Summer 2018

New developments of dimension reduction

Lei Huo

Department: Mathematics and Statistics

NEW DEVELOPMENTS OF DIMENSION REDUCTION

by

LEI HUO

A DISSERTATION

Presented to the Graduate Faculty of the

MISSOURI UNIVERSITY OF SCIENCE AND TECHNOLOGY

In Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

in

MATHEMATICS WITH A STATISTICS EMPHASIS

2018

Approved by

Dr. Xuerong Wen, Advisor
Dr. V.A. Samaranayake
Dr. Akim Adekpedjou
Dr. Gayla R. Olbricht
Dr. Wei Jiang

PUBLICATION DISSERTATION OPTION

This dissertation consists of the following two articles which have been accepted or submitted for publication, as follows:

Paper I: Pages 9-37 have been published by Journal of Nonparametric Statistics.

Paper II: Pages 38-63 have been submitted to Journal of Statistical Planning and Inference.

ABSTRACT

Variable selection becomes more crucial than before, since high dimensional data are frequently seen in many research areas. Many model-based variable selection methods have been developed. However, the performance might be poor when the model is misspecified. Sufficient dimension reduction (SDR, Li 1991; Cook 1998) provides a general framework for model-free variable selection methods.

In this thesis, we first propose a novel model-free variable selection method to deal with multi-population data by incorporating the grouping information. Theoretical properties of our proposed method are also presented. Simulation studies show that our new method significantly improves the selection performance compared with those ignoring the grouping information. In the second part of this dissertation, we apply partial SDR method to conduct conditional model-free variable (feature) screening for ultra-high dimensional data, when researchers have prior information regarding the importance of certain predictors based on experience or previous investigations. Comparing to the state of art conditional screening method, conditional sure independence screening (CSIS; Barut, Fan and Verhasselt, 2016), our method greatly outperforms CSIS for nonlinear models. The sure screening consistency property of our proposed method is also established.

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Xuerong Meggie Wen, for her guidance and assistance in my research. I would also like to thank Dr. Zhou Yu for his advice in my research. Also I want to extend my gratitude to Dr. V. A. Samaranayake for serving on my committee and his great help in and out of the classroom. I would like to thank Dr. Robert Paige for the courses he provided which I gained a lot from. I also greatly appreciate Dr. Wei Jiang's advice on my comprehensive exam and dissertation. I want to express my appreciation towards Dr. Akim Adekpedjou and Dr. Gayla Olbricht for serving on my committee.

Finally, I am thankful of the support that my family and friends have provided throughout my pursuit of the doctorate degree. Especially, I want to thank my wife Yuqing Su for being supportive all the time and for her unconditional love. Also I want to thank my parents and parents-in-law for their great love and help in my life.

TABLE OF CONTENTS

Page

SECTION

LIST OF TABLES

Table Page

PAPER I

PAPER II

SECTION

1. INTRODUCTION

1.1. HIGH DIMENSIONAL DATA

In many statistical applications, researchers need to extract important information from high dimensional data, where the dimension p is much larger than the sample size n. The problems are frequently seen in genomics, biomedical imaging, functional MRI, tomography, tumor classifications, signal processing, image analysis, and finance. For instance, researchers in biomedical area often need to use microarrays or proteomics datasets, which consist of only several hundred samples but with thousands of genes, to do tumor classification or to predict certain clinical prognosis such as injury scores and survival time. Tremendous amount of new financial products have been created, as a new era of financial markets have been introduced by the development of technology and trade globalization. High dimensional statistical problems frequently arise in estimating the covariance matrices of the returns of assets during optimizing the performance of a portfolio. Statistical analysis of high dimensional data is generally acknowledged as an important challenge to traditional statistics. There is little doubt that high dimensional data analysis will be the most important topic of statistics in the 21st century. Please refer to Donoho (2000) and Fan and Li (2007) for overviews of statistical challenges with high dimensional data.

Dimension reduction is fundamental to information extraction from high dimensional data. It has two different branches: feature extraction and variable selection. To reduce the dimension, feature extraction generates new features or variables by combining the original ones. It is preferable in applications such as image analysis, signal processing, and information retrieval, where model accuracy is more important than model interpretability

(Boln-Canedo *et al.*, 2015). Variable selection achieves dimension reduction by identifying significant ones from all variables. It is frequently applied in text mining, genetics analysis, sensor data processing and so on, where the original variables are important for model interpretation and information extraction (Boln-Canedo *et al.*, 2015). In this dissertation, we will focus on variable selection.

1.2. VARIABLE SELECTION

A huge amount of variable selection procedures have been proposed in literature. However, traditional variable selection procedures such as C_p, AIC and BIC are infeasible for high dimensional data because of the expensive computational costs. Innovative variable selection procedures are needed for high dimensional data analysis. Many methods have been developed in recent years to extract the important variables effectively from high dimensional data. Tibshirani (1996) proposed the least absolute shrinkage and selection operator (LASSO), which is an l_1 penalized least squared method, for linear models. It minimizes the residual sum of squares with the sum of the absolute values of the coefficients less than a constant. Certain coefficients are forced to be set to zero through this procedure. Many variants of LASSO have been proposed to make it more useful in different applications such as adaptive LASSO (Zou, 2006) and group LASSO (Yuan and Lin, 2006). Fan and Li (2001) proposed penalized likelihood approach which can be applied to generalized linear models. Many variable selection procedures such as the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li, 2001) and the procedures mentioned above can be considered as members of the family of penalized likelihood approach.

There are many other variable selection procedures proposed based on different models. For example, variable selections for Cox's proportional hazards model and frailty model were studied by Fan and Li (2002); Efron *et al.* (2004) proposed least-angle regression (LARS); Candes and Tao (2007) proposed the Dantzig selector which is a solution to an l_1-

regularization problem. However, in many applications, the models underlying are complex or unknown. These procedures can give biased results when the models are mis-specified. To avoid this problem, model-free variable selection methods are desired.

1.3. MODEL-FREE VARIABLE SELECTION

The concept of model-free variable selection was proposed by Li *et al.* (2005). It aims to find the important predictors without the full knowledge of the underlying model structure. Hence, it can avoid the problem due to the model misspecification. Let $\mathbf{X} = (X_1, \cdots, X_p)^T$ be the p-dimensional predictor, and Y be the scalar response. Let $\mathcal{I} = \{1, 2, \ldots, p\}$ denote the complete index set. Model-free variable selection considers the problem seeking the index set $\mathcal{A} \subset \mathcal{I}$ such that

$$Y \perp\!\!\!\perp \mathbf{X}_{\mathcal{A}^c} | \mathbf{X}_{\mathcal{A}}, \tag{1.1}$$

where $\perp\!\!\!\perp$ means independent and $\mathbf{X}_{\mathcal{A}} = \{X_i : i \in \mathcal{A}\}$. Ideally, the smallest subset $\mathcal{A}$, where only the indices of active predictors are included, can be identified. The existence and uniqueness of such a set $\mathcal{A}$ have been discussed in Yin and Hilafu (2015).

Model-free variable selection can be derived from the perspective of sufficient dimension reduction (SDR) (Li 1991; Cook 1998), which aims to find a set of linear combinations of $\mathbf{X}$, say $\boldsymbol{\beta}^T\mathbf{X}$, such that

$$Y \perp\!\!\!\perp \mathbf{X} | \boldsymbol{\beta}^T\mathbf{X}, \tag{1.2}$$

Where $\boldsymbol{\beta}$ is a $p \times d$ matrix with $d \leq p$. The column space of $\boldsymbol{\beta}$ is called a dimension reduction space. The central subspace, $\mathcal{S}_{Y|\mathbf{X}}$, is the smallest dimension reduction space. As pointed out by Bondell and Li (2009), the general framework of sufficient dimension reduction is very useful for model-free variable selection since **no** pre-specified underlying models between the response and the predictors are required. Many SDR methods have been proposed in literature such as sliced inverse regression (SIR) (Li, 1991), sliced average variance estimator (SAVE) (Cook and Weisberg, 1991), minimum average variance estimators (MAVE) (Xia *et al.*, 2002), directional regression (DR) (Li and Wang, 2007), and likelihood acquired directions (LAD) (Cook and Forzani, 2009).

Many model-free variable selection procedures based on SDR have been proposed in literature. They can be summarized into two branches: shrinkage selection procedures and hypothesis testing ones. Ni *et al.* (2005) proposed the shrinkage sliced inverse regression (SIR) estimators by integrating SIR with LASSO. A unified approach was proposed by Li (2007) through combining SDR and shrinkage estimation to produce sparse estimators of the central subspace. Chen *et al.* (2010) proposed coordinate-independent sparse dimension reduction (CISE) by imposing a subspace-oriented penalty. Other shrinkage selection procedures include sparse SIR (Li and Nachtsheim, 2006) and regularized SIR (Li and Yin, 2008). Unlike the traditional SDR methods, these shrinkage selection procedures can achieve feature extraction and variable selection simultaneously.

Model-free variable selection through SDR can be also considered as a hypothesis testing problem. As in Yu *et al.* (2016), without loss of generality, we assume that the active index set $\mathscr{A} = \{1, \ldots, q\}$. Then (1.1) is equivalent to the following hypothesis testing within the framework of sufficient dimension reduction:

$$P_{\mathcal{H}}\mathcal{S}_{Y|\mathbf{X}} = \mathcal{O}_p, \tag{1.3}$$

where $P(.)$ denotes the projection operator with respect to the standard inner product, $\mathcal{H} = \text{Span}\{(\mathbf{0}_{(p-q)\times q}, \mathbf{I}_{p-q})^T\}$ is the subspace of the predictor space, corresponding to the coordinates of the inactive predictors, and $\mathcal{O}_p$ is the origin in $\mathbb{R}^p$. Hence, now we successfully transform the original variable selection problem (1.1) to a testing hypothesis problem (1.3) , which enables us to set up the connection between variable selection and sufficient dimension reduction. Based on SIR, Cook (2004) proposed marginal coordinate hypothesis test to check the contribution of predictors. Shao *et al.* (2007) studied the similar test on SAVE. Li *et al.* (2005) proposed the gridded chi-squared test. However, these shrinkage and test methods are often not suitable for high dimensional data where $n < p$.

To deal with $n > p$ situation, Zhong *et al.* (2012) proposed correlation pursuit (COP) based on SIR. Unfortunately, it inherits the limitations of SIR in the sense that it also might miss important predictors which are linked to the response through quadratic functions or interactions. For example, active predictors, which are linked to the response through quadratic functions or interactions, may be missed. Furthermore, COP involves the estimation of the dimension of $\mathcal{S}_{Y|\mathbf{X}}$, which could be challenging for high dimensional data. Yin and Hilafu (2015) proposed a sequential method, which transforms the original problem to the traditional $n < p$ problem by partitioning the original data into pieces. However, there might be some issues with implementations of their method since different partitions of the predictors might lead to different results. Recently, Yu *et al.* (2016) developed a novel general framework of model-free variable selection for $n > p$ situation, the *trace pursuit* method, which could be combined with many existing sufficient dimension reduction methods. Their method provides a versatile framework for variable selection via stepwise trace pursuit (STP), which can be viewed as a model-free counterpart of the classical stepwise regression. Mimicing the forward regression in linear model, the forward trace pursuit (FTP) was proposed to conduct the initial variable screening.

All these procedures mentioned above are based on single population data. However, in practice, researchers often need to deal with data from different groups, such as different genders and regions. It would be desirable to incorporate those grouping information into the variable selection procedure, since it might be related to both the response and the predictors. In Paper I, we extend the trace pursuit method to data with multiple groups. Our simulation studies suggest that the selection performances could be greatly improved with the utilization of the grouping information. Specifically, the underfit (omission of significant variables) rate is greatly reduced, while the correct fit rate is significantly improved.

1.4. CONDITIONAL SCREENING

Working on data sets with high or even ultra-high dimensional structure is very common in different research areas, such as genomics, neuroscience and finance. Here ultra-high dimension means dimension p increases with an exponential rate of sample size n. A common assumption for this kind of data is that only a small number of predictors actually contribute to the response, and it is called sparsity assumption. In consideration of the expensive time cost, researchers usually prefer to use a fast screening procedure first to reduce the dimension of data, then do variable selection through more sophisticated procedures. Fan and Lv (2008) proposed the sure independence screening (SIS) procedure for linear models through ranking the marginal correlation between the response variable and each individual predictor. SIS has the so-called *sure screening property* (Fan and Lv, 2008), in the sense that as $n \to \infty$, the important predictors are guaranteed to be retained in the model with probability tending to 1. Fan *et al.* (2009) and Fan and Song (2010) extended SIS to generalized linear models. Fan *et al.* (2011) further extended SIS to additive models and proposed nonparametric independence screening (NIS) using nonparametric marginal ranking. Wang (2012) investigated forward regression (FR) for high dimensional data. Many other variable screening procedures have been developed, such as Xue and Zou (2011), Zhao and Li (2012), and Chang *et al.* (2013).

However, all the variable screening procedures mentioned above are model-based, such as linear models and generalized linear models. The performance would be poor if the model is mis-specified. To avoid the restriction of specification of the model structure, statisticians proposed many model-free variable screening methods, where model-free variable screening means the variable screening procedure works without knowledge of the link function between Y and $\mathbf{X}$. For example, Zhu *et al.* (2011) proposed a sure independent ranking and screening (SIRS), Li *et al.* (2012) proposed a sure independence screening procedure based on the distance correlation (DC-SIS), He *et al.* (2013) proposed a quantile-adaptive model-free screening framework, which estimated marginal quantile regression nonparametrically using B-spline approximation, and Mai and Zou (2015) proposed the fused Kolmogorov filter approach, which performs feature screening for the data with many types of predictors and response. There are also some model-free variable screening procedures developed for discriminant analysis with high dimensional data, such as Mai and Zou (2013), Cui *et al.* (2014), and Pan *et al.* (2016).

As we discussed before, Yu et al. (2016) recently proposed a novel model-free feature screening method, the *forward trace pursuit (FTP)*, based on the framework of sufficient dimension reduction. It was showed that FTP can work with different sufficient dimension reduction methods, such as SIR (Li, 1991), SAVE (Cook and Weisberg, 1991), and DR (Li and Wang, 2007). The screening consistency property of the SIR-based FTP was also established.

As discussed in the existing literatures such as Fan and Lv (2008), Zhu *et al.* (2011), and Barut *et al.* (2016), the simple variable screening procedures are heavily influenced by the correlations among predictors. When the correlations among predictors are high, these procedures may raise false positives, where the inactive predictors are mistakenly screened in as active ones, and also false negatives, where the active predictors are mistakenly screened out as inactive ones. Unfortunately, as mentioned in Hall and Li (1993) and Fan and Lv (2008), there always exist spurious correlations among predictors with growing dimensionality p. Hence, this problem is unavoidable for high dimensional data analysis. To obtain the sure screening property, some restrictions are needed on the correlation structure among predictors for variable screening procedures.

Also, in many applications, researchers have some prior knowledge that certain predictors are important from experience or previous research work, such as the treatment effects in biological studies and market risk factors in financial studies. To fully utilize this prior information and also to relieve the influence of high correlation among predictors, Barut *et al.* (2016) proposed sure independence screening (CSIS), which performed variable screening on the rest of predictors conditioning on the known ones. Through simulation studies and real data analysis, Barut *et al.* (2016) showed that CSIS could greatly improve the screening performance compared with SIS Fan and Lv (2008). Compared with SIS, CSIS makes it possible to identify those significant hidden predictors whose contributions might otherwise get canceled out due to the correlations with other predictors. Also, when there are high correlations among significant predictors and insignificant ones, CSIS can help to reduce the number of false negatives.

However, CSIS was proposed for generalized linear models. The misspecification of model structure might corrupt the performance of the variable screening procedure. To address this issue, we propose a model-free conditional screening method via sufficient dimension reduction in Paper II. Specifically, our method is based on the partial sufficient dimension reduction procedure proposed by Feng *et al.* (2013). In numerical studies, we compare the performance of our proposed method with CSIS using the true model coverage rate (CR, the rate of all the significant predictors being selected), the average model size (MS), the average false positive rate (FP), and the average false negative rate (FN). Comparing to CSIS, our proposed method can produce screening results with smaller model sizes, similar or better coverage rates, smaller false positive rates and/or false negative rates when the model structure is nonlinear, which is often the case in real data applications.

PAPER

I. TRACE PURSUIT VARIABLE SELECTION FOR MULTI-POPULATION DATA

Lei Huo[1], Xuerong Meggie Wen[1], and Zhou Yu[2]

[1]Department of Mathematics and Statistics,

Missouri University of Science and Technology, MO 65409, U.S.A.

email: wenx@mst.edu

[2]School of Finance and Statistics,

East China Normal University, Shanghai, China

ABSTRACT

Variable selection is a very important tool when dealing with high dimensional data. However, most popular variable selection methods are model-based, which might provide misleading results when the model assumption is not satisfied. Sufficient dimension reduction provides a general framework for model-free variable selection methods. In this paper, we propose a model-free variable selection method via sufficient dimension reduction, which incorporates the grouping information into the selection procedure for multi-population data. Theoretical properties of our selection methods are also discussed. Simulation studies suggest that our method greatly outperforms those ignoring the grouping information.

Keywords: Trace Pursuit; Variable Selection; Partial Central Subspace; Sufficient Dimension Reduction.

1. INTRODUCTION

The importance of variable selection becomes more critical nowadays since modern scientific innovations allow scientists to collect massive and high-dimensional data at a rapid rate. Often the dimensions of the predictors (p) may greatly surpass the relative small sample size (n). Many methods have been developed in recent years to extract the significant variables effectively under the so called $n < p$ context. However, most of the popular variable selection methods, such as nonnegative garrotte (Breiman, 1995), LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001), adaptive LASSO (Zou, 2006), group LASSO (Yuan and Lin, 2006), Dantzig selector (Candes and Tao, 2007), and MCP (Zhang, 2010), are model-based, where a linear model or generalized linear model is assumed. Such methods might generate biased results if the underlying modeling assumption is violated, which is typically the case for complex or unknown models. Hence, *model-free* variable selection method, which does not require the full knowledge of the underlying true model, is called for.

Let $\mathbf{X} = (X_1, \cdots, X_p)^T$ be the p-dimensional predictor, and Y be the scalar response. Let $\mathcal{I} = \{1, 2, \ldots, p\}$ denote the complete index set. Model-free variable selection aims to identify the index set $\mathcal{A} \subset \mathcal{I}$ such that

$$Y \perp\!\!\!\perp \mathbf{X}_{\mathcal{A}^c} | \mathbf{X}_{\mathcal{A}},$$

where $\mathcal{A}^c$ is the complement set of $\mathcal{A}$, and $\mathbf{X}_{\mathcal{A}} = \{X_i : i \in \mathcal{A}\}$. The goal here is to identify the smallest $\mathbf{X}_{\mathcal{A}}$ which contains all the active predictors. Yin and Hilafu (2015) gave a detailed discussion of the existence and uniqueness of such a set $\mathcal{A}$. As pointed out by Bondell and Li (2009), the general framework of sufficient dimension reduction (Li 1991; Cook 1998) is very useful for model-free variable selection since **no** pre-specified underlying models between the response and the predictors are required.

When $n > p$, Ni *et al.* (2005), Li and Nachtsheim (2006), and Li and Yin (2008) proposed model-free variable selections by reformulating sufficient dimension reduction as a penalized regression problem. Li (2007) proposed a unified approach combining SDR and shrinkage estimation to produce sparse estimators of the central subspace. Wang and Zhu (2015) proposed a distribution-weighted lasso method for the single-index model. Chen *et al.* (2010) proposed coordinate-independent sparse dimension reduction (CISE) imposing a subspace-oriented penalty. However, none of those model-free variable selections can deal with variable selection when $n < p$. Such situations do arise in many high dimensional data sets in bioinformatics, machine learning and pattern recognition. Recently, Yin and Hilafu (2015) proposed a sequential method which transforms the original problem to the regular $n < p$ one, by decomposing the original data into pieces. However, there might be some issues with implementations of their method since different partitions of the predictors might lead to different results. Yu *et al.* (2016) developed a novel model-free variable selection method under the $n < p$ context, the *trace pursuit* method, which could be combined with many existing sufficient dimension reduction methods. Their method provides a versatile framework for variable selection via stepwise trace pursuit (STP), which can be viewed as a model-free counterpart of the classical stepwise regression.

However, in practice, we often deal with situations where the data came from different groups, say, males or females. It would be desirable to incorporate those grouping information into the variable selection procedure, since it might be related to both the response and the predictors. In this paper, we extend the trace pursuit method to data with multiple groups. Our simulation studies suggest that the selection performances could be greatly improved with the utilization of the grouping information. Specifically, the underfit (omission of significant variables) rate is greatly reduced, while the correct fit rate is significantly improved.

The rest of this article is organized as follows. We first give a brief introduction of sufficient dimension reduction methods and trace pursuit method for a single population in Section 2. In Section 3, we present our new estimation method in details; and also discuss its related asymptotic properties. We illustrate the performance of our methods via simulation studies in Section 4. Brief conclusions and a discussion on future research directions are given in Section 5.

2. SUFFICIENT DIMENSION REDUCTION FOR A SINGLE POPULATION

For regression problems $Y|\mathbf{X}$ within a single population, Li (1991) and Cook (1998) proposed sufficient dimension reduction that aims at reducing the dimension of $\mathbf{X}$ while preserving the regression relationship between Y and $\mathbf{X}$ without requiring a parametric model. Specifically, the scope of sufficient dimension reduction is to seek a set of linear combinations of $\mathbf{X}$, say $\boldsymbol{\beta}^T\mathbf{X}$, where $\boldsymbol{\beta}$ is a $p \times d$ matrix with $d \leq p$, such that

$$Y \perp\!\!\!\perp \mathbf{X}|\boldsymbol{\beta}^T\mathbf{X}.$$

The column space of $\boldsymbol{\beta}$ is then called a dimension reduction space, and the smallest dimension reduction space is defined as the central subspace, denoted by $\mathcal{S}_{Y|\mathbf{X}}$. It is the intersection of all dimension reduction spaces. The goal of sufficient dimension reduction is to make inferences about the central subspace and its dimension d, which is called the structural dimension of the regression. Subsequent modeling and prediction can be built upon those d reduced directions.

Sufficient dimension reduction has received considerable interests in recent years due to the ubiquity of large high-dimension data sets which are now more readily available than in the past. Many methods have been developed, including sliced inverse regression (SIR; Li 1991), sliced average variance estimation (SAVE; Cook and Weisberg 1991), minimum average variance estimation (MAVE; Xia *et al.* 2002), directional regression (DR; Li and

Wang 2007), likelihood acquired directions (LAD; Cook and Forzani 2009), cumulative slicing estimation (CUME; Zhu *et al.* 2010), dimension reduction for special-structured $\mathbf{X}$ (Li *et al.*, 2010), nonlinear sufficient dimension reduction (Lee *et al.*, 2013), sufficient dimension reduction via a semiparametric approach (Ma and Zhu 2012, 2013) and many others.

We now briefly review the most widely used sufficient dimension reduction method, SIR (Li, 1991). Let $\mathbf{\Sigma} = \mathrm{Cov}(\mathbf{X})$ denote the marginal covariance matrix of $\mathbf{X}$, $\boldsymbol{\mu} = \mathrm{E}(\mathbf{X})$, and let $\mathbf{Z} = \mathbf{\Sigma}^{-\frac{1}{2}}(\mathbf{X}-\mathrm{E}(\mathbf{X}))$ be the standardized predictor. By the invariance property (Cook, 1998), we have $\mathcal{S}_{Y|\mathbf{X}} = \mathbf{\Sigma}^{-\frac{1}{2}}\mathcal{S}_{Y|\mathbf{Z}}$, where $\mathcal{S}_{Y|\mathbf{Z}}$ is the central subspace for the regression of $Y|\mathbf{Z}$. Unlike traditional regression modeling, sufficient dimension reduction methods, rely on an assumption about the marginal distribution of $\mathbf{Z}$ instead of the conditional distribution of $Y|\mathbf{Z}$. The so-called *linearity condition* requires that $\mathrm{E}(\mathbf{Z}|\boldsymbol{\rho}^T\mathbf{Z})$ be a linear function of $\boldsymbol{\rho}^T\mathbf{Z}$, where the columns of the $p \times d$ matrix $\boldsymbol{\rho}$ form an orthonormal basis for $\mathcal{S}_{Y|\mathbf{Z}}$. For more detailed discussions of the linearity condition (LM condition), please see Feng *et al.* (2013).

The linearity condition connects the central subspace with the inverse regression of $\mathbf{Z}$ on Y. Li (1991) showed that $\mathrm{E}(\mathbf{Z}|Y) \in \mathcal{S}_{Y|\mathbf{Z}}$ when it holds. When Y is continuous, Li (1991) proposed estimating $\mathrm{E}(\mathbf{Z}|Y)$ by replacing Y with a discrete version constructed by partitioning the range of Y into H fixed non-overlapping slices $s_1, \ldots, s_H$. Let $p_h = \Pr\{Y \in s_h\}$, $\mathbf{m}_h = \mathrm{E}(\mathbf{Z}|Y \in s_h)$, $\mathbf{M}_{sir} = \sum_{h=1}^{H} p_h\mathbf{m}_h\mathbf{m}_h^T$. Li (1991) showed that the eigenvectors corresponding to the d nonzero eigenvalues of $\mathbf{M}_{sir}$ form a basis of $\mathcal{S}_{Y|\mathbf{Z}}$.

Let $\widehat{\mathbf{M}}_{sir}$ denote a consistent estimate of $\mathbf{M}_{sir}$, SIR made use of the span of the eigenvectors corresponding to the d largest eigenvalues of $\widehat{\mathbf{M}}_{sir}$ to estimate $\mathrm{Span}(\mathbf{M}_{sir})$. The eigenvalues provide a test statistic for hypotheses on the structural dimension, and the eigenvectors can be linearly transformed back to the $\mathbf{X}$-scale to form a basis for $\mathcal{S}_{Y|\mathbf{X}}$. This is the so called spectral decomposition approach (Wen and Cook, 2009), since it is based on a spectral decomposition of the sample kernel matrix $\widehat{\mathbf{M}}_{sir}$. SAVE (Cook and Weisberg,

1991) and DR (Li and Wang, 2007) took the same spectral decomposition approach via different kernel matrices: $\mathbf{M}_{save} = \mathrm{E}\{I_p - \mathrm{Var}(\mathbf{Z}|Y)\}^2$, and $\mathbf{M}_{dr} = 2\mathrm{E}\{\mathrm{E}^2(\mathbf{Z}\mathbf{Z}^T|Y)\} + 2\mathrm{E}^2\{\mathrm{E}(\mathbf{Z}|Y)\mathrm{E}(\mathbf{Z}^T|Y)\} + 2\mathrm{E}\{\mathrm{E}(\mathbf{Z}^T|Y)\mathrm{E}(\mathbf{Z}|Y)\}\mathrm{E}\{\mathrm{E}(\mathbf{Z}|Y)\mathrm{E}(\mathbf{Z}^T|Y)\} - 2I_p$. SAVE and DR require a constant conditional variance condition ($\mathrm{Var}(\mathbf{Z}|\boldsymbol{\rho}^T\mathbf{Z})$ is nonrandom) in addition to the linearity condition.

3. TRACE PURSUIT VARIABLE SELECTION FOR MULTIPLE GROUPS

3.1. The Test Statistics. For easy of exposition, we follow Yu *et al.* (2016) to assume that $\mathcal{A} = \{1, \ldots, q\}$. Then (1) is equivalent to the following hypothesis testing within the framework of sufficient dimension reduction:

$$P_{\mathcal{H}}\mathcal{S}_{Y|\mathbf{X}} = O_p, \tag{3.1}$$

where $P(.)$ denotes the projection operator with respect to the standard inner product, $\mathcal{H} = \mathrm{Span}\{(\mathbf{0}_{(p-q)\times q}, \mathbf{I}_{p-q})^T\}$ is the subspace of the predictor space, corresponding to the coordinates of the inactive predictors, and O_p is the origin in $\mathbb{R}^p$. Cook (2004) first proposed a test for testing hypothesis of (3.1) based on a generalized least square rederivation of the SIR estimator for $\mathcal{S}_{Y|\mathbf{X}}$. Shao et al. (2007) and many others also considered (3.1) based on other estimators of $\mathcal{S}_{Y|\mathbf{X}}$. However, all those tests will not be applicable when $n < p$, due to the difficulty of obtaining a sensible initial estimator for $\mathcal{S}_{Y|\mathbf{X}}$. Zhong *et al.* (2012) and Jiang and Liu (2013) tackled testing (3.1) via sliced inverse regression (SIR) method. However, both methods require the estimation of the rank of $\mathcal{S}_{Y|\mathbf{X}}$ (the so-called order determination), which is a very challenging problem when $n < p$. Yu *et al.* (2016) proposed a novel trace pursuit approach to conduct model-free variable selection via sufficient dimension reduction approach for $n < p$, which successfully circumvents the need of order determination. However, as we discussed in Section 1, none of those methods took the grouping information into consideration for data from multiple groups. In this section, we extend the trace pursuit method to deal with this specific issue. As Yu *et al.* (2016) pointed out, the trace pursuit method can be combined with many commonly used sufficient dimension reduction methods. We will propose our method with SIR in this article, since the methodology can be extended to SAVE and DR similarly. In the numerical studies, we provide simulation results via all three methods.

We first introduce the concept of partial central subspace which was proposed by Chiaromonte *et al.* (2002) when the predictor is a mixture of a p-dimensional continuous vector $\mathbf{X}$ and a categorical variable W, and the dimension reduction was focused on $\mathbf{X}$ alone. The *partial central subspace* ($\mathcal{S}_{Y|\mathbf{X}}^{(W)}$) is defined as the intersection of all subspaces Span($\boldsymbol{\beta}$) satisfying

$$Y \perp\!\!\!\perp \mathbf{X} \mid (\boldsymbol{\beta}^T\mathbf{X}, W),$$

where $W \in \{1, \ldots, K\}$ is a categorical predictor (or group indicator). Let $(\mathbf{X}^w, Y^w)$ denote a generic pair of $(\mathbf{X}, Y)$ for the w-th group, $\boldsymbol{\Sigma}_w = \mathrm{Var}(\mathbf{X}_w)$, and $\mathbf{Z}_w = \boldsymbol{\Sigma}_w^{-\frac{1}{2}}(\mathbf{X}_w - \boldsymbol{\mu}_w)$. Let $\mathcal{S}_{Y_w|\mathbf{X}_w}$ be the central subspace for the regression of $Y^w|\mathbf{X}^w$. The following equation (Chiaromonte *et al.*, 2002) connects the partial central subspace with the conditional central subspaces:

$$\mathcal{S}_{Y|\mathbf{X}}^{(W)} = \sum_{w=1}^{K} \mathcal{S}_{Y_w|\mathbf{X}_w}. \tag{3.2}$$

Equation (3.2) is the key to the connection between the partial central subspace and the conditional central subspaces. It showed how we can obtain an estimate of the partial central subspace through the conditional central subspaces. Partial SIR (Chiaromonte *et al.*, 2002), Partial OPIRE (Wen and Cook, 2007), and PDEE (Feng *et al.*, 2013) were all developed to estimate the partial central subspace based on Equation (3.2). Equation (3.2) also suggests that $\mathcal{S}_{Y|\mathbf{X}}^{(W)}$ contains each conditional central subspace $\mathcal{S}_{Y_w|\mathbf{X}_w}$.

For multiple population data, the original testing problem (1) becomes

$$Y \perp\!\!\!\perp \mathbf{X}_{\mathcal{A}^c} | (\mathbf{X}_{\mathcal{A}}, W), \tag{3.3}$$

where W is the group indicator. Adopting the concept of partial central subspace, (3.3) is equivalent to testing:

$$H_o : P_{\mathcal{H}} \mathcal{S}_{Y|\mathbf{X}}^{(W)} = O_p,$$

versus not H_o.

Within group w, without loss of generality, we assume that $\mathrm{E}(\mathbf{X}_w = 0)$. Partition the range of Y_w into H_w fixed non-overlapping slices $s_1, \ldots, s_{Hw}$. Let $p_w = \Pr(W = w)$, $p_{hw} = \Pr\{Y_w \in s_{hw}\}$, $\mathbf{U}_{hw} = \mathrm{E}(\mathbf{X}_w | Y_w \in s_{hw})$. Based on (3.2), we can hence construct the kernel matrix for SIR as $\mathbf{M} = \sum_{w=1}^{K} p_w \mathbf{\Sigma}_w^{-\frac{1}{2}} (\sum_{h=1}^{Hw} p_{hw} \mathbf{U}_{hw} \mathbf{U}_{hw}^T) \mathbf{\Sigma}_w^{-\frac{1}{2}}$. For any index set $\mathcal{F}$, denote $\mathbf{X}_{\mathcal{F}} = \{X_i : i \in \mathcal{F}\}$, $\mathrm{Var}(\mathbf{X}_{\mathcal{F}} | W = w) = \mathbf{\Sigma}_{w_{\mathcal{F}}}$, and $\mathbf{U}_{\mathcal{F},hw} = \mathrm{E}(\mathbf{X}_{\mathcal{F}} | Y_w \in s_{hw}, W = w)$. Define $\mathbf{M}_{\mathcal{F}} = \sum_{w=1}^{K} p_w \mathbf{\Sigma}_{w_{\mathcal{F}}}^{-\frac{1}{2}} (\sum_{h=1}^{Hw} p_{hw} \mathbf{U}_{\mathcal{F},hw} \mathbf{U}_{\mathcal{F},hw}^T) \mathbf{\Sigma}_{w_{\mathcal{F}}}^{-\frac{1}{2}}$, we have the following proposition.

Proposition 1 *Assuming the linearity condition for* $\mathbf{X}$ *within each group, then for any index set* $\mathcal{F}$ *such that* $\mathcal{A} \subseteq \mathcal{F} \subseteq \mathcal{I}$, *we have* $\mathrm{tr}(\mathbf{M}_{\mathcal{A}}) = \mathrm{tr}(\mathbf{M}_{\mathcal{F}}) = \mathrm{tr}(\mathbf{M}_{\mathcal{I}})$, *where* $\mathcal{A}$ *denotes the active index set such that* $Y \perp\!\!\!\perp \mathbf{X}_{\mathcal{A}^c} | (\mathbf{X}_{\mathcal{A}}, W)$, *and* I_s *denotes the full index set.*

The proof of Proposition 1 is provided in the appendix. It suggests that for all the sets satisfying $\mathcal{F} \supseteq \mathcal{A}$, $\mathrm{tr}(\mathbf{M}_{\mathcal{F}})$ will be the same as $\mathrm{tr}(\mathbf{M}_{\mathcal{A}})$. Hence, assuming that $\mathbf{X}_{\mathcal{F}}$ is already in the model, then for any $X_j \notin \mathbf{X}_{\mathcal{F}}$, we can use the differences between $\mathrm{tr}(\mathbf{M}_{\mathcal{F} \cup j})$ and $\mathrm{tr}(\mathbf{M}_{\mathcal{F}})$ to test the contribution of the additional variable X_j to the regression of Y versus $(\mathbf{X}, W)$.

Assuming a subset linearity condition for any $X_j \notin \mathbf{X}_{\mathcal{F}}$, which requires that $\mathrm{E}(X_j | \mathbf{X}_{\mathcal{F}}, W = w)$ is a linear function of $\mathbf{X}_{\mathcal{F}}$ within each group w, the following theorem provides a way to calculate the trace differences: $\mathrm{tr}(\mathbf{M}_{\mathcal{F} \cup j}) - \mathrm{tr}(\mathbf{M}_{\mathcal{F}})$.

Theorem 1 *Assuming a subset linearity condition defined as above, then for any* $\mathcal{F} \subset \mathcal{I}$, *and* $j \in \mathcal{F}^c$, *we have*

- *If* $\mathcal{A} \subseteq \mathcal{F}$, *then* $\mathrm{tr}(\mathbf{M}_{\mathcal{F} \cup j}) - \mathrm{tr}(\mathbf{M}_{\mathcal{F}}) = 0$.

- *If* $\mathcal{A} \nsubseteq \mathcal{F}$*, then* $\mathrm{tr}(\mathbf{M}_{\mathcal{F}\cup j}) - \mathrm{tr}(\mathbf{M}_{\mathcal{F}}) = \sum_{w=1}^{K} p_w\big(\sum_{h=1}^{Hw} p_{hw}\boldsymbol{\gamma}^2_{j|w\mathcal{F},hw}\big)$ *where* $\boldsymbol{\gamma}_{j|w\mathcal{F},hw} = \mathrm{E}(\boldsymbol{\gamma}_{j|\mathcal{F}}|Y \in s_{h_w}, W = w)$ *with* $X_{j|\mathcal{F}} = X_j - \mathrm{E}(X_j|\mathbf{X}_{\mathcal{F}})$*,* $\sigma^2_{j|\mathcal{F}} = \mathrm{Var}(X_j|\mathcal{F})$*, and* $\boldsymbol{\gamma}_{j|\mathcal{F}} = X_{j|\mathcal{F}}/\sigma_{j|\mathcal{F}}$.

Let $(Y_{wi}, \mathbf{X}_{wi})$, $i = 1, \ldots, n_w$ be a simple random sample of size n_w from the wth group $(Y_w, \mathbf{X}_w)$ for $w = 1, \ldots, K$. Let $\bar{\mathbf{X}}_w = \frac{1}{n_w}\sum_{i=1}^{n_w} \mathbf{X}_{wi}$, and $\widehat{\boldsymbol{\Sigma}}_w = \frac{1}{n_w}\sum_{i=1}^{n_w}(\mathbf{X}_{wi} - \bar{\mathbf{X}}_w)(\mathbf{X}_{wi} - \bar{\mathbf{X}}_w)^T$. $\bar{\mathbf{X}}_{w\mathcal{F}}$ and $\widehat{\boldsymbol{\Sigma}}_{w\mathcal{F}}$ can be defined similarly. Let n_{hw} denote the total number of data points in the hth slice within group w. Let $\hat{p}_w = \frac{n_w}{n}$, where $n = n_1 + \cdots + n_K$. Let $\hat{p}_{hw} = \frac{n_{hw}}{n_w}$, the sample proportion of data points in the hth slice within group w. Let $\widehat{\mathbf{U}}_{\mathcal{F},hw} = 1/n_{hw} \sum_{i:Y_{wi}\in s_{hw}} (\mathbf{X}_{wi,\mathcal{F}} - \bar{\mathbf{X}}_{w\mathcal{F}})$. We can construct $\widehat{\mathbf{M}}_{\mathcal{F}}$, the sample version of $\mathbf{M}_{\mathcal{F}}$, as $\sum_{w=1}^{K} \hat{p}_w \widehat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}_{w\mathcal{F}}\big(\sum_{h=1}^{H_w} \hat{p}_{hw}\widehat{\mathbf{U}}_{\mathcal{F},hw}\widehat{\mathbf{U}}^T_{\mathcal{F},hw}\big)\widehat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}_{w\mathcal{F}}$.

Let $T_{j|\mathcal{F}} = n\big(\mathrm{tr}(\widehat{\mathbf{M}}_{\mathcal{F}\cup j}) - \mathrm{tr}(\widehat{\mathbf{M}}_{\mathcal{F}})\big)$ be the test statistic for hypothesis (3.3). Theorem 1 can be used to calculate $T_{j|\mathcal{F}}$, with p_w, p_{hw} and $\boldsymbol{\gamma}_{j|\mathcal{F}}$ being estimated using their corresponding sample versions. The asymptotic distribution of $T_{j|\mathcal{F}}$ is given in the following theorem.

Theorem 2 *Let* $(Y_{wi}, \mathbf{X}_{wi})$*,* $j = 1, \ldots, n_w$ *be a simple random sample with finite fourth moments of size* n_w *from the wth group* $(Y_w, \mathbf{X}_w)$ *for* $w = 1, \ldots, K$*. Assuming the subset linearity condition as in Theorem 1, and* $|\mathcal{F}|$ *is fixed when n goes to infinity, then under* $H_o : Y \perp\!\!\!\perp \mathbf{X}_j|(\mathbf{X}_{\mathcal{F}}, W)$*,* $j \in \mathcal{F}^c$*, we have:*

$$T_{j|\mathcal{F}} \longrightarrow \sum_{i=1}^{H} \omega^2_{j|\mathcal{F},i}\ \chi^2_1,$$

where $H = H_1 + \cdots + H_K$ *is the total number of slices,* $\omega_{j|\mathcal{F},1} \geq \cdots \geq \omega_{j|\mathcal{F},H}$ *are the eigenvalues of* $\boldsymbol{\Omega}_{j|\mathcal{F}}$ *as defined in the Appendix.*

3.2. The Selection Procedure. Following Yu *et al.* (2016), we use the forward trace pursuit (FTP) and the stepwise trace pursuit (STP) procedures to select the active variables. Specifically we use FTP to serve as a screening tool, and STP to refine the selection. Yu *et al.* (2016) call this selection method the hybrid trace pursuit (HTP) procedure. Below are the algorithms for FTP and STP procedures respectively.

Forward trace pursuit

1) Let $\mathcal{F}_0 = \varnothing$.

2) At the kth $(k \geq 1)$ *iteration, find* a_k *such that*

$$a_k = \underset{j \in \mathcal{F}_{k-1}^c}{\operatorname{argmax}} \operatorname{tr}(\widehat{\mathbf{M}}_{\mathcal{F}_{k-1} \cup j}).$$

3) Repeating 2) n times, to obtain a sequence of n nested index sets. Denote the solution path as $S = \{\mathcal{F}_k : 1 \leq k \leq n\}$, *where* $\mathcal{F}_k = \{a_1, \ldots, a_k\}$.

Stepwise trace pursuit

1) Let $\mathcal{F}_0 = \varnothing$.

2) Forward addition: Find $a_{\mathcal{F}}$ *such that*

$$a_{\mathcal{F}} = \underset{j \in \mathcal{F}^c}{\operatorname{argmax}} \operatorname{tr}(\widehat{\mathbf{M}}_{\mathcal{F} \cup j}).$$

If $T_{a_{\mathcal{F}}|\mathcal{F}}$ *is greater than a pre-specified cut off value* c_1, *then update* $\mathcal{F}$ *to be* $\mathcal{F} \cup a_{\mathcal{F}}$.

3) Backward deletion: Find $d_{\mathcal{F}}$ *such that*

$$d_{\mathcal{F}} = \underset{j \in \mathcal{F}^c}{\operatorname{argmax}} \operatorname{tr}(\widehat{\mathbf{M}}_{\mathcal{F} \setminus j}).$$

If $T_{d_{\mathcal{F}}|\mathcal{F} \setminus d_{\mathcal{F}}}$ *is less than a pre-specified cut off value* c_2, *then update* $\mathcal{F}$ *to be* $\mathcal{F} \setminus d_{\mathcal{F}}$.

4) Repeat 2) and 3) until no predictors can be added or deleted.

We now discuss the theoretical properties of our procedures. Assume $\mathrm{Var}\{\mathrm{E}(\mathbf{Z}_w|Y \in s_{h_w})\}$ has q_w nonzero eigenvalues $\lambda_{w1} \geq \cdots \geq \lambda_{wq_w}$ with corresponding eigenvectors $\boldsymbol{\eta}_{w1}, \ldots, \boldsymbol{\eta}_{wq_w}$, where $w = 1, \ldots, K$. Let $\boldsymbol{\beta}_{wi} = \boldsymbol{\Sigma}_w^{-1/2}\boldsymbol{\eta}_{wi}$ for $i = 1, \ldots, q_w$ and $w = 1, \ldots, K$. Let $\beta_{wi,j}$ be the jth elements of $\boldsymbol{\beta}_{wi}$, $j = 1, \ldots, p$. Define $\beta_{min} = \min_{\substack{w=1,\ldots,K \\ j\in\mathcal{A}}} \left\{ \sqrt{\sum_{i=1}^{q_w} \beta_{wi,j}^2} \right\}$. Let $\lambda_0 = \min_{w=1,\ldots,K}\{\lambda_{wq_w}\}$, $\lambda_{max} = \max_{w=1,\ldots,K}\{\lambda_{max}(\boldsymbol{\Sigma}_w)\}$ and $\lambda_{min} = \min_{w=1,\ldots,K}\{\lambda_{min}(\boldsymbol{\Sigma}_w)\}$, where $\lambda_{max}(\boldsymbol{\Sigma}_w)$ and $\lambda_{min}(\boldsymbol{\Sigma}_w)$ are the largest and the smallest eigenvalues of $\boldsymbol{\Sigma}_w$.

Proposition 2 *Assuming $Span\{\boldsymbol{\beta}_{w1}, \ldots, \boldsymbol{\beta}_{wq_w}\} = \mathcal{S}_{Y_w|\mathbf{X}_w}$ and the subset linearity condition as in Theorem 1, then for any index set $\mathcal{F}$ such that $\mathcal{F}^c \cap \mathcal{A} \neq \emptyset$, we have*

$$\max_{j\in\mathcal{F}^c\cup\mathcal{A}}\{\mathrm{tr}(\mathbf{M}_{\mathcal{F}\cup j} - \mathrm{tr}(\mathbf{M}_{\mathcal{F}})\} \geq \lambda_0\lambda_{min}\lambda_{max}^{-1}\beta_{min}.$$

The above proposition suggests that when $\mathcal{F}$ does not contain $\mathcal{A}$, the maximum value of $\mathrm{tr}(\mathbf{M}_{\mathcal{F}\cup j}) - \mathrm{tr}(\mathbf{M}_{\mathcal{F}})$ is greater than 0. The proof is given in the Appendix.

We assume the following condition for the selection consistency for STP procedure:

Condition 1 *Assuming that there exist $\alpha > 0$ and $0 < \theta < 1/2$ such that*

$$\min_{\mathcal{F}:\mathcal{F}^c\cap\mathcal{A}\neq\emptyset}\ \max_{j\in\mathcal{F}^c\cup\mathcal{A}}\{\mathrm{tr}(\mathbf{M}_{\mathcal{F}\cup j} - \mathrm{tr}(\mathbf{M}_{\mathcal{F}})\} \geq \alpha n^{-\theta}$$

Theorem 3 *Let $(Y_{wi}, \mathbf{X}_{wi})$, $i = 1, \ldots, n_w$ be a simple random sample with finite fourth moments of size n_w from the wth group $(Y_w, \mathbf{X}_w)$ for $w = 1, \ldots, K$. Let c_1 and c_2 be two constants such that $0 < c_1 < 1/2\alpha n^{1-\theta}$ and $c_2 > An^{1-\theta}$ for any $A > 0$. Assuming the subset linearity condition and Condition 1, then*

$$\lim_{n\to\infty} \Pr(\min_{\mathcal{F}:\mathcal{F}^c\cap\mathcal{A}\neq\emptyset}\ \max_{j\in\mathcal{F}^c\cup\mathcal{A}} T_{j|\mathcal{F}} > c_1) = 1,$$

and

$$\lim_{n\to\infty} \Pr(\max_{\mathcal{F}:\mathcal{F}^c\cap\mathcal{A}=\emptyset}\ \min_{j\in\mathcal{F}} T_{j|\{\mathcal{F}/j\}} < c_2) = 1$$

Theorem 3 provides the selection consistency result for the STP method. It suggests that the addition step will not stop til all significant predictors are included, and the deletion step will continue until all insignificant predictors are removed.

We need the following conditions for the consistency of the FTP procedure.

Condition 2

a. $\mathbf{X}_w$ *follows a multinormal distribution for* $w = 1, \dots, K$.

b. There exist $\gamma_1 > 0$ *and* $\gamma_2 > 0$ *such that* $\gamma_1 < \lambda_{min} < \lambda_{max} < \gamma_2$.

c. There exist constants α_1, θ_1 *and* θ_2 *such that* $\log p \leq \alpha_1 n^{\theta_1}$, $|\mathcal{A}| \leq \alpha_1 n^{\theta_2}$ *and* $2\theta + \theta_1 + \theta_2 < 1$, *where* θ *is a constant from Condition 1.*

Follow Chen and Chen (2008) and define the modified BIC criterion

$$\text{BIC}(\mathcal{F}) = -\text{log}\left\{\text{tr}(\widehat{\mathbf{M}}_{\mathcal{F}})\right\} + n^{-1}|\mathcal{F}|(\text{log}n + 2\text{log}p).$$

Theorem 4 *Assume Condition 1 and Condition 2 hold true, then we have*

$$\Pr(\mathcal{A} \subset \mathcal{F}_{\hat{m}}) \longrightarrow 1,$$

as $n \longrightarrow \infty$ *and* $p \longrightarrow \infty$, *where* $\hat{m} = \underset{1 \leq k \leq n}{\text{argmin}}\, BIC(\mathcal{F}_k)$, *and* $\mathcal{F}_k$ *is defined in the FTP procedure.*

Hence Theorem 4 guarantees the selection consistency for FTP procedure.

4. NUMERICAL STUDIES

In this section, we compare the performance of our method with Yu *et al.* (2016). We summarize our results over 50 replications for each simulation study. We studied the performance of our proposed tests via SIR, SAVE and DR with different choices of p. Throughout our simulation studies, the number of slices is set as $h = 4$, the sample size is

$n = 400$. Following Yu *et al.* (2016), the under fitted count (UF), the correctly fitted count (CF), the over fitted count (OF), and the average model size (MS) are used to evaluate the performances of different methods.

Model **I**. We first consider the following model

$$Y = \begin{cases} \text{sign}(X_1 + X_p)\exp(X_2 + X_{p-1}) + \epsilon_1, & W = 0; \\ \text{sign}(X_1 - X_p)\exp(X_2 + X_{p-1}) + \epsilon_2, & W = 1. \end{cases}$$

$\mathbf{X} = (X_1, \ldots, X_p) \sim N(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma} = (\sigma_{ij}) = \rho^{|i-j|}$, and $\epsilon_i \sim N(0, 0.2)$, for $i = 1, 2$. We considered uncorrelated predictors ($\rho = 0$), and correlated predictors with $\rho = 0.5$. W is generated independently with $\mathbf{X}$ from Bernoulli($\frac{1}{2}$) distribution. Hence we have two populations ($W = 2$), and the active predictors are X_1, X_2, X_{p-1} and X_p for both populations. Yu *et al.* (2016) also considered this model with a single population. For uncorrelated predictors case, Table 4.1 showed the great improvement of correct selection rates when the grouping information is considered. For example, when $p = 2000$, our method via SIR and DR both select the correct predictors all the time (CF rate 100%, while the single population method proposed by Yu *et al.* (2016) always underfits. SAVE based methods are expected to fail since for this model the predictors are linked to the response through monotone functions. Table 4.2 tells the same story with correlated predictors.

Model **II**. We then consider a variant of Model I with W being generated from Bernoulli(0.7) distribution, and all the other model configurations are the same as Model I. Table 4.3 reported the simulation results with uncorrelated and correlated predictors for SIR-based methods. We observed a similar trend as that of Model I. The utilization of grouping information has greatly improved the correct selection rates. Unreported simulation results suggest that SAVE-based and DR-based methods provide similar performance as that of Model I.

Table 1. Selection Performances (50 Runs) for Model **I** with $\rho = 0$

p		Multi-SIR	SIR	Multi-$SAVE$	$SAVE$	Multi-DR	DR
100	MS	4	3	6.6	2.24	4.04	9.08
	UF	0	50	35	50	0	46
	CF	50	0	0	0	48	0
	OF	0	0	15	0	2	4
1000	MS	4.06	3	9	2.06	4.12	10.68
	UF	0	50	48	50	0	50
	CF	48	0	0	0	45	0
	OF	2	0	2	0	5	0
2000	MS	4	3	8.6	2.1	4	10.5
	UF	0	50	49	50	0	50
	CF	50	0	0	0	50	0
	OF	0	0	1	0	0	0

Table 2. Selection Performances (50 Runs) for Model **I** with $\rho = 0.5$

p		Multi-SIR	SIR	Multi-$SAVE$	$SAVE$	Multi-DR	DR
100	MS	4.02	3	6.22	2.16	4.02	7.12
	UF	0	50	24	50	0	44
	CF	49	0	0	0	49	0
	OF	1	0	26	0	1	6
1000	MS	4.08	3	8.8	2.06	4.14	9.22
	UF	0	50	24	50	0	49
	CF	47	0	0	0	45	0
	OF	3	0	26	0	5	1
2000	MS	4	3	8.46	2.14	4	9.22
	UF	0	50	49	50	0	48
	CF	50	0	0	0	50	0
	OF	0	0	1	0	0	2

Model **III**. We now consider a model where Y depends on quadratic functions X_1^2, X_2^2, X_{p-1}^2 and X_p^2. $\mathbf{X}$ and ϵ's are generated the same way as in Model I. Due to the model structure, SAVE-based methods are expected to perform well, while SIR-based methods are expected to fail. Table 4.4 and 4.5 report the performances of the multiple group and single

Table 3. Selection Performances (50 Runs) for Model **II** with $W \sim Bin(0.7)$

		$\rho = 0$				$\rho = 0.5$			
p	Method	MS	UF	CF	OF	MS	UF	CF	OF
100	SIR	3.12	44	6	0	3.22	39	11	0
	M-SIR	4	0	50	0	4	0	50	0
1000	SIR	3.04	48	2	0	3.08	46	4	0
	M-SIR	4	0	50	0	4	0	50	0
2000	SIR	3.08	46	4	0	3	50	0	0
	M-SIR	4	0	50	0	4	0	50	0

group selection methods for Model III. Again, the incorporation of grouping information greatly improves the correct selection rates. Also, it seems that DR performs well for both models, as suggested by the literature.

$$Y = \begin{cases} 2X_1^2X_p^2 - 2X_2^2X_{p-1}^2 + \epsilon_1, & W = 0; \\ 2X_1^2X_p^2 + 2X_2^2X_{p-1}^2 + \epsilon_2, & W = 1. \end{cases}$$

Model **IV**. Model IV is generated in a similar way as that of Yu *et al.* (2016). Again, **X**, W and ϵ's are generated the same way as in Model I. As suggested by Yu *et al.* (2016), this model is specially constructed to favor DR-based methods. As shown in Table 4.6 and 4.7, the multiple population selection methods again dominate over the single population selection method. For example, with $p = 1000$ and $\rho = 0.5$, the average model size for DR-based multiple population selection method is 4.06, which is slightly greater than the true model size 4; while the average model size yielded by DR-based single population selection method is 9.14.

$$Y = \begin{cases} X_1^4 - X_p^4 + \exp(0.8X_2 + 0.6X_{p-1}) + \epsilon_1, & W = 0; \\ X_1^4 + X_p^4 + \exp(0.8X_2 - 0.6X_{p-1}) + \epsilon_2, & W = 1. \end{cases}$$

Table 4. Selection Performances (50 Runs) for Model **III** with $\rho = 0$

p		Multi-SIR	SIR	Multi-$SAVE$	$SAVE$	Multi-DR	DR
100	MS	5.84	6.54	4.18	6.18	4.22	8.94
	UF	50	50	6	28	13	19
	CF	0	0	36	2	29	0
	OF	0	0	8	20	8	31
1000	MS	4.38	6.82	3.84	4.82	4.1	10.1
	UF	50	50	26	43	23	40
	CF	0	0	22	1	20	0
	OF	0	0	2	6	7	10
2000	MS	4.2	6.42	4.02	4.76	3.62	10.54
	UF	50	50	32	48	33	39
	CF	0	0	15	0	17	0
	OF	0	0	3	2	0	11

Table 5. Selection Performances (50 Runs) for Model **III** with $\rho = 0.5$

p		Multi-SIR	SIR	Multi-$SAVE$	$SAVE$	Multi-DR	DR
100	MS	5.9	5.68	4.02	5.02	4.08	10.92
	UF	50	50	12	29	6	23
	CF	0	0	31	4	39	0
	OF	0	0	7	17	5	27
1000	MS	4.64	6.62	4.2	4.54	4.22	9.78
	UF	50	50	15	44	20	46
	CF	0	0	25	2	21	0
	OF	0	0	10	4	9	4
2000	MS	4.06	6.5	3.86	4.22	4	9.42
	UF	50	50	35	49	34	47
	CF	0	0	14	0	6	0
	OF	0	0	1	1	10	3

Model **V**. Model V is generated as the following:

$$Y = \begin{cases} \mathrm{sign}(X_1 + X_p)\exp(X_2 + X_{p-1}) + \epsilon_1, & W = 0; \\ \exp(X_2 + X_{p-1}) + \epsilon_2, & W = 1. \end{cases}$$

Table 6. Selection Performances (50 Runs) for Model **IV** with $\rho = 0$

p		Multi-SIR	SIR	Multi-$SAVE$	$SAVE$	Multi-DR	DR
100	MS	3.44	2.84	4.34	4.72	4.08	10.72
	UF	50	50	44	45	4	32
	CF	0	0	5	4	37	0
	OF	0	0	1	1	9	18
1000	MS	2.34	2.38	4.94	4.26	4.12	10.24
	UF	50	50	50	50	12	45
	CF	0	0	0	0	25	0
	OF	0	0	0	0	13	5
2000	MS	2.12	2.14	5.06	4.2	3.6	9.8
	UF	50	50	49	50	27	47
	CF	0	0	1	0	20	0
	OF	0	0	0	0	3	3

Table 7. Selection Performances (50 Runs) for Model **IV** with $\rho = 0.5$

p		Multi-SIR	SIR	Multi-$SAVE$	$SAVE$	Multi-DR	DR
100	MS	3.64	3.62	3.88	4.8	4.14	9.16
	UF	47	50	34	42	4	37
	CF	2	0	12	5	38	0
	OF	1	0	4	3	8	13
1000	MS	2.32	3.04	5	4.56	4.06	9.14
	UF	50	50	46	49	11	47
	CF	0	0	2	1	29	0
	OF	0	0	2	0	10	3
2000	MS	2.18	3.2	4.58	4.34	3.74	8.82
	UF	50	50	50	50	23	49
	CF	0	0	0	0	23	0
	OF	0	0	0	0	4	21

The $\mathbf{X}$, W, and ϵ_i, $i = 1, 2$ are all generated the same as in Model I. Notice that population one and two now has different active sets: X_1, X_2, X_{p-1}, X_p for population one; and X_2, X_{p-1} for population two, though the active set in population one consists of that of population

Table 8. Selection Performances (50 Runs) for Model **V**

		$\rho = 0$				$\rho = 0.5$			
p	Method	MS	UF	CF	OF	MS	UF	CF	OF
100	SIR	3.48	44	5	1	3.3	44	5	1
	M-SIR	5.02	0	39	11	4.2	0	41	9
1000	SIR	3.28	49	0	1	3.22	49	1	0
	M-SIR	4.04	1	46	3	4	1	48	1
2000	SIR	3.22	50	0	0	3.08	50	0	0
	M-SIR	4	1	48	1	4	3	45	2

two. Table 4.8 showed that our multiple population selection method greatly improves the correct fit rate. For example, with $p = 2000$ and $\rho = 0$, the correct fit rate is $48/50$ for selections via multi-SIR, and $0/50$ for SIR-based method.

Model **VI**. Model VI is considered to investigate the performance of our method when each population consists of its unique active variables. Model VI is generated similarly as Model I except for Y, which is generated as:

$$Y = \begin{cases} \operatorname{sign}(X_1 + X_p)\exp(X_3 + X_{p-2}) + \epsilon_1, & W = 0; \\ \operatorname{sign}(X_2 + X_{p-1})\exp(X_3 + X_{p-2}) + \epsilon_2, & W = 1. \end{cases}$$

Hence the active sets for population one and two are X_1, X_3, X_{p-2}, X_p and X_2, X_3, X_{p-1}, X_p respectively. The current model size is 6. Table 4.9 showed the our multiple population selection method still outperforms single population selection method. For example, when $p = 2000$ and $\rho = 0$, the average model size for our method is 5.38, which is much closer to the true model size (6) comparing to 3.42 from the single population method.

Table 9. Selection Performances (50 Runs) for Model **VI**

		$\rho = 0$				$\rho = 0.5$			
p	Method	MS	UF	CF	OF	MS	UF	CF	OF
100	SIR	5.48	24	25	1	4.36	50	0	0
	M-SIR	5.86	7	43	0	5.7	14	36	0
1000	SIR	3.86	49	1	0	4.02	50	0	0
	M-SIR	5.44	27	23	0	5.34	29	21	0
2000	SIR	3.42	50	0	0	4	50	0	0
	M-SIR	5.38	30	20	0	4.98	37	13	0

5. CONCLUSIONS AND DISCUSSION

Sufficient dimension reduction provides a general framework for model-free variable selections. However, few of the current variable selection methods consider the grouping information when dealing with data from multi-populations. In this paper, we propose a model-free variable selection method for $n < p$ multi-population data, which fully utilizes the grouping information. Simulation studies show that our method provides superior performance comparing to those ignoring the grouping information.

APPENDIX

Proof of Proposition 1:

Assume $\mathrm{Var}\{\mathrm{E}(\mathbf{Z}_w|Y \in s_{h_w})\}$ has q_w nonzero eigenvalues $\lambda_{w1} \geq \cdots \geq \lambda_{wq_w}$ with corresponding eigenvectors $\boldsymbol{\eta}_{w1}, \ldots, \boldsymbol{\eta}_{wq_w}$, where $w = 1, \ldots, K$. Let $\boldsymbol{\beta}_{wi} = \boldsymbol{\Sigma}_w^{-1/2}\boldsymbol{\eta}_{wi}$ for $i = 1, \ldots, q_w$ and $w = 1, \ldots, K$.

Note that $\mathbf{M} = \sum_{w=1}^{K}\sum_{i=1}^{q_w} p_w\lambda_{wi}\boldsymbol{\eta}_{wi}\boldsymbol{\eta}_{wi}^{\top} = \sum_{w=1}^{K}\sum_{i=1}^{q_w} p_w\lambda_{wi}\boldsymbol{\Sigma}_w^{1/2}\boldsymbol{\beta}_{wi}\boldsymbol{\beta}_{wi}^{\top}\boldsymbol{\Sigma}_w^{1/2}$, we have $\mathrm{tr}(\mathbf{M}) = \mathrm{tr}(\sum_{w=1}^{K} p_w\boldsymbol{\Sigma}_w \sum_{i=1}^{q_w}\lambda_{wi}\boldsymbol{\beta}_{wi}\boldsymbol{\beta}_{wi}^{\top})$.

Under the linearity condition for $\mathbf{X}$ within each group, we know $\boldsymbol{\beta}_{wi} \in \mathcal{S}_{Y_w|\mathbf{X}_w}$. Define $\boldsymbol{\beta}_{wi,\mathcal{A}} = \{\beta_{wi,j} : j \in \mathcal{A}\}$ and $\boldsymbol{\beta}_{wi,\mathcal{A}^c} = \{\beta_{wi,j} : j \in \mathcal{A}^c\}$, where $w = 1, \ldots, K$. Since $Y \perp\!\!\!\perp \mathbf{X}_{\mathcal{A}^c}|(\mathbf{X}_{\mathcal{A}}, W)$, $\boldsymbol{\beta}_{wi,\mathcal{A}^c} = \mathbf{0}$ for all $w \in \{1, \ldots, K\}$. Therefore, $\mathrm{tr}(\mathbf{M})$ can be rewritten as $\mathrm{tr}(\sum_{w=1}^{K} p_w\boldsymbol{\Sigma}_{w,\mathcal{A}}\sum_{i=1}^{q_w}\lambda_{wi}\boldsymbol{\beta}_{wi,\mathcal{A}}\boldsymbol{\beta}_{wi,\mathcal{A}}^{\top})$

Recall that $\mathcal{A} = \{1, \ldots, q\}$, so

$$\mathrm{Var}(\mathrm{E}(\mathbf{X}|Y)|W = w) = \boldsymbol{\Sigma}_w(\sum_{i=1}^{q_w}\lambda_{wi}\boldsymbol{\beta}_{wi}\boldsymbol{\beta}_{wi}^{\top})\boldsymbol{\Sigma}_w$$

$$= \begin{pmatrix} \boldsymbol{\Sigma}_{w,\mathcal{A}} & \boldsymbol{\Sigma}_{w,\mathcal{A}\mathcal{A}^c} \\ \boldsymbol{\Sigma}_{w,\mathcal{A}^c\mathcal{A}} & \boldsymbol{\Sigma}_{w,\mathcal{A}^c} \end{pmatrix} \begin{pmatrix} \sum_{i=1}^{q_w}\lambda_{wi}\boldsymbol{\beta}_{wi,\mathcal{A}}\boldsymbol{\beta}_{wi,\mathcal{A}}^{\top} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{\Sigma}_{w,\mathcal{A}} & \boldsymbol{\Sigma}_{w,\mathcal{A}\mathcal{A}^c} \\ \boldsymbol{\Sigma}_{w,\mathcal{A}^c\mathcal{A}} & \boldsymbol{\Sigma}_{w,\mathcal{A}^c} \end{pmatrix},$$

where $\boldsymbol{\Sigma}_{w,\mathcal{A}} = \mathrm{Var}(\mathbf{X}_{\mathcal{A}}|W = w)$, $\boldsymbol{\Sigma}_{w,\mathcal{A}^c} = \mathrm{Var}(\mathbf{X}_{\mathcal{A}^c}|W = w)$, and $\boldsymbol{\Sigma}_{w,\mathcal{A}\mathcal{A}^c} = \mathrm{Cov}(\mathbf{X}_{\mathcal{A}}, \mathbf{X}_{\mathcal{A}^c}| W = w)$. Hence,

$$\mathrm{Var}(\mathrm{E}(\mathbf{X}_{\mathcal{A}}|Y)|W = w) = \boldsymbol{\Sigma}_{w,\mathcal{A}}\sum_{i=1}^{q_w}\lambda_{wi}\boldsymbol{\beta}_{wi,\mathcal{A}}\boldsymbol{\beta}_{wi,\mathcal{A}}^{\top}\boldsymbol{\Sigma}_{w,\mathcal{A}},$$

and

$$\mathbf{M}_{\mathcal{A}} = \sum_{w=1}^{K} p_w\boldsymbol{\Sigma}_{w,\mathcal{A}}^{-1/2}\mathrm{Cov}(\mathrm{E}(\mathbf{X}_{\mathcal{A}}|Y)|w)\boldsymbol{\Sigma}_{w,\mathcal{A}}^{-1/2} = \sum_{w=1}^{K} p_w\boldsymbol{\Sigma}_{w,\mathcal{A}}^{1/2}\sum_{i=1}^{q_w}\lambda_{wi}\boldsymbol{\beta}_{wi,\mathcal{A}}\boldsymbol{\beta}_{wi,\mathcal{A}}^{\top}\boldsymbol{\Sigma}_{w,\mathcal{A}}^{1/2}.$$

Based on these results, we have $\mathrm{tr}(\mathbf{M}_{\mathcal{A}}) = \mathrm{tr}(\mathbf{M}_{\mathcal{I}})$. Similarly, we can prove $\mathrm{tr}(\mathbf{M}_{\mathcal{F}}) = \mathrm{tr}(\mathbf{M}_{\mathcal{A}})$ for any $\mathcal{F}$ such that $\mathcal{A} \subset \mathcal{F}$. □

Proof of Theorem 1:

i) Since $\mathcal{A} \subseteq \mathcal{F}$, $\mathcal{A} \subseteq \mathcal{F} \cup j$. From Proposition 1, it is easy to show that $\mathrm{tr}(\mathbf{M}_{\mathcal{F}\cup j}) - \mathrm{tr}(\mathbf{M}_{\mathcal{F}}) = \mathrm{tr}(\mathbf{M}_{\mathcal{A}}) - \mathrm{tr}(\mathbf{M}_{\mathcal{A}}) = 0$.

ii) If the subset linearity condition holds in each group, then $X_{wj|\mathcal{F}} = X_{wj} - \mathrm{E}(X_{wj}|\mathbf{X}_{w\mathcal{F}}) = X_{wj} - \mathbf{\Sigma}_{w,j\mathcal{F}}^{T}\mathbf{\Sigma}_{w\mathcal{F}}^{-1}\mathbf{X}_{w\mathcal{F}}$ for any $w \in \{1, \ldots, K\}$, where $\mathbf{\Sigma}_{w,j\mathcal{F}} = \mathrm{Cov}(X_j, \mathbf{X}_{\mathcal{F}}| W = w)$. We construct two matrices $\mathbf{P}_w$ and $\mathbf{V}_w$ as

$$\mathbf{P}_w = \begin{pmatrix} \mathbf{I}_{|\mathcal{F}|} & \mathbf{0} \\ -\mathbf{\Sigma}_{w,j\mathcal{F}}^{T}\mathbf{\Sigma}_{w\mathcal{F}}^{-1} & 1 \end{pmatrix} \textit{ and } \mathbf{V}_w = \begin{pmatrix} \mathbf{\Sigma}_{w\mathcal{F}} & \mathbf{0} \\ \mathbf{0} & \sigma_{w,j|\mathcal{F}}^2 \end{pmatrix}.$$

where $|\mathcal{F}|$ is the cardinality of $\mathcal{F}$ and $\sigma_{w,j|\mathcal{F}}^2$ is $\sigma_{j|\mathcal{F}}^2$ in group w. Note that $\mathrm{Cov}(\mathbf{X}_{w\mathcal{F}}, X_{wj|\mathcal{F}})$ $= 0$, then we have $\mathrm{Var}(\mathbf{P}_w\mathbf{X}_{w,\mathcal{F}\cup j}) = \mathbf{P}_w\mathbf{\Sigma}_{w,\mathcal{F}\cup j}\mathbf{P}_w^{\top} = \mathbf{V}_w$ and $\mathbf{\Sigma}_{w,\mathcal{F}\cup j} = \mathbf{P}_w^{\top}\mathbf{V}_w^{-1}\mathbf{P}_w$.

We can rewrite $\mathbf{M}_{\mathcal{F}\cup j}$ as

$$\begin{aligned}
\mathbf{M}_{\mathcal{F}\cup j} &= \mathrm{E}[\mathrm{Cov}(\mathrm{E}(\mathbf{Z}_{\mathcal{F}\cup j}|Y)|W)] \\
&= \mathrm{E}[\mathbf{\Sigma}_{w,\mathcal{F}\cup j}^{-1/2}\mathrm{Cov}(\mathrm{E}(\mathbf{X}_{\mathcal{F}\cup j}|Y)|W)\mathbf{\Sigma}_{w,\mathcal{F}\cup j}^{-1/2}] \\
&= \mathrm{E}[\mathbf{P}_w^{\top}\mathbf{V}_w^{-1/2}\mathbf{P}_w\mathrm{Cov}(\mathrm{E}(\mathbf{X}_{,\mathcal{F}\cup j}|Y)|W)\mathbf{P}_w^{\top}\mathbf{V}_w^{-1/2}\mathbf{P}_w] \\
&= \sum_{w=1}^{K} p_w\mathbf{P}_w^{\top}\mathbf{V}_w^{-1/2}\Big(\sum_{h=1}^{Hw} p_{hw}\mathbf{P}_w\mathbf{U}_{hw,\mathcal{F}\cup j}\mathbf{U}_{hw,\mathcal{F}\cup j}^{\top}\mathbf{P}_w^{\top}\Big)\mathbf{V}_w^{-1/2}\mathbf{P}_w
\end{aligned}$$

Because $\mathbf{P}_w\mathbf{U}_{hw,\mathcal{F}\cup j} = (\mathbf{U}^\top_{hw,\mathcal{F}}, \mathrm{E}(X_{j|\mathcal{F}}|Y_w \in s_{hw}, W = w))^\top$, then we have

$$\begin{aligned}
\mathrm{tr}(\mathbf{M}_{\mathcal{F}\cup j}) &= \mathrm{tr}\big(\sum_{w=1}^{K} p_w \mathbf{P}_w^\top \mathbf{V}_w^{-1/2}\big(\sum_{h=1}^{Hw} p_{hw}\mathbf{P}_w\mathbf{U}_{hw,\mathcal{F}\cup j}\mathbf{U}^\top_{hw,\mathcal{F}\cup j}\mathbf{P}_w^\top\big)\mathbf{V}_w^{-1/2}\mathbf{P}_w\big)\\
&= \mathrm{tr}\big(\sum_{w=1}^{K} p_w \mathbf{V}_w^{-1}\big(\sum_{h=1}^{Hw} p_{hw}\mathbf{P}_w\mathbf{U}_{hw,\mathcal{F}\cup j}\mathbf{U}^\top_{hw,\mathcal{F}\cup j}\mathbf{P}_w^\top\big)\big)\\
&= \mathrm{tr}\big(\sum_{w=1}^{K} p_w \mathbf{\Sigma}^{-1}_{w\mathcal{F}}\big(\sum_{h=1}^{Hw} p_{hw}\mathbf{U}_{\mathcal{F}w,h}\mathbf{U}^T_{\mathcal{F}w,h}\big)\big)\\
&\quad + \sum_{w=1}^{K}\sum_{h=1}^{Hw} p_w p_{hw}\mathrm{E}^2(X_{j|\mathcal{F}}/\sigma_{j|\mathcal{F}}|Y_w \in s_{hw}, W = w)
\end{aligned}$$

Hence, $\mathrm{tr}(\mathbf{M}_{\mathcal{F}\cup j}) - \mathrm{tr}(\mathbf{M}_{\mathcal{F}}) = \sum_{w=1}^{K} p_w\big(\sum_{h=1}^{Hw} p_{hw}\boldsymbol{\gamma}^2_{j|w\mathcal{F},hw}\big)$ □

Proof of Theorem 2:

For any $w \in 1, \ldots, K$, we define F_w as the joint distribution of $(\mathbf{X}_w, Y_w)$ and F_{nw} as the empirical distribution for random sample $(Y_{wj}, \mathbf{X}_{wj})$, $j = 1, \ldots, n_w$. Let $\mathcal{G}$ be a real or matrix valued functional. Based on Frechet derivative and the regularity conditions in Fernholz (1983), we know that $\mathcal{G}(F_{nw})$ satisfies

$$\mathcal{G}(F_{nw}) = \mathcal{G}(F_w) + \mathrm{E}_n[\mathcal{G}^\star(F_w)] + O_p(n_w^{-1}), \tag{5.1}$$

where $\mathcal{G}(F_w)$ is fixed for each group, and $\mathrm{E}_n[\mathcal{G}^\star(F_w)] = O_p(n_w^{-1/2})$ as $\mathrm{E}[\mathcal{G}^\star(F_w)] = 0$. Let $R_{hw} = I(Y_w \in s_{hw})$, $\mu_{j,hw} = \mathrm{E}(X_j|Y_w \in s_{hw}, W = w)$ and $\nu_{wj|\mathcal{F}} = \mathbf{\Sigma}^{-1}_{w,\mathcal{F}}\mathbf{\Sigma}^\top_{w,j\mathcal{F}}$. To prove Theorem 2, we need the results in Lemma 1 in the following.

Lemma 1 *If the conditions in 2 holds and H_o is true, then $\hat{\mathbf{\Sigma}}_{w,\mathcal{F}}$, $\hat{\mathbf{\Sigma}}^{-1}_{w,\mathcal{F}}$, $\hat{\mathbf{U}}_{\mathcal{F}w,h}$, $\hat{\nu}_{wj|\mathcal{F}}$,$\mu_{j,hw}$ and $\hat{\boldsymbol{\gamma}}_{j|\mathcal{F}w,hw}$ have expansions in the form (5.1) with $\mathbf{\Sigma}_{w,\mathcal{F}}$, $\mathbf{\Sigma}^{-1}_{w,\mathcal{F}}$, $\mathbf{U}_{\mathcal{F}w,h}$, $\nu_{wj|\mathcal{F}}$, $\hat{\mu}_{j,hw}$ or $\boldsymbol{\gamma}_{j|\mathcal{F}w,hw}$ as substitutes for $\mathcal{G}(F_w)$, and $\mathbf{\Sigma}^\star_{w,\mathcal{F}} = \mathbf{X}_{w,\mathcal{F}}\mathbf{X}^\top_{w,\mathcal{F}}$, $(\mathbf{\Sigma}^{-1}_{w,\mathcal{F}})^\star = -\mathbf{\Sigma}^{-1}_{w,\mathcal{F}}\mathbf{\Sigma}^\star_{w,\mathcal{F}}\mathbf{\Sigma}^{-1}_{w,\mathcal{F}}$, $\mathbf{U}^\star_{\mathcal{F}w,h} = (\mathbf{X}_{w,\mathcal{F}} - \mathbf{U}_{\mathcal{F}w,h})R_{hw}/p_{hw} - \mathbf{X}_{w,\mathcal{F}}$, $\nu^\star_{wj|\mathcal{F}} = \mathbf{\Sigma}^{-1}_{w,\mathcal{F}}\big((X_{wj}|\mathbf{X}_{w\mathcal{F}} - \mathrm{E}((X_{wj}|\mathbf{X}_{w\mathcal{F}})\big) + (\mathbf{\Sigma}^{-1}_{w,\mathcal{F}})^\star\mathrm{E}((X_{wj}|\mathbf{X}_{w\mathcal{F}})$, $\mu^\star_{j,hw} = (X_{w,j} - \mathbf{U}_{jw,h})R_{hw}/p_{hw} - X_{w,j}$ or $\boldsymbol{\gamma}^\star_{j|\mathcal{F}w,hw} = \big(\mu^\star_{j,hw} - (\nu^\star_{wj|\mathcal{F}})^\top\mathbf{U}_{\mathcal{F}w,h} - \nu^\top_{wj|\mathcal{F}}\mathbf{U}^\star_{\mathcal{F}w,h}\big)/\sigma_{w,j|\mathcal{F}}$ as substitutes for $\mathcal{G}^\star(F_w)$*

Since the proof is similar to Yu et al. (2016), we omit the proof for Lemma 1.

Let $\hat{\mathbf{L}}_{j|\mathcal{F},w} = (\hat{p}_w^{1/2}\hat{p}_{\{hw=1\}}^{1/2}\hat{\boldsymbol{\gamma}}_{j|\mathcal{F}w,1}, \ldots, \hat{p}_w^{1/2}\hat{p}_{\{hw=Hw\}}^{1/2}\hat{\boldsymbol{\gamma}}_{j|\mathcal{F}w,Hw})^\top$ and $\hat{\mathbf{L}}_{j|\mathcal{F}} = (\hat{\mathbf{L}}_{j|\mathcal{F},1}^\top, \ldots, \hat{\mathbf{L}}_{j|\mathcal{F},K}^\top)^\top$. Based on Lemma 1, we define $\boldsymbol{\Omega}_{j|\mathcal{F}} = \mathrm{E}\big(\mathbf{L}_{j|\mathcal{F},1}^\star(\mathbf{L}_{j|\mathcal{F},1}^\star)^\top\big)$, $(\mathbf{L}_{j|\mathcal{F},w})^\star = (p_w^{1/2}p_{\{hw=1\}}^{1/2}\boldsymbol{\gamma}_{j|\mathcal{F}w,1}^\star, \ldots, p_w^{1/2}p_{\{hw=Hw\}}^{1/2}\hat{\boldsymbol{\gamma}}_{j|\mathcal{F}w,Hw}^\star)^\top$ and $(\mathbf{L}_{j|\mathcal{F}})^\star = \big((\mathbf{L}_{j|\mathcal{F},1}^\star)^\top, \ldots, (\mathbf{L}_{j|\mathcal{F},K}^\star)^\top\big)^\top$. Then we have $T_{j|\mathcal{F}} = n(\hat{\mathbf{L}}_{j|\mathcal{F}})^\top\hat{\mathbf{L}}_{j|\mathcal{F}}$. Under H_0, we have

$$\hat{\mathbf{L}}_{j|\mathcal{F}} = \mathbf{L}_{j|\mathcal{F}} + \mathrm{E}_n\big((\mathbf{L}_{j|\mathcal{F}})^\star\big) + o_p(n^{-1/2}),$$

Then the result in Theorem 2 follows directly. □

Proof of Proposition 2:

Without loss of generality, we assume that $(\mathbf{X}_{\mathcal{F}}^\top, X_j)$ are the first $|\mathcal{F}| + 1$ elements of $\mathbf{X}^T$ in the proof. Recall that $\mathrm{Var}(\mathrm{E}(\mathbf{X}|Y)|W = w) = \boldsymbol{\Sigma}_w(\sum_{i=1}^{q_w}\lambda_{wi}\boldsymbol{\beta}_{wi}\boldsymbol{\beta}_{wi}^\top)\boldsymbol{\Sigma}_w$ and $\mathrm{tr}(\mathbf{M}_{\mathcal{F}\cup j}) - \mathrm{tr}(\mathbf{M}_{\mathcal{F}}) = \sum\limits_{w=1}^{K} p_w\big(\sum\limits_{h=1}^{Hw} p_{hw}\boldsymbol{\gamma}_{j|\mathcal{F}w,hw}^2\big)$, then we have

$$\begin{aligned}
&\sigma_{w,j|\mathcal{F}}^2\big(\sum_{h=1}^{Hw} p_{hw}\boldsymbol{\gamma}_{j|\mathcal{F}w,hw}^2\big) = \mathrm{Var}(\mathrm{E}(X_{j|\mathcal{F}}|\mathbf{Y})|W = w)\\
=&\big(-\boldsymbol{\Sigma}_{w,j\mathcal{F}}\boldsymbol{\Sigma}_{w\mathcal{F}}^{-1}, 1\big)\mathbf{A}\mathrm{Var}(\mathrm{E}(\mathbf{X}|Y)|W = w)\mathbf{A}^\top\big(-\boldsymbol{\Sigma}_{w,j\mathcal{F}}\boldsymbol{\Sigma}_{w\mathcal{F}}^{-1}, 1\big)^\top \qquad (5.2)\\
=&\big(-\boldsymbol{\Sigma}_{w,j\mathcal{F}}\boldsymbol{\Sigma}_{w\mathcal{F}}^{-1}, 1\big)\mathbf{A}\boldsymbol{\Sigma}_w\big(\sum_{i=1}^{q_w}\lambda_{wi}\boldsymbol{\beta}_{wi}\boldsymbol{\beta}_{wi}^\top\big)\boldsymbol{\Sigma}_w\mathbf{A}^\top\big(-\boldsymbol{\Sigma}_{w,j\mathcal{F}}\boldsymbol{\Sigma}_{w\mathcal{F}}^{-1}, 1\big)^\top
\end{aligned}$$

where $\mathbf{A} = (\mathbf{I}_{|\mathcal{F}|+1}, \mathbf{0}_{(|\mathcal{F}|+1)(p-|\mathcal{F}|-1)})$. Note that $\big(\boldsymbol{\Sigma}_{w,j\mathcal{F}} - \boldsymbol{\Sigma}_{w,j\mathcal{F}}\boldsymbol{\Sigma}_{w\mathcal{F}}^{-1}\boldsymbol{\Sigma}_{w\mathcal{F}}\big) = 0$, then we obtain

$$\big(-\boldsymbol{\Sigma}_{w,j\mathcal{F}}\boldsymbol{\Sigma}_{w\mathcal{F}}^{-1}, 1\big)\mathbf{A}\boldsymbol{\Sigma}_w\boldsymbol{\beta}_{wi} = \big(\boldsymbol{\Sigma}_{w,j\mathcal{F}^c} - \boldsymbol{\Sigma}_{w,j\mathcal{F}}\boldsymbol{\Sigma}_{w\mathcal{F}}^{-1}\boldsymbol{\Sigma}_{w\mathcal{F}\mathcal{F}^c}, 1\big)\boldsymbol{\beta}_{wi,\mathcal{F}^c}$$

Recall that $\boldsymbol{\beta}_{wi,\mathcal{A}^c} = \mathbf{0}$ for all $w \in \{1, \ldots, K\}$. Let $\tilde{\mathcal{F}} = \mathcal{F}^c \cap \mathcal{A}$, then it follows

$$\big(-\boldsymbol{\Sigma}_{w,j\mathcal{F}}\boldsymbol{\Sigma}_{w\mathcal{F}}^{-1}, 1\big)\mathbf{A}\boldsymbol{\Sigma}_w\boldsymbol{\beta}_{wi} = \big(\boldsymbol{\Sigma}_{wj\tilde{\mathcal{F}}} - \boldsymbol{\Sigma}_{w,j\mathcal{F}}\boldsymbol{\Sigma}_{w\mathcal{F}}^{-1}\boldsymbol{\Sigma}_{w\mathcal{F}\tilde{\mathcal{F}}}, 1\big)\boldsymbol{\beta}_{wi,\tilde{\mathcal{F}}}$$

From this equation and 5.2, we can obtain that

$$\sigma_{w,j|\mathcal{F}}\big(\sum_{h=1}^{Hw} p_{hw}\boldsymbol{\gamma}^2_{j|\mathcal{F}w,hw}\big) = \sum_{i=1}^{q_w} \lambda_{wi}\{\big(\mathbf{\Sigma}_{wj\tilde{\mathcal{F}}} - \mathbf{\Sigma}_{w,j\mathcal{F}}\mathbf{\Sigma}^{-1}_{w\mathcal{F}}\mathbf{\Sigma}_{w\mathcal{F}\tilde{\mathcal{F}}}, 1\big)\boldsymbol{\beta}_{wi,\tilde{\mathcal{F}}}\}^2$$

Note that

$$\begin{aligned}&\sum_{j\in\tilde{\mathcal{F}}}\{\big(\mathbf{\Sigma}_{wj\tilde{\mathcal{F}}} - \mathbf{\Sigma}_{w,j\mathcal{F}}\mathbf{\Sigma}^{-1}_{w\mathcal{F}}\mathbf{\Sigma}_{w\mathcal{F}\tilde{\mathcal{F}}}, 1\big)\boldsymbol{\beta}_{wi,\tilde{\mathcal{F}}}\}^2\\ =&\boldsymbol{\beta}_{wi,\tilde{\mathcal{F}}}{}^\top\big(\mathbf{\Sigma}_{w,\tilde{\mathcal{F}}} - \mathbf{\Sigma}_{w,\tilde{\mathcal{F}}\ \mathcal{F}}\mathbf{\Sigma}^{-1}_{w\mathcal{F}}\mathbf{\Sigma}_{w\mathcal{F}\tilde{\mathcal{F}}}, 1\big)\boldsymbol{\beta}_{wi,\tilde{\mathcal{F}}}\end{aligned}$$

and

$$\begin{aligned}&\lambda_{min}\big(\mathbf{\Sigma}_{w,\tilde{\mathcal{F}}} - \mathbf{\Sigma}_{w,\tilde{\mathcal{F}}\ \mathcal{F}}\mathbf{\Sigma}^{-1}_{w\mathcal{F}}\mathbf{\Sigma}_{w\mathcal{F}\tilde{\mathcal{F}}}, 1\big)\\ =&\lambda^{-1}_{max}\{\big(\mathbf{\Sigma}_{w,\tilde{\mathcal{F}}} - \mathbf{\Sigma}_{w,\tilde{\mathcal{F}}\ \mathcal{F}}\mathbf{\Sigma}^{-1}_{w\mathcal{F}}\mathbf{\Sigma}_{w\mathcal{F}\tilde{\mathcal{F}}}, 1\big)^{-1}\} \geq \lambda^{-1}_{max}(\mathbf{\Sigma}_w) = \lambda_{min}(\mathbf{\Sigma}_w)\end{aligned}$$

for any $w \in \{1, \ldots, K\}$, then it follows

$$\begin{aligned}&\max_{j\in\mathcal{F}^c\cap\mathcal{A}} \sigma_{w,j|\mathcal{F}}\big(\sum_{h=1}^{Hw} p_{hw}\boldsymbol{\gamma}^2_{j|\mathcal{F}w,hw}\big)\\ \geq&|\mathcal{F}^c\cap\mathcal{A}|^{-1}\sum_{j\in\tilde{\mathcal{F}}}\{\big(\mathbf{\Sigma}_{wj\tilde{\mathcal{F}}} - \mathbf{\Sigma}_{w,j\mathcal{F}}\mathbf{\Sigma}^{-1}_{w\mathcal{F}}\mathbf{\Sigma}_{w\mathcal{F}\tilde{\mathcal{F}}}, 1\big)\boldsymbol{\beta}_{wi,\tilde{\mathcal{F}}}\}^2\\ =&|\mathcal{F}^c\cap\mathcal{A}|^{-1}\sum_{i=1}^{q_w}\lambda_{wi}\boldsymbol{\beta}_{wi,\tilde{\mathcal{F}}}{}^\top\big(\mathbf{\Sigma}_{w,\tilde{\mathcal{F}}} - \mathbf{\Sigma}_{w,\tilde{\mathcal{F}}\ \mathcal{F}}\mathbf{\Sigma}^{-1}_{w\mathcal{F}}\mathbf{\Sigma}_{w\mathcal{F}\tilde{\mathcal{F}}}, 1\big)\boldsymbol{\beta}_{wi,\tilde{\mathcal{F}}}\\ \geq&|\mathcal{F}^c\cap\mathcal{A}|^{-1}\sum_{i=1}^{q_w}\lambda_{wi}\lambda_{min}\big(\mathbf{\Sigma}_{w,\tilde{\mathcal{F}}} - \mathbf{\Sigma}_{w,\tilde{\mathcal{F}}\ \mathcal{F}}\mathbf{\Sigma}^{-1}_{w\mathcal{F}}\mathbf{\Sigma}_{w\mathcal{F}\tilde{\mathcal{F}}}, 1\big)\boldsymbol{\beta}_{wi,\tilde{\mathcal{F}}}{}^\top\boldsymbol{\beta}_{wi,\tilde{\mathcal{F}}}\\ \geq&\lambda_{w,q_w}\lambda_{min}(\mathbf{\Sigma}_w)^2\beta_{min}\end{aligned}$$

Because $\sigma_{w,j|\mathcal{F}} \leq \sigma_{j|\mathcal{F}} \leq \mathrm{Var}(X_j) \leq \lambda_{max}$, then

$$\begin{aligned}&\max_{j\in\mathcal{F}^c\cap\mathcal{A}}\big(\mathrm{tr}(\mathbf{M}_{\mathcal{F}\cup j}) - \mathrm{tr}(\mathbf{M}_{\mathcal{F}})\big) = \max_{j\in\mathcal{F}^c\cap\mathcal{A}}\sum_{w=1}^{K} p_w\big(\sum_{h=1}^{Hw} p_{hw}\boldsymbol{\gamma}^2_{j|\mathcal{F}w,hw}\big)\\ \geq&\sum_{w=1}^{K} p_w\sigma^{-2}_{w,j|\mathcal{F}}\max_{j\in\mathcal{F}^c\cap\mathcal{A}}\sigma_{w,j|\mathcal{F}}\big(\sum_{h=1}^{Hw} p_{hw}\boldsymbol{\gamma}^2_{j|\mathcal{F}w,hw}\big)\\ \geq&\sum_{w=1}^{K} p_w\sigma^{-2}_{w,j|\mathcal{F}}\lambda_{w,q_w}\lambda_{min}(\mathbf{\Sigma}_w)^2\beta_{min} \geq \lambda_q\lambda_{min}\lambda^{-1}_{max}\beta_{min} \quad \square\end{aligned}$$

Proof of Theorem 3:

i) Let $\Delta = \alpha n^{-\theta} - n^{-1}c_1 > 0$. Since t $0 < c_1 < (1/2)\alpha n^{1-\theta}$, we have $\Delta = O_p(n^{-\theta})$. Because $\left(\mathrm{tr}(\widehat{\mathbf{M}}_{\mathcal{F}\cup j}) - \mathrm{tr}(\widehat{\mathbf{M}}_{\mathcal{F}})\right) - \left(\mathrm{tr}(\mathbf{M}_{\mathcal{F}\cup j}) - \mathrm{tr}(\mathbf{M}_{\mathcal{F}})\right) = O_p(n^{-1/2})$ as $\mathcal{F}^c \cap \mathcal{A} \neq \varnothing$ and $0 < c_1 < 1/2$,

$$\max_{\mathcal{F}:\mathcal{F}^c\cap\mathcal{A}\neq\varnothing}\max_{j\in\mathcal{F}^c\cap\mathcal{A}}\left[\left(\mathrm{tr}(\widehat{\mathbf{M}}_{\mathcal{F}\cup j}) - \mathrm{tr}(\widehat{\mathbf{M}}_{\mathcal{F}})\right) - \left(\mathrm{tr}(\mathbf{M}_{\mathcal{F}\cup j}) - \mathrm{tr}(\mathbf{M}_{\mathcal{F}})\right)\right] < \Delta$$

with probability 1, as n goes to infinity. Hence,

$$\begin{aligned}
&\min_{\mathcal{F}:\mathcal{F}^c\cap\mathcal{A}\neq\varnothing}\max_{j\in\mathcal{F}^c\cap\mathcal{A}}\left(\mathrm{tr}(\widehat{\mathbf{M}}_{\mathcal{F}\cup j}) - \mathrm{tr}(\widehat{\mathbf{M}}_{\mathcal{F}})\right)\\
>&\min_{\mathcal{F}:\mathcal{F}^c\cap\mathcal{A}\neq\varnothing}\max_{j\in\mathcal{F}^c\cap\mathcal{A}}\left[\left(\mathrm{tr}(\mathbf{M}_{\mathcal{F}\cup j}) - \mathrm{tr}(\mathbf{M}_{\mathcal{F}})\right)\right.\\
&-\max_{\mathcal{F}:\mathcal{F}^c\cap\mathcal{A}\neq\varnothing}\max_{j\in\mathcal{F}^c\cap\mathcal{A}}\left[\left(\mathrm{tr}(\widehat{\mathbf{M}}_{\mathcal{F}\cup j}) - \mathrm{tr}(\widehat{\mathbf{M}}_{\mathcal{F}})\right) - \left(\mathrm{tr}(\mathbf{M}_{\mathcal{F}\cup j}) - \mathrm{tr}(\mathbf{M}_{\mathcal{F}})\right)\right]\\
>&\alpha n^{-\theta} - \Delta = n^{-1}c_1
\end{aligned}$$

It is easy to obtain that $\lim_{n\to\infty}\Pr(\min_{\mathcal{F}:\mathcal{F}^c\cap\mathcal{A}\neq\varnothing}\max_{j\in\mathcal{F}^c\cap\mathcal{A}} T_{j|\mathcal{F}} > c_1) = 1$.

ii) It is obvious that $\mathcal{A} \subset \mathcal{F}$ as $\mathcal{F}^c \cap \mathcal{A} = \varnothing$. There are two different situations for j. One is $j \in \mathcal{A}$, the other one is $j \in \mathcal{F} \setminus \mathcal{A}$. If $j \in \mathcal{A}$, we can have $T_{j|\{\mathcal{F}\setminus j\}} > (1/2)\alpha n^{1-\theta}$ with probability 1 based on the proof before. If $j \in \mathcal{F} \setminus \mathcal{A}$, we know $T_{j|\{\mathcal{F}\setminus j\}}$ follows a weighted χ^2_1 distribution from Theorem 2. Then $T_{j|\{\mathcal{F}\setminus j\}}$is O_p and asymptotically smaller than $(1/2)\alpha n^{1-\theta}$. Hence, $\min_{j\in\mathcal{F}} T_{j|\{\mathcal{F}\setminus j\}} < c_2 = O_p < An^{1-\theta}$ for $\theta < 1$ and $A > 0$. It follows that $\lim_{n\to\infty}\Pr(\max_{\mathcal{F}:\mathcal{F}^c\cap\mathcal{A}=\varnothing}\min_{j\in\mathcal{F}} T_{j|\{\mathcal{F}\setminus j\}} < c_2) = 1$ □

Proof of Theorem 4:

Let $R_{w,j|\mathcal{F}} = \mathrm{Var}(\mathrm{E}(X_{j|\mathcal{F}}|\mathbf{Y})|W = w)$ and $\widehat{R}_{w,j|\mathcal{F}}$ be the estimate for $R_{w,j|\mathcal{F}}$. We can derive that

$$\mathrm{tr}(\mathbf{M}_{\mathcal{F}\cup j}) - \mathrm{tr}(\mathbf{M}_{\mathcal{F}}) = \sum_{w=1}^{w=k} p_w \sigma^2_{w,j|\mathcal{F}} R_{w,j|\mathcal{F}}$$

and

$$\mathrm{tr}(\widehat{\mathbf{M}}_{\mathcal{F}\cup j}) - \mathrm{tr}(\widehat{\mathbf{M}}_{\mathcal{F}}) = \sum_{w=1}^{w=K} \hat{p}_w \hat{\sigma}^{-2}_{w,j|\mathcal{F}} \widehat{R}_{w,j|\mathcal{F}}.$$

Suppose that $|\mathcal{F}| = O(n^{\theta+\theta_2})$. From Lemma 7 in Yu et al. (2016), we know that$|\widehat{R}_{w,j|\mathcal{F}} - R_{w,j|\mathcal{F}}| \leq D_0|\mathcal{F}|\sqrt{\log p/n}$ with probability tending to 1, where D_0 is some constant. Since $\hat{p}_w - \hat{p}_w = O_P(n^{-1/2})$ and

$$|\hat{p}_w \widehat{R}_{w,j|\mathcal{F}} - p_w R_{w,j|\mathcal{F}}| \leq |\hat{p}_w(\widehat{R}_{w,j|\mathcal{F}} - R_{w,j|\mathcal{F}})| + |(\hat{p}_w - p_w)R_{w,j|\mathcal{F}}|,$$

there exists some constant D_1 such that

$$|\hat{p}_w \widehat{R}_{w,j|\mathcal{F}} - p_w R_{w,j|\mathcal{F}}| \leq D_1|\mathcal{F}|\sqrt{\log p/n},$$

with probability tending to 1. Based on the proof of Lemma 3 in Jiang and Liu (2013) and Lemma 6 in Yu et al. (2016), we have that $|\hat{\sigma}^2_{w,j|\mathcal{F}} - \sigma^2_{w,j|\mathcal{F}}| = O_p(|\mathcal{F}|\sqrt{\log p/n})$. It follows that $\hat{\sigma}^{-2}_{w,j|\mathcal{F}} \geq \sigma^{-2}_{w,j|\mathcal{F}}$. based on the proof of Theorem 5.1 in Yu et al. (2016), we can know that $\Pr(\mathcal{A} \subset \mathcal{F}_{2H\alpha^{-1}An^{\theta+\theta 2}}) \longrightarrow 1$, as $n \longrightarrow \infty$ and $p \longrightarrow \infty$. Define $k_0 = \min_{1\leq k\leq n}\{k : \mathcal{A} \in \mathcal{F}_k\}$, then $k_0 \leq 2H\alpha^{-1}An^{\theta+\theta 2}$. The conclusion is easy to be proved based the proof of Theorem 2 in Wang (2009), and we omit the details. □

REFERENCES

Bondell, H. D. and Li, L., ‘Shrinkage inverse regression estimation for model-free variable selection,’ Journal of the Royal Statistical Society, Ser. B, 2009, **71**, pp. 287–299.

Breiman, L., ‘Better subset regression using the nonnegative garrote,’ Technometrics, 1995, **37**, pp. 373–384.

Candes, E. and Tao, T., ‘2007,’ The Dantzig selector: statistical estimation when p is much larger than n., 2007, **35**, pp. 2313–2351.

Chen, X., Zou, C., and Cook, R. D., ‘Coordinate-independent sparse sufficient dimension reduction and variable selection,’ The Annals of Statistics, 2010, **38**, pp. 3696–3723.

Chiaromonte, F., Cook, R., and Li, B., ‘Sufficient dimension reduction in regressions with categorical predictors,’ The Annals of Statistics, 2002, **30**, pp. 475–497.

Cook, R. and Forzani, B., ‘Likelihood-based sufficient dimension reduction,’ Journal of the American Statistical Association, 2009, **104**, pp. 197–208.

Cook, R. D., *Regression Graphics*, Wiley, New York, 1998.

Cook, R. D., ‘Testing predictor contributions in sufficient dimension reduction,’ The Annals of Statistics, 2004, **32**, pp. 1062–1092.

Cook, R. D. and Weisberg, S., ‘Discussion of “sliced inverse regression for dimension reduction”,’ Journal of the American Statistical Association, 1991, **86**, pp. 328–332.

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

Feng, Z., Wen, X., Yu, Z., and Zhu, L.-X., ‘On partial sufficient dimension reduction with applications to partially linear multi-index models,’ Journal of the American Statistical Association, 2013, **108**, pp. 237–246.

Jiang, B. and Liu, J. S., ‘Sliced inverse regression with variable selection and interaction detection,’ Manuscript, 2013.

Lee, K., Li, B., and Chiaromonte, F., ‘A general theory for nonlinear sufficient dimension reduction: Formulation and estimation,’ The Annals of Statistics, 2013, **6**, pp. 3182–3210.

Li, B., Kim, M. K., and Altman, N., ‘On dimension folding of matrix or array valued statistical objects,’ The Annals of Statistics, 2010, **38**, pp. 1097–1121.

Li, B. and Wang, S., ‘On directional regression for dimension reduction,’ Journal of the American Statistical Association, 2007, **102**, pp. 997–1008.

Li, K. C., ‘Sliced inverse regression for dimension reduction (with discussion),’ Journal of the American Statistical Association, 1991, **86**, pp. 316–342.

Li, L., ‘Sparse sufficient dimension reduction,’ Biometrika, 2007, **94**, pp. 603–613.

Li, L. and Nachtsheim, C., ‘Sparse sliced inverse regression,’ Technometrics, 2006, **48**, pp. 503–510.

Li, L. and Yin, X., ‘Sliced inverse regression with regularization,’ Biometrics, 2008, **64**, pp. 124–131.

Ma, Y. and Zhu, L., ‘A semiparametric approach to dimension reduction,’ Journal of the American Statistical Association, 2012, **107**, pp. 168–179.

Ma, Y. and Zhu, L., ‘Efficient estimation in sufficient dimension reduction,’ The Annals of Statistics, 2013, **41**, pp. 250–268.

Ni, L., Cook, R. D., and Tsai, C. L., ‘A note on shrinkage sliced inverse regression,’ Biometrika, 2005, **92**, pp. 242–247.

Tibshirani, R., ‘Regression shrinkage and selection via the lasso,’ Journal of the Royal Statistical Society, Ser. B, 1996, **58**, pp. 267–288.

Wang, H., ‘Forward regression for ultra-high dimensional variable screening,’ Journal of the American Statistical Association, 2009, **104**, pp. 1512–1524.

Wang, T. and Zhu, L., ‘A distribution-based lasso for a general single-index model,’ Science China Mathematics, 2015, **58**, pp. 109–130.

Wen, X. and Cook, R. D., ‘Optimal sufficient dimension reduction for regressions with categorical predictors,’ Journal of Statistical Planning and Inference, 2007, **137**, pp. 1961–1978.

Wen, X. and Cook, R. D., ‘New approaches to model-free dimension reduction for bivariate regression,’ Journal of Statistical Planning and Inference, 2009, **139**, pp. 734–748.

Xia, Y., Tong, H., Li, W. K., and Zhu, L., ‘An adaptive estimation of dimension reduction space,’ Journal of the Royal Statistical Society, Ser. B, 2002, **64**, pp. 363–410.

Yin, X. and Hilafu, H., ‘Sequential sufficient dimension reduction for large p, small n problems,’ Journal of the Royal Statistical Society, Ser. B, 2015, **77**, pp. 879–892.

Yu, Z., Dong, Y., and Zhu, L., ‘Trace pursuit: a general framework for model-free variable selection,’ Journal of the American Statistical Association, 2016, **111**, pp. 813–821.

Yuan, M. and Lin, Y., ‘Model selection and estimation in regression with grouped variables,’ Journal of the Royal Statistical Society, Ser. B, 2006, **68**, pp. 49–67.

Zhang, C., ‘Nearly unbiased variable selection under minimax concave penalty,’ The Annals of Statistics, 2010, **38**, pp. 894–942.

Zhong, W., Zhang, T., Zhu, M., and Liu, J. S., ‘Correlation pursuit: forward stepwise variable selection for index models,’ Journal of the Royal Statistical Society, Ser. B, 2012, **74**, pp. 849–870.

Zhu, L. P., Wang, T., Zhu, L. X., and Ferré, L., ‘Sufficient dimension reduction through discretization-expectation estimation,’ Biometrika, 2010, **97**, pp. 295–304.

Zou, H., ‘The adaptive lasso and its oracle properties,’ Journal of the American Statistical Association, 2006, **101**, pp. 1418–1429.

II. A MODEL-FREE CONDITIONAL SCREENING APPROACH VIA SUFFICIENT DIMENSION REDUCTION

Lei Huo[1], Xuerong Meggie Wen[1], and Zhou Yu[2]

[1]Department of Mathematics and Statistics,

Missouri University of Science and Technology, MO 65409, U.S.A.

email: wenx@mst.edu

[2]School of Finance and Statistics,

East China Normal University, Shanghai, China

ABSTRACT

In many applications, conditional variable screening arises when researchers have some prior information regarding the importance of certain predictors, such as the treatment effects in biological studies and market risk factors in financial studies. It is natural to consider feature screening methods conditioning on these known important predictors. Barut, Fan and Verhasselt (2016) proposed conditional sure independence screening (CSIS) to address this issue under the context of generalized linear models. While CSIS outperforms the marginal screening method when few of the factors are known to be important and/or significant correlations among some of the factors exist, unfortunately, CSIS is model-based and might fail when the models are mis-specified. We propose a model-free conditional screening method under the framework of sufficient dimension reduction (SDR, Li 1991; Cook 1998) for ultra-high dimensional statistical problems. Numerical studies show that our method can easily beat CSIS for nonlinear models, and performs comparable to CSIS for (generalized) linear models. The sure screening consistency property for our method is also proved.

KEY WORDS: Conditional Screening; Trace Pursuit; Variable Selection; Sufficient Dimension Reduction.

1. INTRODUCTION

Researchers in many different fields, such as economics and finance, need to analyze high dimensional data, where the number of predictors p is frequently huge compared with the sample size n. Most traditional statistical methods failed when p is large. Also, with high dimensional data, it is often reasonable to assume only a small number of predictors actually contribute to the response (sparsity assumption). Hence, dimension reduction or feature selection is often conducted as the first step of data analysis. Estimation accuracy and model interpretability can be greatly improved in the subsequent analysis by effectively identifying the important predictors first. Fan and Lv (2008) proposed the sure independence screening (SIS), which is a feature screening procedure for linear models by ranking the marginal correlations between the response and each individual predictor. SIS has the so-called *sure screening property* (Fan and Lv, 2008), in the sense that as $n \to \infty$, the important predictors are guaranteed to be retained in the model with probability tending to 1, even for ultra-high dimensional predictor space, where p can diverge at an exponential rate of n. SIS was extended to generalized linear models in Fan and Song (2010). Fan *et al.* (2011) proposed nonparametric independence screening (NIS) for nonparametric models with additive structure using nonparametric marginal ranking. Many other feature screening methodologies have been developed, such as Xue and Zou (2011), Wang (2012), Zhao and Li (2012), and Chang *et al.* (2013).

However, all the aforementioned procedures are model-based and might yield poor performance when the models are mis-specified. Motivated by this fact, model-free feature screening procedures, which can identify the important predictors without specifying the model structure, were developed. To list a few, Zhu *et al.* (2011) proposed a sure independent ranking and screening (SIRS), Lin *et al.* (2013) proposed a nonparametric ranking feature screening (NRS) using the function-correlation between the response and predictors, He *et al.* (2013) proposed quantile-adaptive model-free screening through the marginal quantile regression, Mai and Zou (2015) proposed the fused Kolmogorov filter approach, which performs feature screening for the data with many types of predictors and response. For discriminant analysis with high dimensional data, model-free feature screening has been studied by Mai and Zou (2013), Cui *et al.* (2014), and Pan *et al.* (2016).

The performance of these feature screening procedures is heavily influenced by the correlations among the predictors, as mentioned in Fan and Lv (2008), Zhu *et al.* (2011), and Barut *et al.* (2016). As Barut *et al.* (2016) pointed out, the correlations among predictors might cause false positives (where the unimportant predictors are mistakenly considered as important ones through the screening procedure), and/or false negatives (where the important predictors are screened out as the unimportant ones). Unfortunately, the correlations among predictors are unavoidable for high dimensional data analysis (Hall and Li 1993; Fan and Lv 2008), since spurious correlations among predictors always exist as p diverges. To obtain the sure screening property, feature screening procedures usually need to impose some restrictions on the correlation structure among predictors.

One possible way to alleviate the above problem is to consider conditional screening method, since researchers in many applications have some prior information regarding the importance of certain predictors, such as the treatment effects in biological studies and market risk factors in financial studies, it is natural to consider feature screening methods conditioning on these known important predictors. For example, consider the leukemia data studied by Golub *et al.* (1999), Barut *et al.* (2016) and others, where gene expression data from 72 patients with two types of acute leukemia, acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML) were collected. Gene expression levels were measured for 7129 genes. Golub *et al.* (1999) described that two genes, Zyxin and Transcriptional activator hSNF2b, had empirically high correlations for the difference between people with AML and ALL. Barut *et al.* (2016) proposed a conditional screening method called *conditional sure independence screening (CSIS)* to conduct screening in the presence of the known set of predictors. They applied CSIS to the aforementioned leukemia data conditioning on the two genes, and were able to select TCRD (T-cell receptor delta locus) which had not been previously detected. Numerical studies also showed that, compared with SIS, CSIS makes it possible to identify those significant hidden predictors whose contributions might otherwise get canceled out due to the correlations with other predictors. Also, when there are high correlations among significant predictors and insignificant ones, CSIS can help to reduce the number of false negatives.

Although CSIS improves the performance of the screening procedure by using prior information, it is still a model-based screening procedure for generalized linear models and it might fail when the model assumption is not satisfied. To address this issue, we propose a model-free conditional screening method via sufficient dimension reduction in this article. Specifically, our method is based on the partial sufficient dimension reduction procedure proposed by Feng *et al.* (2013). The rest of this paper is organized as follows. In Section 2, we briefly review partial sufficient dimension reduction. We then propose our model-free conditional screening method and discuss its properties in Section 3. Numerical studies and real data analysis are provided in Section 4. A brief discussion and conclusion are given in Section 5. We defer all proofs to the Appendix.

2. PARTIAL SUFFICIENT DIMENSION REDUCTION

In this section, we give a brief introduction to partial sufficient dimension reduction since our model-free conditional screening method is based on it. For a regression problem, partial sufficient dimension reduction arises when one considers the predictive role of all predictors but limits dimension reduction to a subset of the predictors. Those predictors on which dimension reduction is performed are referred to as the predictors of primary interest, and the rest of predictors are referred to as the predictors of secondary interest. Partial dimension reduction is of practical use, since in many applications, some predictors play a particular role and must be shielded from the dimension reduction process. Considering the leukemia data discussed in Section 1, the two predictors (genes), Zyxin and Transcriptional activator hSNF2b, are the predictors of “secondary interest", since prior knowledge indicated that further dimension reduction should be conducted on other predictors (genes) while conditioning on these two predictors.

Let Y be a univariate random response, $\mathbf{X} = \{X_1, X_2, \ldots, X_p\} \in \mathbf{R}^p$ be a vector of continuous predictors of primary interest, and $\mathbf{W} = \{W_1, W_2, \ldots, W_q\} \in \mathbf{R}^q$ be a vector of predictors of secondary interest. The aim of partial sufficient dimension reduction is to find the partial central subspace $\mathcal{S}_{Y|\mathbf{X}}^{(W)}$, which is the intersection of all subspaces $\mathcal{S}$ such that

$$Y \perp\!\!\!\perp \mathbf{X} \mid (P_{\mathcal{S}}\mathbf{X}, \mathbf{W}),$$

where $\perp\!\!\!\perp$ stands for independence and $P_{\mathcal{S}}$ is the orthogonal projection on subspace $\mathcal{S}$. The concept of partial central subspace was first proposed by Chiaromonte *et al.* (2002) to deal with dimension reductions for regressions with a mixture of continuous and categorical predictors where the dimension reduction procedure focused on continuous predictors. Although it expands the scope of sufficient dimension reduction with practical applications, the method developed by Chiaromonte *et al.* (2002) is only limited to situations where $\mathbf{W}$ is categorical, and is difficult to be extended to cases with continuous $\mathbf{W}$. Hilafu and Wu (2017) proposed partial projective resampling dimension reduction (PPR-DR) to estimate the partial central subspace for any type of $\mathbf{W}$ by changing the role of $\mathbf{W}$ from predictor to the response variable. However, the subspace they estimated is larger than the partial central subspace when $\mathbf{W}$ is not independent with $\mathbf{X}$ given $P_{\mathcal{S}_{Y|\mathbf{X}}^{(W)}}\mathbf{X}$.

Feng *et al.* (2013) proposed partial discretization-expectation estimation (PDEE) to estimate the partial central subspace $\mathcal{S}_{Y|\mathbf{X}}^{(W)}$ when $\mathbf{W}$ is continuous, upon which our model-free conditional screening method is based. A brief review of PDEE is given below. First, the continuous $\mathbf{W}$ is discretized into a set of binary variables by defining $\mathbf{W}(\mathbf{T}) = (I_{\{W_1 \le T_1\}}, I_{\{W_2 \le T_2\}}, \ldots, I_{\{W_q \le T_q\}})$, where $\mathbf{T} = \{T_1, T_2, \ldots, T_q\} \in \mathbf{R}^q$ is an independent copy of $\mathbf{W}$ with support of $\mathbf{R}_{\mathbf{T}}^q$, and $I_{\{W_i \le T_i\}}$ is an indicator function taking value 1 for $W_i \le T_i$, and

0 otherwise, for $i = 1, \ldots, q$. Then, let $\mathcal{S}_{Y|\mathbf{X}}^{\mathbf{W}(\mathbf{t})}$ be the partial central subspace of $Y|(\mathbf{X}, \mathbf{W}(\mathbf{t}))$, for $\mathbf{T} = \mathbf{t} \in \mathbf{R}_{\mathbf{T}}^{q}$, Feng *et al.* (2013) showed that

$$\mathcal{S}_{Y|\mathbf{X}}^{(W)} = \bigcup_{\mathbf{t} \in \mathbf{R}_{\mathbf{T}}^{q}} \mathcal{S}_{Y|\mathbf{X}}^{\mathbf{W}(\mathbf{t})}. \tag{2.1}$$

Hence, an estimate of $\mathcal{S}_{Y|\mathbf{X}}^{(W)}$ can be obtained via those $\mathcal{S}_{Y|\mathbf{X}}^{\mathbf{W}(\mathbf{t})}$.

For simplicity, $(Y, \mathbf{X})|\mathbf{W}(\mathbf{t})$ is denoted as $(\mathbf{X}^{\mathbf{t}}, Y^{\mathbf{t}})$ for any fixed $\mathbf{t} \in \mathbf{R}_{\mathbf{T}}^{q}$. We can construct kernel matrices $\mathbf{M}(\mathbf{t})$ such that $\text{Span}\{\mathbf{M}(\mathbf{t})\} = \mathcal{S}_{Y|\mathbf{X}}^{\mathbf{W}(\mathbf{t})}$ to infer about the partial central subspace $\mathcal{S}_{Y|\mathbf{X}}^{\mathbf{W}(\mathbf{t})}$. Notice that (2.1) not only provides a general framework for estimating the partial central subspace, it can also be combined with many different sufficient dimension reduction methods by choosing different kernel matrices $\mathbf{M}(\mathbf{t})$. The following are the kernel matrices of the three most popular sufficient dimension reduction methods:

$$\begin{aligned}
\text{SIR:} \quad & \mathbf{M}(\mathbf{t}) = \mathbf{\Sigma}_{\mathbf{t}}^{-1}\text{Var}\{\text{E}(\mathbf{X}^{\mathbf{t}}|Y^{\mathbf{t}})\}\mathbf{\Sigma}_{\mathbf{t}}^{-1}; \\
\text{SAVE:} \quad & \mathbf{M}(\mathbf{t}) = \mathbf{\Sigma}_{\mathbf{t}}^{-1}\text{E}\{\mathbf{\Sigma}_{\mathbf{t}} - \text{Var}(\mathbf{X}^{\mathbf{t}}|Y^{\mathbf{t}})\}^{2}\mathbf{\Sigma}_{\mathbf{t}}^{-1}; \\
\text{DR:} \quad & \mathbf{M}(\mathbf{t}) = \mathbf{\Sigma}_{\mathbf{t}}^{-1}\text{E}\{2\mathbf{\Sigma}_{\mathbf{t}} - \text{E}\big((\widetilde{\mathbf{X}^{\mathbf{t}}} - \mathbf{X}^{\mathbf{t}})(\widetilde{\mathbf{X}^{\mathbf{t}}} - \mathbf{X}^{\mathbf{t}})^{T}|Y^{\mathbf{t}}, \widetilde{Y^{\mathbf{t}}}\big)\}^{2}\mathbf{\Sigma}_{\mathbf{t}}^{-1},
\end{aligned}$$

where $\mathbf{\Sigma}_{\mathbf{t}} = \text{Var}(\mathbf{X}^{\mathbf{t}})$, and $(\widetilde{Y^{\mathbf{t}}}, \widetilde{\mathbf{X}^{\mathbf{t}}})$ is an independent copy of $(Y^{\mathbf{t}}, \mathbf{X}^{\mathbf{t}})$. Interested readers may refer to Li and Dong (2009) and Li *et al.* (2010) for further details.

The following conditions are commonly used in sufficient dimension reduction area to ensure that $\text{Span}\{\mathbf{M}(\mathbf{t})\} = \mathcal{S}_{Y|\mathbf{X}}^{\mathbf{W}(\mathbf{t})}$ holds for the above choices of $\mathbf{M}(\mathbf{t})$.

Condition 3 *For any* $\mathbf{t} \in \mathbf{R}_{\mathbf{T}}^{q}$, *we assume that*

(a) $\text{E}(\mathbf{X}^{\mathbf{t}}|P_{\mathcal{S}_{Y|\mathbf{X}}^{\mathbf{W}(\mathbf{t})}}\mathbf{X}^{\mathbf{t}})$ *is linear combination of* $P_{\mathcal{S}_{Y|\mathbf{X}}^{\mathbf{W}(\mathbf{t})}}\mathbf{X}^{\mathbf{t}}$*;*

(b) $\text{Var}(\mathbf{X}^{\mathbf{t}}|P_{\mathcal{S}_{Y|\mathbf{X}}^{\mathbf{W}(\mathbf{t})}}\mathbf{X}^{\mathbf{t}})$ *is nonrandom.*

Condition 3(a) is also called the linear conditional mean (LCM) assumption, while condition 3(b) is the constant conditional variance (CCV) assumption. Both conditions hold for normally distributed $\mathbf{X}$. When $\mathbf{X}$ is not normally distributed, please refer to Cook and Nachtsheim (1994), Li and Dong (2009), Dong and Li (2010) for possible options. SIR (Li, 1991) only requires 3(a), while SAVE (Cook and Weisberg, 1991) and DR (Li and Wang, 2007) need both conditions.

Feng *et al.* (2013) showed that it suffices to take the expectation over the aforementioned random vector $\mathbf{T}$ (an independent copy of $\mathbf{W}$) to obtain the target matrix $\mathbf{M} = \mathrm{E}\{\mathbf{M}(\mathbf{T})\}$ such that $\mathrm{Span}\{\mathbf{M}\} = \mathcal{S}_{Y|\mathbf{X}}^{(W)}$.

3. CONDITIONAL SCREENING THROUGH TRACE PURSUIT

For model-free conditional screening, we set $\mathbf{W}$ as the set of predictors which should be retained in the model based on the prior knowledge, and perform feature screening on $\mathbf{X}$ while conditioning on $\mathbf{W}$. We seek the smallest active index set $\mathcal{A}$ such that

$$Y \perp\!\!\!\perp \mathbf{X}_{\mathcal{A}^c} | (\mathbf{X}_{\mathcal{A}}, \mathbf{W}), \tag{3.1}$$

where $\mathcal{A}^c$ is the complement set of $\mathcal{A}$ with respective to the index set $\mathcal{I} = \{1, \ldots, p\}$. From (3.1), it is obvious that $\mathbf{X}_{\mathcal{A}}$ just includes all important predictors for predicting Y given $\mathbf{W}$. Without loss of generality, we may assume the active index set $\mathcal{A} = \{1, \ldots, K\}$ for ease of exposition. We can see that (3.1) is equivalent to $P_{\mathcal{H}}\mathcal{S}_{Y|\mathbf{X}}^{(W)} = O_p$, where $\mathcal{H} = \mathrm{Span}\{(\mathbf{0}_{(p-K)\times K}, \mathbf{I}_{p-K})^T\}$ is the subspace of the primary predictor space, corresponding to the coordinates of the inactive predictors, and O_p is the origin in $\mathbb{R}^p$.

Cook (2004) first considered variable selection via a testing hypothesis approach by testing $Y \perp\!\!\!\perp \mathbf{X}_{\mathcal{A}^c} | \mathbf{X}_{\mathcal{A}}$, when the predictors are treated indiscriminately. Under the context of the regression of Y versus $\mathbf{X}$, Cook (2004) proposed a test for testing hypothesis of $Y \perp\!\!\!\perp \mathbf{X}_{\mathcal{A}^c} | \mathbf{X}_{\mathcal{A}}$ based on a generalized least square rederivation of the SIR estimator for

$\mathcal{S}_{Y|\mathbf{X}}$. Shao *et al.* (2007) and many others also investigated the same testing problem based on other estimators of $\mathcal{S}_{Y|\mathbf{X}}$. Zhong *et al.* (2012) and Jiang and Liu (2013) tackled the problem when $n < p$ via sliced inverse regression (SIR) method. However, both methods require the estimation of the rank of $\mathcal{S}_{Y|\mathbf{X}}$ (the so-called order determination), which is a very challenging problem when $n < p$. The trace pursuit approach proposed by Yu *et al.* (2016) successfully circumvents the need of order determination to conduct model-free variable selection via sufficient dimension reduction approach for $n < p$. In this article, we will conduct conditional variable screening via testing approach (3.1) from the partial sufficient dimension reduction perspective. We give a detailed discussion of our method using SIR (Li, 1991), though we can extend our approach to other sufficient dimension reduction methods such as SAVE (Cook and Weisberg, 1991) and DR (Li and Wang, 2007) by using different kernel matrices $\mathbf{M}$.

Let $\boldsymbol{\mu}_{\mathbf{t}} = \mathrm{E}(\mathbf{X}^t)$, $\mathbf{Z}^t = \boldsymbol{\Sigma}_{\mathbf{t}}^{-1/2}(\mathbf{X}^t - \boldsymbol{\mu}_t)$ and denote the $\mathbf{Z}$-scaled central space as $\mathcal{S}_{Y|\mathbf{Z}}^{\mathbf{W}(\mathbf{t})}$. By the so called invariance property (Cook, 1998), we have $\mathcal{S}_{Y|\mathbf{X}}^{\mathbf{W}(\mathbf{t})} = \boldsymbol{\Sigma}_{\mathbf{t}}^{-1/2}\mathcal{S}_{Y|\mathbf{Z}}^{\mathbf{W}(\mathbf{t})}$. We will work with the $\mathbf{Z}$-scaled central spaces first in the following discussions. For any given $\mathbf{t} \in \mathbf{R}_{\mathbf{T}}^{q}$, partition the range of $Y^{\mathbf{t}}$ into $H_{\mathbf{t}}$ non-overlapping slices $J_1^{\mathbf{t}}, \ldots, J_{H_{\mathbf{t}}}^{\mathbf{t}}$. Let $p_{h_{\mathbf{t}}} = \Pr(Y^{\mathbf{t}} \in J_{h_{\mathbf{t}}}^{\mathbf{t}})$, $\mathbf{U}_{h_{\mathbf{t}}} = \mathrm{E}(\mathbf{X}^t|\mathbf{Y}^t \in J_{h_{\mathbf{t}}}^{\mathbf{t}}) - \boldsymbol{\mu}_{\mathbf{t}}$, then the SIR-based $\mathbf{Z}$-scaled kernel matrix $\mathbf{M} = \mathrm{E}\{\mathbf{M}(t)\} = \mathrm{E}\big\{\boldsymbol{\Sigma}_{\mathbf{t}}^{-1/2}(\sum_{h_{\mathbf{t}}=1}^{H_{\mathbf{t}}} p_{h_{\mathbf{t}}}\mathbf{U}_{h_{\mathbf{t}}}\mathbf{U}_{h_{\mathbf{t}}}^{\top})\boldsymbol{\Sigma}_{\mathbf{t}}^{-1/2}\big\}$. Notice that for easy of exposition, with a slight abuse of notation, we keep using the same notation $\mathbf{M}$, for $\mathbf{Z}$-scaled kernel matrices as the $\mathbf{X}$-scaled ones, which were previously discussed in Section 2. For any index set $\mathcal{F}$, we denote $\mathbf{X}_{\mathcal{F}}^{\mathbf{t}} = \{X_i^{\mathbf{t}}, i \in \mathcal{F}\}$, $\boldsymbol{\mu}_{\mathcal{F},\mathbf{t}} = \mathrm{E}(\mathbf{X}_{\mathcal{F}}^t)$, $\mathbf{U}_{\mathcal{F},h_{\mathbf{t}}} = \mathrm{E}(\mathbf{X}_{\mathcal{F}}^t|\mathbf{Y}^t \in J_{h_{\mathbf{t}}}^{\mathbf{t}}) - \boldsymbol{\mu}_{\mathcal{F},\mathbf{t}}$ and $\boldsymbol{\Sigma}_{\mathcal{F},\mathbf{t}} = \mathrm{Var}(\mathbf{X}_{\mathcal{F}}^{\mathbf{t}})$. Moreover, we define $\mathbf{M}_{\mathcal{F}}(\mathbf{t}) = \boldsymbol{\Sigma}_{\mathcal{F},\mathbf{t}}^{-1/2}(\sum_{h_{\mathbf{t}}=1}^{H_{\mathbf{t}}} p_{h_{\mathbf{t}}}\mathbf{U}_{\mathcal{F},h_{\mathbf{t}}}\mathbf{U}_{\mathcal{F},h_{\mathbf{t}}}^{\top})\boldsymbol{\Sigma}_{\mathcal{F},\mathbf{t}}^{-1/2}$ and $\mathbf{M}_{\mathcal{F}} = \mathrm{E}(\mathbf{M}_{\mathcal{F}}(\mathbf{t}))$, then we have the following proposition.

Proposition 3 *Suppose Condition 3 holds, then for any index set* $\mathcal{F}$ *such that* $\mathcal{A} \subseteq \mathcal{F} \subseteq \mathcal{I}$, *we have* $\mathrm{tr}(\mathbf{M}_{\mathcal{A}}) = \mathrm{tr}(\mathbf{M}_{\mathcal{F}}) = \mathrm{tr}(\mathbf{M}_{\mathcal{I}})$.

Proposition 3 shows that $\mathrm{tr}(\mathbf{M}_{\mathcal{F}})$ can be used to capture the strength of the relationship between Y and $\mathbf{X}$ given $\mathbf{W}$. If $\mathcal{A}$ is a subset of $\mathcal{F}$, then the kernel matrix $\mathbf{M}_{\mathcal{F}}$ has the same trace as $\mathbf{M}_{\mathcal{A}}$. Denote $\mathcal{F} \cup j$ as the index set consisting of j and all the indices in $\mathcal{F}$. Suppose we already have the index set $\mathcal{F}$ selected in the model, and $\mathcal{F}$ does not contain $\mathcal{A}$, based on the following theorem, we can use the difference between $\mathrm{tr}(\mathbf{M}_{\mathcal{F}\cup j})$ and $\mathrm{tr}(\mathbf{M}_{\mathcal{F}})$ to measure the contribution of the additional X_j to Y given $(\mathbf{X}_{\mathcal{F}}, \mathbf{W})$.

Theorem 5 *For any* $\mathbf{t} \in \mathbf{R}^q_{\mathbf{T}}$, *suppose that we have*

$$\mathrm{E}(X^{\mathbf{t}}_j | \mathbf{X}^{\mathbf{t}}_{\mathcal{F}})\ is\ a\ linear\ function\ of\ \mathbf{X}^{\mathbf{t}}_{\mathcal{F}},\ for\ any\ j \notin \mathcal{F}\ and\ \mathcal{F} \subseteq \mathcal{I}.$$

Then

- *If* $\mathcal{A} \subseteq \mathcal{F}$, *then* $\mathrm{tr}(\mathbf{M}_{\mathcal{F}\cup j}) - \mathrm{tr}(\mathbf{M}_{\mathcal{F}}) = 0$.
- *If* $\mathcal{A} \nsubseteq \mathcal{F}$, *then* $\mathrm{tr}(\mathbf{M}_{\mathcal{F}\cup j}) - \mathrm{tr}(\mathbf{M}_{\mathcal{F}}) = \mathrm{E}_{\mathbf{T}}\big(\sum_{h_{\mathbf{t}}=1}^{H_{\mathbf{t}}} p_{h_{\mathbf{t}}} (\boldsymbol{\gamma}^{\mathbf{t}}_{j|\mathcal{F},h_{\mathbf{t}}})^2\big)$, *where* $\mu^{\mathbf{t}}_{j|\mathcal{F}} = \mathrm{E}(\boldsymbol{\gamma}_{j|\mathcal{F}} | \mathbf{T} = \mathbf{t})$ *and* $\boldsymbol{\gamma}^{\mathbf{t}}_{j|\mathcal{F},h_{\mathbf{t}}} = \mathrm{E}(\boldsymbol{\gamma}_{j|\mathcal{F}} | Y \in J_{h_{\mathbf{t}}}, \mathbf{T} = \mathbf{t}) - \mu^{\mathbf{t}}_{j|\mathcal{F}}$ *with* $X_{j|\mathcal{F}} = \mathbf{X}_j - \mathrm{E}(X_j | \mathbf{X}_{\mathcal{F}})$, $\sigma^2_{j|\mathcal{F}} = \mathrm{Var}(X_{j|\mathcal{F}})$, *and* $\boldsymbol{\gamma}_{j|\mathcal{F}} = X_{j|\mathcal{F}} / \sigma_{j|\mathcal{F}}$.

Condition 5 is parallel to Condition 3 (a). When $\mathbf{X}^{\mathbf{t}}$ follows an elliptical contour distribution for any $\mathbf{t}$, both conditions are satisfied. The first part of Theorem 5 shows that the trace difference between $\mathbf{M}_{\mathcal{F}\cup j}$ and $\mathbf{M}_{\mathcal{F}}$ is 0, when the active set $\mathcal{A}$ is already included in the set $\mathcal{F}$. The second part provides a formula to calculate the trace difference, when the set $\mathcal{F}$ does not include all the active predictors.

For the derivation of the asymptotic consistency of our method, we hence assume that $\boldsymbol{\Sigma}_{\mathbf{t}} = \boldsymbol{\Sigma}$, for any $\mathbf{t} \in \mathbf{R}^q_{\mathbf{T}}$. Although simulation studies suggest that our method still performs well in applications where this "homogeneous variance condition" does not hold.

Suppose that $d = \dim(\mathcal{S}_{Y|\mathbf{Z}}^{(W)}) = \dim(\mathcal{S}_{Y|\mathbf{X}}^{(W)})$, and let $\lambda_1 \geq, \cdots, \geq \lambda_d$ be the nonzero eigenvalues for $\mathbf{M}$ and $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_d$ be the corresponding eigenvectors. Denote $\boldsymbol{\beta}_i = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\eta}_i = (\beta_{i,1}, \ldots, \beta_{i,p})^\top$, for $i = 1, \ldots, d$. Under Condition 3, we have $\mathrm{Span}\{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d\} = \mathcal{S}_{Y|\mathbf{X}}^{(W)}$. Furthermore, we define $\beta_{min}^2 = \min_{j\in\mathcal{A}} \sum_{i=1}^d \beta_{i,j}^2$, where λ_{min} and λ_{max} are the smallest and the largest eigenvalues of $\boldsymbol{\Sigma}$ respectively.

Proposition 4 *Suppose that condition 5 in Theorem 5 holds, for any $\mathcal{F}$ which $\mathcal{A} \nsubseteq \mathcal{F}$ we have*

$$\max_{j\in\mathcal{A}\cap\mathcal{F}^c} \left(\mathrm{tr}(\mathbf{M}_{\mathcal{F}\cup j}) - \mathrm{tr}(\mathbf{M}_{\mathcal{F}})\right) \geq \lambda_d \lambda_{max}^{-1} \lambda_{min} \beta_{min}^2,$$

Under the sufficient dimension reduction framework, we know $Y \perp\!\!\!\perp \mathbf{X}|(\boldsymbol{\beta}_1^T\mathbf{X}, \ldots, \boldsymbol{\beta}_d^T\mathbf{X}, \mathbf{W})$. Since $\mathcal{A}$ is the smallest active index set such that $Y \perp\!\!\!\perp \mathbf{X}|(\mathbf{X}_{\mathcal{A}}, \mathbf{W})$, then $\sum_{i=1}^d \beta_{i,j}^2 > 0$ for any $j \in \mathcal{A}$. Hence, for any $\mathcal{F}$ which does not include all the active predictors, the maximum difference between $\mathbf{M}_{\mathcal{F}\cup j}$ and $\mathbf{M}_{\mathcal{F}}$ over $j \in \mathcal{F}^c \cap \mathcal{A}$ is larger than 0 based on the result in Proposition 4.

Let $(\mathbf{X}_i, Y_i, \mathbf{W}_i)$, $i = 1, \ldots, n$ be simple random sample of size n. Follow Feng *et al.* (2013), for easy of implementation, we choose l_n different $\mathbf{t}_m$'s of which l_n is of order $O(n)$ and use $n_{\mathbf{t}_m}$ to denote the subsample size for a given $\mathbf{t}_m$. Then we can rewrite the sample as $(\mathbf{X}_i^{\mathbf{t}_m}, Y_i^{\mathbf{t}_m})$, $i = 1, \ldots, n_{\mathbf{t}_m}$ for a given $\mathbf{t}_m$. Let $\widehat{\mathbf{M}}(\mathbf{t}_m)$ be the sample estimate of $\mathbf{M}(\mathbf{t}_m)$, then we can estimate $\mathbf{M}$ using $\widehat{\mathbf{M}} = \frac{1}{l_n}\sum_{m=1}^{l_n} \widehat{\mathbf{M}}(\mathbf{t}_m)$. Follow the SIR-based forward trace pursuit algorithm in Yu *et al.* (2016), the screening procedure starts with an empty index set $\mathcal{F}_0$, then, each time, add the index which maximize the difference between the traces of successive kernel matrices to the working set, until we obtain a working index set with n indices. Hence, we obtain a sequence of n nested working index sets $\mathcal{F}_1, \ldots, \mathcal{F}_n$. To select a model from this sequence of nested working index sets, we use the modified BIC criterion defined in Yu *et al.* (2016):

$$\mathrm{BIC}(\mathcal{F}) = -\log\left\{\mathrm{tr}(\widehat{\mathbf{M}}_{\mathcal{F}})\right\} + n^{-1}|\mathcal{F}|(\log n + 2\log p),$$

where $|\mathcal{F}|$ denotes the cardinality of set $\mathcal{F}$.

To obtain the sure screening property of conditional forward trace pursuit based on SIR, we need the following conditions.

Condition 4

a. There exist some constants $\alpha_0 > 0$ *and* $0 < b_0 < 1/2$ *such that*

$$\min_{\mathcal{F}:\ \mathcal{A} \not\subseteq \mathcal{F}} \ \max_{j \in \mathcal{A} \cap \mathcal{F}^c} \big(\mathrm{tr}(\mathbf{M}_{\mathcal{F} \cup j}) - \mathrm{tr}(\mathbf{M}_{\mathcal{F}})\big) \geq a_0 n^{-b_0}.$$

b. $\mathbf{X}$ *and* $\mathbf{X}^{\mathbf{t}}$ *follows multi-normal distributions for any* $\mathbf{t} \in \mathbf{R}^q_{\mathbf{T}}$.

c. There exist $c_1 > 0$ *and* $c_2 > 0$ *such that* $c_1 < \lambda_{min} < \lambda_{max} < c_2$.

d. There exist constants a_1, b_1 *and* b_2 *such that* $log(p) \leq a_1 n^{\theta_1}$, $|\mathcal{A}| \leq a_1 n^{b_2}$ *and* $2b_0 + b_1 + 2b_2 < 1$.

e. There exists constant b_3 *such that* $l_n = O(n^{b_3})$ *and* $n_{\mathbf{t}_m} = O(n^{1-b_3})$ *for any* $\mathbf{t}_m$ *among the* l_n *points where* $0.5(1 - c_3) < b_3 < 1 - c_3$.

Motivated by the conclusion in Proposition 4, we assume that Condition 4 (a) holds. Condition 4 (b) and (c) are common for variable screening of high dimensional data. Assuming Condition 4 (b) and (c), Wang (2009) studied the sure screening property of forward linear regression. Condition 4 (d) allows the dimension p and the number of important predictors to go to infinity as sample size n goes to infinity. We assume Condition 4 (e) to guarantee that it is not too sparse for each subsample and $\widehat{\mathbf{M}}(\mathbf{t}_m)$ is $\sqrt{n}$ consistent estimator of $\mathbf{M}(\mathbf{t}_m)$ for $m = 1, \ldots, l_n$.

Theorem 6 *Assume Condition 1 and Condition 2 hold, then we have*

$$\Pr(\mathcal{A} \subset \mathcal{F}_{\hat{m}}) \longrightarrow 1,$$

as $n \longrightarrow \infty$ *and* $p \longrightarrow \infty$, *where* $\hat{m} = \operatorname*{argmin}_{1 \leq k \leq n} BIC(\mathcal{F}_k)$.

Theorem 6 shows that our conditional forward trace pursuit method based on SIR has the desired sure screening property.

4. NUMERICAL STUDIES

Table 1. Results for Model I, II and III

Model	Method	CR	MS	FP	FN
I	CFTP-SIR	1	8.3	0.0037	0
	CFTP-SAVE	0	31.8	0.0159	1
	CFTP-DR	1	32.3	0.0157	0
	CSIS	1	859	0.4303	0
II	CFTP-SIR	1	13	0.0065	0
	CFTP-SAVE	0	30	0.0150	1
	CFTP-DR	1	33	0.0165	0
	CSIS	0.1	16.5	0.0082	0.9
III	CFTP-SIR	1	11.2	0.005	0
	CFTP-SAVE	0	32.3	0.016	1
	CFTP-DR	1	33.3	0.016	0
	CSIS	0.18	10.5	0.052	0.82

4.1. Simulation Studies. In this part, we compare the screening performance of our conditional forward trace pursuit (CFTP) method with CSIS (Barut *et al.*, 2016). Based on 100 repetitions, we evaluate the performance using the true model coverage rate (CR, the rate of all the significant predictors being selected), the average model size (MS), the average false positive rate (FP), and the average false negative rate (FN). For CSIS, we use random decoupling, which was discussed in (Barut *et al.*, 2016), to select the thresholding parameters and determine the model size for Model I-VI; while for Model VII-IX, $[n/\log(n)]$ is used as the model size since those provided by random decoupling method would be too small.

The following models are considered.

$$
\begin{aligned}
\text{(I) } Y &= 3W_1 + 3W_2 + 3W_3 + 3W_4 + 3W_5 - 7.5X_1 + \epsilon \\
\text{(II) } Y &= (3W_1 + 3W_2 + 3W_3 + 3W_4 + 3W_5 - 7.5X_1 + \epsilon)^2 \\
\text{(III) } Y &= \exp(3W_1 + 3W_2 + 3W_3 + 3W_4 + 3W_5 - 7.5X_1) + \epsilon \\
\text{(IV) } Y &= 5W + 2X_p + \epsilon \\
\text{(V) } Y &= (5W + 2X_p)^2 + \epsilon \\
\text{(VI) } Y &= \exp(5W + 2X_p) + (5W + 2X_p)^3 + \epsilon \\
\text{(VII) } Y &= 8W_1 - 6W_2 + 5W_3 + (X_1 + X_p)^2 + \epsilon \\
\text{(VIII) } Y &= 2W_1 - 1.5W_2 + \exp(X_{p-1}) + 2X_p^4 + \epsilon \\
\text{(IX) } Y &= \text{sign}(W_1 - W_2)\exp(X_1 + X_2 + X_{p-1} + X_p) + \epsilon
\end{aligned}
$$

We set the sample size $n = 400$ for all models. The random error ϵ follows a standard normal distribution $N(0, 1)$ and is independent with $\mathbf{W}$ and $\mathbf{X}$. For Model I, II and III, we generate $[\mathbf{W}^\top, \mathbf{X}^\top]^\top$ from $N(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma} = 0.5\mathbf{I_{q+p}} + \mathbf{0.5J_{q+p}}$, $q = 5$, $p + q = 2000$. We use $\mathbf{I_p}$ to denote the p-dimensional identity matrix, and $\mathbf{J}_p$ is the $p \times p$ square matrix of all ones. Model I was also studied in Barut *et al.* (2016) to show that the conditional screening can recover the hidden significant predictors since $\text{Cov}(Y, X_1) = 0$ under the setting in this model. For Model IV, V and VI, $[\mathbf{W}, \mathbf{X}]$ are also generated from multivariate normal distribution with zero mean vector. In these three models, we set $q = 1$, $p + q = 2000$, $X_1, \ldots, X_{p-1}$ and W are all correlated with each other with correlation coefficient of 0.8, while X_p is independent with all of them. Under this setting, we have $\text{Cov}(Y, X_i) = 4$ for $i = 1, \ldots, p - 1$, and $\text{Cov}(Y, X_p) = 2$ for Model IV. Barut *et al.* (2016) discussed a similar model and show that conditional screening can reduce the false negative rate. In Model VII, W_i, $i = 1, 2, 3$, are independently generated from $U[0, 1]$, and $\mathbf{X}$ follows $N(\mathbf{0}, \mathbf{\Sigma})$ with elements $\sigma_{i,j} = \rho^{|i-j|}$ for $i,\ j = 1,\ldots,p$ and $p = 2000$. For Model VIII, $[\mathbf{W}^\top, \mathbf{X}^\top]^\top$ are generated from $N(\mathbf{0}, \mathbf{\Sigma})$, where $\sigma_{i,j} = \rho^{|i-j|}$, $q = 2$, $p+q = 2000$. In Model IX, Var($\mathbf{X}|W$) is dependent on $\mathbf{W}$, which violates the homogeneous variance assumption. Here W_1 and W_2 are independently generated from $U[0,1]$, and $\mathbf{X}$ is generated from $N(\mathbf{0}, \mathbf{\Sigma})$. As in Model VII, we set $\sigma_{i,j} = \rho^{|i-j|}$ for $i,\ j = 1,\ldots,p$ and $p = 2000$. However, in this model, we consider $\rho = \rho^\star$ which takes two different values depending on the difference between W_1 and W_2: $\rho^\star = 0$ if $W_1 - W_2 > 0$, and $\rho^\star = 0.5$ otherwise.

Table 4.1 compares the performance of our method with CSIS for Model I–III. As expected, the SAVE based method does not perform well as it could not deal with linear trends well (Cook and Forzani, 2009). However, both SIR and DR based conditional forward trace pursuit methods outperform CSIS: the true model coverage rates provided by our methods are 1, which means that our method can always select all the significant predictors; the false positive rate and false negative rate are also much smaller than those of CSIS; the average model sizes are also much smaller than those of CSIS. For example, for Model III, CR and FN from CSIS are 0.18 and 0.82 respectively, comparing with 1 (the closer to one the better) and 0 (the smaller the better) from our method.

Table 2. Results for Model IV, V and VI

Model	Method	CR	MS	FP	FN
IV	CFTP-SIR	1	9.75	0.0044	0
	CFTP-SAVE	0	27	0.0135	1
	CFTP-DR	0.97	28.6	0.0138	0.03
	CSIS	1	221	0.1106	0
V	CFTP-SIR	1	9.2	0.0041	0
	CFTP-SAVE	0.04	27.1	0.0135	0.96
	CFTP-DR	1	28.1	0.0136	0
	CSIS	0.01	223.94	0.1121	0.99
VI	CFTP-SIR	1	9.1	0.0041	0
	CFTP-SAVE	0	27.1	0.0135	1
	CFTP-DR	1	28.3	0.0137	0
	CSIS	0.13	209.05	0.1046	0.87

Table 4.2 gives simulation results for Model IV–VI. Still, CSIS is outperformed by our SIR and DR based methods. Our methods can provide screening results with much smaller model sizes, similar or better coverage rates, smaller false positive rates and/or false negative rates for all three models. The nonlinear model structure does not affect the performance of our screening method, however it adversely affects the performance of CSIS greatly for Model V and VI. Results for Model VII and VIII are given on Table 4.3. Model VII has a quadratic structure in the mean function where SAVE is expected to perform well, which agrees with the simulation results. For Model VIII, DR based method dominates all the other methods.

Table 3. Results for Model VII and VIII

		$\rho = 0$				$\rho = 0.5$			
Model	Method	CR	MS	FP	FN	CR	MS	FP	FN
VII	CFTP-SIR	0	15.5	0.0078	1	0	15.15	0.0076	0.975
	CFTP-SAVE	1	30.6	0.0143	0	1	30.3	0.0142	0
	CFTP-DR	0.94	33.6	0.0159	0.060	1	33.6	0.0158	0
	CSIS	0	67	0.0333	0.885	0	67	0.0333	0.865
VIII	CFTP-SIR	0.03	11.9	0.0050	0.475	0.10	12.36	0.0056	0.450
	CFTP-SAVE	0.20	27.5	0.0132	0.400	0.08	27.3	0.0132	0.465
	CFTP-DR	1	32.4	0.0152	0	1	32.2	0.0151	0
	CSIS	0	67	0.0332	0.810	0	67	0.0332	0.790

Simulation results for Model IX with different correlation structures are shown on Table 4.4. We discussed before, when $\rho = \rho*$, the homogeneous variance assumption is violated. As we can see, both SIR and DR based methods still outperform CSIS. Though DR based method does not perform as well as SIR based method since the constant variance condition does not hold for this model. The false negative rates for SIR based method, DR based method, and CSIS are 0, 0.075, and 0.455 respectively; while the coverage rates for the three methods are 1, 0.83 and 0.21 respectively. CSIS mistakenly screens out some of the significant predictors frequently. All our simulation results suggest that DR based conditional forward trace pursuit method is the most robust screening method, while SIR based method most of the time provides the best screening performance. We suggest to use SIR based screening method first, and use DR based method as a complement.

Table 4. Results for Model IX

ρ	Method	CR	MS	FP	FN
$\rho = \rho^{\star}$	CFTP-SIR	1	10.3	0.0032	0
	CFTP-SAVE	0	30.9	0.0155	1
	CFTP-DR	0.83	33.2	0.0148	0.075
	CSIS	0.21	67	0.0325	0.455

4.2. Real Data Analysis. In this section, we consider the aforementioned leukemia data set which was first studied by Golub *et al.* (1999) and has become a benchmark in many gene expression studies. The dataset consists of 72 samples and gene expression level of 7129 genes in two types of acute leukemias, acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). There are 38 (27 ALL and 11 AML) training samples and 34 (20 ALL and 14 AML) testing samples. Our goal is to select related genes and classify future patients to the two leukemia types based on those genes.

We standardized the gene expression dataset by centering and scaling each array with mean 0 and standard deviation 1. The proposed conditional screening method and CSIS are performed based on the following three different choices of **W**.

- $\mathbf{W}_1$={X95735, D26156};
- $\mathbf{W}_2$={X95735, M27783};
- $\mathbf{W}_3$={X95735, MD88422}.

The genes X95735 (Zyxin) and D26156 (Transcriptional activator hSNF2b) in $\mathbf{W}_1$ have empirically high correlations for the difference between patients with AML and ALL and were used in Barut *et al.* (2016). The genes X95735 and M27783 (ELA2 Elastatse 2, neutrophil) in $\mathbf{W}_2$ are the two top ranked genes from marginal screening SIS. For $\mathbf{W}_3$, the genes X95735 and MD88422 (CYSTATIN A) were identified in Hong *et al.* (2016). To compare with CSIS, we first perform our conditional forward trace pursuit method to select genes based on the training samples given $\mathbf{W}_i,\ i = 1, 2, 3$ respectively. Next, we establish a classification rule through the logistic model based on the genes being selected, and apply this rule to the testing samples. The results are shown on Table 5.

Table 5. Results for Model V

	$\mathbf{W}_1$		$\mathbf{W}_2$		$\mathbf{W}_3$	
Method	Train Err	Test Err	Train Err	Test Err	Train Err	Test Err
CSIS	0/38	2/34	1/38	5/34	0/38	2/34
CFTP-SIR	0/38	1/34	0/38	5/34	0/38	3/34
CFTP-SAVE	0/38	3/34	0/38	5/34	0/38	3/34
CFTP-DR	0/38	3/34	0/38	5/34	0/38	3/34

Conditioning on {X95735, D26156} ($\mathbf{W}_1$), we identified another gene Z32765 (GB DEF = CD36 gene exon 15) using SIR-based conditional trace pursuit method. Armesilla *et al.* (1996) showed that Gene CD36 was associated with acute myeloid leukemia. The classification rule based on these three genes can achieve $0/38$ training error rate and $1/34$ testing error rate.

5. CONCLUSIONS

In this paper, we proposed a model-free conditional screening method to fully utilize the prior information regarding the importance of certain predictors. Comparing to CSIS developed by Barut, Fan and Verhasselt (2016), our method outperforms CSIS when the model structure is nonlinear, and is comparable to CSIS for generalized linear model. Numerical studies suggest that our methods can provide screening results with much smaller model sizes, similar or better coverage rates, smaller false positive rates and/or false negative rates for nonlinear models.

APPENDIX

Proof of Proposition 3:

For any given $\mathbf{t}$, we denote $\lambda_1^{\mathbf{t}} \geq \cdots \geq \lambda_{d_{\mathbf{t}}}^{\mathbf{t}}$ as the nonzero eigenvalues for $\mathbf{M}(\mathbf{t})$ and $\boldsymbol{\eta}_1(\mathbf{t}), \ldots, \boldsymbol{\eta}_{d_{\mathbf{t}}}(\mathbf{t})$ as the corresponding eigenvectors. Let $\boldsymbol{\beta}_i(\mathbf{t}) = \boldsymbol{\Sigma}_{\mathbf{t}}^{-1/2}\boldsymbol{\eta}_i(\mathbf{t}) = (\beta_{i,1}(\mathbf{t}), \ldots, \beta_{i,p}(\mathbf{t}))^\top$ for $i = 1, \ldots, d_{\mathbf{t}}$. Since $Y \perp\!\!\!\perp \mathbf{X}_{\mathcal{A}^c}|(\mathbf{X}_{\mathcal{A}}, W)$, then we have $\beta_{i,j}(\mathbf{t}) = 0$, for any $j \in \mathcal{A}^c$. Recall that $\mathcal{A} = \{1, \ldots, K\}$. Define $\boldsymbol{\beta}_{\mathcal{A},i}(\mathbf{t}) = (\beta_{i,1}, \ldots, \beta_{i,K})^\top$ and $\boldsymbol{\beta}_{\mathcal{A}^c,i}(\mathbf{t}) = (\beta_{i,K+1}, \ldots, \beta_{i,p})^\top$, then $\boldsymbol{\beta}_{i,\mathcal{A}^c}(\mathbf{t}) = \mathbf{0}$.
Note that $\mathbf{M}(\mathbf{t}) = \sum\limits_{i=1}^{d_{\mathbf{t}}} \lambda_i^{\mathbf{t}}\boldsymbol{\eta}_i(\mathbf{t})\boldsymbol{\eta}_i(\mathbf{t})^\top = \boldsymbol{\Sigma}_{\mathbf{t}}^{1/2}\big(\sum_{i=1}^{d_{\mathbf{t}}} \lambda_i^{\mathbf{t}}\boldsymbol{\beta}_i(\mathbf{t})\boldsymbol{\beta}_i(\mathbf{t})^\top\big)\boldsymbol{\Sigma}_{\mathbf{t}}^{1/2}$, then we have

$$\mathrm{tr}(\mathbf{M}(\mathbf{t})) = \mathrm{tr}\big\{\boldsymbol{\Sigma}_{\mathbf{t}}\big(\sum_{i=1}^{d_{\mathbf{t}}} \lambda_i^{\mathbf{t}}\boldsymbol{\beta}_i(\mathbf{t})\boldsymbol{\beta}_i(\mathbf{t})^\top\big)\big\} = \mathrm{tr}\big\{\boldsymbol{\Sigma}_{\mathbf{t},\mathcal{A}}\big(\sum_{i=1}^{d_{\mathbf{t}}} \lambda_i^{\mathbf{t}}\boldsymbol{\beta}_{\mathcal{A},i}(\mathbf{t})\boldsymbol{\beta}_{\mathcal{A},i}(\mathbf{t})^\top\big)\big\}. \tag{A.1}$$

Since $\mathbf{M}_{\mathcal{A}}(\mathbf{t}) = \mathrm{Var}\{\mathrm{E}(\mathbf{Z}_{\mathcal{A}}^{\mathbf{t}}|Y^{\mathbf{t}} \in J_{h_{\mathbf{t}}}^{\mathbf{t}})\} = \boldsymbol{\Sigma}_{\mathcal{A},\mathbf{t}}^{-1/2}\mathrm{Var}\{\mathrm{E}(\mathbf{X}_{\mathcal{A}}^{\mathbf{t}}|Y^{\mathbf{t}} \in J_{h_{\mathbf{t}}}^{\mathbf{t}})\}\boldsymbol{\Sigma}_{\mathcal{A},\mathbf{t}}^{-1/2}$, we have

$$\mathrm{tr}(\mathbf{M}_{\mathcal{A}}(\mathbf{t})) = \mathrm{tr}\big\{\boldsymbol{\Sigma}_{\mathcal{A},\mathbf{t}}^{-1}\mathrm{Var}\{\mathrm{E}(\mathbf{X}_{\mathcal{A}}^{\mathbf{t}}|Y^{\mathbf{t}} \in J_{h_{\mathbf{t}}}^{\mathbf{t}})\}\big\}. \tag{A.2}$$

Note that

$$\begin{aligned}
&\mathrm{Var}\{\mathrm{E}(\mathbf{X}^{\mathbf{t}}|Y^{\mathbf{t}} \in J_{h_{\mathbf{t}}}^{\mathbf{t}})\} = \boldsymbol{\Sigma}_{\mathbf{t}}^{1/2}\mathbf{M}(\mathbf{t})\boldsymbol{\Sigma}_{\mathbf{t}}^{1/2} = \boldsymbol{\Sigma}_{\mathbf{t}}\big(\sum_{i=1}^{d_{\mathbf{t}}} \lambda_i^{\mathbf{t}}\boldsymbol{\beta}_i(\mathbf{t})\boldsymbol{\beta}_i(\mathbf{t})^{\mathbf{t}}\big)\boldsymbol{\Sigma}_{\mathbf{t}} \\
=&\begin{pmatrix} \boldsymbol{\Sigma}_{\mathcal{A},\mathbf{t}} & \boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}^c,\mathbf{t}} \\ \boldsymbol{\Sigma}_{\mathcal{A}^c\mathcal{A},\mathbf{t}} & \boldsymbol{\Sigma}_{\mathcal{A}^c,\mathbf{t}} \end{pmatrix}\begin{pmatrix} \sum_{i=1}^{d_{\mathbf{t}}} \lambda_i^{\mathbf{t}}\boldsymbol{\beta}_{\mathcal{A},i}(\mathbf{t})\boldsymbol{\beta}_{\mathcal{A},i}(\mathbf{t})^\top & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} \boldsymbol{\Sigma}_{\mathcal{A},\mathbf{t}} & \boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}^c,\mathbf{t}} \\ \boldsymbol{\Sigma}_{\mathcal{A}^c\mathcal{A},\mathbf{t}} & \boldsymbol{\Sigma}_{\mathcal{A}^c,\mathbf{t}} \end{pmatrix}
\end{aligned} \tag{A.3}$$

From A.3, it is obvious that $\mathrm{Var}\{\mathrm{E}(\mathbf{X}_{\mathcal{A}}^{\mathbf{t}}|Y^{\mathbf{t}} \in J_{h_{\mathbf{t}}}^{\mathbf{t}})\} = \boldsymbol{\Sigma}_{\mathcal{A},\mathbf{t}}\big(\sum_{i=1}^{d_{\mathbf{t}}} \lambda_i^{\mathbf{t}}\boldsymbol{\beta}_{\mathcal{A},i}(\mathbf{t})\boldsymbol{\beta}_{\mathcal{A},i}(\mathbf{t})^{\mathbf{t}}\big)\boldsymbol{\Sigma}_{\mathcal{A},\mathbf{t}}$. Combined with A.1 and A.2, we have $\mathrm{tr}(\mathbf{M}_{\mathcal{A}}(\mathbf{t})) = \mathrm{tr}(\mathbf{M}_{\mathcal{I}}(\mathbf{t}))$. Similarly, we have $\mathrm{tr}(\mathbf{M}_{\mathcal{A}}(\mathbf{t})) = \mathrm{tr}(\mathbf{M}_{\mathcal{F}}(\mathbf{t}))$ for any $\mathcal{F}$ such that $\mathcal{A} \subseteq \mathcal{F}$. Then the conclusion follows. □

Proof of Theorem 5:

From Proposition 3, we know that $\mathrm{tr}(\mathbf{M}_{\mathcal{A}}) = \mathrm{tr}(\mathbf{M}_{\mathcal{F}})$ for any $\mathcal{F}$ such that $\mathcal{A} \subseteq \mathcal{F}$. Then the first part of Theorem 5 follows.

For any fixed $\mathbf{t}$, if Condition 5 holds, we have $\mathrm{E}(x_j^{\mathbf{t}}|\mathbf{X}_{\mathcal{F}}^{\mathbf{t}}) = \mathrm{Cov}(\mathbf{X}_{\mathcal{F}}^{\mathbf{t}}, X_j^{\mathbf{t}})\boldsymbol{\Sigma}_{\mathcal{F},\mathbf{t}}^{-1}\mathbf{X}_{\mathcal{F}}^{\mathbf{t}}$. Let $|\mathcal{F}|$ denote as the cardinality of $\mathcal{F}$, then we construct two $(|\mathcal{F}|+1)\times(|\mathcal{F}|+1)$ matrices $\mathbf{A}_{\mathbf{t}}$ and $\mathbf{C}_{\mathbf{t}}$ as

$$\mathbf{A}_{\mathbf{t}} = \begin{pmatrix} \mathbf{I}_{|\mathcal{F}|} & \mathbf{0} \\ \mathrm{Cov}(\mathbf{X}_{\mathcal{F}}^{\mathbf{t}}, X_j^{\mathbf{t}})\boldsymbol{\Sigma}_{\mathcal{F},\mathbf{t}}^{-1} & 1 \end{pmatrix} \; \textit{and} \; \mathbf{C}_{\mathbf{t}} = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathcal{F},\mathbf{t}} & \mathbf{0} \\ \mathbf{0} & \sigma_{j|\mathcal{F},\mathbf{t}}^2 \end{pmatrix}.$$

where $\sigma_{j|\mathcal{F},\mathbf{t}}^2 = \mathrm{Var}(X_{j|\mathcal{F}}^{\mathbf{t}})$ with $X_{j|\mathcal{F}}^{\mathbf{t}} = \mathbf{X}_j^{\mathbf{t}} - \mathrm{E}(X_j^{\mathbf{t}}|\mathbf{X}_{\mathcal{F}}^{\mathbf{t}})$. Then we have that

$$\mathbf{A}_{\mathbf{t}}\mathbf{X}_{\mathcal{F}\cup j}^{\mathbf{t}} = \begin{pmatrix} \mathbf{X}_{\mathcal{F}}^{\mathbf{t}} \\ X_{j|\mathcal{F}}^{\mathbf{t}} \end{pmatrix} \; \textit{and} \; \mathbf{A}_{\mathbf{t}}\mathbf{U}_{\mathcal{F}\cup j,h_{\mathbf{t}}} = \begin{pmatrix} \mathbf{U}_{\mathcal{F},h_{\mathbf{t}}} \\ \mathrm{E}(X_{j|\mathcal{F}}^{\mathbf{t}}|\mathbf{Y}^t \in J_{h_{\mathbf{t}}}^{\mathbf{t}}) - \mathrm{E}(X_{j|\mathcal{F}}^{\mathbf{t}}) \end{pmatrix}$$

From the definition of $X_{j|\mathcal{F}}^{\mathbf{t}}$, it is obvious that $\mathrm{Cov}(X_{j|\mathcal{F}}^{\mathbf{t}}, \mathbf{X}_{\mathcal{F}}^{\mathbf{t}}) = \mathbf{0}$. Then we have $\mathrm{Var}(\mathbf{A}_{\mathbf{t}}\mathbf{X}_{\mathcal{F}\cup j}^{\mathbf{t}}) = \mathbf{A}_{\mathbf{t}}\boldsymbol{\Sigma}_{\mathcal{F}\cup j,\mathbf{t}}\mathbf{A}_{\mathbf{t}}^{\top} = \mathbf{C}_{\mathbf{t}}$. Therefore, we have $\boldsymbol{\Sigma}_{\mathcal{F}\cup j,\mathbf{t}}^{-1} = \mathbf{A}_{\mathbf{t}}\mathbf{C}_{\mathbf{t}}^{-1}\mathbf{A}_{\mathbf{t}}^{\top}$. Then we can rewrite $\mathrm{tr}(\mathbf{M}_{\mathcal{F}\cup j}(\mathbf{t}))$ as

$$\begin{aligned}
\mathrm{tr}(\mathbf{M}_{\mathcal{F}\cup j}(\mathbf{t})) &= \mathrm{tr}\big\{\boldsymbol{\Sigma}_{\mathcal{F}\cup j,\mathbf{t}}^{-1/2}(\sum_{h_{\mathbf{t}}=1}^{H_{\mathbf{t}}} p_{h_{\mathbf{t}}}\mathbf{U}_{\mathcal{F}\cup j,h_{\mathbf{t}}}\mathbf{U}_{\mathcal{F}\cup j,h_{\mathbf{t}}}^{\top})\boldsymbol{\Sigma}_{\mathcal{F}\cup j,\mathbf{t}}^{-1/2}\big\} \\
&= \mathrm{tr}\big\{\boldsymbol{\Sigma}_{\mathcal{F}\cup j,\mathbf{t}}^{-1}(\sum_{h_{\mathbf{t}}=1}^{H_{\mathbf{t}}} p_{h_{\mathbf{t}}}\mathbf{U}_{\mathcal{F}\cup j,h_{\mathbf{t}}}\mathbf{U}_{\mathcal{F}\cup j,h_{\mathbf{t}}}^{\top})\big\} \\
&= \mathrm{tr}\big\{\mathbf{C}_{\mathbf{t}}^{-1}\big(\sum_{h_{\mathbf{t}}=1}^{H_{\mathbf{t}}} p_{h_{\mathbf{t}}}(\mathbf{A}_{\mathbf{t}}\mathbf{U}_{\mathcal{F}\cup j,h_{\mathbf{t}}})(\mathbf{A}_{\mathbf{t}}\mathbf{U}_{\mathcal{F}\cup j,h_{\mathbf{t}}}^{\top})\big)\big\} \\
&= \mathrm{tr}\big\{\boldsymbol{\Sigma}_{\mathcal{F},\mathbf{t}}^{-1}(\sum_{h_{\mathbf{t}}=1}^{H_{\mathbf{t}}} p_{h_{\mathbf{t}}}\mathbf{U}_{\mathcal{F},h_{\mathbf{t}}}\mathbf{U}_{\mathcal{F},h_{\mathbf{t}}}^{\top})\big\} + \sum_{h_{\mathbf{t}}=1}^{H_{\mathbf{t}}} p_{h_{\mathbf{t}}}(\boldsymbol{\gamma}_{j|\mathcal{F},h_{\mathbf{t}}}^{\mathbf{t}})^2
\end{aligned}$$

Then we have $\mathrm{tr}(\mathbf{M}_{\mathcal{F}\cup j}) - \mathrm{tr}(\mathbf{M}_{\mathcal{F}}) = \mathrm{E}_{\mathbf{T}}\{\mathrm{tr}(\mathbf{M}_{\mathcal{F}\cup j}(t)) - \mathrm{tr}(\mathbf{M}_{\mathcal{F}}(\mathbf{t}))\} = \mathrm{E}_{\mathbf{T}}\big(\sum_{h_{\mathbf{t}}=1}^{H_{\mathbf{t}}} p_{h_{\mathbf{t}}}(\boldsymbol{\gamma}_{j|\mathcal{F},h_{\mathbf{t}}}^{\mathbf{t}})^2\big)$ □

Proof of Proposition 4:

Denote $\boldsymbol{\Sigma}_{\mathcal{F}_1\mathcal{F}_2,\mathbf{t}} = \mathrm{Cov}(\mathbf{X}_{\mathcal{F}_1}^{\mathbf{t}}, \mathbf{X}_{\mathcal{F}_2}^{\mathbf{t}})$ and $\boldsymbol{\Sigma}_{\mathcal{F}_1\mathcal{F}_2,\mathbf{t}} = \mathrm{Cov}(\mathbf{X}_{\mathcal{F}_1}, \mathbf{X}_{\mathcal{F}_2})$ for any $\mathcal{F}_1, \mathcal{F}_2 \subseteq \mathcal{I}$. Since we suppose $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{\mathbf{t}}$, then we have that $\boldsymbol{\Sigma}_{\mathcal{F}_1\mathcal{F}_2,\mathbf{t}} = \boldsymbol{\Sigma}_{\mathcal{F}_1\mathcal{F}_2,\mathbf{t}}$. For any $j \in \mathcal{F}^c \cap \mathcal{A}$, we have

$$\begin{aligned}\sigma^2_{j|\mathcal{F}}\big(\mathrm{tr}(\mathbf{M}_{\mathcal{F}\cup j})-\mathrm{tr}(\mathbf{M}_{\mathcal{F}})\big) &= \mathrm{E}_{\mathbf{T}}\{\mathrm{Var}(\mathrm{E}(X_{j|\mathcal{F}}|\mathbf{Y}))\}\\ =&\big(-\boldsymbol{\Sigma}_{\mathcal{F}j}\boldsymbol{\Sigma}^{-1}_{\mathcal{F}},1\big)\mathbf{P}\mathrm{E}_{\mathbf{T}}\{\mathrm{Var}(\mathrm{E}(\mathbf{X}^t|\mathbf{Y}^t\in J^{\mathbf{t}}_{h_{\mathbf{t}}}))\}\mathbf{P}^\top\big(-\boldsymbol{\Sigma}_{\mathcal{F}j}\boldsymbol{\Sigma}^{-1}_{\mathcal{F}},1\big)^\top,\end{aligned} \tag{A.4}$$

For simplicity, we suppose the first $|\mathcal{F}|+1$ elements of $\mathbf{X}$ is $(\mathbf{X}_{\mathcal{F}},X_j)^\top$, then $\mathbf{P}$ in 5.2 can be denoted as $\mathbf{P}=(\mathbf{I}_{|\mathcal{F}|+1},\mathbf{0}_{(|\mathcal{F}|+1)(p-|\mathcal{F}|-1)})$. Since $\mathbf{M}=\mathrm{E}_{\mathbf{T}}\{\mathrm{Var}(\mathrm{E}(\mathbf{Z}^{\mathbf{t}}|Y^{\mathbf{t}}\in J^{\mathbf{t}}_{h_{\mathbf{t}}}))\}=\sum_{i=1}^d\lambda_i\boldsymbol{\eta}_i\boldsymbol{\eta}_i^\top$, and $\boldsymbol{\beta}_i=\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\eta}_i$, then we have

$$\mathrm{E}_{\mathbf{T}}\{\mathrm{Var}(\mathrm{E}(\mathbf{X}^t|\mathbf{Y}^t\in J^{\mathbf{t}}_{h_{\mathbf{t}}}))\}=\boldsymbol{\Sigma}^{1/2}(\sum_{i=1}^d\lambda_i\boldsymbol{\eta}_i\boldsymbol{\eta}_i^\top)\boldsymbol{\Sigma}^{1/2}=\boldsymbol{\Sigma}(\sum_{i=1}^d\lambda_i\boldsymbol{\beta}_i\boldsymbol{\beta}_i^\top)\boldsymbol{\Sigma}. \tag{A.5}$$

It follows that

$$\begin{aligned}&\big(-\boldsymbol{\Sigma}_{\mathcal{F}j}\boldsymbol{\Sigma}^{-1}_{\mathcal{F}},1\big)\mathbf{P}\boldsymbol{\Sigma}\boldsymbol{\beta}_i\\ =&\big(-\boldsymbol{\Sigma}_{\mathcal{F}j}\boldsymbol{\Sigma}^{-1}_{\mathcal{F}},1\big)\mathbf{P}\boldsymbol{\Sigma}_{(\mathcal{F}\cup j)\mathcal{I}}\boldsymbol{\beta}_i\\ =&\big(\boldsymbol{\Sigma}_{j\mathcal{I}}-\boldsymbol{\Sigma}_{j\mathcal{F}}\boldsymbol{\Sigma}^{-1}_{\mathcal{F}}\boldsymbol{\Sigma}_{\mathcal{F}\mathcal{I}}\big)\boldsymbol{\beta}_i\end{aligned}$$

Let $\boldsymbol{\beta}_{i,\mathcal{F}}=\{\beta_{i,j},j\in\mathcal{F}\}$. Since $\big(\boldsymbol{\Sigma}_{j\mathcal{I}}-\boldsymbol{\Sigma}_{j\mathcal{F}}\boldsymbol{\Sigma}^{-1}_{\mathcal{F}}\boldsymbol{\Sigma}_{\mathcal{F}\mathcal{F}}\big)\boldsymbol{\beta}_i=0$ and $\boldsymbol{\beta}_{i,\mathcal{F}^c\cap\mathcal{I}^c}=0$, we have

$$\begin{aligned}\big(-\boldsymbol{\Sigma}_{\mathcal{F}j}\boldsymbol{\Sigma}^{-1}_{\mathcal{F}},1\big)\mathbf{P}\boldsymbol{\Sigma}\boldsymbol{\beta}_i &= \big(\boldsymbol{\Sigma}_{j\mathcal{F}^c}-\boldsymbol{\Sigma}_{j\mathcal{F}}\boldsymbol{\Sigma}^{-1}_{\mathcal{F}}\boldsymbol{\Sigma}_{\mathcal{F}\mathcal{F}^c}\big)\boldsymbol{\beta}_{i,\mathcal{F}^c}\\ =&\big(\boldsymbol{\Sigma}_{j(\mathcal{F}^c\cap\mathcal{A})}-\boldsymbol{\Sigma}_{j\mathcal{F}}\boldsymbol{\Sigma}^{-1}_{\mathcal{F}}\boldsymbol{\Sigma}_{\mathcal{F}(\mathcal{F}^c\cap\mathcal{A})}\big)\boldsymbol{\beta}_{i,\mathcal{F}^c\cap\mathcal{A}},\end{aligned} \tag{A.6}$$

for any $i=1,\dots,d$. From A.4, A.5 and A.6, it follows that

$$\sigma^2_{j|\mathcal{F}}\big(\mathrm{tr}(\mathbf{M}_{\mathcal{F}\cup j})-\mathrm{tr}(\mathbf{M}_{\mathcal{F}})\big)=\sum_{i=1}^d\lambda_i\{\big(\boldsymbol{\Sigma}_{j(\mathcal{F}^c\cap\mathcal{A})}-\boldsymbol{\Sigma}_{j\mathcal{F}}\boldsymbol{\Sigma}^{-1}_{\mathcal{F}}\boldsymbol{\Sigma}_{\mathcal{F}(\mathcal{F}^c\cap\mathcal{A})}\big)\boldsymbol{\beta}_{i,\mathcal{F}^c\cap\mathcal{A}}\}^2$$

Note that

$$\begin{aligned}&\sum_{j\in\mathcal{F}^c}\{\big(\boldsymbol{\Sigma}_{j(\mathcal{F}^c\cap\mathcal{A})}-\boldsymbol{\Sigma}_{j\mathcal{F}}\boldsymbol{\Sigma}^{-1}_{\mathcal{F}}\boldsymbol{\Sigma}_{\mathcal{F}(\mathcal{F}^c\cap\mathcal{A})}\big)\boldsymbol{\beta}_{i,\mathcal{F}^c\cap\mathcal{A}}\}^2\\ =&\boldsymbol{\beta}^\top_{i,\mathcal{F}^c\cap\mathcal{A}}\big(\boldsymbol{\Sigma}_{(\mathcal{F}^c\cap\mathcal{A})}-\boldsymbol{\Sigma}_{(\mathcal{F}^c\cap\mathcal{A})\mathcal{F}}\boldsymbol{\Sigma}^{-1}_{\mathcal{F}}\boldsymbol{\Sigma}_{\mathcal{F}(\mathcal{F}^c\cap\mathcal{A})}\big)^2\boldsymbol{\beta}_{i,\mathcal{F}^c\cap\mathcal{A}},\end{aligned}$$

and

$$\begin{aligned}
&\lambda_{\min}\big(\mathbf{\Sigma}_{(\mathcal{F}^c\cap\mathcal{A})}-\mathbf{\Sigma}_{(\mathcal{F}^c\cap\mathcal{A})\mathcal{F}}\mathbf{\Sigma}_{\mathcal{F}}^{-1}\mathbf{\Sigma}_{\mathcal{F}(\mathcal{F}^c\cap\mathcal{A})}\big)\\
=&\lambda_{\max}^{-1}\{\big(\mathbf{\Sigma}_{(\mathcal{F}^c\cap\mathcal{A})}-\mathbf{\Sigma}_{(\mathcal{F}^c\cap\mathcal{A})\mathcal{F}}\mathbf{\Sigma}_{\mathcal{F}}^{-1}\mathbf{\Sigma}_{\mathcal{F}(\mathcal{F}^c\cap\mathcal{A})}\big)^{-1}\}\\
\geq&\lambda_{\max}^{-1}(\mathbf{\Sigma}^{-1})=\lambda_{min}
\end{aligned}$$

Then we have that

$$\begin{aligned}
&\max_{j\in\mathcal{F}^c\cap\mathcal{A}}\sigma_{j|\mathcal{F}}^2\big(\mathrm{tr}(\mathbf{M}_{\mathcal{F}\cup j})-\mathrm{tr}(\mathbf{M}_{\mathcal{F}})\big)\geq|\mathcal{F}^c\cap\mathcal{A}|^{-1}\sum_{j\in\mathcal{F}^c\cap\mathcal{A}}[\sigma_{j|\mathcal{F}}^2\big(\mathrm{tr}(\mathbf{M}_{\mathcal{F}\cup j})-\mathrm{tr}(\mathbf{M}_{\mathcal{F}})\big)]\\
=&|\mathcal{F}^c\cap\mathcal{A}|^{-1}\sum_{i=1}^{d}\lambda_i\boldsymbol{\beta}_{i,\mathcal{F}^c\cap\mathcal{A}}^{\top}\big(\mathbf{\Sigma}_{(\mathcal{F}^c\cap\mathcal{A})}-\mathbf{\Sigma}_{(\mathcal{F}^c\cap\mathcal{A})\mathcal{F}}\mathbf{\Sigma}_{\mathcal{F}}^{-1}\mathbf{\Sigma}_{\mathcal{F}(\mathcal{F}^c\cap\mathcal{A})}\big)^2\boldsymbol{\beta}_{i,\mathcal{F}^c\cap\mathcal{A}}\\
\geq&|\mathcal{F}^c\cap\mathcal{A}|^{-1}\sum_{i=1}^{d}\lambda_i\lambda_{\min}^2\big(\mathbf{\Sigma}_{(\mathcal{F}^c\cap\mathcal{A})}-\mathbf{\Sigma}_{(\mathcal{F}^c\cap\mathcal{A})\mathcal{F}}\mathbf{\Sigma}_{\mathcal{F}}^{-1}\mathbf{\Sigma}_{\mathcal{F}(\mathcal{F}^c\cap\mathcal{A})}\big)^2\boldsymbol{\beta}_{i,\mathcal{F}^c\cap\mathcal{A}}^{\top}\boldsymbol{\beta}_{i,\mathcal{F}^c\cap\mathcal{A}}\\
\geq&\lambda_d\lambda_{min}|\mathcal{F}^c\cap\mathcal{A}|^{-1}\sum_{i=1}^{d}\boldsymbol{\beta}_{i,\mathcal{F}^c\cap\mathcal{A}}^{\top}\boldsymbol{\beta}_{i,\mathcal{F}^c\cap\mathcal{A}}\geq\lambda_d\lambda_{max}^{-1}\lambda_{min}\beta_{min}^2.\quad\Box
\end{aligned}$$

To prove Theorem 6, we need the following lemmas.

Lemma 2 *Let* $\widetilde{\mathbf{M}}=1/l_n\sum_{m=1}^{l_n}\mathbf{M}(\mathbf{t}_m)$*,* $\psi_{h_\mathbf{t}}=p_{h_\mathbf{t}}^{-1/2}(I(Y^\mathbf{t}\in J_{h_\mathbf{t}}^\mathbf{t}))$ *and* $\zeta_{h_\mathbf{t}}=\mathbf{\Sigma}_\mathbf{t}^{-1}\mathrm{E}(\mathbf{X}^\mathbf{t}\psi_{h_\mathbf{t}})$*, then we have* $\mathrm{tr}(\widetilde{\mathbf{M}})=(H-1)-\frac{1}{l_n}\sum_{m=1}^{l_n}\sum_{h_{\mathbf{t}_m}=1}^{H_{\mathbf{t}_m}}\mathrm{E}(\psi_{h_{\mathbf{t}_m}}-\zeta_{h_{\mathbf{t}_m}}^{\top}\mathbf{X}^{\mathbf{t}_m})^2$*, where* $H=1/l_n\sum_{m=1}^{l_n}H_{\mathbf{t}_m}$*.*

Proof of Lemma 2: For any $\mathbf{t}_m$, $m=1,\ldots,l_n$ and $h_{\mathbf{t}_m}$, $h_{\mathbf{t}_m}=1,\ldots,H_{\mathbf{t}_m}$,

$$\begin{aligned}
&\mathrm{E}(\psi_{h_\mathbf{t}}-\zeta_{h_{\mathbf{t}_m}}^{\top}\mathbf{X}^{\mathbf{t}_m})^2=\mathrm{E}(\psi_{h_\mathbf{t}}^2)-2\mathrm{E}(\psi_{h_{\mathbf{t}_m}}\zeta_{h_{\mathbf{t}_m}}^{\top}\mathbf{X}^{\mathbf{t}_m})+\mathrm{E}((\zeta_{h_{\mathbf{t}_m}}^{\top}\mathbf{X}^{\mathbf{t}_m}\mathbf{X}^{\mathbf{t}_m\top}\zeta_{h_{\mathbf{t}_m}})\\
=&\mathrm{E}(\psi_{h_\mathbf{t}}^2)-\zeta_{h_{\mathbf{t}_m}}^{\top}\mathbf{\Sigma}_\mathbf{t}\zeta_{h_{\mathbf{t}_m}}=(1-p_{h_\mathbf{t}})-p_{h_\mathbf{t}}^{-1}\mathrm{E}\{\mathbf{Z}^{\mathbf{t}_m\top}I(Y^\mathbf{t}\in J_{h_\mathbf{t}}^\mathbf{t})\}\mathrm{E}\{\mathbf{Z}^{\mathbf{t}_m}I(Y^\mathbf{t}\in J_{h_\mathbf{t}}^\mathbf{t})\}
\end{aligned}$$

Then we have that

$$\begin{aligned}
\mathrm{tr}(\widetilde{\mathbf{M}})=1/l_n\sum_{m=1}^{l_n}\mathrm{tr}(\mathbf{M}(\mathbf{t}_m))=&1/l_n\sum_{m=1}^{l_n}\sum_{h_{\mathbf{t}_m}=1}^{H_{\mathbf{t}m}}p_{h_\mathbf{t}}^{-1}\mathrm{E}\{\mathbf{Z}^{\mathbf{t}_m\top}I(Y^\mathbf{t}\in J_{h_\mathbf{t}}^\mathbf{t})\}\mathrm{E}\{\mathbf{Z}^{\mathbf{t}_m}I(Y^\mathbf{t}\in J_{h_\mathbf{t}}^\mathbf{t})\}\\
=&(H-1)-\frac{1}{l_n}\sum_{m=1}^{l_n}\sum_{h_{\mathbf{t}_m}=1}^{H_{\mathbf{t}m}}\mathrm{E}(\psi_{h_{\mathbf{t}_m}}-\zeta_{h_{\mathbf{t}_m}}^{\top}\mathbf{X}^{\mathbf{t}_m})^2.\Box
\end{aligned}$$

Lemma 3 *Let* $D_{\mathcal{F}\cup j} = \mathrm{tr}(\mathbf{M}_{\mathcal{F}\cup j}) - \mathrm{tr}(\mathbf{M}_{\mathcal{F}})$ *and* $\widehat{D}_{\mathcal{F}\cup j} = \mathrm{tr}(\widehat{\mathbf{M}}_{\mathcal{F}\cup j}) - \mathrm{tr}(\widehat{\mathbf{M}}_{\mathcal{F}})$. *Suppose* $|\mathcal{F}| = O(n^{b_0+b_2})$ *and Condition 4 holds, there exists some constant* d_0 *such that* $\widehat{D}_{\mathcal{F}\cup j} - D_{\mathcal{F}\cup j} \leq d_0|\mathcal{F}|\sqrt{\log p/n^{1-b_3}}$ *with probability tending to* 1.

Proof of Lemma 3: Define $\widetilde{D}_{F\cup j} = \mathrm{tr}(\widetilde{\mathbf{M}}_{\mathcal{F}\cup j}) - \mathrm{tr}(\widetilde{\mathbf{M}}_{\mathcal{F}})$, then we have that

$$\widehat{D}_{\mathcal{F}\cup j} - D_{\mathcal{F}\cup j} = [\widehat{D}_{\mathcal{F}\cup j} - \widetilde{D}_{F\cup j}] + [\widetilde{D}_{F\cup j} - D_{\mathcal{F}\cup j}]$$

Form Lemma 7 in Yu et al. (2016), we know that $\widehat{\mathrm{Var}}\{\mathrm{E}(X^{\mathbf{t}_m}_{j|\mathcal{F}})\} - \mathrm{Var}\{\mathrm{E}(X^{\mathbf{t}_m}_{j|\mathcal{F}})\} = O\big(|\mathcal{F}|\sqrt{\log p/n^{1-b_3}}\big)$ for any given $\mathbf{t}_m$, $m = 1, \dots, l_n$. Furthermore, from the proof of Lemma 3 in Jiang and Liu (2013), we have $\hat{\sigma}^2_{j|\mathcal{F},\mathbf{t}} - \sigma^2_{j|\mathcal{F},\mathbf{t}} = O\big(|\mathcal{F}|\sqrt{\log p/n^{1-b_3}}\big)$. Then we have that

$$\begin{aligned}
&\big\{\mathrm{tr}(\widehat{\mathbf{M}}_{\mathcal{F}\cup j}(\mathbf{t}_m)) - \mathrm{tr}(\widehat{\mathbf{M}}_{\mathcal{F}}(\mathbf{t}_m))\big\} - \big\{\mathrm{tr}(\widetilde{\mathbf{M}}_{\mathcal{F}\cup j}(\mathbf{t}_m)) - \mathrm{tr}(\widetilde{\mathbf{M}}_{\mathcal{F}}(\mathbf{t}_m))\big\} \\
=&\hat{\sigma}^2_{j|\mathcal{F},\mathbf{t}}\widehat{\mathrm{Var}}\{\mathrm{E}(X^{\mathbf{t}_m}_{j|\mathcal{F}})\} - \sigma^2_{j|\mathcal{F},\mathbf{t}}\mathrm{Var}\{\mathrm{E}(X^{\mathbf{t}_m}_{j|\mathcal{F}})\} \\
=&\big\{\hat{\sigma}^2_{j|\mathcal{F},\mathbf{t}}\widehat{\mathrm{Var}}\{\mathrm{E}(X^{\mathbf{t}_m}_{j|\mathcal{F}})\} - \hat{\sigma}^2_{j|\mathcal{F},\mathbf{t}}\mathrm{Var}\{\mathrm{E}(X^{\mathbf{t}_m}_{j|\mathcal{F}})\}\big\} \\
&- \big\{\hat{\sigma}^2_{j|\mathcal{F},\mathbf{t}}\mathrm{Var}\{\mathrm{E}(X^{\mathbf{t}_m}_{j|\mathcal{F}})\} - \sigma^2_{j|\mathcal{F},\mathbf{t}}\mathrm{Var}\{\mathrm{E}(X^{\mathbf{t}_m}_{j|\mathcal{F}})\}\big\} \\
=&O\big(|\mathcal{F}|\sqrt{\log p/n^{1-b_3}}\big) + O\big(|\mathcal{F}|\sqrt{\log p/n^{1-b_3}}\big) = O\big(|\mathcal{F}|\sqrt{\log p/n^{1-b_3}}\big).
\end{aligned}$$

Hence, we have that

$$\begin{aligned}
&\widehat{D}_{\mathcal{F}\cup j} - \widetilde{D}_{F\cup j} \\
=&\frac{1}{l_n}\sum_{m=1}^{l_n}\Big[\big\{\mathrm{tr}(\widehat{\mathbf{M}}_{\mathcal{F}\cup j}(\mathbf{t}_m)) - \mathrm{tr}(\widehat{\mathbf{M}}_{\mathcal{F}}(\mathbf{t}_m))\big\} - \big\{\mathrm{tr}(\widetilde{\mathbf{M}}_{\mathcal{F}\cup j}(\mathbf{t}_m)) - \mathrm{tr}(\widetilde{\mathbf{M}}_{\mathcal{F}}(\mathbf{t}_m))\big\}\Big] \\
=&\frac{1}{l_n}\sum_{m=1}^{l_n} O\big(|\mathcal{F}|\sqrt{\log p/n^{1-b_3}}\big) = O\big(|\mathcal{F}|\sqrt{\log p/n^{1-b_3}}\big).
\end{aligned}$$

From this, it is obvious that there exists some constant d_0 such that $\widehat{D}_{\mathcal{F}\cup j} - D_{\mathcal{F}\cup j} \leq d_0|\mathcal{F}|\sqrt{\log p/n^{1-b_3}}$ with probability tending to 1. □

Proof of Theorem 6:

Firstly , we prove that CFTP method can select in all $|\mathcal{A}|$ important predictors within $[2Ha_0^{-1}a_1n^{b_0+b_2}]$ steps by showing that at least one important predictor in the model within $[2Ha_0^{-1}n^{b_0}]$ steps since $|\mathcal{A}| \leq a_1n^{b_2}$ under Condition 4. Without loss of generality, we just show that at least one important is selected in the model within the first $[2Ha_0^{-1}n^{b_0}]$.

Recall that $\mathcal{F}_k$ is the index set after kth step, we let $Q(k) = \mathrm{tr}(\widehat{\mathbf{M}}_{\mathcal{F}_k}) - \mathrm{tr}(\widehat{\mathbf{M}}_{\mathcal{F}_{k-1}})$. We assume that no important is selected in the model within the first $[2Ha_0^{-1}n^{b_0}]$ steps. From lemma 3 and Condition 4, we have that

$$\begin{aligned} Q(k) &\geq 2^{-1}\big(\mathrm{tr}(\mathbf{M}_{\mathcal{F}_k}) - \mathrm{tr}(\mathbf{M}_{\mathcal{F}_{k-1}}) - d_0|\mathcal{F}_k|\sqrt{\log p/n^{1-b_3}}\big) \\ &\geq 2^{-1}\big(a_0n^{-b_0} - d_02Ha_0^{-1}a_1n^{b_0+b_2}\sqrt{\log p/n^{1-b_3}}\big) \rightarrow 2^{-1}a_0n^{-b_0} \end{aligned}$$

if $\mathcal{F}_k \cap \mathcal{A} = \varnothing$ for any $k = 1, \ldots, [2Ha_0^{-1}n^{b_0}]$.

Hence, we have that

$$\sum_{k=1}^{[2Ha_0^{-1}n^{b_0}]} Q(k) \geq [2Ha_0^{-1}n^{b_0}] \times 2^{-1}a_0n^{-b_0} \geq H.$$

However, from Lemma 2, we know

$$\sum_{k=1}^{[2Ha_0^{-1}n^{b_0}]} Q(k) = \mathrm{tr}(\widehat{\mathbf{M}}_{\mathcal{F}_{[2Ha_0^{-1}n^{b_0}]}}) \leq H - 1.$$

Therefore, this implies at least one important predictor is selected in the model within the first $[2Ha_0^{-1}n^{b_0}]$ steps.

Moreover, follow the proof of Theorem 5.2 in Yu *et al.* (2016) and the proof of Theorem 2 in Wang (2009), it is easy to prove that $\Pr(\mathcal{A} \subset \mathcal{F}_{\hat{m}}) \rightarrow 1$, and the details are omitted. □

REFERENCES

Armesilla, A. L., Calvo, D., and Vega, M. A., 'Structural and functional characterization of the human cd36 gene promoter,' Journal of Biological Chemistry, 1996, **271**, pp. 7781–7787.

Barut, E., Fan, J., and Verhasselt, A., 'Conditional sure independence screening,' Journal of the American Statistical Association, 2016, **111**, pp. 1266–1277.

Chang, J., Tang, C., and Wu, Y., 'Marginal empirical likelihood and sure independence feature screening,' The Annals of Statistics, 2013, **41**, pp. 1693–2262.

Chiaromonte, F., Cook, R., and Li, B., 'Sufficient dimension reduction in regressions with categorical predictors,' The Annals of Statistics, 2002, **30**, pp. 475–497.

Cook, R. and Forzani, B., 'Likelihood-based sufficient dimension reduction,' Journal of the American Statistical Association, 2009, **104**, pp. 197–208.

Cook, R. D., *Regression Graphics*, Wiley, New York, 1998.

Cook, R. D., 'Testing predictor contributions in sufficient dimension reduction,' The Annals of Statistics, 2004, **32**, pp. 1062–1092.

Cook, R. D. and Nachtsheim, C., 'Reweighting to achieve elliptically contoured covariates in regression,' Journal of the American Statistical Association, 1994, **89**, pp. 592–599.

Cook, R. D. and Weisberg, S., 'Discussion of "sliced inverse regression for dimension reduction",' Journal of the American Statistical Association, 1991, **86**, pp. 328–332.

Cui, H., Li, R., and Zhong, W., 'Model-free feature screening for ultrahigh dimensional discriminant analysis,' Journal of the American Statistical Association, 2014, **110**, pp. 630–641.

Dong, Y. and Li, B., 'Dimension reduction for non-elliptically distributed predictors: second-order methods,' Biometrika, 2010, **97**, pp. 279–294.

Fan, J., Feng, Y., and Song, R., 'Nonparametric independence screening in sparse ultrahigh-dimensional additive models,' Journal of the American Statistical Association, 2011, **106**, pp. 544–5570.

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

Fan, J. and Song, R., 'Sure independence screening in generalized linear models with np-dimensionality,' The Annals of Statistics, 2010, **38**, pp. 3567–3604.

Feng, Z., Wen, X., Yu, Z., and Zhu, L.-X., ‘On partial sufficient dimension reduction with applications to partially linear multi-index models,’ Journal of the American Statistical Association, 2013, **108**, pp. 237–246.

Golub, T., Slonim, D., Tamyo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., and Caligiuri, M., ‘Molecular classification of cancer: class discovery and class prediction by gene expression monitoring,’ Science, 1999, **286**, pp. 531–536.

Hall, P. and Li, K., ‘On almost linearity of low dimensional projection from high dimensional data,’ The Annals of Statistics, 1993, **21**, pp. 867–889.

He, X., Wang, L., and Hong, H., ‘Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data,’ The Annals of Statistics, 2013, **41**, pp. 342–369.

Hilafu, H. and Wu, W., ‘Partial projective resampling method for dimension reduction: With applications to partially linear models,’ Computational Statistics and Data Analysis, 2017, **109**, pp. 1–14.

Hong, H. G., Wang, L., and He, X., ‘A data-driven approach to conditional screening of high-dimensional variables,’ STAT, 2016, **5**, pp. 200–212.

Jiang, B. and Liu, J. S., ‘Sliced inverse regression with variable selection and interaction detection,’ Manuscript, 2013.

Li, B. and Dong, Y., ‘Dimension reduction for non-elliptically distributed predictors,’ The Annals of Statistics, 2009, **37**, pp. 1272–1298.

Li, B., Kim, M. K., and Altman, N., ‘On dimension folding of matrix or array valued statistical objects,’ The Annals of Statistics, 2010, **38**, pp. 1097–1121.

Li, B. and Wang, S., ‘On directional regression for dimension reduction,’ Journal of the American Statistical Association, 2007, **102**, pp. 997–1008.

Li, K. C., ‘Sliced inverse regression for dimension reduction (with discussion),’ Journal of the American Statistical Association, 1991, **86**, pp. 316–342.

Lin, L., Sun, J., and Zhu, L., ‘Nonparametric feature screening,’ Computational Statistics and Data Analysis, 2013, **67**, pp. 162–174.

Mai, Q. and Zou, H., ‘The kolmogorov filter for variable screening in high-dimensional binary classification.’ Biometrika, 2013, **100**, pp. 229–234.

Mai, Q. and Zou, H., ‘The fused kolmogorov filter: a nonparametric model-free screening method,’ The Annals of Statistics, 2015, **43**, pp. 1471–1497.

Pan, R., Wang, H., and Li, R., ‘Ultrahigh dimensional multi-class linear discriminant analysis by pairwise sure independence screening’,’ Journal of the American Statistical Association, 2016, **111**, pp. 169–179.

Shao, Y., Cook, R. D., and Weisberg, S., ‘Marginal tests with sliced average variance estimation,’ Biometrika, 2007, **94**, pp. 285–296.

Wang, H., ‘Forward regression for ultra-high dimensional variable screening,’ Journal of the American Statistical Association, 2009, **104**, pp. 1512–1524.

Wang, H., ‘Factor profiled sure independence screening,’ Biometrika, 2012, **99**, pp. 15–28.

Xue, L. and Zou, H., ‘Sure independence screening and compressed random sensing,’ Biometrika, 2011, **98**, pp. 371–380.

Yu, Z., Dong, Y., and Zhu, L., ‘Trace pursuit: a general framework for model-free variable selection,’ Journal of the American Statistical Association, 2016, **111**, pp. 813–821.

Zhao, S. and Li, Y., ‘Sure screening for estimating equations in ultra-high dimensions,’ Manuscript, 2012.

Zhong, W., Zhang, T., Zhu, M., and Liu, J. S., ‘Correlation pursuit: forward stepwise variable selection for index models,’ Journal of the Royal Statistical Society, Ser. B, 2012, **74**, pp. 849–870.

Zhu, L., Li, L., Li, R., and Zhu, L.-X., ‘Model-free feature screening for ultrahigh dimensional data,’ Journal of American Statistical Association, 2011, **106**, pp. 1464–1475.

SECTION

2. SUMMARY AND CONCLUSIONS

Many model-free variable selection prodecures have been developed under the framework of sufficient dimension reduction. However, none of these existing methods considered the grouping information when dealing with multiple population data. In paper I, we propose a novel model-free variable selection method for $n < p$ multi-population data. Unlike the existing methods, our method makes full use of the grouping information, which greatly improves the selection performance. Simulation studies have shown that our method could easily beat those ignoring the grouping information.

In Paper II, a model-free conditional screening method was proposed, in order to conduct conditional variable screening for ultrahigh dimension data, when prior information regarding certain predictors are available. Our method outperforms the state of the art method, CSIS, proposed by Barut, Fan and Verhasselt (2016) with nonlinear model structure, and is comparable to CSIS with generalized linear model. Simulation studies indicate that the proposed methods can provide screening results with much smaller model sizes, similar or better coverage rates, smaller false positive rates and/or false negative rates for nonlinear models. A real data analysis is also provided to illustrate the performance of our method.

In summary, in our first paper, we studied the dimension reduction problem for high dimensional data ($n < p$) from multiple populations using variable selection, and proposed a model-free method through sufficient dimension reduction framework. In our second paper, we developed a model free conditional screening method for high or ultra-high dimensional data to reduce the dimension to a reasonable size, when certain predictors need to be retained in the model based on prior information.

REFERENCES

Armesilla, A. L., Calvo, D., and Vega, M. A., 'Structural and functional characterization of the human cd36 gene promoter,' Journal of Biological Chemistry, 1996, **271**, pp. 7781–7787.

Barut, E., Fan, J., and Verhasselt, A., 'Conditional sure independence screening,' Journal of the American Statistical Association, 2016, **111**, pp. 1266–1277.

Boln-Canedo, V., Snchez-Maroo, N., and Alonso-Betanzos, A., *Feature Selection for High-Dimensional Data*, Springer Publishing Company, Incorporated, 1st edition, 2015, ISBN 3319218573, 9783319218571.

Bondell, H. D. and Li, L., 'Shrinkage inverse regression estimation for model-free variable selection,' Journal of the Royal Statistical Society, Ser. B, 2009, **71**, pp. 287–299.

Breiman, L., 'Better subset regression using the nonnegative garrote,' Technometrics, 1995, **37**, pp. 373–384.

Candes, E. and Tao, T., '2007,' The Dantzig selector: statistical estimation when p is much larger than n., 2007, **35**, pp. 2313–2351.

Chang, J., Tang, C., and Wu, Y., 'Marginal empirical likelihood and sure independence feature screening,' The Annals of Statistics, 2013, **41**, pp. 1693–2262.

Chen, X., Zou, C., and Cook, R. D., 'Coordinate-independent sparse sufficient dimension reduction and variable selection,' The Annals of Statistics, 2010, **38**, pp. 3696–3723.

Chiaromonte, F., Cook, R., and Li, B., 'Sufficient dimension reduction in regressions with categorical predictors,' The Annals of Statistics, 2002, **30**, pp. 475–497.

Cook, R. and Forzani, B., 'Likelihood-based sufficient dimension reduction,' Journal of the American Statistical Association, 2009, **104**, pp. 197–208.

Cook, R. D., *Regression Graphics*, Wiley, New York, 1998.

Cook, R. D., 'Testing predictor contributions in sufficient dimension reduction,' The Annals of Statistics, 2004, **32**, pp. 1062–1092.

Cook, R. D. and Nachtsheim, C., 'Reweighting to achieve elliptically contoured covariates in regression,' Journal of the American Statistical Association, 1994, **89**, pp. 592–599.

Cook, R. D. and Weisberg, S., 'Discussion of "sliced inverse regression for dimension reduction",' Journal of the American Statistical Association, 1991, **86**, pp. 328–332.

Cui, H., Li, R., and Zhong, W., 'Model-free feature screening for ultrahigh dimensional discriminant analysis,' Journal of the American Statistical Association, 2014, **110**, pp. 630–641.

Dong, Y. and Li, B., 'Dimension reduction for non-elliptically distributed predictors: second-order methods,' Biometrika, 2010, **97**, pp. 279–294.

Donoho, D. L., 'High-dimensional data analysis: The curses and blessings of dimensionality,' in 'AMS CONFERENCE ON MATH CHALLENGES OF THE 21ST CENTURY,' 2000 .

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R., 'Least angle regression (with discussion),' The Annals of Statistics, 2004, **32**, pp. 407–499.

Fan, J., Feng, Y., and Song, R., 'Nonparametric independence screening in sparse ultrahigh-dimensional additive models,' Journal of the American Statistical Association, 2011, **106**, pp. 544–5570.

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

Fan, J. and Li, R., 'Variable selection for cox's proportional hazards model and frailty model,' The Annals of Statistics, 2002, **30**, pp. 74–99.

Fan, J. and Li, R., 'Statistical challenges with high dimensionality: feature selection in knowledge discovery,' in 'Proceedings of the International Congress of Mathematicians Madrid, August 22-30, 2006,' 2007 pp. 595–622.

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

Fan, J., Samworth, R., and Wu, Y., 'Ultrahigh dimensional variable selection:beyond the linear model,' Journal of Machine Learning Research, 2009, **10**, pp. 1829–1853.

Fan, J. and Song, R., 'Sure independence screening in generalized linear models with np-dimensionality,' The Annals of Statistics, 2010, **38**, pp. 3567–3604.

Feng, Z., Wen, X., Yu, Z., and Zhu, L.-X., 'On partial sufficient dimension reduction with applications to partially linear multi-index models,' Journal of the American Statistical Association, 2013, **108**, pp. 237–246.

Golub, T., Slonim, D., Tamyo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., and Caligiuri, M., 'Molecular classification of cancer: class discovery and class prediction by gene expression monitoring,' Science, 1999, **286**, pp. 531–536.

Hall, P. and Li, K., 'On almost linearity of low dimensional projection from high dimensional data,' The Annals of Statistics, 1993, **21**, pp. 867–889.

He, X., Wang, L., and Hong, H., 'Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data,' The Annals of Statistics, 2013, **41**, pp. 342–369.

Hilafu, H. and Wu, W., 'Partial projective resampling method for dimension reduction: With applications to partially linear models,' Computational Statistics and Data Analysis, 2017, **109**, pp. 1–14.

Hong, H. G., Wang, L., and He, X., 'A data-driven approach to conditional screening of high-dimensional variables,' STAT, 2016, **5**, pp. 200–212.

Jiang, B. and Liu, J. S., 'Sliced inverse regression with variable selection and interaction detection,' Manuscript, 2013.

Lee, K., Li, B., and Chiaromonte, F., 'A general theory for nonlinear sufficient dimension reduction: Formulation and estimation,' The Annals of Statistics, 2013, **6**, pp. 3182–3210.

Li, B. and Dong, Y., 'Dimension reduction for non-elliptically distributed predictors,' The Annals of Statistics, 2009, **37**, pp. 1272–1298.

Li, B., Kim, M. K., and Altman, N., 'On dimension folding of matrix or array valued statistical objects,' The Annals of Statistics, 2010, **38**, pp. 1097–1121.

Li, B. and Wang, S., 'On directional regression for dimension reduction,' Journal of the American Statistical Association, 2007, **102**, pp. 997–1008.

Li, K. C., 'Sliced inverse regression for dimension reduction (with discussion),' Journal of the American Statistical Association, 1991, **86**, pp. 316–342.

Li, L., 'Sparse sufficient dimension reduction,' Biometrika, 2007, **94**, pp. 603–613.

Li, L., Cook, R., , and Nachtsheim, C., 'Model-free variable selection,' Journal of the Royal Statistical Society. Series B: Statistical Methodology, 2005, **67**, pp. 285–299.

Li, L. and Nachtsheim, C., 'Sparse sliced inverse regression,' Technometrics, 2006, **48**, pp. 503–510.

Li, L. and Yin, X., 'Sliced inverse regression with regularization,' Biometrics, 2008, **64**, pp. 124–131.

Li, R., Zhong, W., and Zhu, L., 'Feature screening via distance correlation learning,' Journal of the American Statistical Association, 2012, **107**, pp. 1129–1139.

Lin, L., Sun, J., and Zhu, L., 'Nonparametric feature screening,' Computational Statistics and Data Analysis, 2013, **67**, pp. 162–174.

Ma, Y. and Zhu, L., ‘A semiparametric approach to dimension reduction,’ Journal of the American Statistical Association, 2012, **107**, pp. 168–179.

Ma, Y. and Zhu, L., ‘Efficient estimation in sufficient dimension reduction,’ The Annals of Statistics, 2013, **41**, pp. 250–268.

Mai, Q. and Zou, H., ‘The kolmogorov filter for variable screening in high-dimensional binary classification.’ Biometrika, 2013, **100**, pp. 229–234.

Mai, Q. and Zou, H., ‘The fused kolmogorov filter: a nonparametric model-free screening method,’ The Annals of Statistics, 2015, **43**, pp. 1471–1497.

Ni, L., Cook, R. D., and Tsai, C. L., ‘A note on shrinkage sliced inverse regression,’ Biometrika, 2005, **92**, pp. 242–247.

Pan, R., Wang, H., and Li, R., ‘Ultrahigh dimensional multi-class linear discriminant analysis by pairwise sure independence screening’,’ Journal of the American Statistical Association, 2016, **111**, pp. 169–179.

Shao, Y., Cook, R. D., and Weisberg, S., ‘Marginal tests with sliced average variance estimation,’ Biometrika, 2007, **94**, pp. 285–296.

Tibshirani, R., ‘Regression shrinkage and selection via the lasso,’ Journal of the Royal Statistical Society, Ser. B, 1996, **58**, pp. 267–288.

Wang, H., ‘Forward regression for ultra-high dimensional variable screening,’ Journal of the American Statistical Association, 2009, **104**, pp. 1512–1524.

Wang, H., ‘Factor profiled sure independence screening,’ Biometrika, 2012, **99**, pp. 15–28.

Wang, T. and Zhu, L., ‘A distribution-based lasso for a general single-index model,’ Science China Mathematics, 2015, **58**, pp. 109–130.

Wen, X. and Cook, R. D., ‘Optimal sufficient dimension reduction for regressions with categorical predictors,’ Journal of Statistical Planning and Inference, 2007, **137**, pp. 1961–1978.

Wen, X. and Cook, R. D., ‘New approaches to model-free dimension reduction for bivariate regression,’ Journal of Statistical Planning and Inference, 2009, **139**, pp. 734–748.

Xia, Y., Tong, H., Li, W. K., and Zhu, L., ‘An adaptive estimation of dimension reduction space,’ Journal of the Royal Statistical Society, Ser. B, 2002, **64**, pp. 363–410.

Xue, L. and Zou, H., ‘Sure independence screening and compressed random sensing,’ Biometrika, 2011, **98**, pp. 371–380.

Yin, X. and Hilafu, H., ‘Sequential sufficient dimension reduction for large p, small n problems,’ Journal of the Royal Statistical Society, Ser. B, 2015, **77**, pp. 879–892.

Yu, Z., Dong, Y., and Zhu, L., ‘Trace pursuit: a general framework for model-free variable selection,’ Journal of the American Statistical Association, 2016, **111**, pp. 813–821.

Yuan, M. and Lin, Y., ‘Model selection and estimation in regression with grouped variables,’ Journal of the Royal Statistical Society, Ser. B, 2006, **68**, pp. 49–67.

Zhang, C., ‘Nearly unbiased variable selection under minimax concave penalty,’ The Annals of Statistics, 2010, **38**, pp. 894–942.

Zhao, S. and Li, Y., ‘Sure screening for estimating equations in ultra-high dimensions,’ Manuscript, 2012.

Zhong, W., Zhang, T., Zhu, M., and Liu, J. S., ‘Correlation pursuit: forward stepwise variable selection for index models,’ Journal of the Royal Statistical Society, Ser. B, 2012, **74**, pp. 849–870.

Zhu, L., Li, L., Li, R., and Zhu, L.-X., ‘Model-free feature screening for ultrahigh dimensional data,’ Journal of American Statistical Association, 2011, **106**, pp. 1464–1475.

Zhu, L. P., Wang, T., Zhu, L. X., and Ferré, L., ‘Sufficient dimension reduction through discretization-expectation estimation,’ Biometrika, 2010, **97**, pp. 295–304.

Zou, H., ‘The adaptive lasso and its oracle properties,’ Journal of the American Statistical Association, 2006, **101**, pp. 1418–1429.

VITA

In July 2010, Lei Huo graduated from Shandong Agricultural University with a B.S. in Computational Mathematics. In 2014 July, he finished the courses of Master's program in Statistics in Shandong University. In July 2018, he received a Doctor of Philosophy in Mathematics with a Statistics Emphasis from Missouri University of Science and Technology.