Statistica Sinica **26** (2016), 1159-1174
doi:http://dx.doi.org/10.5705/ss.2014.111

MODEL-FREE COORDINATE TEST AND VARIABLE SELECTION VIA DIRECTIONAL REGRESSION

Zhou Yu and Yuexiao Dong

East China Normal University and Temple University

Abstract: We consider the paradigm of marginal coordinate tests (Cook (2004)) for model-free variable selection. To combine the strength of existing tests based on sliced inverse regression and sliced average variance estimation, we design a new marginal coordinate test via directional regression. Given the method-specific marginal coordinate test statistics, a maximum ratio criteria is proposed to facilitate model-free variable selection. Under mild conditions, the variable-selection consistency is guaranteed when n goes to ∞.

Key words and phrases: Marginal coordinate test, maximum ratio criteria, sliced average variance estimation, sliced inverse regression.

1. Introduction

While most variable selection approaches are based on some model, and selecting active predictors is part of the overall process of fitting a model based on the data, one can argue the importance of identifying the explanatory variables that have detectable effects without assuming any model. See, for example, Cox and Snell (1974). Let $\mathbf{X} = (X_1, \ldots, X_p)^{\tau}$ be the p-dimensional predictor and Y be the univariate response. Model-free variable selection aims to select a subset of the predictors without assuming the functional form between Y and $\mathbf{X}$, such that the selected predictors are sufficient to predict $F(Y|\mathbf{X})$, the conditional distribution of Y given $\mathbf{X}$. Let $\mathcal{I} = \{1, \ldots, p\}$ be the full index set. The active set $\mathcal{A}$ and the the inactive set $\mathcal{A}^c$ are defined as

$$\mathcal{A} = \{k \in \mathcal{I} : F(Y \mid \mathbf{X}) \text{ functionally depends on } X_k\} \text{ and}$$
$$\mathcal{A}^c = \{k \in \mathcal{I} : F(Y \mid \mathbf{X}) \text{ does not functionally depend on } X_k\}.$$

Let $\mathbf{X}_{\mathcal{A}} = \{X_i : i \in \mathcal{A}\}$ denote the vector that contains all the active predictors. Then we have that Y is independent of $\mathbf{X}$ conditioning on $\mathbf{X}_{\mathcal{A}}$. We write $Y \perp\!\!\!\perp \mathbf{X}|\mathbf{X}_{\mathcal{A}}$, where "$\perp\!\!\!\perp$" means independency.

The hypothesis testing paradigm for model-free variable selection was first introduced in Cook (2004). To test wether the ith predictor is active or not, one can consider

$$H_0 : i \in \mathcal{A}^c \text{ versus } H_a : i \in \mathcal{A}. \tag{1.1}$$

This is known as the marginal coordinate hypothesis in Cook (2004). Several tests for (1.1) have been introduced in the literature. See, for example, Cook (2004) and Shao, Cook, and Weisberg (2007). Based on a chosen test for (1.1), Li, Cook, and Nachtsheim (2005) suggested a sequential test approach for model-free variable selection. This procedure can be viewed as an adaptation of the standard normal theory backward elimination procedure, and it replaces the t-test with the marginal coordinate test. In the presence of nonlinear link functions between Y and $\mathbf{X}$, Li, Cook, and Nachtsheim (2005) demonstrated that the sequential test procedure performs better than model-based variable selection methods such as AIC, BIC, and the classical backward elimination based on the t-test.

We address two limitations of this paradigm. First, the existing tests for (1.1) are based on sliced inverse regression (SIR) (Li (1991)) and sliced average variance estimation (SAVE) (Cook and Weisberg (1991)). As two popular methods for sufficient dimension reduction (Li (1991); Cook (1998)), the shortcomings of SIR and SAVE are well-known in the literature: SIR does not work well when the link function between the response and the predictor is symmetric, and SAVE may not be very effective with monotone link functions. These shortcomings could potentially be inherited by their respective marginal coordinate tests. This motivates us to propose the marginal coordinate test based on directional regression. Directional regression (Li and Wang (2007)) is an effective sufficient dimension reduction method that inherently combines SIR and SAVE. Not surprisingly, the newly designed directional regression-based test can be viewed as an approximate combination of the SIR-based test in Cook (2004) and the SAVE-based test in Shao, Cook, and Weisberg (2007), and we expect it to work well across a wide range of link functions.

The second limitation of this paradigm is the need of implementing sequential tests for model-free variable selection. Without loss of generality, denote T_i as the test statistic for $H_0 : i \in \mathcal{A}^c$ versus $H_a : i \in \mathcal{A}$ based on SIR, SAVE, or directional regression. The asymptotic null distribution of the method-specific T_i is a sum of weighted $\chi^2(1)$ distributions under mild conditions, where $\chi^2(1)$ denotes a chi-square distribution with one degree of freedom, and the unknown weights have to be estimated in practice. Implementing sequential tests based on T_i can be time-consuming, especially when the sample size n is large. Furthermore, it is not clear whether the backward elimination procedure is variable-selection consistent when n goes to ∞. This leads to the second contribution of the paper: a novel maximum ratio criteria (MRC) for model-free variable selection. This criteria is quite general, and can be combined with marginal coordinate tests based on any of SIR, SAVE and directional regression. We can bypass approximation of the asymptotic null distribution of T_i, as only the ratios of consecutively ranked test statistics T_i are needed. The variable-selection consistency of the MRC procedure is theoretically justified under mild conditions.

The rest of the paper is organized as follows. In Section 2, we briefly review existing marginal coordinate tests in the sufficient dimension reduction literature. The new test based on directional regression is proposed in Section 3. Maximum ratio criteria is introduced in Section 4 to facilitate model-free variable selection. We report numerical studies in Section 5 and conclude the paper with some discussions in Section 6.

2. Review of Existing Marginal Coordinate Tests

The marginal coordinate hypothesis $H_0 : i \in \mathcal{A}^c$ versus $H_a : i \in \mathcal{A}$ is closely related to the concept of sufficient dimension reduction. Sufficient dimension reduction (Li (1991); Cook (1998)) aims to find linear combinations of $\mathbf{X}$, such that Y is independent of $\mathbf{X}$ given these linear combinations. Specifically, the goal is to identify $\boldsymbol{\beta} \in \mathbb{R}^{p\times d}$ with the minimum column space satisfying $Y \perp\!\!\!\perp \mathbf{X}|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}$ for some $d < p$. This minimum column space is then referred to as the central space of Y versus $\mathbf{X}$ and is denoted by $\mathcal{S}_{Y|\mathbf{X}}$. Under mild regularity conditions (Yin, Li, and Cook (2008)), the central space exists and is unique. We say that $\boldsymbol{\beta}$ is the basis of the $\mathbf{X}$-scale central space with $\mathrm{Span}(\boldsymbol{\beta}) = \mathcal{S}_{Y|\mathbf{X}}$. Here $\mathrm{Span}(\boldsymbol{\beta})$ denotes the column space of $\boldsymbol{\beta}$.

The central space has an important invariance property. Let $E(\mathbf{X}) = \boldsymbol{\mu}$, $\mathrm{Var}\,(\mathbf{X}) = \boldsymbol{\Sigma}$, and $\mathbf{Z} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{X} - \boldsymbol{\mu})$. Let $\mathcal{S}_{Y|\mathbf{Z}}$ denote the central space of Y versus the standardized predictor $\mathbf{Z}$. Suppose $\mathcal{S}_{Y|\mathbf{Z}}$ has basis $\boldsymbol{\eta}$ such that $\mathrm{Span}(\boldsymbol{\eta}) = \mathcal{S}_{Y|\mathbf{Z}}$. The invariance property states that $\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\eta}$. This invariance property can facilitate our discussions about the marginal coordinate tests. As we will see, one can work with the $\mathbf{Z}$-scale central space estimation and achieve the $\mathbf{X}$-scale coordinate test due to the invariance property.

For $i = 1, \ldots, p$, let $\mathbf{e}_i \in \mathbb{R}^p$, where the ith element of $\mathbf{e}_i$ is 1 and all other elements are zero. For $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ denoting the bases for the $\mathbf{X}$-scale central space $\mathcal{S}_{Y|\mathbf{X}}$ and the $\mathbf{Z}$-scale central space $\mathcal{S}_{Y|\mathbf{Z}}$, respectively, we have $i \in \mathcal{A}^c$ if and only if $\mathbf{e}_i^{\mathrm{T}}\boldsymbol{\beta} = \mathbf{0}$. See Lemma A.2 in the Appendix. It follows that testing $H_0 : i \in \mathcal{A}^c$ versus $H_a : i \in \mathcal{A}$ in (1.1) is equivalent to testing $H_0 : \mathbf{e}_i^{\mathrm{T}}\boldsymbol{\beta} = \mathbf{0}$ versus $H_a : \mathbf{e}_i^{\mathrm{T}}\boldsymbol{\beta} \neq \mathbf{0}$. Due to invariance, we consider

$$H_0 : \mathbf{e}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\eta} = \mathbf{0} \text{ versus } H_a : \mathbf{e}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\eta} \neq \mathbf{0}. \tag{2.1}$$

Here $\boldsymbol{\eta} \in \mathbb{R}^{p\times d}$ can be replaced with any matrix that has the same column space as $\boldsymbol{\eta}$. The key to the marginal coordinate test thus becomes estimation of the $\mathbf{Z}$-scale central space $\mathcal{S}_{Y|\mathbf{Z}} = \mathrm{Span}(\boldsymbol{\eta})$.

Let $\{J_1, \ldots, J_H\}$ be a measurable partition of the sample space of Y. Let $R_h = I(Y \in J_h)$ be the indicator function of Y belonging to the hth slice. For $h = 1, \ldots, H$, write $p_h = E(R_h)$, $\mathbf{U}_h = \boldsymbol{\Sigma}^{1/2}E(\mathbf{Z}R_h)$, and $\mathbf{V}_h = \boldsymbol{\Sigma}^{1/2}E(\mathbf{Z}\mathbf{Z}^{\mathrm{T}}R_h)\boldsymbol{\Sigma}^{1/2}$.

Define the SIR and SAVE kernel matrices as $\mathbf{M}^{\mathrm{SIR}} = \sum_{h=1}^{H} p_h^{-1}\boldsymbol{\Sigma}^{-1/2}\mathbf{U}_h\mathbf{U}_h^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}$ and $\mathbf{M}^{\mathrm{SAVE}} = \sum_{h=1}^{H} p_h\mathbf{B}_h^2$, where $\mathbf{B}_h = \boldsymbol{\Sigma}^{-1/2}(p_h^{-1}\mathbf{V}_h - \boldsymbol{\Sigma} - p_h^{-2}\mathbf{U}_h\mathbf{U}_h^{\mathrm{T}})\boldsymbol{\Sigma}^{-1/2}$. We list some common assumptions in the sufficient dimension reduction literature as follows.

(C1) Linear conditional mean condition: $E(\mathbf{Z}|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{Z})$ is linear in $\boldsymbol{\eta}^{\mathrm{T}}\mathbf{Z}$.

(C2) Constant conditional variance condition: $\mathrm{Var}\,(\mathbf{Z}|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{Z})$ is a constant.

(C3) Coverage condition for SIR: $\mathrm{Span}\{\boldsymbol{\Sigma}^{-1/2}\mathbf{U}_h : h = 1, \ldots, H\} = \mathcal{S}_{Y|\mathbf{Z}}$.

(C4) Coverage condition for SAVE: for any nonzero $\boldsymbol{\eta} \in \mathcal{S}_{Y|\mathbf{Z}}$, at least one of $\mathrm{Var}\,\{E(\boldsymbol{\eta}^{\mathrm{T}}\mathbf{Z}|Y)\} > 0$ and $\mathrm{Var}\,\{\mathrm{Var}\,(\boldsymbol{\eta}^{\mathrm{T}}\mathbf{Z}|Y)\} > 0$ holds true.

Condition (C1) holds true if $\mathbf{Z}$ has an elliptically-contoured distribution. Both (C1) and (C2) are guaranteed when $\mathbf{Z}$ is normally distributed. More discussions about (C1) and (C2) can be found in Li and Dong (2009), Dong and Li (2010). Refer to Cook (2004) about (C3). The coverage condition (C4) for SAVE has been discussed in Shao, Cook, and Weisberg (2007), and more insights are provided in Li and Wang (2007). Coverage condition (C4) is generally considered to be very mild. Meanwhile, coverage condition (C3) can be violated in the case when the link function between Y and $\mathbf{Z}$ is symmetric.

Under (C1), Li (1991) showed that $\mathrm{Span}(\mathbf{M}^{\mathrm{SIR}}) \subseteq \mathcal{S}_{Y|\mathbf{Z}}$. With the additional (C3), we have $\mathrm{Span}(\mathbf{M}^{\mathrm{SIR}}) = \mathcal{S}_{Y|\mathbf{Z}}$. From the discussions following (2.1), we know $\mathbf{e}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\mathbf{M}^{\mathrm{SIR}}\boldsymbol{\Sigma}^{-1/2}\mathbf{e}_i = 0$ for $i \in \mathcal{A}^c$. Given an i.i.d. sample $(\mathbf{X}_1, Y_1)$, …, $(\mathbf{X}_n, Y_n)$, we can calculate the sample estimators $\hat{p}_h$, $\hat{\mathbf{U}}_h$, $\hat{\mathbf{V}}_h$, $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\Sigma}}$ and $\hat{\mathbf{M}}^{\mathrm{SIR}}$. For example, $\hat{\mathbf{U}}_h = n^{-1}\sum_{i=1}^{n}(\mathbf{X}_i - \hat{\boldsymbol{\mu}})I(Y_i \in J_h)$ and $\hat{\mathbf{M}}^{\mathrm{SIR}} = \sum_{h=1}^{H}\hat{p}_h^{-1}\hat{\boldsymbol{\Sigma}}^{-1/2}\hat{\mathbf{U}}_h\hat{\mathbf{U}}_h^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}^{-1/2}$. The SIR based test statistic for (1.1) is then n times the sample estimator of $\mathbf{e}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\mathbf{M}^{\mathrm{SIR}}\boldsymbol{\Sigma}^{-1/2}\mathbf{e}_i$,

$$T_i^{\mathrm{SIR}} = n\sum_{h=1}^{H}\mathbf{e}_i^{\mathrm{T}}\hat{p}_h^{-1}\hat{\boldsymbol{\Sigma}}^{-1}\hat{\mathbf{U}}_h\hat{\mathbf{U}}_h^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{e}_i. \tag{2.2}$$

Cook (2004) showed that T_i^{SIR} has an asymptotic distribution that is the sum of weighted $\chi^2(1)$ under $H_0 : i \in \mathcal{A}^c$.

Cook and Weisberg (1991) proved that $\mathrm{Span}(\mathbf{M}^{\mathrm{SAVE}}) \subseteq \mathcal{S}_{Y|\mathbf{Z}}$ under (C1) and (C2). With the additional coverage condition (C4), we have $\mathrm{Span}(\mathbf{M}^{\mathrm{SAVE}}) = \mathrm{Span}(\sum_{h=1}^{H} p_h\mathbf{B}_h^2) = \mathcal{S}_{Y|\mathbf{Z}}$. It follows that for $i \in \mathcal{A}^c$, we have $\sum_{h=1}^{H} p_h(\mathbf{e}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\mathbf{B}_h\boldsymbol{\Sigma}^{-1/2}\mathbf{e}_i)^2 = 0$. At the sample level, Shao, Cook, and Weisberg (2007) developed a SAVE-based test statistic for (1.1):

$$T_i^{\mathrm{SAVE}} = n\sum_{h=1}^{H}\hat{p}_h\{\mathbf{e}_i^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}^{-1}(\hat{p}_h^{-1}\hat{\mathbf{V}}_h - \hat{\boldsymbol{\Sigma}} - \hat{p}_h^{-2}\hat{\mathbf{U}}_h\hat{\mathbf{U}}_h^{\mathrm{T}})\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{e}_i\}^2. \tag{2.3}$$

The asymptotic distribution of T_i^{SAVE} under $H_0 : i \in \mathcal{A}^c$ is also a sum of weighted $\chi^2(1)$ distributions.

3. A New Test Based on Directional Regression

Directional regression (Li and Wang (2007)) is a popular sufficient dimension reduction method that implicitly synthesizes SIR and SAVE. Let $\mathbf{A}_{hk} = E\{(\mathbf{Z} - \tilde{\mathbf{Z}})(\mathbf{Z} - \tilde{\mathbf{Z}})^{\mathrm{T}} | Y \in J_h, \tilde{Y} \in J_k\}$, where $(\tilde{\mathbf{Z}}, \tilde{Y})$ is an independent copy of $(\mathbf{Z}, Y)$. The kernel matrix for directional regression is defined as $\mathbf{M}^{\mathrm{DR}} = \sum_{h=1}^{H}\sum_{k=1}^{H} p_h p_k (2\mathbf{I}_p - \mathbf{A}_{hk})^2$. Then we have $\mathrm{Span}(\mathbf{M}^{\mathrm{DR}}) \subseteq \mathcal{S}_{Y|\mathbf{Z}}$ (Li and Wang (2007)). The kernel matrices $\mathbf{M}^{\mathrm{DR}}$ and $\mathbf{M}^{\mathrm{SAVE}}$ have similar forms. This similarity motivates us to mimic the development of the SAVE-based test, and consider a population level quantity for directional regression,

$$\tau_i = \sum_{h=1}^{H}\sum_{k=1}^{H} p_h p_k \{\mathbf{e}_i^{\mathrm{T}} \boldsymbol{\Sigma}^{-1/2} (2\mathbf{I}_p - \mathbf{A}_{hk}) \boldsymbol{\Sigma}^{-1/2} \mathbf{e}_i\}^2. \tag{3.1}$$

Obviously $\tau_i \geq 0$ from the definition. Furthermore, we have

Proposition 1. *If* (C1), (C2) *and* (C4) *hold, then* $\tau_i = 0$ *if and only if* $i \in \mathcal{A}^c$.

The next result greatly simplifies the sample level calculation.

Proposition 2. τ_i *in* (3.1) *can be reexpressed as* $\tau_i = 2\tau_{i,1} + 4\tau_{i,2}^2$, *where* $\tau_{i,1} = \sum_{h=1}^{H} p_h \{\mathbf{e}_i^{T} \boldsymbol{\Sigma}^{-1} (p_h^{-1}\mathbf{V}_h - \boldsymbol{\Sigma}) \boldsymbol{\Sigma}^{-1} \mathbf{e}_i\}^2$ *and* $\tau_{i,2} = \sum_{h=1}^{H} p_h^{-1} \mathbf{e}_i^{T} \boldsymbol{\Sigma}^{-1} \mathbf{U}_h \mathbf{U}_h^{T} \boldsymbol{\Sigma}^{-1} \mathbf{e}_i$.

The sample estimators of $\tau_{i,1}$ and $\tau_{i,2}$ are $\hat{\tau}_{i,1} = \sum_{h=1}^{H} \hat{p}_h \{\mathbf{e}_i^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}^{-1} (\hat{p}_h^{-1} \hat{\mathbf{V}}_h - \hat{\boldsymbol{\Sigma}}) \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{e}_i\}^2$ and $\hat{\tau}_{i,2} = \sum_{h=1}^{H} \hat{p}_h^{-1} \mathbf{e}_i^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}^{-1} \hat{\mathbf{U}}_h \hat{\mathbf{U}}_h^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{e}_i$ respectively. Upon closer examination, we note that $n\hat{\tau}_{i,2}$ is exactly T_i^{SIR} defined in (2.2), and $n\hat{\tau}_{i,1}$ bears close resemblance to T_i^{SAVE} defined in (2.3).

For $h = 1, \ldots, 2H$, let $\chi_h^2(1)$ be i.i.d. random variables with $\chi^2(1)$ distribution. Let

$$\ell_{ih} = p_h^{1/2} \mathbf{e}_i^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (p_h^{-1} \mathbf{V}_h^* - p_h^{-2} p_h^* \mathbf{V}_h - \boldsymbol{\Sigma}^*) \boldsymbol{\Sigma}^{-1} \mathbf{e}_i$$

and

$$g_{ih} = p_h^{-1/2} \mathbf{e}_i^{\mathrm{T}} (\boldsymbol{\Sigma}^{-1})^* \mathbf{U}_h + p_h^{-1/2} \mathbf{e}_i^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \mathbf{U}_h^*,$$

where p_h^*, $\mathbf{U}_h^*$, $\mathbf{V}_h^*$, $\boldsymbol{\Sigma}^*$ and $(\boldsymbol{\Sigma}^{-1})^*$ are defined in the proof of Lemma 1. Denote the eigenvalues of $\mathrm{Cov}(\mathbf{L}_i)$ and $\mathrm{Cov}(\mathbf{G}_i)$ by $\iota_{i1}, \ldots, \iota_{iH}$ and $\gamma_{i1}, \ldots, \gamma_{iH}$, respectively, where $\mathbf{L}_i = (\ell_{i1}, \ldots, \ell_{iH})^{\mathrm{T}}$ and $\mathbf{G}_i = (g_{i1}, \ldots, g_{iH})^{\mathrm{T}}$. Let $\xrightarrow{D}$ denote convergence in distribution and $\xrightarrow{P}$ denote convergence in probability.

Lemma 1. *Suppose the entries of* $E(\mathbf{Z}|Y)$ *and* $E(\mathbf{Z}\mathbf{Z}^T|Y)$ *have finite second moments. If* (C1) *and* (C2) *hold, then* $n\hat{\tau}_{i,1} \xrightarrow{D} \sum_{h=1}^{H} \iota_{ih}\chi_h^2(1)$ *and* $n\hat{\tau}_{i,2} \xrightarrow{D} \sum_{h=1}^{H} \gamma_{ih}\chi_h^2(1)$ *under* $H_0 : i \in \mathcal{A}^c$. *If* (C3) *and* (C4) *hold in addition, then* $\hat{\tau}_{i,1} \xrightarrow{P} \tau_{i,1}$ *and* $\hat{\tau}_{i,2} \xrightarrow{P} \tau_{i,2}$ *with* $\tau_{i,1} > 0$ *and* $\tau_{i,2} > 0$ *under* $H_a : i \in \mathcal{A}$.

Since $\tau_i = 2\tau_{i,1} + 4\tau_{i,2}^2$, simply multiplying its sample version by n leads to $2n\hat{\tau}_{i,1} + 4n\hat{\tau}_{i,2}^2$. Under $H_0 : i \in \mathcal{A}^c$, we know $n\hat{\tau}_{i,1} = O_P(1)$ and $n\hat{\tau}_{i,2}^2 = O_P(n^{-1})$ according to the first part of Lemma 1. Thus $2n\hat{\tau}_{i,1} + 4n\hat{\tau}_{i,2}^2$ has the same asymptotic distribution as $2n\hat{\tau}_{i,1}$. To keep the balance between the two terms, we replace $4n\hat{\tau}_{i,2}^2$ with $4n\hat{\tau}_{i,2}$, and define the directional regression-based marginal coordinate test statistic $T_i^{\mathrm{DR}} = 2n\hat{\tau}_{i,1} + 4n\hat{\tau}_{i,2}$. Here T_i^{DR} can be viewed as a hybrid between the SIR-based test in Cook (2004) and the SAVE-based test in Shao, Cook, and Weisberg (2007). Since (C3) for SIR can be easily violated when the link function between Y and $\mathbf{X}$ is symmetric, and SAVE may not be as effective when the link function between Y and $\mathbf{X}$ is monotone, we expect the directional regression-based test to be a safe alternative to the existing tests in Cook (2004) and Shao, Cook, and Weisberg (2007), and that it will work well across a wide range of link functions.

Theorem 1. *Suppose the entries of* $E(\mathbf{Z}|Y)$ *and* $E(\mathbf{Z}\mathbf{Z}^T|Y)$ *have finite second moments, and that* (C1) *and* (C2) *hold. For* $i \in \mathcal{A}^c$, *we have* $T_i^{\mathrm{DR}} \xrightarrow{D} 2\sum_{h=1}^{2H} \omega_{ih}\chi_h^2(1)$, *where* $\{\omega_{ih}, h = 1, \ldots, 2H\}$ *are the eigenvalues of* $\mathrm{Cov}(\mathbf{W}_i)$, *and* $\mathbf{W}_i = (\ell_{i1}, \ldots, \ell_{iH}, \sqrt{2}g_{i1}, \ldots, \sqrt{2}g_{iH})^T$.

Thus under $H_0 : i \in \mathcal{A}^c$, T_i^{DR} has the same distribution as the sum of weighted $\chi^2(1)$ distributions. The unknown weights ω_{ih} can be replaced with their consistent sample estimators in practice.

4. The Maximum Ratio Criteria

We propose a novel maximum ratio criteria (MRC) to facilitate estimation of the active set $\mathcal{A}$. MRC can be combined with marginal coordinate tests based on SIR, SAVE or directional regression. Let T_i be a method-specific test statistic for $H_0 : i \in \mathcal{A}^c$ versus $H_a : i \in \mathcal{A}$. In all such tests, we reject H_0 and conclude the ith predictor is active when T_i is larger than a certain threshold, or when the corresponding p-value is smaller than a prespecified nominal level α. Let $T_{(1)} > T_{(2)} > \cdots > T_{(p)}$ be the ordered test statistics for all p predictors. Consider the ratio of consecutively ranked statistics $r_i = T_{(i)}/T_{(i+1)}$. Lemma 1 reveals that $T_i = O_P(n)$ for $i \in \mathcal{A}$ and $T_i = O_P(1)$ for $i \in \mathcal{A}^c$. Due to the different order of magnitude for T_i depending on $i \in \mathcal{A}^c$ or $i \in \mathcal{A}$, we expect that for large n, the top-ranked test statistics belong to the active predictors, while the bottom-ranked statistics correspond to the inactive predictors. Furthermore, we expect

the ratio r_i to be maximized when $T_{(i)}$ corresponds to the active predictor with the smallest test statistic, and $T_{(i+1)}$ corresponds to the inactive predictor with the largest test statistic.

To formalize this idea, we assume the cardinality of the active set is $|\mathcal{A}| = q$ with $q < p$. Let u_i be the subscript of the statistic such that it matches the ith order statistic, or $T_{u_i} = T_{(i)}$. The estimator of q is thus

$$\hat{q} = \arg \max_{i=1,\ldots,p-1} \frac{T_{(i)}}{T_{(i+1)}},$$

and the corresponding estimator of the active set is $\hat{\mathcal{A}} = \{u_1, u_2, \ldots, u_{\hat{q}}\}$. We remark that MRC is not applicable with $q = p$, when all the predictors are active.

Theorem 2. *Suppose the entries of* $E(\mathbf{Z}|Y)$ *and* $E(\mathbf{Z}\mathbf{Z}^{\tau}|Y)$ *have finite second moments, and* (C1) *through* (C4) *hold. Then* $P(\hat{\mathcal{A}} = \mathcal{A}) \to 1$ *as* $n \to \infty$.

Theorem 2 applies to MRC with T_i^{SIR}, T_i^{SAVE}, or T_i^{DR}. For the SIR based procedure to be variable-selection consistent, we only need (C1) and (C3). For the SAVE-based or the directional regression-based procedure to be variable-selection consistent, we need (C1), (C2) and (C4). While a formal marginal coordinate test based on T_i involves estimation of unknown weights in the weighted χ^2 distribution, calculating the ratios of the test statistics suffices for the MRC. An additional benefit is that we avoid selecting the significance level α, as is required by the existing sequential test procedure.

5. Numerical Analysis

The finite sample performances of the proposed marginal coordinate test based on T_i^{DR} were compared with existing marginal coordinate tests. The MRC procedure for model-free variable selection was examined through both synthetic data sets and the Iris data (Fisher (1936)).

5.1. Simulation studies

First we used synthetic data to demonstrate the effectiveness of the marginal coordinate test and variable selection via directional regression. Consider the models

$$\begin{aligned}
\mathrm{I}: &\quad Y = 3\sin(X_1) + 3\sin(X_p) + .1\epsilon,\\
\mathrm{II}: &\quad Y = \mathrm{sgn}(X_1 + X_p)\exp(X_2 + X_{p-1}) + .1\epsilon,\\
\mathrm{III}: &\quad Y = (X_1 + X_p)^2 + X_2 + X_{p-1} + (X_3 + 1)^2\epsilon.
\end{aligned}$$

Here $\mathbf{X} = (X_1, \ldots, X_p)^{\mathrm{T}}$ is multivariate normal with mean $\boldsymbol{\mu} = \mathbf{0}$. The predictors are correlated, and the covariance between X_i and X_j is $.5^{|i-j|}$ for $1 \le i, j \le p$.

Table 1. Marginal coordinate tests. Based on 1,000 repetitions, frequencies of rejecting $H_0 : i \in \mathcal{A}^c$ with nominal 5% tests are reported.

Model	n	Method	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
I	100	SIR	**1**	0.086	0.084	0.081	0.090	0.086	0.079	0.082	0.087	**1**
		SAVE	**0.306**	0.026	0.017	0.019	0.027	0.025	0.031	0.022	0.025	**0.310**
		DR	**0.959**	0.024	0.027	0.026	0.033	0.035	0.038	0.026	0.023	**0.949**
	400	SIR	**1**	0.060	0.052	0.064	0.058	0.060	0.067	0.062	0.058	**1**
		SAVE	**0.997**	0.040	0.044	0.036	0.040	0.043	0.035	0.037	0.041	**0.999**
		DR	**1**	0.033	0.042	0.035	0.038	0.046	0.039	0.037	0.049	**1**
II	100	SIR	**1**	**1**	0.074	0.089	0.078	0.078	0.086	0.094	**1**	**0.999**
		SAVE	**0.088**	**0.084**	0.034	0.031	0.024	0.026	0.024	0.020	**0.088**	**0.103**
		DR	**0.893**	**0.906**	0.034	0.034	0.027	0.025	0.023	0.026	**0.913**	**0.870**
	400	SIR	**1**	**1**	0.047	0.049	0.053	0.063	0.056	0.059	**1**	**1**
		SAVE	**0.927**	**0.982**	0.051	0.033	0.034	0.036	0.052	0.037	**0.986**	**0.917**
		DR	**1**	**1**	0.044	0.031	0.036	0.038	0.057	0.043	**1**	**1**
III	100	SIR	**0.162**	**0.610**	**0.469**	0.084	0.087	0.084	0.086	0.087	**0.625**	**0.160**
		SAVE	**0.229**	**0.022**	**0.035**	0.024	0.05	0.020	0.016	0.026	**0.014**	**0.216**
		DR	**0.253**	**0.228**	**0.185**	0.029	0.032	0.031	0.026	0.028	**0.240**	**0.248**
	400	SIR	**0.408**	**0.997**	**0.976**	0.062	0.058	0.063	0.062	0.063	**0.999**	**0.446**
		SAVE	**0.970**	**0.105**	**0.178**	0.035	0.032	0.042	0.042	0.037	**0.102**	**0.972**
		DR	**0.982**	**0.975**	**0.881**	0.043	0.034	0.050	0.043	0.046	**0.967**	**0.978**

The variance of X_i is 1 for $i = 1, \ldots, p$. The error term ϵ is standard normal and is independent of $\mathbf{X}$. The active sets $\mathcal{A}$ for these models are $\{1, p\}$, $\{1, 2, p-1, p\}$, and $\{1, 2, 3, p-1, p\}$, respectively. Each simulation run was based on 1,000 repetitions, and we fixed the number of slices to be $H = 5$ in each repetition. The choice of $H = 10$ led to similar results.

We compared the performance of the directional regression-based marginal coordinate test with existing tests. For the test statistic based on directional regression for testing $H_0 : i \in \mathcal{A}^c$ v.s. $H_a : i \in \mathcal{A}$, we have from Theorem 1 that $T_i^{\mathrm{DR}} \xrightarrow{D} 2\sum_{h=1}^{2H} \omega_{ih}\chi_h^2(1)$. Let $\mathbf{K}_i = (\hat{\omega}_{i1}, \hat{\omega}_{i2}, \ldots, \hat{\omega}_{is})^T$, where $s = 2H$ and $\hat{\omega}_{ih}$ is the sample estimator of ω_{ih} for $h = 1, \ldots, s$. Let $\mathbf{C}$ be an $m \times s$-dimensional matrix of i.i.d. $\chi^2(1)$ realizations. Then $2\mathbf{C}\mathbf{K}_i$ is an m-dimensional vector of i.i.d. realizations of $2\sum_{h=1}^{2H} \hat{\omega}_{ih}\chi_h^2(1)$. The proportion of these m realizations larger than T_i^{DR} is then the approximated p-value for testing H_0. We reject H_0 if the approximated p-value is smaller than the nominal level α. We set $m = 1000$ and $\alpha = .05$ in all settings.

In Table 1, we report the frequencies that $H_0 : i \in \mathcal{A}^c$ is rejected for each predictor X_i. We fixed the predictor dimension at $p = 10$, and considered sample sizes $n = 100$, 400. For all three models, the frequencies for predictors belonging to the inactive set $\mathcal{A}^c$ correspond to the estimated levels, and we want them to be close to the nominal level 0.05. On the other hand, the frequencies for predictors

in the active set $\mathcal{A}$ are the estimated powers, and we want them to be close to 1. For easy reference, the boldfaced entries in Table 1 correspond to the estimated powers. There are some common trends across all three models. The estimated nominal levels are never too far away from the true nominal level, as they range from 0.017 to 0.090. As sample size increases from 100 to 400, the estimated levels for all three methods become closer to 0.05. The SIR-based test seems to be more liberal compared with the SAVE-based and the directional regression-based tests. The main difference arises from the estimated powers of the tests. Not surprisingly, the powers of all three tests improve as sample size increases. The SIR-based test has large powers for Models I and II, but has small powers for testing X_1 and X_{10} in Model III even with $n = 400$. This is because X_1 and X_{10} appear in a quadratic link function in Model III, and the coverage condition (C3) for SIR is violated in this case. The SAVE-based test has poor powers when $n = 100$. In most instances, the estimated powers for SAVE increase dramatically when $n = 400$. However, SAVE still has poor powers in Model III for the two linear terms X_2 and X_9, as well as for X_3 which appears in the error term. Our proposed directional regression-based test has decent powers for Models I and II with relatively small sample size $n = 100$, and achieves large powers for all three models when $n = 400$.

The performances of maximum ratio criteria for variable selection are summarized in Table 2. The frequencies of predictors X_1, X_2, X_3, X_{p-1} and X_p being selected are reported based on 1,000 repetitions. Let $\hat{\mathcal{A}}_{(i)}$ be the estimated active set in the ith repetition. We also report the average model size $MS = \sum_{i=1}^{1,000} |\hat{\mathcal{A}}_{(i)}|/1,000$, the underfitted count $UF = \sum_{i=1}^{1,000} I(\mathcal{A} \not\subseteq \hat{\mathcal{A}}_{(i)})$, the correctly-fitted count $CF = \sum_{i=1}^{1,000} I(\mathcal{A} = \hat{\mathcal{A}}_{(i)})$, and the overfitted count $OF = \sum_{i=1}^{1,000} I(\mathcal{A} \subset \hat{\mathcal{A}}_{(i)})$. We consider two combinations of sample size and predictor dimensionality here: $n = 400$ with $p = 10$, or n =1,600 with $p = 40$.

Since the three models have different active predictor sets, we use boldfaced entries to denote the truly active predictors. For a variable selection procedure to work well, not only do we need these boldfaced entries to be close to 1, but the average model size also has to be close to the true number of predictors $|\mathcal{A}|$, which are 2, 4, and 5 for the three models, respectively. In the case when the average model size is different from $|\mathcal{A}|$, we prefer the procedure to overfit rather than to underfit, as missing truly active predictors oftentimes has more severe effect on the model building. We see from Table 2 that MRC works well with all three methods in Models I and II. In particular, MRC based on directional regression achieves perfect selection for Models I and II when n =1,600 and $p = 40$. This justifies our theoretical finding in Theorem 2. In the challenging case of Model III, MRC based on SIR tends to miss X_1 and X_p, while MRC based on SAVE tends to miss X_2, X_3 and X_{p-1}. MRC based on directional regression does a very good

Table 2. Maximum ratio criteria (MRC) for variable selection. Based on 1,000 repetitions, frequencies of X_1, X_2, X_3, X_{p-1} and X_p being selected, the average model size (MS), the underfitted count (UF), the correctly-fitted count (CF), and the overfitted count (OF) are reported.

	Model	Method	X_1	X_2	X_3	X_{p-1}	X_p	MS	UF	CF	OF
$n = 400$ & $p = 10$	I	SIR	**1**	0.015	0.013	0.017	**1**	2.126	0	982	18
		SAVE	**0.993**	0.197	0.216	0.217	**0.996**	3.687	11	641	348
		DR	**1**	0.019	0.014	0.020	**1**	2.151	0	969	31
	II	SIR	**1**	**1**	0.006	**1**	**1**	4.062	0	987	13
		SAVE	**0.980**	**0.989**	0.435	**0.994**	**0.981**	6.491	31	243	726
		DR	**0.999**	**1**	0.038	**1**	**1**	4.268	1	914	85
	III	SIR	**0.527**	**0.969**	**0.899**	**0.970**	**0.525**	5.312	543	46	411
		SAVE	**0.893**	**0.360**	**0.380**	**0.362**	**0.909**	4.173	688	4	308
		DR	**0.902**	**0.961**	**0.920**	**0.970**	**0.908**	6.260	172	260	568
$n = 1,600$ & $p = 40$	I	SIR	**1**	0.002	0.002	0.002	**1**	2.074	0	998	2
		SAVE	**1**	0.023	0.023	0.024	**1**	2.888	0	976	24
		DR	**1**	0	0	0	**1**	2	0	1000	0
	II	SIR	**1**	**1**	0.001	**1**	**1**	4.035	0	999	1
		SAVE	**1**	**1**	0.076	**1**	**1**	6.761	0	916	84
		DR	**1**	**1**	0	**1**	**1**	4	0	1000	0
	III	SIR	**0.272**	**0.999**	**0.985**	**1**	**0.280**	8.264	756	104	14
		SAVE	**0.998**	**0.068**	**0.078**	**0.067**	**0.998**	4.477	933	0	67
		DR	**0.999**	**1**	**0.998**	**1**	**0.999**	5.373	4	983	13

job with Model III: selecting all five active predictors with high probability, it does not suffer from serious underfitting, and has modest overfitting. In the case of $n = 1,600$ and $p = 40$, performing a single test $H_0 : i \in \mathcal{A}^c$ versus $H_a : i \in \mathcal{A}$ is very cumbersome, not to mention a full sequential test through backward elimination. Performing variable selection through MRC with 1,000 repetitions, on the other hand, can be done on a personal computer within several minutes.

5.2. Analysis of the Iris data

For the data analysis, we used the Iris data (Fisher (1936)) that can be downloaded from the UCI machine learning repository. This data contains three classes of irises (Setosa, Versicolour, and Virginica), with 50 observations for each class. The four predictors are sepal length, sepal width, petal length, and petal width. We used this analysis to demonstrate the effectiveness of our proposals with discrete response. Denote the marginally standardized predictors as X_1, X_2, X_3, and X_4 respectively. Without fitting a classification model, we want to decide which predictors are sufficient to predict the response Y, the class label of the Iris. One possibility is to implement the backward elimination procedure (Li, Cook, and Nachtsheim (2005)) with our directional regression test. This

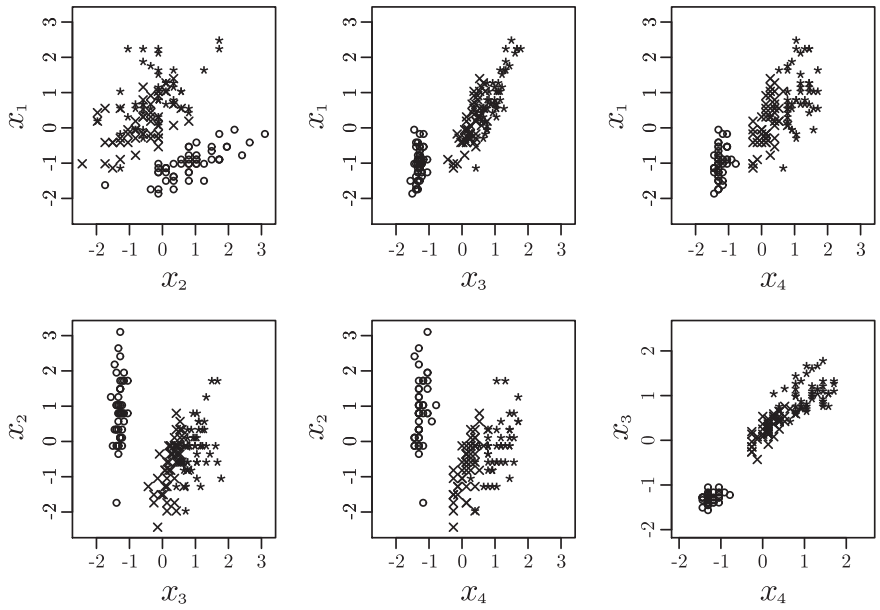

Figure 1. The scatterplots of all possible pairs of predictors. ∘ denotes Iris Setosa; × denotes Iris Versicolour, and ⋆ denotes Iris Virginica.

procedure can be viewed as an adaptation of the classical backward elimination procedure in linear models. In the first iteration, we performed the marginal coordinate test for $H_0 : i \in \mathcal{A}^c$ versus $H_a : i \in \mathcal{A}$ based on T_i^{DR}. The p-values were 0.0538, 0.0601, 0.0221, and 0 for $i = 1, 2, 3, 4$, respectively. We see that X_2 is the least significant predictor in terms of predicting Y. At nominal level $\alpha = .05$, we conclude X_2 is inactive and eliminate X_2 from further analysis. In the second iteration, we recalculated the p-values for X_1, X_3, and X_4, obtaining 0.1280, 0.0118, and 0. Dropping X_1 because $0.1280 > \alpha$, the next iteration found that the p-values for X_3 and X_4 were both 0, and no variables could be further dropped from the model. We conclude X_3 and X_4 are sufficient to predict Y.

The selection procedure through MRC is much more straightforward. We do not need any iterations or the calculation of any p-values. The test statistics for $H_0 : i \in \mathcal{A}^c$ versus $H_a : i \in \mathcal{A}$ were calculated as $T_1^{\mathrm{DR}} = 1,841$, $T_2^{\mathrm{DR}} = 197$, $T_3^{\mathrm{DR}} = 77,232$, and $T_4^{\mathrm{DR}} = 74,428$. The ratios of the consecutively ranked test statistics are $r_1 = 1.038$, $r_2 = 40.42$, and $r_3 = 9.339$. Thus $\hat{q} = \arg\max_i r_i = 2$, and the predictors corresponding to the top-two ranked test statistics are taken to be active. Again we conclude X_3 and X_4 are sufficient to predict Y. The scatterplots between pairs of predictors are listed in Figure 1. Among all the possible pairs, we see that X_3 and X_4 provide the clearest separation for the three Iris classes.

6. Discussion

We provide a simple maximum ratio criteria to facilitate model-free variable selection. A related method in the sufficient dimension reduction literature is maximal eigenvalue ratio criterion, proposed in Luo, Wang, and Tsai (2009), to determine the dimensionality of the central space $\mathcal{S}_{Y|\mathbf{X}}$. Although the maximum ratio criteria can be readily combined with existing marginal coordinate tests based on SIR or SAVE, it yields the best variable selection performance when we couple it with our directional regression-based test. The proposed test statistic for directional regression can be viewed as a hybrid between the SIR-based test (Cook (2004)) and the SAVE-based test (Shao, Cook, and Weisberg (2007)). More general combination methods can be studied along the lines of Ye and Weiss (2003). Due to the involvement of such terms as $\mathbf{\Sigma}^{-1}$, our proposal is not directly applicable when $p > n$. Cook, Li, and Chiaromonte (2007) studied sufficient dimension reduction without matrix inversion. Developments towards testing predictor contribution without matrix inversion warrant further investigation.

Acknowledgement

The research of the first author is supported by National Natural Science Foundation of China (No.11201151 and No.11571111), the 111 project (B14019), the program of Shanghai Subject Chief Scientist (14XD1401600), and the Shanghai rising star program (16QA1401700). We sincerely thank an associate editor and two anonymous referees for giving useful comments that led to a much-improved presentation of the paper.

Appendix: proofs

Recall that $\text{Span}(\boldsymbol{\beta}) = \mathcal{S}_{Y|\mathbf{X}}$ and $\mathcal{I} = \{1, \ldots, p\}$ denotes the full index set. Suppose $\mathcal{D} \subseteq \mathcal{I}$ is an index set and $\mathcal{D}^c$ is the complement of $\mathcal{D}$ in $\mathcal{I}$. Let $\mathbf{X}_\mathcal{D} = \{X_i : i \in \mathcal{D}\}$ denote the vector that contains all the predictors from index set $\mathcal{D}$. Let $\mathbf{I}_{\mathcal{D}^c} = \text{diag}\{d_1, \ldots, d_p\}$ be the $p \times p$ dimensional diagonal matrix with $d_i = 1$ for $i \in \mathcal{D}^c$ and $d_i = 0$ for $i \in \mathcal{D}$.

Lemma A.1. $\mathbf{I}_{\mathcal{D}^c}\boldsymbol{\beta} = \mathbf{0}$ if and only if $Y \perp\!\!\!\perp \mathbf{X}|\mathbf{X}_\mathcal{D}$.

Proof of Lemma A.1. First the "only if" part. Let $\mathbf{I}_\mathcal{D} = \text{diag}\{d_1, \ldots, d_p\}$ be a $p \times p$ dimensional diagonal matrix, where $d_i = 1$ for $i \in \mathcal{D}$, and $d_i = 0$ for $i \in \mathcal{D}^c$. Then we have $\mathbf{I}_{\mathcal{D}^c} + \mathbf{I}_\mathcal{D} = \mathbf{I}_p$. Together with $\mathbf{I}_{\mathcal{D}^c}\boldsymbol{\beta} = \mathbf{0}$, we have $\boldsymbol{\beta}^\text{T}\mathbf{X} = \boldsymbol{\beta}^\text{T}\mathbf{I}_p\mathbf{X} = \boldsymbol{\beta}^\text{T}\mathbf{I}_\mathcal{D}\mathbf{X}$. From the definition of $\mathcal{S}_{Y|\mathbf{X}}$, we have $Y \perp\!\!\!\perp \mathbf{X}|\boldsymbol{\beta}^\text{T}\mathbf{X}$, which is $Y \perp\!\!\!\perp \mathbf{X}|\boldsymbol{\beta}^\text{T}\mathbf{I}_\mathcal{D}\mathbf{X}$. As $\mathbf{I}_\mathcal{D}\mathbf{X}$ involves only 0 and elements in $\mathbf{X}_\mathcal{D}$, we have $Y \perp\!\!\!\perp \mathbf{X}|\mathbf{X}_\mathcal{D}$. For the "if" part, $Y \perp\!\!\!\perp \mathbf{X}|\mathbf{X}_\mathcal{D}$ implies $Y \perp\!\!\!\perp \mathbf{X}|\mathbf{I}_\mathcal{D}\mathbf{X}$. Since $\boldsymbol{\beta}$ has the smallest columns space among all those satisfying $Y \perp\!\!\!\perp \mathbf{X}|\boldsymbol{\beta}^\text{T}\mathbf{X}$, we have $\mathcal{S}_{Y|\mathbf{X}} = \text{Span}(\boldsymbol{\beta}) \subseteq \text{Span}(\mathbf{I}_\mathcal{D})$. It follows immediately that $\mathbf{I}_{\mathcal{D}^c}\boldsymbol{\beta} = \mathbf{0}$.

Lemma A.2. $i \in \mathcal{A}^c$ if and only if $\mathbf{e}_i^{\mathrm{T}}\boldsymbol{\beta} = \mathbf{0}$.

Proof of Lemma A.2. First consider the "only if" part. By definition, $Y \perp\!\!\!\perp \mathbf{X}|\mathbf{X}_{\mathcal{A}}$. It follows from Lemma A.1 that $\mathbf{I}_{\mathcal{A}^c}\boldsymbol{\beta} = \mathbf{0}$. For $i \in \mathcal{A}^c$, the ith row of $\mathbf{I}_{\mathcal{A}^c}$ is $\mathbf{e}_i^{\mathrm{T}}$. Thus we have $\mathbf{e}_i^{\mathrm{T}}\boldsymbol{\beta} = \mathbf{0}$. Now consider the "if" part. Take $\mathbf{I}_{\{i\}} = \mathrm{diag}\{\mathbf{e}_i\}$. Then $\mathbf{e}_i^{\mathrm{T}}\boldsymbol{\beta} = \mathbf{0}$ guarantees that $\mathbf{I}_{\{i\}}\boldsymbol{\beta} = \mathbf{0}$. Let $\mathcal{E} = \{1, \ldots, i-1, i+1, \ldots, p\}$. It follows from Lemma A.1 that $Y \perp\!\!\!\perp \mathbf{X}|\mathbf{X}_{\mathcal{E}}$. By the definition of the active set $\mathcal{A}$, we know $\mathcal{A} \subseteq \mathcal{E}$ and $i \in \mathcal{A}^c$.

Proof of Proposition 1. Li and Wang (2007) proved that (C1), (C2), and (C4) guarantee $\mathrm{Span}(\mathbf{M}^{\mathrm{DR}}) = \mathcal{S}_{Y|\mathbf{Z}}$. By the invariance law of the central space, we have $\mathrm{Span}(\boldsymbol{\Sigma}^{-1/2}\mathbf{M}^{\mathrm{DR}}\boldsymbol{\Sigma}^{-1/2}) = \mathcal{S}_{Y|\mathbf{X}}$. If $i \in \mathcal{A}^c$, we know from Lemma A.2 that $\mathbf{e}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\mathbf{M}^{\mathrm{DR}}\boldsymbol{\Sigma}^{-1/2}\mathbf{e}_i = 0$. Because $p_h p_k > 0$, it follows that $\mathbf{e}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}(2\mathbf{I}_p - \mathbf{A}_{hk}) = \mathbf{0}$ for any h and k. From the definition of τ_i in (3.1), we have $\tau_i = 0$ if $i \in \mathcal{A}^c$.

Because $\tau_i \geq 0$, to prove that $\tau_i = 0$ must lead to $i \in \mathcal{A}^c$, all we need to show is $i \in \mathcal{A}$ indicates $\tau_i > 0$. Condition (C4) guarantees that $\mathrm{Span}\{(2\mathbf{I}_p - \mathbf{A}_{hk})^2 : h, k = 1, \ldots, H\} = \mathcal{S}_{Y|\mathbf{Z}}$, which in turn implies $\mathrm{Span}\{2\mathbf{I}_p - \mathbf{A}_{hk} : h, k = 1, \ldots, H\} = \mathcal{S}_{Y|\mathbf{Z}}$. From the invariance law of the central space, $\mathbf{e}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}(2\mathbf{I}_p - \mathbf{A}_{hk})\boldsymbol{\Sigma}^{-1/2}\mathbf{e}_i > 0$ for at least one set of h and k if $i \in \mathcal{A}$. Otherwise we get a contradiction to the "only if" part of Lemma A.2. As a result, we have $\tau_i \geq p_h p_k\{\mathbf{e}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}(2\mathbf{I}_p - \mathbf{A}_{hk})\boldsymbol{\Sigma}^{-1/2}\mathbf{e}_i\}^2 > 0$ as long as $i \in \mathcal{A}$.

Proof of Proposition 2. Let $\mathbf{A}(Y, \tilde{Y}) = E\{(\mathbf{Z}-\tilde{\mathbf{Z}})(\mathbf{Z}-\tilde{\mathbf{Z}})^{\mathrm{T}}|Y, \tilde{Y}\}$ and $a_i(Y, \tilde{Y}) = \mathbf{e}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}(2\mathbf{I}_p - \mathbf{A}(Y, \tilde{Y}))\boldsymbol{\Sigma}^{-1/2}\mathbf{e}_i$. Then τ_i in (3.1) is the discretized version of $E\{a_i^2(Y, \tilde{Y})\}$. Similarly, $\tau_{i,1}$ and $\tau_{i,2}^2$ are the discretized version of $E[\{\mathbf{e}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}(E(\mathbf{Z}\mathbf{Z}^{\mathrm{T}}|Y) - \mathbf{I}_p)\boldsymbol{\Sigma}^{-1/2}\mathbf{e}_i\}^2]$ and $[E\{\mathbf{e}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}E(\mathbf{Z}|Y)E^{\mathrm{T}}(\mathbf{Z}|Y)\boldsymbol{\Sigma}^{-1/2}\mathbf{e}_i\}]^2$, respectively. The former can be rewritten as

$$\begin{aligned} &E[\{\mathbf{e}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}(E(\mathbf{Z}\mathbf{Z}^{\mathrm{T}}|Y) - \mathbf{I}_p)\boldsymbol{\Sigma}^{-1/2}\mathbf{e}_i\}^2] \\ &\quad = E[\{\mathbf{e}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}E(\mathbf{Z}\mathbf{Z}^{\mathrm{T}}|Y)\boldsymbol{\Sigma}^{-1/2}\mathbf{e}_i\}^2] - (\mathbf{e}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{e}_i)^2. \end{aligned}$$

All we need is to prove that

$$\begin{aligned} E\{a_i^2(Y, \tilde{Y})\} &= 2E[\{\mathbf{e}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}E(\mathbf{Z}\mathbf{Z}^{\mathrm{T}}|Y)\boldsymbol{\Sigma}^{-1/2}\mathbf{e}_i\}^2] - 2(\mathbf{e}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{e}_i)^2 \\ &\quad +4[E\{\mathbf{e}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}E(\mathbf{Z}|Y)E^{\mathrm{T}}(\mathbf{Z}|Y)\boldsymbol{\Sigma}^{-1/2}\mathbf{e}_i\}]^2. \end{aligned} \tag{A.1}$$

Let $c_i(Y, \tilde{Y}) = \mathbf{e}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\mathbf{A}(Y, \tilde{Y})\boldsymbol{\Sigma}^{-1/2}\mathbf{e}_i$. Then $E\{a_i^2(Y, \tilde{Y})\}$ can be expressed as

$$E\{a_i^2(Y, \tilde{Y})\} = E\{c_i^2(Y, \tilde{Y})\} - 4(\mathbf{e}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{e}_i)^2. \tag{A.2}$$

It is easy to check that $c_i(Y, \tilde{Y}) = b_i(Y, \tilde{Y}) + b_i(\tilde{Y}, Y)$, where

$$\begin{aligned} b_i(Y, \tilde{Y}) &= \mathbf{e}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}E(\mathbf{Z}\mathbf{Z}^{\mathrm{T}}|Y)\boldsymbol{\Sigma}^{-1/2}\mathbf{e}_i - \mathbf{e}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}E(\mathbf{Z}|Y)E^{\mathrm{T}}(\tilde{\mathbf{Z}}|\tilde{Y})\boldsymbol{\Sigma}^{-1/2}\mathbf{e}_i, \\ b_i(\tilde{Y}, Y) &= \mathbf{e}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}E^{\mathrm{T}}(\tilde{\mathbf{Z}}\tilde{\mathbf{Z}}|\tilde{Y})\boldsymbol{\Sigma}^{-1/2}\mathbf{e}_i - \mathbf{e}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}E(\tilde{\mathbf{Z}}|\tilde{Y})E^{\mathrm{T}}(\mathbf{Z}|Y)\boldsymbol{\Sigma}^{-1/2}\mathbf{e}_i. \end{aligned}$$

Plug them into (A.2) and get

$$E\{a_i^2(Y,\tilde{Y})\} = 2E\{b_i^2(Y,\tilde{Y})\} + 2E\{b_i(Y,\tilde{Y})b_i(\tilde{Y},Y)\} - 4(\mathbf{e}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{e}_i)^2. \quad \text{(A.3)}$$

It can be shown that $E\{b_i^2(Y,\tilde{Y})\} = b_{1i} - b_{2i} - b_{3i} + b_{4i}$, where

$$b_{1i} = E[\{\mathbf{e}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}E(\mathbf{Z}\mathbf{Z}^{\mathrm{T}}|Y)\boldsymbol{\Sigma}^{-1/2}\mathbf{e}_i\}^2],$$
$$b_{2i} = E\{\mathbf{e}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}E(\mathbf{Z}\mathbf{Z}^{\mathrm{T}}|Y)\boldsymbol{\Sigma}^{-1/2}\mathbf{e}_i\mathbf{e}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}E(\mathbf{Z}|Y)E^{\mathrm{T}}(\tilde{\mathbf{Z}}|\tilde{Y})\boldsymbol{\Sigma}^{-1/2}\mathbf{e}_i\},$$
$$b_{3i} = E\{\mathbf{e}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}E(\mathbf{Z}|Y)E^{\mathrm{T}}(\tilde{\mathbf{Z}}|\tilde{Y})\boldsymbol{\Sigma}^{-1/2}\mathbf{e}_i\mathbf{e}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}E(\mathbf{Z}\mathbf{Z}^{\mathrm{T}}|Y)\boldsymbol{\Sigma}^{-1/2}\mathbf{e}_i\}, \text{ and}$$
$$b_{4i} = E\{\mathbf{e}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}E(\mathbf{Z}|Y)E^{\mathrm{T}}(\tilde{\mathbf{Z}}|\tilde{Y})\boldsymbol{\Sigma}^{-1/2}\mathbf{e}_i\mathbf{e}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}E(\mathbf{Z}|Y)E^{\mathrm{T}}(\tilde{\mathbf{Z}}|\tilde{Y})\boldsymbol{\Sigma}^{-1/2}\mathbf{e}_i\}.$$

Because $(\mathbf{Z},Y) \perp\!\!\!\perp (\tilde{\mathbf{Z}},\tilde{Y})$ and $E(\mathbf{Z}) = \mathbf{0}$, we have $b_{2i} = b_{3i} = 0$, and b_{4i} can be simplified as $b_{4i} = E^2\{\mathbf{e}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}E(\mathbf{Z}|Y)E^{\mathrm{T}}(\mathbf{Z}|Y)\boldsymbol{\Sigma}^{-1/2}\mathbf{e}_i\}$. We also have

$$E\{b_i(Y,\tilde{Y})b_i(\tilde{Y},Y)\} = (\mathbf{e}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{e}_i)^2 + E^2\{\mathbf{e}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}E(\mathbf{Z}|Y)E^{\mathrm{T}}(\mathbf{Z}|Y)\boldsymbol{\Sigma}^{-1/2}\mathbf{e}_i\}.$$

Plug them into (A.3) to get (A.1) as the desired result.

Proof of Lemma 1. For the notion of Frechet derivative, see, for example, Fernholz (1983). Let F be the joint distribution of $(\mathbf{X},Y)$ and F_n be the empirical distribution based on the i.i.d. sample $(\mathbf{X}_1,Y_1),\ldots,(\mathbf{X}_n,Y_n)$. Let G be a real or matrix-valued functional. Then the sample estimator $G(F_n)$ can be expanded as $G(F_n) = G(F) + E_n\{G^*(F)\} + O_p(n^{-1})$, where $E_n\{G^*(F)\} = O_p(n^{-1/2})$. We refer to $G^*(F)$ as the Frechet derivative of $G(F)$. Following Li and Wang (2007), we have

$$p_h^* = R_h - p_h, \boldsymbol{\Sigma}^* = \mathbf{X}\mathbf{X}^{\mathrm{T}} - E(\mathbf{X}\mathbf{X}^{\mathrm{T}}) - (\mathbf{X}-\boldsymbol{\mu})\boldsymbol{\mu}^{\mathrm{T}} - \boldsymbol{\mu}(\mathbf{X}-\boldsymbol{\mu})^{\mathrm{T}},$$
$$\mathbf{U}_h^* = \mathbf{X}R_h - E(\mathbf{X}R_h) - (\mathbf{X}-\boldsymbol{\mu})p_h - \boldsymbol{\mu}(R_h - p_h), (\boldsymbol{\Sigma}^{-1})^* = -\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}^*\boldsymbol{\Sigma}^{-1}, \text{ and}$$
$$\begin{aligned}\mathbf{V}_h^* = {} & \mathbf{X}\mathbf{X}^{\mathrm{T}}R_h - E(\mathbf{X}\mathbf{X}^{\mathrm{T}}R_h) + (R_h - p_h)\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + p_h(\mathbf{X}-\boldsymbol{\mu})\boldsymbol{\mu}^{\mathrm{T}} + p_h\boldsymbol{\mu}(\mathbf{X}-\boldsymbol{\mu})^{\mathrm{T}} \\ & -\{\mathbf{X} - E(\mathbf{X}R_h)\}\boldsymbol{\mu}^{\mathrm{T}} - (\mathbf{X}-\boldsymbol{\mu})E^{\mathrm{T}}(\mathbf{X}R_h) - \boldsymbol{\mu}\{\mathbf{X} - E(\mathbf{X}R_h)\}^{\mathrm{T}} \\ & -E(\mathbf{X}R_h)(\mathbf{X}-\boldsymbol{\mu})^{\mathrm{T}}.\end{aligned}$$

Take $\mathbf{t}_{1i} = (t_{1i,1},\ldots,t_{1i,H})^{\mathrm{T}}$ and $\hat{\mathbf{t}}_{1i} = (\hat{t}_{1i,1},\ldots,\hat{t}_{1i,H})^{\mathrm{T}}$, where, for $i = 1,\ldots,H$,

$$t_{1i,h} = p_h^{1/2}\mathbf{e}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(p_h^{-1}\mathbf{V}_h - \boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}\mathbf{e}_i \text{ and } \hat{t}_{1i,h} = \hat{p}_h^{1/2}\mathbf{e}_i^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}^{-1}(\hat{p}_h^{-1}\hat{\mathbf{V}}_h - \hat{\boldsymbol{\Sigma}})\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{e}_i.$$

The asymptotic expansion of $\hat{\mathbf{t}}_{1i}$ is $\hat{\mathbf{t}}_{1i} = \mathbf{t}_{1i} + E_n(\mathbf{t}_{1i}^*) + O_p(n^{-1})$, where $E_n(\mathbf{t}_{1i}^*) = O_p(n^{-1/2})$. For the Frechet derivative of $t_{1i,h}$, we get

$$\begin{aligned}(t_{1i,h})^* = {} & (p_h^{1/2})^*\mathbf{e}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(p_h^{-1}\mathbf{V}_h - \boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}\mathbf{e}_i + p_h^{1/2}\mathbf{e}_i^{\mathrm{T}}(\boldsymbol{\Sigma}^{-1})^*(p_h^{-1}\mathbf{V}_h - \boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}\mathbf{e}_i \\ & + p_h^{1/2}\mathbf{e}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(p_h^{-1}\mathbf{V}_h - \boldsymbol{\Sigma})(\boldsymbol{\Sigma}^{-1})^*\mathbf{e}_i + p_h^{1/2}\mathbf{e}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(p_h^{-1}\mathbf{V}_h - \boldsymbol{\Sigma})^*\boldsymbol{\Sigma}^{-1}\mathbf{e}_i. \quad \text{(A.4)}\end{aligned}$$

Conditions (C1) and (C2) imply that for $i \in \mathcal{A}^c$,

$$\mathbf{e}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(p_h^{-1}\mathbf{V}_h - \boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1} = \mathbf{e}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\{E(\mathbf{Z}\mathbf{Z}^{\mathrm{T}}|Y) - \mathbf{I}_p\}\boldsymbol{\Sigma}^{-1/2} = \mathbf{0}.$$

Thus the first three terms on the right hand side of (A.4) are all 0. Together with the fact that $(p_h^{-1}\mathbf{V}_h - \boldsymbol{\Sigma})^* = p_h^{-1}\mathbf{V}_h^* - p_h^{-2}p_h^*\mathbf{V}_h - \boldsymbol{\Sigma}^*$, we have $(t_{1i,h})^* = \ell_{ih}$, where $\ell_{ih} = p_h^{1/2}\mathbf{e}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(p_h^{-1}\mathbf{V}_h^* - p_h^{-2}p_h^*\mathbf{V}_h - \boldsymbol{\Sigma}^*)\boldsymbol{\Sigma}^{-1}\mathbf{e}_i$ was defined above Lemma 1. It follows that $\mathbf{t}_{1i}^* = \mathbf{L}_i = (\ell_{i1}, \ldots, \ell_{iH})^{\mathrm{T}}$. By the Central Limit Theorem, we have $\sqrt{n}(\hat{\mathbf{t}}_{1i} - \mathbf{t}_{1i}) \xrightarrow{D} N(\mathbf{0}, \mathrm{Cov}(\mathbf{L}_i))$. Here $\tau_{i,1} = \mathbf{t}_{1i}^{\mathrm{T}}\mathbf{t}_{1i}$ and $\hat{\tau}_{i,1} = \hat{\mathbf{t}}_{1i}^{\mathrm{T}}\hat{\mathbf{t}}_{1i}$. Because $\mathbf{t}_{1i} = \mathbf{0}$ for $i \in \mathcal{A}^c$ under (C1) and (C2), it follows that $n\hat{\tau}_{i,1} = n\hat{\mathbf{t}}_{1i}^{\mathrm{T}}\hat{\mathbf{t}}_{1i} \xrightarrow{D} \sum_{h=1}^{H} \iota_{ih}\chi_h^2(1)$.

Let $\mathbf{t}_{2i} = (t_{2i,1}, \ldots, t_{2i,H})^{\mathrm{T}}$ and $\hat{\mathbf{t}}_{2i} = (\hat{t}_{2i,1}, \ldots, \hat{t}_{2i,H})^{\mathrm{T}}$, where $t_{2i,h} = p_h^{-1/2}\mathbf{e}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{U}_h$ and $\hat{t}_{2i,h} = \hat{p}_h^{-1/2}\mathbf{e}_i^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}^{-1}\hat{\mathbf{U}}_h$. The Frechet derivative of $t_{2i,h}$ is $(t_{2i,h})^* = (p_h^{-1/2})^*\mathbf{e}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{U}_h + p_h^{-1/2}\mathbf{e}_i^{\mathrm{T}}(\boldsymbol{\Sigma}^{-1})^*\mathbf{U}_h + p_h^{-1/2}\mathbf{e}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{U}_h^*$. Because $\mathbf{e}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{U}_h = 0$ for $i \in \mathcal{A}^c$, we have $(t_{2i,h})^* = g_{ih} = p_h^{-1/2}\mathbf{e}_i^{\mathrm{T}}(\boldsymbol{\Sigma}^{-1})^*\mathbf{U}_h + p_h^{-1/2}\mathbf{e}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{U}_h^*$. It follows that $\mathbf{t}_{2i}^* = \mathbf{G}_i = (g_{i1}, \ldots, g_{iH})^{\mathrm{T}}$ and $\sqrt{n}(\hat{\mathbf{t}}_{2i} - \mathbf{t}_{2i}) \xrightarrow{D} N(\mathbf{0}, \mathrm{Cov}(\mathbf{G}_i))$. Here $\tau_{i,2} = \mathbf{t}_{2i}^{\mathrm{T}}\mathbf{t}_{2i}$ and $\hat{\tau}_{i,2} = \hat{\mathbf{t}}_{2i}^{\mathrm{T}}\hat{\mathbf{t}}_{2i}$. Because $\mathbf{t}_{2i} = \mathbf{0}$ for $i \in \mathcal{A}^c$, we have $n\hat{\tau}_{i,2} \xrightarrow{D} \sum_{h=1}^{H} \gamma_{ih}\chi_h^2(1)$.

For the second part, $\hat{\tau}_{i,1} \xrightarrow{P} \tau_{i,1}$ and $\hat{\tau}_{i,2} \xrightarrow{P} \tau_{i,2}$ by the Weak Law of Large Numbers. It remains to show that $\tau_{i,1} > 0$ and $\tau_{i,2} > 0$ for $i \in \mathcal{A}$. Following proofs similar to that of the second part of Proposition 1, we know $\tau_{i,1} > 0$ is guaranteed by (C1), (C2), and (C4). Meanwhile, $\tau_{i,2} > 0$ is guaranteed by (C3).

Proof of Theorem 1. We have seen that $\hat{\tau}_{i,1} = \hat{\mathbf{t}}_{1i}^{\mathrm{T}}\hat{\mathbf{t}}_{1i}$ and $\hat{\tau}_{i,2} = \hat{\mathbf{t}}_{2i}^{\mathrm{T}}\hat{\mathbf{t}}_{2i}$. Now $\hat{\tau}_{i,1} + 2\hat{\tau}_{i,2} = (\hat{\mathbf{t}}_{1i}^{\mathrm{T}}, \sqrt{2}\hat{\mathbf{t}}_{2i}^{\mathrm{T}})^{\mathrm{T}}(\hat{\mathbf{t}}_{1i}, \sqrt{2}\hat{\mathbf{t}}_{2i})$ and $\mathbf{W}_i = (\mathbf{L}_i^{\mathrm{T}}, \sqrt{2}\mathbf{G}_i^{\mathrm{T}})^{\mathrm{T}}$. The result follows directly from the proof of Lemma 1.

Proof of Theorem 2. Without loss of generality, we show the result based on directional regression, and the test statistic T_i exclusively refers to T_i^{DR}. It is straightforward to get a parallel proof for T_i^{SIR} or T_i^{SAVE}. Write $a \asymp b$ if $a = O_p(b)$ and $b = O_p(a)$. Theorem 1 implies that $T_i \asymp 1$ for $i \in \mathcal{A}^c$. The second part of Lemma 1 guarantees that $T_i/n \asymp 1$ for $i \in \mathcal{A}$. Together we get

$$\lim_{n\to\infty} P(T_i > T_j) = 1 \text{ for any } i \in \mathcal{A}, j \in \mathcal{A}^c. \tag{A.5}$$

From the definition of u_i, we have $T_{u_1} > T_{u_2} > \cdots > T_{u_p}$. Here u_i depends on n as $T_{u_i} = T_{(i)}$ changes as n changes. Because $|\mathcal{A}| = q$, as $n \to \infty$, the top ranked q test statistics $T_{u_1}, \ldots, T_{u_q}$ must come from the active predictors, and the inactive predictors must correspond to the remaining test statistics $T_{u_{q+1}}, \ldots, T_{u_p}$. Otherwise we get a contradiction to (A.5).

By definition $\hat{\mathcal{A}} = \{u_1, u_2, \ldots, u_{\hat{q}}\}$, so it remains to show that $\lim_{n\to\infty} P(\hat{q} = q) = 1$. For $i = 1, \ldots, q-1$, we have $T_{u_i} \in \mathcal{A}$ and $T_{u_{i+1}} \in \mathcal{A}$. Thus $T_{u_i}/n \asymp 1$ and $T_{u_{i+1}}/n \asymp 1$. It follows that $T_{u_i}/T_{u_{i+1}} \asymp 1$ and $\max_{i=1,\ldots,q-1} T_{u_i}/T_{u_{i+1}} \asymp 1$. Similarly for $i = q+1, \ldots, p-1$, we have $T_{u_i} \in \mathcal{A}^c$ and

$T_{u_{i+1}} \in \mathcal{A}^c$. Thus $T_{u_i} \asymp 1$ and $T_{u_{i+1}} \asymp 1$. It follows that $T_{u_i}/T_{u_{i+1}} \asymp 1$ and $\max_{i=q+1,\ldots,p-1} T_{u_i}/T_{u_{i+1}} \asymp 1$. On the other hand, for $i = q$, we have $T_{u_q} \in \mathcal{A}$ and $T_{u_{q+1}} \in \mathcal{A}^c$. Thus $T_{u_q}/n \asymp 1$ and $T_{u_{q+1}} \asymp 1$. It follows that $T_{u_q}/(nT_{u_{q+1}}) \asymp 1$. As a result, $\lim_{n\to\infty} P(T_{u_q}/T_{u_{q+1}} > \max_{i=1,\ldots,q-1,q+1,\ldots,p-1} T_{u_i}/T_{u_{i+1}}) = 1$. Consequently $\lim_{n\to\infty} P(q = \arg\max_{i=1,\ldots,p-1} T_{u_i}/T_{u_{i+1}}) = 1$.

References

Cook, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions through Graphics*. John Wiley, New York.

Cook, R. D. (2004). Testing predictor contributions in sufficient dimension reduction. *Ann. Statist.* **32**, 1062-1092.

Cook, R. D., Li, B. and Chiaromonte, F. (2007). Dimension reduction in regression without matrix inversion. *Biometrika* **94**, 569-584.

Cook, R. D. and Weisberg, S. (1991). Comment on "Sliced inverse regression for dimension reduction," by K. C. Li, *J. Amer. Statist. Assoc.* **86**, 316-342.

Cox, D. R. and Snell, E. J. (1974). The choice of variables in observational studies. *J. Roy. Statist. Soc. Ser. C* **23**, 51-59.

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

Fernholz, L. T. (1983). *Von Mises Calculus for Statistical Functionals*. Springer, New York.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annual Eugenics* **7**, Part II, 179-188.

Li, B. and Dong, Y. (2009). Dimension reduction for non-elliptically distributed predictors. *Ann. Statist.* **37**, 1272-1298.

Li, B. and Wang, S. (2007). On directional regression for dimension reduction. *J. Amer. Statist. Assoc.* **102**, 997-1008.

Li, K. C. (1991). Sliced inverse regression for dimension reduction (with comments). *J. Amer. Statist. Assoc.* **86**, 316-342.

Li, L., Cook, R. D. and Nachtsheim, C. J. (2005). Model-free variable selection. *J. Roy. Statist. Soc. Ser. B* **67**, 285-299.

Luo, R., Wang, H. and Tsai, C. L. (2009). Contour projected dimension reduction. *Ann. Statist.* **37**, 3743-3778.

Shao, Y., Cook, R. D. and Weisberg, S. (2007). Marginal tests with sliced average variance estimation. *Biometrika* **94**, 285-296.

Ye, Z. and Weiss, R. (2003). Using the bootstrap to select one of a new class of dimension reduction methods. *J. Amer. Statist. Assoc.* **98**, 968-979.

Yin, X., Li, B. and Cook, R. D. (2008). Successive direction extraction for estimating the central subspace in a multiple-index regression. *J. Multivariate Anal.* **99**, 1733-1757.

School of Statistics, East China Normal University, Shanghai 200241, P. R. China.

E-mail: zyu@stat.ecnu.edu.cn

Department of Statistics, Temple University, Philadelphia, PA 19122, U.S.A.

E-mail: ydong@temple.edu

(Received March 2014; accepted September 2015)