Taylor & Francis
Taylor & Francis Group

Journal of Nonparametric Statistics

ISSN: 1048-5252 (Print) 1029-0311 (Online) Journal homepage: https://www.tandfonline.com/loi/gnst20

Trace pursuit variable selection for multi-population data

Lei Huo, Xuerong Meggie Wen & Zhou Yu

To cite this article: Lei Huo, Xuerong Meggie Wen & Zhou Yu (2018) Trace pursuit variable selection for multi-population data, Journal of Nonparametric Statistics, 30:2, 430-447, DOI: 10.1080/10485252.2018.1430364

To link to this article: https://doi.org/10.1080/10485252.2018.1430364

JOURNAL OF NONPARAMETRIC STATISTICS, 2018
VOL. 30, NO. 2, 430–447
https://doi.org/10.1080/10485252.2018.1430364

Taylor & Francis
Taylor & Francis Group

Trace pursuit variable selection for multi-population data

Lei Huo[a], Xuerong Meggie Wen[a] and Zhou Yu[b]

[a]Department of Mathematics and Statistics, Missouri University of Science and Technology, Rolla, MO, USA; [b]School of Finance and Statistics, East China Normal University, Shanghai, People's Republic of 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.

ARTICLE HISTORY
Received 12 May 2017
Accepted 11 January 2018

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 generalised linear model is assumed. Such methods might generate biased results if the underlying modelling 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 $\mathbf{X} = (X_1, \ldots, X_p)^{\mathrm{T}}$ 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

$$Y \perp\!\!\!\perp \mathbf{X}_{\mathcal{A}^c} \mid \mathbf{X}_{\mathcal{A}}, \tag{1}$$

CONTACT Xuerong Meggie Wen wenx@mst.edu Department of Mathematics and Statistics, Missouri University of Science and Technology, Rolla, MO 65409, USA

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, Cook, and Tsai (2005), Li and Nachtsheim (2006) and Li and Yin (2008) proposed model-free variable selections by reformulating sufficient dimension reduction as a penalised 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, Zou, and Cook (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, Dong, and Zhu (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 utilisation 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 organised 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 $Y \mid \mathbf{X}$ 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 $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}$, where $\boldsymbol{\beta}$ is a $p \times d$ matrix with $d \leq p$, such that

$$Y \perp\!\!\!\perp \mathbf{X} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}. \tag{2}$$

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 $\mathcal{S}_{Y \mid \mathbf{X}}$. 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 modelling 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, Tong, Li, and Zhu 2002), directional regression (DR; Li and Wang 2007), likelihood acquired directions (LAD; Cook and Farzani, 2009) , cumulative slicing estimation (CUME; Zhu, Wang, Zhu, and Ferré 2010), dimension reduction for special-structured $\mathbf{X}$ (Li, Kim, and Altman 2010), nonlinear sufficient dimension reduction (Lee, Li, and Chiaromonte 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 $\boldsymbol{\Sigma} = \mathrm{Cov}(\mathbf{X})$ denote the marginal covariance matrix of $\mathbf{X}$, $\boldsymbol{\mu} = \mathrm{E}(\mathbf{X})$, and let $\mathbf{Z} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{X} - \mathrm{E}(\mathbf{X}))$ be the standardized predictor. By the invariance property (Cook 1998), we have $\mathcal{S}_{Y \mid \mathbf{X}} = \boldsymbol{\Sigma}^{-1/2}\mathcal{S}_{Y \mid \mathbf{Z}}$, where $\mathcal{S}_{Y \mid \mathbf{Z}}$ is the central subspace for the regression of $Y \mid \mathbf{Z}$. Unlike traditional regression modelling, sufficient dimension reduction methods rely on an assumption about the marginal distribution of $\mathbf{Z}$ instead of the conditional distribution of $Y \mid \mathbf{Z}$. The so-called *linearity condition* requires that $\mathrm{E}(\mathbf{Z} \mid \boldsymbol{\rho}^{\mathrm{T}}\mathbf{Z})$ be a linear function of $\boldsymbol{\rho}^{\mathrm{T}}\mathbf{Z}$, where the columns of the $p \times d$ matrix $\boldsymbol{\rho}$ form an orthonormal basis for $\mathcal{S}_{Y \mid \mathbf{Z}}$. For more detailed discussions of the linearity condition (LM condition), please see Feng, Wen, Yu, and Zhu (2013) .

The linearity condition connects the central subspace with the inverse regression of $\mathbf{Z}$ on Y . Li (1991) showed that $\mathrm{E}(\mathbf{Z} \mid Y) \in \mathcal{S}_{Y \mid \mathbf{Z}}$ when it holds. When Y is continuous, Li (1991) proposed estimating $\mathrm{E}(\mathbf{Z} \mid Y)$ 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 $p_h = \Pr\{Y \in s_h\}$, $\mathbf{m}_h = \mathrm{E}(\mathbf{Z} \mid Y \in s_h)$, $\mathbf{M}_{sir} = \sum_{h=1}^{H} p_h \mathbf{m}_h \mathbf{m}_h^{\mathrm{T}}$. Li (1991) showed that the eigenvectors corresponding to the d nonzero eigenvalues of $\mathbf{M}_{sir}$ form a basis of $\mathcal{S}_{Y \mid \mathbf{Z}}$.

Let $\hat{\mathbf{M}}_{sir}$ denote a consistent estimate of $\mathbf{M}_{sir}$, SIR made use of the span of the eigenvectors corresponding to the d largest eigenvalues of $\hat{\mathbf{M}}_{sir}$ 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 $\mathcal{S}_{Y \mid \mathbf{X}}$. 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 $\hat{\mathbf{M}}_{sir}$. SAVE (Cook and Weisberg 1991) and DR (Li and Wang 2007) took the same spectral decomposition approach via different kernel matrices: $\mathbf{M}_{save} = \mathrm{E}\{I_p - \mathrm{Var}(\mathbf{Z} \mid Y)\}^2$ and $\mathbf{M}_{dr} = 2\mathrm{E}\{\mathrm{E}^2(\mathbf{Z}\mathbf{Z}^{\mathrm{T}} \mid Y)\} + 2\mathrm{E}^2\{\mathrm{E}(\mathbf{Z} \mid Y)\mathrm{E}(\mathbf{Z}^{\mathrm{T}} \mid Y)\} + 2\mathrm{E}\{\mathrm{E}(\mathbf{Z}^{\mathrm{T}} \mid Y)\mathrm{E}(\mathbf{Z} \mid Y)\}\mathrm{E}\{\mathrm{E}(\mathbf{Z} \mid Y)\mathrm{E}(\mathbf{Z}^{\mathrm{T}} \mid Y)\} - 2I_p$. SAVE and

DR require a constant conditional variance condition ($\mathrm{Var}(\mathbf{Z} \mid \boldsymbol{\rho}^{\mathrm{T}}\mathbf{Z})$ is nonrandom) in addition to the linearity condition.

3. Trace pursuit variable selection for multiple groups

3.1. The test statistics

For ease 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:

$$P_{\mathcal{H}}\mathcal{S}_{Y \mid \mathbf{X}} = \mathcal{O}_p, \tag{3}$$

where $P(.)$ denotes the projection operator with respect to the standard inner product, $\mathcal{H} = \mathrm{Span}\{(\mathbf{0}_{(p-q)\times q}, \mathbf{I}_{p-q})^{\mathrm{T}}\}$ is the subspace of the predictor space, corresponding to the coordinates of the inactive predictors, and $\mathcal{O}_p$ is the origin in $\mathbb{R}^p$. Cook (2004) first proposed a test for testing hypothesis of (3) based on a generalised least square rederivation of the SIR estimator for $\mathcal{S}_{Y \mid \mathbf{X}}$. Shao, Cook, and Weisberg (2009) and many others also considered (3) based on other estimators of $\mathcal{S}_{Y \mid \mathbf{X}}$. However, all those tests will not be applicable when $n < p$, due to the difficulty of obtaining a sensible initial estimator for $\mathcal{S}_{Y \mid \mathbf{X}}$. Zhong, Zhang, Zhu, and Liu (2012) and Jiang and Liu (2014) tackled testing (3) via SIR method. However, both methods require the estimation of the rank of $\mathcal{S}_{Y \mid \mathbf{X}}$ (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, Cook, and Li (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* ($\mathcal{S}^{(W)}_{Y \mid \mathbf{X}}$) is defined as the intersection of all subspaces $\mathrm{Span}(\boldsymbol{\beta})$ satisfying

$$Y \perp\!\!\!\perp \mathbf{X} \mid (\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, W), \tag{4}$$

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 wth group, $\boldsymbol{\Sigma}_w = \mathrm{Var}(\mathbf{X}_w)$, and $\mathbf{Z}_w = \boldsymbol{\Sigma}_w{}^{-1/2}(\mathbf{X}_w - \boldsymbol{\mu}_w)$. Let $\mathcal{S}_{Y_w \mid \mathbf{X}_w}$ be the central subspace for the regression of $Y^w \mid \mathbf{X}^w$. The following equation (Chiaromonte et al. 2002) connects the partial central subspace with the conditional central subspaces:

$$\mathcal{S}^{(W)}_{Y \mid \mathbf{X}} = \sum_{w=1}^{K} \mathcal{S}_{Y_w \mid \mathbf{X}_w}. \tag{5}$$

Equation (5) 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 (5). Equation (5) also suggests that $\mathcal{S}^{(W)}_{Y\mid\mathbf{X}}$ contains each conditional central subspace $\mathcal{S}_{Y_w\mid\mathbf{X}_w}$.

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

$$Y \perp\!\!\!\perp \mathbf{X}_{\mathcal{A}^c} \mid (\mathbf{X}_{\mathcal{A}}, W), \tag{6}$$

where W is the group indicator. Adopting the concept of partial central subspace, (6) is equivalent to testing:

$$H_o : P_{\mathcal{H}}\mathcal{S}^{(W)}_{Y\mid\mathbf{X}} = \mathcal{O}_p, \tag{7}$$

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 $p_w = \Pr(W = w)$, $p_{hw} = \Pr\{Y_w \in s_{hw}\}$, $\mathbf{U}_{hw} = \mathrm{E}(\mathbf{X}_w \mid Y_w \in s_{hw})$. Based on (5), we can hence construct the kernel matrix for SIR as $\mathbf{M} = \sum_{w=1}^{K} p_w \boldsymbol{\Sigma}_w^{-1/2}(\sum_{h=1}^{Hw} p_{hw}\mathbf{U}_{hw}\mathbf{U}_{hw}^{\mathrm{T}})\boldsymbol{\Sigma}_w^{-1/2}$. For any index set $\mathcal{F}$, denote $\mathbf{X}_{\mathcal{F}} = \{X_i : i \in \mathcal{F}\}$, $\mathrm{Var}(\mathbf{X}_{\mathcal{F}} \mid W = w) = \boldsymbol{\Sigma}_{w\mathcal{F}}$ and $\mathbf{U}_{\mathcal{F},hw} = \mathrm{E}(\mathbf{X}_{\mathcal{F}} \mid Y_w \in s_{hw}, W = w)$. Define $\mathbf{M}_{\mathcal{F}} = \sum_{w=1}^{K} p_w \boldsymbol{\Sigma}_{w\mathcal{F}}^{-1/2}(\sum_{h=1}^{Hw} p_{hw}\mathbf{U}_{\mathcal{F},hw}\mathbf{U}_{\mathcal{F},hw}^{\mathrm{T}})\boldsymbol{\Sigma}_{w\mathcal{F}}^{-1/2}$, we have the following proposition.

Proposition 3.1: *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* $Y \perp\!\!\!\perp \mathbf{X}_{\mathcal{A}^c} \mid (\mathbf{X}_{\mathcal{A}}, W)$, *and* I_s *denotes the full index set.*

The proof of Proposition 3.1 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 $\mathrm{E}(X_j \mid \mathbf{X}_{\mathcal{F}}, W = w)$ 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 3.1: *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* $\mathcal{A} \not\subseteq \mathcal{F}$, *then* $\mathrm{tr}(\mathbf{M}_{\mathcal{F}\cup j}) - \mathrm{tr}(\mathbf{M}_{\mathcal{F}}) = \sum_{w=1}^{K} p_w(\sum_{h=1}^{Hw} p_{hw}\boldsymbol{\gamma}^2_{j\mid w\mathcal{F},hw})$, *where* $\boldsymbol{\gamma}_{j\mid w\mathcal{F},hw} = \mathrm{E}(\boldsymbol{\gamma}_{j\mid\mathcal{F}} \mid Y \in s_{h_w}, W = w)$ *with* $X_{j\mid\mathcal{F}} = X_j - \mathrm{E}(X_j \mid \mathbf{X}_{\mathcal{F}})$, $\sigma^2_{j\mid\mathcal{F}} = \mathrm{Var}(X_j \mid \mathcal{F})$ *and* $\boldsymbol{\gamma}_{j\mid\mathcal{F}} = X_{j\mid\mathcal{F}}/\sigma_{j\mid\mathcal{F}}$.

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 $\bar{\mathbf{X}}_w = (1/n_w)\sum_{i=1}^{n_w}\mathbf{X}_{wi}$ and $\hat{\boldsymbol{\Sigma}}_w = (1/n_w)\sum_{i=1}^{n_w}(\mathbf{X}_{wi} -$

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

Let $T_{j\,|\,\mathcal{F}} = n(\mathrm{tr}(\hat{\mathbf{M}}_{\mathcal{F}\cup j}) - \mathrm{tr}(\hat{\mathbf{M}}_{\mathcal{F}}))$ be the test statistic for hypothesis (6). Theorem 3.1 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 3.2: *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* w*th group* $(Y_w, \mathbf{X}_w)$ *for* $w = 1, \ldots, K$. *Assuming the subset linearity condition as in Theorem* 3.1, *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*

$$T_{j|\mathcal{F}} \longrightarrow \sum_{i=1}^{H} \omega_{j\,|\,\mathcal{F},i}^2 \chi_1^2,$$

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* $\mathbf{\Omega}_{j\,|\,\mathcal{F}}$ *as defined in the Appendix.*

3.2. The selection procedure

Following Yu et al. (2016), we use the forward trace pursuit (FTP) and 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

$$a_k = \arg\max_{j \in \mathcal{F}_{k-1}^c} \mathrm{tr}(\hat{\mathbf{M}}_{\mathcal{F}_{k-1}\cup j}).$$

(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

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

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

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

If $T_{d_{\mathcal{F}} \mid \mathcal{F}\setminus d_{\mathcal{F}}}$ 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 $\operatorname{Var}\{\operatorname{E}(\mathbf{Z}_w | Y \in s_{h_w})\}$ 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 $\beta_{\min} = \min_{\substack{w=1,\ldots,K \\ j \in \mathcal{A}}} \{\sqrt{\sum_{i=1}^{q_w} \beta_{wi,j}^2}\}$. 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 smallest eigenvalues of $\boldsymbol{\Sigma}_w$.

Proposition 3.2: *Assuming* $\operatorname{Span}\{\boldsymbol{\beta}_{w1}, \ldots, \boldsymbol{\beta}_{wq_w}\} = \mathcal{S}_{Y_w \mid \mathbf{X}_w}$ *and the subset linearity condition as in Theorem* 3.1, *then for any index set* $\mathcal{F}$ *such that* $\mathcal{F}^c \cap \mathcal{A} \neq \varnothing$, *we have*

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

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

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

Condition 3.1: *Assuming that there exist* $\alpha > 0$ *and* $0 < \theta < 1/2$ *such that*

$$\min_{\mathcal{F}:\mathcal{F}^c \cap \mathcal{A} \neq \varnothing} \max_{j \in \mathcal{F}^c \cup \mathcal{A}} \{\operatorname{tr}(\mathbf{M}_{\mathcal{F}\cup j} - \operatorname{tr}(\mathbf{M}_{\mathcal{F}})\} \geq \alpha n^{-\theta} \tag{8}$$

Theorem 3.3: *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* w*th 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* 3.1, *then*

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

and

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

Theorem 3.3 provides the selection consistency result for the STP method. It suggests that the addition step will not stop till 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 3.2: (a) $\mathbf{X}_w$ *follows a multinormal distribution for* $w = 1, \ldots, K$.

(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* $\log p \le \alpha_1 n^{\theta_1}$, $|\mathcal{A}| \le \alpha_1 n^{\theta_2}$ *and* $2\theta + \theta_1 + \theta_2 < 1$, *where* θ *is a constant from Condition* 3.1.

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

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

Theorem 3.4: *Assume Conditions* 3.1 *and* 3.2 *hold true, then we have*

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

as $n \to \infty$ *and* $p \to \infty$, *where* $\hat{m} = \mathrm{argmin}_{1\le k\le n} BIC(\mathcal{F}_k)$, *and* $\mathcal{F}_k$ *is defined in the FTP procedure.*

Hence Theorem 3.4 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 summarise 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:

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

$\mathbf{X} = (X_1, \ldots, X_p) \sim N(\mathbf{0}, \mathbf{\Sigma})$, 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 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 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 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 utilisation of grouping information has greatly improved the correct selection rates. Unreported simulation results

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

Table 3. Selection performances (50 runs) for Model **II** with $W \sim \text{Bin}(0.7)$.

		$\rho = 0$				$\rho = 0.5$			
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

suggest that SAVE-based and DR-based methods provide similar performance as that of Model I.

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. Tables 4 and 5 report the performances of the multiple group and single 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,

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

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, $\mathbf{X}$, W and ϵ's are generated the same way as in Model I. As suggested by Yu et al., this model is specially constructed to favour DR-based methods. As shown in Tables 6 and 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.

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

Model **V** Model V is generated as follows:

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

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

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

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 have 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 two. Table 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

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

I except for Y , which is generated as

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

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 9 showedour 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.

5. Conclusion 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 utilises the grouping information. Simulation studies show that our method provides superior performance comparing to those ignoring the grouping information.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by National Natural Science Foundation of China [1157111], the program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning, the Shanghai Rising-Star Program [16QA1401700], and the 111 project [B14019].

References

Bondell, H.D., and Li, L. (2009), 'Shrinkage Inverse Regression Estimation for Model-Free Variable Selection', *Journal of the Royal Statistical Society: Series B*, 71, 287–299.

Breiman, L. (1995), 'Better Subset Regression using the Nonnegative Garrote', *Technometrics*, 37, 373–384.

Candes, E., and Tao, T. (2007), 'The Dantzig Selector: Statistical Estimation when p is Much Larger than n', *The Annals of Statistics*, 35, 2313–2351.

Chen, J., and Chen, Z. (2008), 'Extended Bayesian Information Criteria for Model Selection with Large Model Spaces', *Biometrika*, 95, 759–771.

Chen, X., Zou, C., and Cook, R.D. (2010), 'Coordinate-Independent Sparse Sufficient Dimension Reduction and Variable Selection', *The Annals of Statistics*, 38, 3696–3723.
Chiaromonte, F., Cook, R.D., and Li, B. (2002), 'Sufficient Dimensions Reduction in Regressions with Categorical Predictors', *The Annals of Statistics*, 30, 475–497.
Cook, R.D. (1998), *Regression Graphics*, New York: Wiley.
Cook, R.D. (2004), 'Testing Predictor Contributions in Sufficient Dimension Reduction', *The Annals of Statistics*, 32, 1062–1092.
Cook, R.D., and Forzani, B. (2009), 'Likelihood-Based Sufficient Dimension Reduction', *Journal of the American Statistical Association*, 104, 197–208.
Cook, R.D., and Weisberg, S. (1991), 'Discussion of "Sliced Inverse Regression for Dimension Reduction"', *Journal of the American Statistical Association*, 86, 328–332.
Fan, J., and Li, R. (2001), 'Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties', *Journal of the American Statistical Association*, 96, 1348–1360.
Feng, Z., Wen, X.M., Yu, Z., and Zhu, L.-X. (2013), 'On Partial Sufficient Dimension Reduction with Applications to Partially Linear Multi-Index Models', *Journal of the American Statistical Association*, 108, 237–246.
Fernholz, L.T. (1983), *Von Mises Calculus for Statistical Functionals*, New York: Springer.
Jiang, B., and Liu, J.S. (2014), 'Sliced inverse regression with variable selection and interaction detection.' *The Annals of Statistics*, 42, 1751–1786.
Lee, K.Y., Li, B., and Chiaromonte, F. (2013), 'A General Theory for Nonlinear Sufficient Dimension Reduction: Formulation and Estimation', *The Annals of Statistics*, 6, 3182–3210.
Li, K.C. (1991), 'Sliced Inverse Regression for Dimension Reduction (with Discussion)', *Journal of the American Statistical Association*, 86, 316–327.
Li, L. (2007), 'Sparse Sufficient Dimension Reduction', *Biometrika*, 94, 603–613.
Li, L., and Nachtsheim, C.J. (2006), 'Sparse Sliced Inverse Regression', *Technometrics*, 48, 503–510.
Li, B., and Wang, S. (2007), 'On Directional Regression for Dimension Reduction', *Journal of the American Statistical Association*, 102, 997–1008.
Li, L., and Yin, X. (2008), 'Sliced Inverse Regression with Regularizations', *Biometrics*, 64, 124–131.
Li, B., Kim, M.K., and Altman, N. (2010), 'On Dimension Folding of Matrix or Array Valued Statistical Objects', *The Annals of Statistics*, 38, 1097–1121.
Ma, Y., and Zhu, L. (2012), 'A Semiparametric Approach to Dimension Reduction', *Journal of the American Statistical Association*, 107, 168–179.
Ma, Y., and Zhu, L. (2013), 'Efficient Estimation in Sufficient Dimension Reduction', *The Annals of Statistics*, 41, 250–268.
Ni, L., Cook, R.D., and Tsai, C.-L. (2005), 'A Note on Shrinkage Sliced Inverse Regression', *Biometrika*, 92, 242–247.
Shao, Y., Cook, R.D., and Weisberg, S. (2009), 'Partial Central Subspace and Sliced Average Variance Estimation', *Journal of Statistical Planning and Inference*, 139, 952–961.
Tibshirani, R. (1996), 'Regression Shrinkage and Selection via the Lasso', *Journal of the Royal Statistical Society, Series B*, 58, 267–288.
Wang, H. (2009), 'Forward Regression for Ultra-High Dimensional Variable Screening', *Journal of the American Statistical Association*, 104, 1512–1524.
Wang, T., and Zhu, L. (2015), 'A Distribution-based LASSO for a General Single-Index Model', *Science China Mathematics*, 58, 109–130.
Wen, X., and Cook, R.D. (2007), 'Optimal Sufficient Dimension Reductionin Regressions with Categorical Predictors', *Journal of Statistical Planning and Inference*, 137, 1961–1978.
Wen, X., and Cook, R.D. (2009), 'New Approaches to Model-Free Dimension Reduction for Bivariate Regression', *Journal of Statistical Planning and Inference*, 139, 734–748.
Xia, Y., Tong, H., Li, W.K., and Zhu, L.-X. (2002), 'An Adaptive Estimation of Dimension Reduction Space', *Journal of the Royal Statistical Society: Series B*, 64, 363–410.
Yin, X., and Hilafu, H. (2015), 'Sequential Sufficient Dimension Reduction for Large p, Small n Problems', *Journal of the Royal Statistical Society: Series B*, 77, 879–892.

Yu, Z., Dong, Y., and Zhu, L.-X. (2016), 'Trace Pursuit: A General Framework for Model-Free Variable Selection', *Journal of the American Statistical Association*, 111, 813–821.
Yuan, M., and Lin, Y. (2006), 'Model Selection and Estimation in Regression with Grouped Variables', *Journal of the Royal Statistical Society: Series B*, 68, 49–67.
Zhang, C.-H. (2010), 'Nearly Unbiased Variable Selection under Minimax Concave Penalty', *The Annals of Statistics*, 38, 894–942.
Zhong, W., Zhang, T., Zhu, M., and Liu, J.S. (2012), 'Correlation Pursuit: Forward Stepwise Variable Selection for Index Models', *Journal of the Royal Statistical Society: Series B*, 74, 849–870.
Zhu, L.P., Wang, T., Zhu, L.X., and Ferré, L. (2010), 'Sufficient Dimension Reduction through Discretization–Expectation Estimation', *Biometrika*, 97, 295–304.
Zou, H. (2006), 'The Adaptive Lasso and its Oracle Properties', *Journal of the American Statistical Association*, 101, 1418–1429.

Appendix

Proof of Proposition 3.1**.:** Assume $\mathrm{Var}\{\mathrm{E}(\mathbf{Z}_w \mid Y \in s_{h_w})\}$ 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 $\boldsymbol{\beta}_{wi} \in \mathcal{S}_{Y_w \mid \mathbf{X}_w}$. 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 $Y \perp\!\!\!\perp \mathbf{X}_{\mathcal{A}^c} \mid (\mathbf{X}_{\mathcal{A}}, W)$, $\boldsymbol{\beta}_{wi,\mathcal{A}^c} = \mathbf{0}$ 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

$$\begin{aligned}\mathrm{Var}(\mathrm{E}(\mathbf{X} \mid Y) \mid W = w) &= \boldsymbol{\Sigma}_w\left(\sum_{i=1}^{q_w}\lambda_{wi}\boldsymbol{\beta}_{wi}\boldsymbol{\beta}_{wi}^{\top}\right)\boldsymbol{\Sigma}_w \\ &= \begin{pmatrix}\boldsymbol{\Sigma}_{w,\mathcal{A}} & \boldsymbol{\Sigma}_{w,\mathcal{A}\mathcal{A}^c}\\ \boldsymbol{\Sigma}_{w,\mathcal{A}^c\mathcal{A}} & \boldsymbol{\Sigma}_{w,\mathcal{A}^c}\end{pmatrix}\begin{pmatrix}\sum_{i=1}^{q_w}\lambda_{wi}\boldsymbol{\beta}_{wi,\mathcal{A}}\boldsymbol{\beta}_{wi,\mathcal{A}}^{\top} & 0\\ 0 & 0\end{pmatrix}\begin{pmatrix}\boldsymbol{\Sigma}_{w,\mathcal{A}} & \boldsymbol{\Sigma}_{w,\mathcal{A}\mathcal{A}^c}\\ \boldsymbol{\Sigma}_{w,\mathcal{A}^c\mathcal{A}} & \boldsymbol{\Sigma}_{w,\mathcal{A}^c}\end{pmatrix}, \qquad \text{(A1)}\end{aligned}$$

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

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

and

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

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 3.1**.:** (i) Since $\mathcal{A} \subseteq \mathcal{F}$, $\mathcal{A} \subseteq \mathcal{F} \cup j$. From Proposition 3.1, 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 $X_{wj \mid \mathcal{F}} = X_{wj} - \mathrm{E}(X_{wj} \mid \mathbf{X}_{w\mathcal{F}}) = X_{wj} - \boldsymbol{\Sigma}_{w,j\mathcal{F}}^{\mathrm{T}}\boldsymbol{\Sigma}_{w\mathcal{F}}^{-1}\mathbf{X}_{w\mathcal{F}}$ for any $w \in \{1, \ldots, K\}$, where $\boldsymbol{\Sigma}_{w,j\mathcal{F}} = \mathrm{Cov}(X_j, \mathbf{X}_{\mathcal{F}} \mid W = w)$. We construct

two matrices $\mathbf{P}_w$ and $\mathbf{V}_w$ as

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

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 $\mathrm{Var}(\mathbf{P}_w \mathbf{X}_{w,\mathcal{F}\cup j}) = \mathbf{P}_w \boldsymbol{\Sigma}_{w,\mathcal{F}\cup j} \mathbf{P}_w^{\top} = \mathbf{V}_w$ and $\boldsymbol{\Sigma}_{w,\mathcal{F}\cup j} = \mathbf{P}_w^{\top} \mathbf{V}_w^{-1} \mathbf{P}_w$.

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

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

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

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

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

Proof of Theorem 3.2**.:** 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

$$\mathcal{G}(F_{nw}) = \mathcal{G}(F_w) + \mathrm{E}_n[\mathcal{G}^{\star}(F_w)] + \mathcal{O}_p(n_w^{-1}), \tag{A2}$$

where $\mathcal{G}(F_w)$ is fixed for each group, and $\mathrm{E}_n[\mathcal{G}^{\star}(F_w)] = \mathcal{O}_p(n_w^{-1/2})$ as $\mathrm{E}[\mathcal{G}^{\star}(F_w)] = 0$. Let $R_{hw} = I(Y_w \in s_{hw})$, $\mu_{j,hw} = \mathrm{E}(X_j \mid Y_w \in s_{hw}, W = w)$ and $\nu_{wj\,|\,\mathcal{F}} = \boldsymbol{\Sigma}_{w,\mathcal{F}}^{-1} \boldsymbol{\Sigma}_{w,j\mathcal{F}}^{\top}$. To prove Theorem 3.2, we need the results in Lemma A.1 in the following.

Lemma A.1: *If the conditions in* 3.2 *holds and H_o is true, then $\hat{\boldsymbol{\Sigma}}_{w,\mathcal{F}}$, $\hat{\boldsymbol{\Sigma}}_{w,\mathcal{F}}^{-1}$, $\hat{\mathbf{U}}_{\mathcal{F}w,h}$, $\hat{\nu}_{wj\,|\,\mathcal{F}}$, $\mu_{j,hw}$ and $\hat{\boldsymbol{\gamma}}_{j\,|\,\mathcal{F}w,hw}$ have expansions in the form* (A.2) *with $\boldsymbol{\Sigma}_{w,\mathcal{F}}$, $\boldsymbol{\Sigma}_{w,\mathcal{F}}^{-1}$, $\mathbf{U}_{\mathcal{F}w,h}$, $\nu_{wj\,|\,\mathcal{F}}$, $\hat{\mu}_{j,hw}$ or $\boldsymbol{\gamma}_{j\,|\,\mathcal{F}w,hw}$ as substitutes for $\mathcal{G}(F_w)$, and $\boldsymbol{\Sigma}_{w,\mathcal{F}}^{\star} = \mathbf{X}_{w,\mathcal{F}} \mathbf{X}_{w,\mathcal{F}}^{\top}$, $(\boldsymbol{\Sigma}_{w,\mathcal{F}}^{-1})^{\star} = -\boldsymbol{\Sigma}_{w,\mathcal{F}}^{-1} \boldsymbol{\Sigma}_{w,\mathcal{F}}^{\star} \boldsymbol{\Sigma}_{w,\mathcal{F}}^{-1}$, $\mathbf{U}_{\mathcal{F}w,h}^{\star} = (\mathbf{X}_{w,\mathcal{F}} - \mathbf{U}_{\mathcal{F}w,h}) R_{hw} / p_{hw} - \mathbf{X}_{w,\mathcal{F}}$, $\nu_{wj\,|\,\mathcal{F}}^{\star} = \boldsymbol{\Sigma}_{w,\mathcal{F}}^{-1} (X_{wj} \mid \mathbf{X}_{w\mathcal{F}} - \mathrm{E}(X_{wj} \mid \mathbf{X}_{w\mathcal{F}})) + (\boldsymbol{\Sigma}_{w,\mathcal{F}}^{-1})^{\star} \mathrm{E}(X_{wj} \mid \mathbf{X}_{w\mathcal{F}})$, $\mu_{j,hw}^{\star} = (X_{w,j} - \mathbf{U}_{jw,h}) R_{hw} / p_{hw} - X_{w,j}$ or $\boldsymbol{\gamma}_{j\,|\,\mathcal{F}w,hw}^{\star} = (\mu_{j,hw}^{\star} - (\nu_{wj\,|\,\mathcal{F}}^{\star})^{\top} \mathbf{U}_{\mathcal{F}w,h} - \nu_{wj\,|\,\mathcal{F}}^{\top} \mathbf{U}_{\mathcal{F}w,h}^{\star}) / \sigma_{w,j\,|\,\mathcal{F}}$ as substitutes for $\mathcal{G}^{\star}(F_w)$.*

Since the proof is similar to Yu et al. (2016), we omit the proof for Lemma A.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 A.1, we define $(\mathbf{L}_{j\,|\,\mathcal{F},w})^\star = (p_w^{1/2}p_{\{hw=1\}}^{1/2}\boldsymbol{\gamma}_{j\,|\,\mathcal{F}w,1}^\star,\ldots,p_w^{1/2}p_{\{hw=Hw\}}^{1/2}\boldsymbol{\gamma}_{j\,|\,\mathcal{F}w,Hw}^\star)^\top$, $(\mathbf{L}_{j\,|\,\mathcal{F}})^\star = (\mathbf{L}_{j\,|\,\mathcal{F},1}^\star)^\top,\ldots,(\mathbf{L}_{j\,|\,\mathcal{F},K}^\star)^\top)^\top$ and $\boldsymbol{\Omega}_{j\,|\,\mathcal{F}} = \mathrm{E}(\mathbf{L}_{j\,|\,\mathcal{F},1}^\star(\mathbf{L}_{j\,|\,\mathcal{F},1}^\star)^\top)$. 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

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

Then the result in Theorem 3.2 follows directly. ■

Proof of Proposition 3.2**.:** Without loss of generality, we assume that $(\mathbf{X}_{\mathcal{F}}^\top, X_j)$ are the first $|\mathcal{F}|+1$ elements of $\mathbf{X}^{\mathrm{T}}$ 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 $\mathrm{tr}(\mathbf{M}_{\mathcal{F}\cup j}) - \mathrm{tr}(\mathbf{M}_{\mathcal{F}}) = \sum_{w=1}^K p_w(\sum_{h=1}^{Hw}p_{hw}\boldsymbol{\gamma}_{j\,|\,\mathcal{F}w,hw}^2)$, then we have

$$\begin{aligned}
\sigma_{w,j\,|\,\mathcal{F}}^2\left(\sum_{h=1}^{Hw}p_{hw}\boldsymbol{\gamma}_{j\,|\,\mathcal{F}w,hw}^2\right) &= \mathrm{Var}(\mathrm{E}(X_{j\,|\,\mathcal{F}}\,|\,\mathbf{Y})\,|\,W=w)\\
&= \left(-\boldsymbol{\Sigma}_{w,j\mathcal{F}}\boldsymbol{\Sigma}_{w\mathcal{F}}^{-1},1\right)\mathbf{A}\mathrm{Var}(\mathrm{E}(\mathbf{X}\,|\,Y)\,|\,W=w)\mathbf{A}^\top\left(-\boldsymbol{\Sigma}_{w,j\mathcal{F}}\boldsymbol{\Sigma}_{w\mathcal{F}}^{-1},1\right)^\top\\
&= \left(-\boldsymbol{\Sigma}_{w,j\mathcal{F}}\boldsymbol{\Sigma}_{w\mathcal{F}}^{-1},1\right)\mathbf{A}\boldsymbol{\Sigma}_w(\sum_{i=1}^{q_w}\lambda_{wi}\boldsymbol{\beta}_{wi}\boldsymbol{\beta}_{wi}^\top)\boldsymbol{\Sigma}_w\mathbf{A}^\top\left(-\boldsymbol{\Sigma}_{w,j\mathcal{F}}\boldsymbol{\Sigma}_{w\mathcal{F}}^{-1},1\right)^\top
\end{aligned}\tag{A3}$$

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

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

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

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

From this equation and A3, we can obtain that

$$\sigma_{w,j\,|\,\mathcal{F}}\left(\sum_{h=1}^{Hw}p_{hw}\boldsymbol{\gamma}_{j\,|\,\mathcal{F}w,hw}^2\right) = \sum_{i=1}^{q_w}\lambda_{wi}\{\left(\boldsymbol{\Sigma}_{wj\tilde{\mathcal{F}}} - \boldsymbol{\Sigma}_{w,j\mathcal{F}}\boldsymbol{\Sigma}_{w\mathcal{F}}^{-1}\boldsymbol{\Sigma}_{w\mathcal{F}\tilde{\mathcal{F}}},1\right)\boldsymbol{\beta}_{wi,\tilde{\mathcal{F}}}\}^2\tag{A4}$$

Note that $\sum_{j\in\tilde{\mathcal{F}}}\{(\boldsymbol{\Sigma}_{wj\tilde{\mathcal{F}}} - \boldsymbol{\Sigma}_{w,j\mathcal{F}}\boldsymbol{\Sigma}_{w\mathcal{F}}^{-1}\boldsymbol{\Sigma}_{w\mathcal{F}\tilde{\mathcal{F}}},1)\boldsymbol{\beta}_{wi,\tilde{\mathcal{F}}}\}^2 = \boldsymbol{\beta}_{wi,\tilde{\mathcal{F}}}{}^\top(\boldsymbol{\Sigma}_{w,\tilde{\mathcal{F}}} - \boldsymbol{\Sigma}_{w,\tilde{\mathcal{F}}\mathcal{F}}\boldsymbol{\Sigma}_{w\mathcal{F}}^{-1}\boldsymbol{\Sigma}_{w\mathcal{F}\tilde{\mathcal{F}}},1)\boldsymbol{\beta}_{wi,\tilde{\mathcal{F}}}$ and $\lambda_{\min}(\boldsymbol{\Sigma}_{w,\tilde{\mathcal{F}}} - \boldsymbol{\Sigma}_{w,\tilde{\mathcal{F}}\mathcal{F}}\boldsymbol{\Sigma}_{w\mathcal{F}}^{-1}\boldsymbol{\Sigma}_{w\mathcal{F}\tilde{\mathcal{F}}},1) = \lambda_{\max}^{-1}\{(\boldsymbol{\Sigma}_{w,\tilde{\mathcal{F}}} - \boldsymbol{\Sigma}_{w,\tilde{\mathcal{F}}\mathcal{F}}\boldsymbol{\Sigma}_{w\mathcal{F}}^{-1}\boldsymbol{\Sigma}_{w\mathcal{F}\tilde{\mathcal{F}}},1)^{-1}\} \geq \lambda_{\max}^{-1}(\boldsymbol{\Sigma}_w) = \lambda_{\min}(\boldsymbol{\Sigma}_w)$ for any $w\in\{1,\ldots,K\}$, then it follows

$$\begin{aligned}
&\max_{j\in\mathcal{F}^c\cap\mathcal{A}}\sigma_{w,j\,|\,\mathcal{F}}\left(\sum_{h=1}^{Hw}p_{hw}\boldsymbol{\gamma}_{j|\mathcal{F}w,hw}^2\right)\\
&\geq |\mathcal{F}^c\cap\mathcal{A}|^{-1}\sum_{j\in\tilde{\mathcal{F}}}\{\left(\boldsymbol{\Sigma}_{wj\tilde{\mathcal{F}}} - \boldsymbol{\Sigma}_{w,j\mathcal{F}}\boldsymbol{\Sigma}_{w\mathcal{F}}^{-1}\boldsymbol{\Sigma}_{w\mathcal{F}\tilde{\mathcal{F}}},1\right)\boldsymbol{\beta}_{wi,\tilde{\mathcal{F}}}\}^2\\
&= |\mathcal{F}^c\cap\mathcal{A}|^{-1}\sum_{i=1}^{q_w}\lambda_{wi}\boldsymbol{\beta}_{wi,\tilde{\mathcal{F}}}{}^\top\left(\boldsymbol{\Sigma}_{w,\tilde{\mathcal{F}}} - \boldsymbol{\Sigma}_{w,\tilde{\mathcal{F}}\mathcal{F}}\boldsymbol{\Sigma}_{w\mathcal{F}}^{-1}\boldsymbol{\Sigma}_{w\mathcal{F}\tilde{\mathcal{F}}},1\right)\boldsymbol{\beta}_{wi,\tilde{\mathcal{F}}}\\
&\geq |\mathcal{F}^c\cap\mathcal{A}|^{-1}\sum_{i=1}^{q_w}\lambda_{wi}\lambda_{\min}\left(\boldsymbol{\Sigma}_{w,\tilde{\mathcal{F}}} - \boldsymbol{\Sigma}_{w,\tilde{\mathcal{F}}\mathcal{F}}\boldsymbol{\Sigma}_{w\mathcal{F}}^{-1}\boldsymbol{\Sigma}_{w\mathcal{F}\tilde{\mathcal{F}}},1\right)\boldsymbol{\beta}_{wi,\tilde{\mathcal{F}}}{}^\top\boldsymbol{\beta}_{wi,\tilde{\mathcal{F}}}\\
&\geq \lambda_{w,q_w}\lambda_{\min}(\boldsymbol{\Sigma}_w)^2\beta_{\min}
\end{aligned}\tag{A5}$$

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

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

■

Proof of Theorem 3.3.**:** (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 $(\mathrm{tr}(\hat{\mathbf{M}}_{\mathcal{F}\cup j}) - \mathrm{tr}(\hat{\mathbf{M}}_{\mathcal{F}})) - (\mathrm{tr}(\mathbf{M}_{\mathcal{F}\cup j}) - \mathrm{tr}(\mathbf{M}_{\mathcal{F}})) = O_p(n^{-1/2})$ as $\mathcal{F}^c \cap \mathcal{A} \neq \varnothing$ and $0 < c_1 < 1/2$,

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

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

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

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 3.2. 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 3.4.**:** Let $R_{w,j|\mathcal{F}} = \mathrm{Var}(\mathrm{E}(X_{j|\mathcal{F}} \mid \mathbf{Y}) \mid W = w)$ and $\hat{R}_{w,j|\mathcal{F}}$ 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 $\mathrm{tr}(\hat{\mathbf{M}}_{\mathcal{F}\cup j}) - \mathrm{tr}(\hat{\mathbf{M}}_{\mathcal{F}}) = \sum_{w=1}^{w=K} \hat{p}_w \hat{\sigma}^{-2}_{w,j|\mathcal{F}} \hat{R}_{w,j|\mathcal{F}}$. Suppose that $|\mathcal{F}| = O(n^{\theta+\theta_2})$. From Lemma 7 in Yu et al. (2016), we know that$|\hat{R}_{w,j|\mathcal{F}} - R_{w,j|\mathcal{F}}| \leq D_0|\mathcal{F}|\sqrt{\log p/n}$ with probability tending to 1, where D_0 is some constant. Since $\hat{p}_w - \hat{p}_w = O_P(n^{-1/2})$ and

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

there exists some constant D_1 such that

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

with probability tending to 1. Based on the proof of Lemma 3 in Jiang and Liu (2014) 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 $\Pr(\mathcal{A} \subset \mathcal{F}_{2H\alpha^{-1}An^{\theta+\theta 2}}) \to 1$, as $n \to \infty$ and $p \to \infty$. 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 on the proof of Theorem 2 in Wang (2009), and we omit the details. ■