

Scenario 2.1: $p = 30$, $d = 3$, $\sigma = 0.1$;

Scenario 2.2: $p = 30$, $d = 3$, $\sigma = 2$;

Scenario 2.3: $p = 400$, $d = 8$, $\sigma = 0.1$.

For each scenario, we generated 100 data sets each with 200 observations (i.e. n=200) and applied MARS, SSIR and COP to each data set. The resulting FP and FN values are reported in Table 2.

For Scenario 2.1, MARS achieved the lowest FN value (0.03), but its FP value was unacceptably high (16.55); SSIR had the lowest FP values, but its FN value was the highest among the three. The FP and FN values of COP were between the extremes. It appears that the performances of SSIR and COP are similar under Scenario 2.1. For Scenario 2.2, COP outperformed SSIR in terms of both FP and FN values. MARS again achieved the lowest FN value (0.32) at the expense of an unacceptable FP value (17.18). Scenario 2.3 is the most challenging among the three scenarios, in which the number of predictors exceeds the number of observations. Both MARS and SSIR broke down under this scenario. However, COP still demonstrated an excellent performance with its FP and FN values reasonably low.

5.3. Heteroscedastic models

In the previous examples, the true predictors affect only the mean response. In this example, we consider the following heteroscedastic model

$$Y=\frac{0.2\varepsilon}{1.5+\sum_{j=1}^{p}\beta_{j,1}X_j}, \quad (33)$$

where $X = (X_1, X_2, \cdots, X_p)'$ follows a p-variate normal distribution with mean zero and covariances $\mathrm{Cov}(X_i, X_j) = \rho^{|i-j|}$ for $1 \le i, j \le p$, ε is independent of X and follows $N(0, 1)$, and $\beta_{j,1} = 1$ for $1 \le j \le 8$ and $=0$ for $j \ge 9$. Note that the central subspace is spanned by $\beta_1 = (\beta_{1,1}, \beta_{2,1}, \ldots, \beta_{p,1})'$ and the number of true predictors is 8. We further specify ρ and p in (33) and consider the following three scenarios,

Scenario 3.1 : $\rho = 0$, $p = 500$;

Scenario 3.2 : $\rho = 0$, $p = 1000$;

Scenario 3.3 : $\rho = 0.3$, $p = 1500$.

For each scenario, we generated 100 data sets each with $n = 1000$ observations and applied MARS, SSIR and COP to the data sets. The FP and FN values of the three methods are listed in Table 3.

Under Scenario 3.1, both SSIR and COP outperformed MARS. The FN value of SSIR (0.99) is less than that of COP (1.21), but the FP value (52.54) is much larger than that of COP (5.71). Under both Scenarios 3.2 and 3.3, in which p is much larger than n, SSIR broke down, but COP still demonstrated excellent performances. The performances of MARS under these two scenarios were fairly poor.

6. Application: Predict Gene Expression from Sequence Using Next-generation Sequencing Data

Embryonic stem cells (ESCs) maintain self-renewal and pluripotency as they have the ability to differentiate into all cell types. To enhance the understanding of the embryonic stem cells development, predictive models, such as regression models, can be constructed in which the gene expression is regarded as the response variable and various features associated with gene-regulating transcription factors (TFs) are taken as the predictors. Examples of such features include motif scores based on position-specific weight matrices of motifs recognized by the TFs Conlon et al. (2003), and ChIP-chip log ratios.

Recently, the emerging next-generation sequencing technologies, in particular, RNA-Seq and ChIP-Seq, offer researchers an unprecedented opportunity to build predictive models for complex biological processes such as gene regulation. Compared to the traditional hybridization-based methods, such as microarray, RNA-Seq and ChIP-Seq provide more accurate quantification of gene expression and TF-DNA binding locations respectively (Mortazavi et al. 2008, Wilhelm et al. 2008, Nagalakshmi et al. 2008, Boyer et al. 2005, Johnson et al. 2007).

To quantify gene expression in RNA-Seq data, one may calculate the RPKM (reads per kilobase of exon region per million mapped reads), which has been shown to be proportional to the gene expression levels (Cloonan et al. 2008). From ChIP-Seq data, Ouyang et al. (2009) proposed a feature named TF association strength (TFAS), which was shown to explain a much higher proportion of gene expression variation than traditional predictors in predictive models. In particular, for each TF, the TFAS for each gene is computed as a weighted sum of the corresponding ChIP-Seq signal strengths, where the weights reflect the proximity of the signal to the gene. We here examine whether we can build a better predictive model for gene expressions by combining both TFASs and motif scores of TF in mouse ESCs.

To achieve this goal, we compiled a data set consisting of gene expressions, TFASs, and motif scores. In this data set, the RPKMs were calculated as gene expression levels from RNA-Seq data in mouse ESCs (Cloonan et al. (2008)). The TFASs of 12 TFs were calculated from the ChIP-Seq experiments in mouse ESCs (Chen et al. (2008)). In addition, we supplement this data set with motif scores of putative TFs of mouse. From the transcription factor database TRANSFAC, we complied a list of 300 TF binding motifs of mouse. For each gene, a matching score was calculated using the scoring system described in Zhong et al. (2005) for each TFBM. The matching score can be considered intuitively as the expected number of occurrences of a TFBM on the gene's promoter region. To build a predictive model in mouse ESC, we treat the gene expression as the response variable and the 12 TFASs as well as the 300 TF motif matching scores as predictors. More precisely, the response is a vector with 12408 entries and the data matrix is a 12408 × 312 matrix with (i, j)th entry representing the TFAS score of of the ith gene's promoter region for TF j if $j \leq 12$; representing the matching score of the ith gene's promoter region for TF j if $j > 12$.

We have applied COP to this data set. The procedure has identified two principal directions and selected in total 42 predictors. The first squared profile correlation is $\lambda_1 = 0.67$, and the second squared profile correlation is $\lambda_2 = 0.20$. Among the 12 TFASs calculated from ChIP-Seq, eight of them were selected by COP. In particular, Oct4 is a well-known master regulator regulating the pluripotency, and Klf4 regulates differentiation (Cai et al. 2010). Evidence also suggests that at these early stages of development, STAT3 activation is required for self-renewal of ESCs (Matsuda et al. 1999). Among the 300 TF motif scores, 34 of them are selected by COP. To further understand what extra information that TF motif scores provide, we annotate the functions of the 34 TFs. It is of interest to note that 24 out of the 34 selected motifs correspond to TFs that are either regulators for development or cancer-related; see Table 4. Since ESCs are in a developmental phase, it is not surprising to have active TFs regulating general development. Some recent evidences suggest that tumor suppressors that control cancer cell proliferation also regulate stem cell self-renewal (Pardal et al. 2005). Thus, a careful study of these cancer-related TFs could lead to a better understanding of the stem cell regulatory network.

7. Discussion

The contribution of the COP procedure to the development of variable selection methodologies for high dimensional regression analysis is two-fold. First, it dos not impose any assumption on the relationship between the response variable and the predictors, and the sufficient dimension reduction framework that the COP procedure relies on includes fully nonparametric models as special cases. Therefore, COP can be considered a model-free variable selection procedure applicable in any high dimensional data analysis. Second, as demonstrated by our simulation studies, the COP procedure can effectively handle hundreds and thousands of predictors, which can be extremely challenging to other existing methods for variable selection beyond linear or parametric models. Like linear stepwise regression, the COP procedure may encounter issues typical to stepwise procedures as discussed in Miller (1984). Nonetheless, we believe that the COP procedure should become an indispensable member of the repository of variable selection tools and recommend its broad use. When a parametric model is postulated for the relationship between the response and the predictor variables and model-specific variable selection methods are available, we recommend to use COP together with these methods as a safeguard against possible model-misidentification. We have implemented the COP procedure in R, and the R package can be downloaded from http://cran.r-project.org/web/packages/COP/ or requested from the authors directly.

As a trade-off, the COP procedure imposes various assumptions on the distribution of the predictors, of which the linearity assumption is the most fundamental and crucial. When the linearity condition is required to hold for any lower dimensional projection, it is equivalent to requiring that the joint distribution of the predictors is elliptically-contoured (Eaton 1986). Hall and Li (1993) establishes the fact that low dimensional projections from high dimensional data approximately satisfy the linearity condition, which to a certain degree alleviates the concern of the linearity assumption and explains why SIR and the COP procedure worked well under mild violation of the assumption. When the linearity condition is heavily violated, data re-weighting schemes such as the Voronoi re-weighting scheme (Cook and Nachtsheim 1994) can be used to correct the violation. We plan to incorporate such schemes into the COP procedure in the future.

When the number of the predictors is extremely large, the performance of the COP procedure can be compromised. This is also the case for variable selection methods under the linear model. Lately, Fan and Lv (2008) advocates a two-step approach to attack the so-called ultra-high dimensionality. The first step is to perform screening to reduce the dimensionality from ultra-high to high or moderately high, and then in the second step, variable selection methods are applied to identify the true predictors. The same approach can be used for variable selection under the SDR framework. More precisely, we can apply the forward COP (FCOP) precedure, which is simply the COP procedure with the deletion step removed, to reduce the dimensionality of a problem from ultra-high to moderely high. The FCOP procedure is much easier to implement and computationally more efficient than the COP procedure. Then, the usual COP procedure is applied to the reduced data to select the true predictors. This approach is currently under investigation and the results will be reported in a future publication.

Acknowledgments

This work was supported by a NIH grant U01 ES016011, a DOE grant from the office of Science (BER) and an NSF grant DMS 1120256 to WZ, a NIH grant R01-HG02518-02 and a NSF grant DMS 1007762 to JL, a NSF grant DMS 0707004 to YZ.

Appendix

A.1 PROOF OF PROPOSITION 2.1

Let $\mathscr{S}^{\perp}(B)$ denote the space of vectors such that for any $\rho \in \mathscr{S}^{\perp}(B)$ and any $\beta \in \mathscr{S}B)$, $\rho'\Sigma\beta = 0$. Let $\mathscr{S}^{\perp}(\tilde{K})$ be the space of vectors such that for any $\rho \in \mathscr{S}^{\perp}(\tilde{K},)$ $\rho'\Sigma\eta_k = 0$ for $k = 1, \cdots, \tilde{K}$. We will show that $\mathscr{S}^{\perp}(B) \subseteq \mathscr{S}^{\perp}(\tilde{K})$, which means, for any $\rho \in \mathscr{S}^{\perp}(B)$ $P(\rho) = 0$. First, because for any T, $T(Y) \perp \eta' X \mid B' X$, then $\mathrm{cov}(T(Y), \eta' X) = E(T(Y)\eta' X) = E(E(T(Y)|B$

$'X)E(\eta' X|B' X))$. Due to the linearity condition, for any $\rho \in \mathscr{S}^{\perp}(B)$, $E(\rho' X|B' X)=c_1\beta_1' X+\cdots+c_K\beta_K' X$, where $c_1, \cdots, c_K$ are linear coefficients. In addition, since $\mathrm{cov}(\rho' X, \beta_k' X)=0$ for $k = 1, \cdots, K$, $E(\rho' X|B' X) = 0$. Consequently, $corr^2(T(Y), \rho' X)=\frac{\mathrm{cov}^2(T(Y), \rho' X)}{\mathrm{var}(T(Y))\mathrm{var}(\rho' X)}=0$, $P(\rho) = 0$ and $\mathscr{S}^{\perp}(B) \subseteq \mathscr{S}^{\perp}(\tilde{K})$. Proposition 2.1 holds.

A.1 PROOF OF THEOREM 3.1

Without loss of generality, we let $\mathscr{A} = \{1, \cdots, d\}$ and $t = d + 1$. Let $X^{(j)}$ be the vector of n i.i.d observations of the jth variable for $j = 1, \cdots, d + 1$. We assume that the predictors have been centered to have zero sample mean. Denote $\mathbf{X}_{n\times j} = (X^{(1)}, \cdots, X^{(j)})$ for $j = d, d + 1$. We let

$$\hat{\mathrm{M}}^{(j)}=\sum_{h=1}^{H}\frac{n_h}{n}(\bar{\mathbf{X}}_h^{(j)})^T\bar{\mathbf{X}}_h^{(j)} \quad for\ j=d, d+1$$

where $\bar{\mathbf{X}}_h^{(j)}$ $(j=d, d+1)$ is the average of the first j variables for those individuals whose responses fall into the hth slice S_h, $h = 1, \cdots, H$. Let n_h be the number of observations in the hth slice, $h = 1, \cdots, H$. Let $\hat{\lambda}_i^{(j)}$ be the ith largest eigenvalue of $\hat{\Sigma}_j^{-1}\hat{\mathrm{M}}^{(j)}$ for $j = d, d + 1$ respectively, where $\hat{\Sigma}_j$ is the sample variance-covariance matrix of $\mathbf{X}_{n\times j}$. It is difficult to see the asymptotic distribution of $\hat{\lambda}_i^{(d+1)} - \hat{\lambda}_i^{(d)}$ for $i = 1, \cdots, K$ directly based on $\hat{\Sigma}_j^{-1}\hat{\mathrm{M}}^{(j)}$ for $j = d$, $d + 1$. We did some transformations such that the transformed $\hat{\Sigma}_j^{-1}\hat{\mathrm{M}}^{(d)}$ (with eigenvalues unchanged) is a sub-matrix of the transformed $\hat{\Sigma}_j^{-1}\hat{\mathrm{M}}^{(d+1)}$.

Let

$$\hat{\gamma}_{n\times 1}=(\hat{\gamma}_1, \cdots, \hat{\gamma}_n)^T=\frac{1}{\hat{\sigma}}[I-\mathbf{X}_{n\times d}(\mathbf{X}_{n\times d}^T\mathbf{X}_{n\times d})^{-1}\mathbf{X}_{n\times d}^T]X^{(d+1)},$$

where $\hat{\sigma}^2$ is the sample variance of $[I-\mathbf{X}_{n\times d}(\mathbf{X}_{n\times d}^T\mathbf{X}_{n\times d})^{-1}\mathbf{X}_{n\times d}^T]X^{(d+1)}$. Denote $\bar{\gamma}_h=n_h^{-1}\sum_{y_i\in S_h}\hat{\gamma}_i$. Let $\gamma = X_{d+1} - E(X_{d+1}|X_1, \cdots, X_d)$, and $\gamma_{n\times 1}$ be the n regression error terms of the n observed X_{d+1} on $X_1, \cdots, X_d$. Then $\gamma_{n\times 1}$ are i.i.d with mean zero and a finite variance. Under the null hypothesis $H_0 : \eta_{d+1,i} = 0$, $i = 1, \cdots, K$ we have $\mathrm{E}(\gamma|y) = E(E(\gamma|X_1, \cdots, X_d)|y) = 0$ for any y. Let $\bar{\gamma}$ be the mean of $\gamma_{n\times 1}$. Then

$$\hat{\gamma}_{n\times 1}=(\hat{\gamma}_1, \cdots, \hat{\gamma}_n)^T=\frac{1}{\hat{\sigma}}[I-\mathbf{X}_{n\times d}(\mathbf{X}_{n\times d}^T\mathbf{X}_{n\times d})^{-1}\mathbf{X}_{n\times d}^T](\gamma_{n\times 1}-\bar{\gamma}).$$

With transformations on $\hat{\Sigma}_j^{-1}\hat{\mathrm{M}}^{(d)}$, we showed that $\hat{\lambda}_i^{(d+1)} - \hat{\lambda}_i^{(d)}$ for $i = 1, \cdots, K$ equals a squared linear combination of $\bar{\gamma}_h$. Thus, we just need to show that $(\bar{\gamma}_1, \cdots, \bar{\gamma}_H)$ converges to a multivariate normal distribution, and we complete the proof. Let $(z_1, \cdots, z_d)'=\Sigma_d^{-/2}(x_1, \cdots, x_d)'$. Define four matrices, $A_{H\times H}$, $B_{H\times d}$, $E_{d\times d}$ and $\Gamma_{H\times d}$, where $A_{H\times H}=\mathrm{diag}\{var(\gamma|y \in S_1), \cdots, var(\gamma|y \in S_H)\}/\sigma^2$; the (h, j)th entry of $B_{H\times d}$ is

$\sqrt{p_h} cov(z_j\gamma, \gamma|y \in S_h)/\sigma^2$, the (j, j')th entry of $E_{d\times d}$ equals $cov(z_{j'}\gamma, z_j\gamma)/\sigma^2$, the $(h, j)th$ entry of $\Gamma_{H\times d}$ is $\sqrt{p_h} E(z_j|y \in S_h)$ and $\sigma^2 = \lim_{n\to} \hat{\sigma}^2 = Var(\gamma)$. Let Υ be a d by d matrix and

$$\Upsilon=\Gamma^T_{H\times d}A_{H\times H}\Gamma_{H\times d}-\Gamma^T_{H\times d}B_{H\times d}\Gamma^T_{H\times d}\Gamma_{H\times d}-\Gamma^T_{H\times d}\Gamma_{H\times d}B^T_{H\times d}\Gamma_{H\times d}+\Gamma^T_{H\times d}\Gamma_{H\times d}E_{d\times d}\Gamma^T_{H\times d}\Gamma_{H\times d}.$$

Define $\tilde{Q}$ to be a $d\times K$ matrix with jth column as $q_j/(\lambda_j^{(d)}(1-\lambda_j^{(d)}))^{1/2}$, where q_j is the the jth eigenvector of the limiting matrix $\lim_{n\to\infty}\hat{\Sigma}_d^{-1/2}\hat{\mathrm{M}}^{(d)}\hat{\Sigma}_d^{-1/2},\ and\ \lambda_j^{(d)}=\lim_{n\to\infty}\hat{\lambda}_j^{(d)}$. Then $W_{Kt} = \tilde{Q}^T\Upsilon\tilde{Q}$.

A.1 PROOF OF Corollary 3.1

With an additional condition that $E(\gamma^2|X_1, \cdots, X_d)$ is constant, we can show that the asymptotic variance matrix of $(\bar{\gamma}_1, \cdots, \bar{\gamma}_H)$ adopts a special form, with which the asymptotic standard chi-quare distribution can be derived.

A.3 PROOF OF THEOREM 3.2

Without loss of generality, we let $\mathscr{A} = \{X_1, \cdots, X_d\}$. Following the notations used in Theorem 3.1, let $\gamma_j = X_j - E(X_j|X_i, i \in \mathscr{A}$ for $j \in \mathscr{A}^c$, and

$$\hat{\gamma}^j=(\hat{\gamma}_{j,1},\cdots,\hat{\gamma}_{j,n})'=\frac{1}{\hat{\sigma}_j}[\mathbf{I}_n-\mathbf{X}_{n\times d}((\mathbf{X}_{n\times d})'\mathbf{X}_{n\times d})^{-1}(\mathbf{X}_{n\times d})']X^{(j)}.$$

Let $\bar{\gamma}_h^j=n_h^{-1}\sum_{y_i\in S_h}\hat{\gamma}_{j,i}$. Similar as in the proof of Theorem 3.1, we basically show $\bar{\gamma}_h^j$ for $j = d + 1, \cdots, p$ and $h = 1, \cdots, H$ converge to a multivariate normal distribution.

A.3 PROOF OF THEOREM 3.3

We use the same notations defined in the proof of Theorem 3.2. Let $\bar{\gamma}_h^j=n_h^{-1}\sum_{y_i\in S_h}\hat{\gamma}_{j,i}$. Let $\hat{\lambda}_k^{(d)}$ be defined as in the proof of Theorem 3.1. First, for any t, $COP_{1:K}^{\mathscr{A}+t} \geq n(\sum_{k=1}^K\hat{\lambda}_k^{(d+1)} - \sum_{k=1}^K\hat{\lambda}_k^{(d)})$, and

$$|\sum_{k=1}^K\hat{\lambda}_k^{(d+1)}-\sum_{k=1}^K\hat{\lambda}_k^{(d)}|\geq|\sum_{k=1}^K(\lambda_k^{(d+1)}-\lambda_k^{(d)})|-|\sum_{k=1}^K(\lambda_k^{(d+1)}-\hat{\lambda}_k^{(d+1)})|-|\sum_{k=1}^K(\lambda_k^{(d)}-\hat{\lambda}_k^{(d)})|.$$

Since X follows a multivariate normal distribution, from Li (1991), $\lambda_k^{(d)}=\lambda_k^{(d+1)}=0$ for $k > K$, then

$$\sum_{k=1}^K(\lambda_k^{(d+1)}-\lambda_k^{(d)})=\lim_{n\to\infty} trace(\tilde{\Omega}^{(d+1)})-\mathrm{trace}(\hat{\Omega}^{(d)})=\lim_{n\to\infty}\sum_{h=1}^H\frac{n_h}{n}(\bar{\gamma}_h^j)^2=\sum_{h=1}^Hp_h\frac{E^2(\gamma_j|y\in S_h)}{\sigma_j^2}.$$

We need to use the two Lemmas 8.1 and 8.2 stated below. The proofs of the two lemmas are omitted here. From Lemma 8.1,

$$\max_{t\in\mathscr{A}^c\cap\mathscr{T}} n(\sum_{k=1}^{K}\hat{\lambda}_k^{(d+1)}-\sum_{k=1}^{K}\hat{\lambda}_k^{(d)}) \geq \varpi\cdot\omega_H\cdot n^{1-\xi_0}\frac{\tau_{\min}^2}{\tau_{\max}}-|\sum_{k=1}^{K}n(\lambda_k^{(d+1)}-\hat{\lambda}_k^{(d+1)})|-|\sum_{k=1}^{K}n(\lambda_k^{(d)}-\hat{\lambda}_k^{(d)})|.$$

Then as long as

$$\max_{\mathscr{A}\subseteq\{1,\cdots,p\}} n|\sum_{k=1}^{K}(\lambda_k^{(d)}-\hat{\lambda}_k^{(d)})| \leq \vartheta n^{1-\xi_0}/2,$$

we have

$$\min_{\mathscr{A}:\mathscr{A}^c\cap\mathscr{T}\neq\varnothing}\ \max_{t\in\mathscr{A}^c\cap\mathscr{T}} COP_{1:K}^{\mathscr{A}+t} \geq \vartheta n^{1-\xi_0}.$$

From Lemma 8.2,

$$P(\max_{\mathscr{A}\subseteq\{1,\cdots,p\}}|\sum_{k=1}^{K}\lambda_k^{(d)}-\sum_{k=1}^{K}\hat{\lambda}_k^{(d)}|>\vartheta n^{-\xi_0}/2) \leq 2Kp(p+1)C_1\ exp\{-C_2 n^{1-2\xi_0}\frac{\tau_{\min}^2\vartheta^2}{256K^2p^2}\}.$$

Under Condition 8, since $p = o(n^{\varrho_0})$ with $2\varrho_0 + 2\xi_0 < 1$, $P(\max_{\mathscr{A}\subseteq\{1,\cdots,p\}}|\sum_{k=1}^{K}\lambda_k^{(d)}-\sum_{k=1}^{K}\hat{\lambda}_k^{(d)}|>\vartheta n^{-\xi_0}/2) \to 0,\ and\ P(\min_{\mathscr{A}:\mathscr{A}^c\cap\mathscr{T}\neq\varnothing}\max_{t\in\mathscr{A}^c\cap\mathscr{T}} COP_{1:K}^{\mathscr{A}+t} \geq \vartheta n^{1-\xi_0}) \to 1$.

A.3 PROOF OF THEOREM 3.4

Since

$$|\sum_{k=1}^{K}\hat{\lambda}_k^{(d+1)}-\sum_{k=1}^{K}\hat{\lambda}_k^{(d)}| \leq |\sum_{k=1}^{K}(\lambda_k^{(d+1)}-\lambda_k^{(d)})|+|\sum_{k=1}^{K}(\lambda_k^{(d+1)}-\hat{\lambda}_k^{(d+1)})|+|\sum_{k=1}^{K}(\lambda_k^{(d)}-\hat{\lambda}_k^{(d)})|,$$

and with $\mathscr{T}\subseteq\mathscr{A}$ $|\sum_{k=1}^{K}(\lambda_k^{(d+1)}-\lambda_k^{(d)})|=0$, then from Lemma 8.2, $P(\max_{\mathscr{A}\subseteq\{1,\cdots,p\}}|\sum_{k=1}^{K}\lambda_k^{(d)}-\sum_{k=1}^{K}\hat{\lambda}_k^{(d)}|>\varepsilon) \to 0$ for $\varepsilon > Cn^{\varrho_0-1/2}$ and Theorem 3.4 holds.

Lemma 8.1. *Under the same conditions as in Theorem 3.3, for any* $\mathscr{A}\subseteq\{1,\cdots,p\}$ *and* $\mathscr{A}^c\cap\mathscr{T}\neq\varnothing$,

$$\max_{j\in\mathscr{A}^c\cap\mathscr{T}}\sum_{h=1}^{H}p_h E^2(\gamma_j|y\in S_h)/\sigma_j^2 \geq \tau_{\min}^2\cdot\varpi\cdot\omega_H\cdot n^{-\xi_0}/\tau_{\max}>0.$$

Lemma 8.2. *Under the same conditions as in Lemma 8.1,*

$$P(\max_{\mathscr{A}\subseteq\{1,\cdots,p\}}|\sum_{k=1}^{K}\lambda_k^{(d)}-\sum_{k=1}^{K}\hat{\lambda}_k^{(d)}|>\varepsilon)\le 2Kp(p+1)C_1\ exp\{-C_2n\frac{\tau_{\min}^2\varepsilon^2}{64K^2p^2}\}.$$

A.3 PROOF OF THEOREM 4.1

For coherence, we use the same notation as defined in the proof of Theorem 3.1. Without loss of generality, let $\mathscr{A} = \{1, \cdots, d\}$ and $t = d + 1$. Under the assumption that $\mathbf{X}_{n\times(d+1)}$ has a multivariate normal distribution, we derive the limiting value of $(\sum_{k=1}^{K}\hat{\lambda}_k^{(d+1)}-\sum_{k=1}^{K}\hat{\lambda}_k^{(d)})$ as $n \to \infty$ for fixed slices. Let $\Xi_{K\times K}$ be the variance-covariance matrix of $\mathbf{v}_K$.

Because $\{X_1, \cdots, X_{d+1}\}$ follow a multivariate normal distribution, we have $\gamma=X_{d+1}-(\rho_0+\sum_{i=1}^{d}\rho_iX_i)\ and\ \gamma\sim N(0,\sigma_{d+1}^2)$ where the ρ_i are the coefficients. Since we assume that the response only depends on K linear combinations of $\mathbf{X}_{n\times(d+1)}$, $\tilde{\Omega}^{(d+1)}$ and $\hat{\Omega}^{(d)}$ have at most K nonzero eigenvalues, and

$$\frac{\sum_{k=1}^{K}\hat{\lambda}_k^{(d+1)}}{\mathrm{trace}(\tilde{\Omega}^{(d+1)})}\xrightarrow{P}1,\ \frac{\sum_{k=1}^{K}\hat{\lambda}_k^{(d)}}{\mathrm{trace}(\hat{\Omega}^{(d)})}\xrightarrow{P}1,\ \frac{\sum_{k=1}^{K}\hat{\lambda}_k^{(d+1)}-\sum_{k=1}^{K}\hat{\lambda}_k^{(d)}}{\mathrm{trace}(\tilde{\Omega}^{(d+1)})-\mathrm{trace}(\hat{\Omega}^{(d)})}\xrightarrow{P}1.$$

Because of the following three results:

1. $\mathrm{trace}(\tilde{\Omega}^{(d+1)})-\mathrm{trace}(\hat{\Omega}^{(d)})=\sum_{h=1}^{H}n_h(\bar{\gamma}_h)^2/n.$
2. $\bar{\gamma}_h\xrightarrow{P}E(\gamma|y\in S_h)/\sigma_{d+1}$, $h = 1, \cdots, H.$
3. Since $E(\gamma|\mathbf{v}_K)=\tilde{\eta}_{t,\mathscr{A}}\Xi_{K\times K}^{-1}\mathbf{v}_K$, then

$$E(\gamma|y\in S_h)=E\,[E(\gamma|\mathbf{v}_K)|y\in S_h]=\tilde{\eta}_{t,\mathscr{A}}^{'}\Xi_{K\times K}^{-1}\mathbf{L}_{H,K}.$$

Combining 1,2,3, we have $\sum_{h=1}^{H}n_h(\bar{\gamma}_h)^2/n\xrightarrow{P}\tilde{\eta}_{t,\mathscr{A}}^{'}\Xi_{K\times K}^{-1}M_{H,K}\Xi_{K\times K}^{-1}\tilde{\eta}_{t,\mathscr{A}}$. Since $\Xi_{K\times K}^{-1}=I_{K\times K}$,

$$\sum_{k=1}^{K}\hat{\lambda}_k^{\mathscr{A}+t}-\sum_{k=1}^{K}\hat{\lambda}_k^{\mathscr{A}}\xrightarrow{P}\frac{1}{\sigma_{d+1}^2}\tilde{\eta}_{t,\mathscr{A}}^{'}M_{H,K}\tilde{\eta}_{t,\mathscr{A}},$$

and Theorem 4.1 holds.

A.3 PROOF OF Proposition 4.1

Note that $\eta' M_{H,K}\eta$ = Var $(E(\eta'\mathbf{v}_K|y\in S_h))$ and

$$Var\ (E(\eta'\mathbf{v}_K|y\in S_{h'}^{'}))=Var\ (E(\eta'\mathbf{v}_K|y\in S_h))+Var\ [Var\ (E(\eta'\mathbf{v}_K|y\in S_{h'}^{'})|S_h)].$$

Thus, Proposition 4.1 holds.

References

Bondell H, Li L. Shrinkage inverse regression estimation for model-free variable selection. J. Roy. Statist. Soc. Ser. B. 2009; 71(1):287–299.

Boyer LA, Lee TI, Cole MF, Johnstone SE, Levine SS, Zucker JP, Guenther MG, Kumar RM, Murray HL, Jenner RG, Gifford DK, Melton DA, Jaenisch R, Young RA. Core transcriptional regulatory circuitry in human embryonic stem cells. Cell. 2005; 122(6):947–956. [PubMed: 16153702]

Cai J, Xie D, Fan Z, Chipperfield H, Marden J, Wong WH, Zhong S. Modeling co-expression across species for complex traits: insights to the difference of human and mouse embryonic stem cells. PLoS Comput. Biol. 2010; 6 e1000707.

Chen C-H, Li K-C. Can SIR be as popular as multiple linear regression? Statistica Sinica. 1998; 8:289–316.

Chen X, Xu H, Yuan P, Fang F, M H, Vega VB, Wong E, Orlov YL, Zhang W, Jiang J, Loh YH, Yeo HC, Yeo ZX, Narang V, Govindarajan KR, Leong B, Shahab A, Ruan Y, Bourque G, Sung WK, Clarke ND, Wei CL, Ng HH. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell. 2008; 133:1106–1117. [PubMed: 18555785]

Cleveland W. Robust locally weighted regression and smoothing scatterplots. J. Amer. Statist. Assoc. 1979; 74(368):829–836.

Cloonan N, Forrest ARR, Kolle G, Gardiner BBA, Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G, Robertson AJ, Perkins AC, Bruce SJ, Lee CC, Ranade SS, Peckham HE, Manning JM, McKernan KJ, Grimmond SM. Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nature Methods. 2008; 5(7):613–619. [PubMed: 18516046]

Conlon E, Liu X, Lieb J, Liu J. Integrating regulatory motif discovery and genome-wide expression analysis. Proceedings of the National Academy of Sciences. 2003; 100(6):3339–3344.

Cook R. Testing predictor contributions in sufficent dimension reduction. Annals of Statistics. 2004; 32(3):1062–1092.

Cook, RD. An Introduction to Regression Graphics. New York: Wiley; 1994.

Cook RD, Nachtsheim CJ. Reweighting to achieve elliptically contoured covariates in regression. J. Amer. Statist. Assoc. 1994; 89:592–599.

Eaton ML. A characterization of spherical distributions. Journal of Multivariate Analysis. 1986; 20:272–276.

Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. The Annals of Statistics. 2004; 32(2):407–499.

Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc. 2001; 96(456):1348–1360.

Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. J. Roy. Statist. Soc. Ser. B. 2008; 70:849–911.

Fan J, Lv J. A selective overview of variable selection in high dimensional feature space. Statistica Sinica. 2010; 20:101–148. [PubMed: 21572976]

Friedman JH. Pathwise coordinate optimization. Annals of applied statistics. 2007; 1(2):302–332.

Friedman JH, Tukey JW. A projection pursuit algorithm for explanatory data analysis. IEEE Trans. Comput. 1974; C-23:881–889.

Fung W, He X, Liu L, Shi P. Dimension reduction based on canonical correlation. Statistica Sinica. 2002; 12(4):1093–1114.

Hall P, Li K-C. On almost linearity of low dimensional projections from high dimensional data. The Annals of Statistics. 1993; 21:867–889.

Huber PJ. Projection pursuit. The Annals of Statistics. 1985; 13:435–475.

Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. Science. 2007; 316(5830):1497–1502. [PubMed: 17540862]

Li K-C. Sliced inverse regression for dimension reduction. J. Amer. Statist. Assoc. 1991; 86:316–327.

Li L. Sparse sufficient dimension reduction. Biometrika. 2007; 94(3):603–613.

Li L, Cook R, Nachtsheim C. Model-free variable selection. J. Roy. Statist. Soc. Ser. B. 2005; 67:285–299.

Matsuda T, Nakamura T, Nakao K, Arai T, Katsuki M, Heike T, Yokota T. STAT3 activation is sufficient to maintain an undifferentiated state of mouse embryonic stem cells. EMBO J. 1999; 18(15):4261–4269. [PubMed: 10428964]

Miller AJ. Selection of subsets of regression variables. J. Roy. Statist. Soc. Ser. A. 1984; 147(3):389–425.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods. 2008; 5(7):621–628. [PubMed: 18516045]

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science. 2008; 320(5881):1344–1349. [PubMed: 18451266]

Ouyang Z, Zhou Q, Wong WH. Chip-seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. Proc. Natl. Acad. Sci. USA. 2009; 106:21521–21526. [PubMed: 19995984]

Pardal R, Molofsky AV, He S, Morrison SJ. Stem cell self-renewal and cancer cell proliferation are regulated by common networks that balance the activation of proto-oncogenes and tumor suppressors. Cold Spring Harb Symp Quant Biol. 2005; 70:177–185. [PubMed: 16869752]

Shao J. An asymptotic theory for linear model selection (with discussion). Statistica Sinica. 1998; 7:221–264.

Tibshirani R. Regression shrinkage and selection via the lasso. J. Roy. Statist. Soc. Ser. B. 1996; 58:267–288.

Wang, h. Forward regression for ultra-high dimensional variable screening. J. Amer. Statist. Assoc. 2009; 104:1512–1524.

Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, Penkett CJ, Rogers J, Bahler J. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature. 2008; 453(7199):1239–U39. [PubMed: 18488015]

Zeng P, Zhu Y. An integral transform method for estimating the central mean and central subspaces. Journal of Multivariate Analysis. 2010; 101(1):271–290.

Zhong W, Zeng P, Ma P, Liu J, Zhu Y. Regularized sliced inverse regression for motif discovery. Bioinformatics. 2005; 21(22):4169–4175. [PubMed: 16166098]

Zhou J, He X. Dimension reduction based on constrained cannonical correlation and variable filtering. Ann. Statist. 2008; 36(4):1649–1668.

Zhu L, Miao B, Peng H. On sliced inverse regression with high-dimensional covariates. J. Amer. Statist. Assoc. 2006; 101:630–643.

Zou H. The adaptive lasso and its oracle properties. J. Amer. Statist. Assoc. 2006; 101(476):1418–1429.

Table 1

Performance comparison under linear models: FP is the average number of irrelevant variables falsely selected by the method, and FN is the average number of true variables falsely excluded by the method; the number in (·) is the standard error of the FP or FN and NA* indicates that the corresponding algorithm broke down.

Methods	***p*** **= 8, *n* = 40, σ = 3, ρ = 0.5**		***p*** **= 1000, *n* = 200, σ = 1, ρ = 0.5**	
	FP **(0, 5)**	**FN** **(0, 3)**	**FP** **(0,990)**	**FN** **(0,10)**
LASSO	0.77(0.093)	0.16(0.037)	8.87(0.586)	0.00(0.000)
SCAD	0.67(0.094)	0.10(0.030)	6.05(0.926)	1.16(0.150)
MARS	4.00(0.059)	0.04(0.020)	30.64(0.165)	0.00(0.000)
SSIR	0.19(0.051)	0.96(0.068)	NA*	NA*
COP	0.71(0.080)	0.56(0.066)	2.28(0.203)	0.75(0.095)

Table 2

Performance comparison under multiple index model: FP is the average number of irrelevant variables falsely selected by the method, and FN is the average number of true variables falsely excluded by the method; the number in (·) is the standard error of the FP or FN and NA* indicates that the corresponding algorithm broke down.

Methods	$\sigma = 0.1, p = 30, d = 3$		$\sigma = 2, p = 30, d = 3$		$\sigma = 0.1, p = 400, d = 8$	
	FP (0, 26)	FN (0, 4)	FP (0, 26)	FN (0–4)	FP (0, 292)	FN (0, 8)
MARS	16.55(0.174)	0.03(0.017)	17.18(0.186)	0.32(0.053)	NA*	NA*
SSIR	0.12(0.033)	0.91(0.029)	4.14(0.288)	1.76(0.115)	NA*	NA*
COP	1.88(0.149)	0.83(0.038)	3.26(0.210)	1.71(0.104)	8.93(0.576)	0.18(0.081)

Table 3

Performance comparison under heteroscedastic model: FP is the average number of irrelevant variables falsely selected by the method, and FN is the average number of true variables falsely excluded by the method; the number in (·) is the standard error of the FP or FN and NA* indicates that the corresponding algorithm broke down.

Methods	$\rho = 0$ $n = 1000, p = 500$		$\rho = 0$ $n = 1000, p = 1000$		$\rho = 0.3$ $n = 1000, p = 1500$	
	FP **(0, 492)**	**FN** **(0, 8)**	**FP** **(0, 992)**	**FN** **(0, 8)**	**FP** **(0, 1492)**	**FN** **(0, 8)**
MARS	212.15(0.428)	4.83(0.116)	230.33(0.372)	6.16(0.129)	236.60(0.524)	6.84(0.126)
SSIR	52.54(1.970)	0.88 (0.149)	NA*	NA*	NA*	NA*
COP	5.79(0.365)	1.21(0.030)	13.14(0.734)	1.29(0.037)	21.36(0.937)	1.5(0.039)

Table 4

Motifs identified

development	COUP-TF, AP2, Sp1, CHOP C/EBpalpha, NF-AT Pax, Pax8, GABP, En1, TTF1 PITX2, NKx2-2, HIXA4, ZF5, PPAR direct repeat 1
cancer	IRF1, EVI1, NF1, GKLF, Whn VDR, POU6F1, Arnt, Cdx2
8 selected TFAS	E2F1, Mycn, ZFx, Klf4 Tcfcp2/1, Oct4, Stat3, Smad1