On dual model-free variable selection with two groups of variables

Ahmad Alothman[a], Yuexiao Dong[a,*], Andreas Artemiou[b]

[a]*Department of Statistical Science, Temple University*
[b]*School of Mathematics, Cardiff University*

Abstract

In the presence of two groups of variables, existing model-free variable selection methods only reduce the dimensionality of the predictors. We extend the popular marginal coordinate hypotheses [3] in the sufficient dimension reduction literature and consider the dual marginal coordinate hypotheses, where the role of the predictor and the response is not important. Motivated by canonical correlation analysis (CCA), we propose a CCA-based test for the dual marginal coordinate hypotheses, and devise a joint backward selection algorithm for dual model-free variable selection. The performances of the proposed test and the variable selection procedure are evaluated through synthetic examples and a real data analysis.

Keywords: Canonical correlation analysis, Dual marginal coordinate hypotheses, Sliced inverse regression, Trace test

1. Introduction

In this paper, we consider dual model-free variable selection with two groups of variables $\mathbf{x} \in \mathbb{R}^p$ and $\mathbf{y} \in \mathbb{R}^q$. As a popular tool for multivariate analysis, classical variable selection aims at identifying important variables among $\mathbf{x}$ for the prediction of $\mathbf{y}$. Most existing variable selection methods are model-based, and consider selecting important predictors under a given parametric or semi-parametric model. Variable selection methods in linear regression include LASSO [17], SCAD [5], the adaptive LASSO [22], and the Dantzig selector [1]. Variable selection in semi-parametric models have been studied in [7, 14, 18]. In multivariate association studies with two sets of random vectors, popular methods such as canonical correlation analysis (CCA) [6] focus on reducing the dimensionality for both sets of variables, where the role of the predictor and the response is not important. This viewpoint motivates us to consider dual variable selection, where the goal is to simultaneously identify the important variables among $\mathbf{x}$ for the prediction of $\mathbf{y}$ and the important variables among $\mathbf{y}$ for the prediction of $\mathbf{x}$.

Unlike model-based procedures in the literature, our proposal is model-free and does not require assuming specific models between $\mathbf{x}$ and $\mathbf{y}$. Existing model-free variable selection methods all focus on selecting important variables among $\mathbf{x}$ for the prediction of $\mathbf{y}$. See, for example, [9, 12, 13, 21]. The aforementioned model-free variable selection methods are closely related to sufficient dimension reduction [2]. An important link between sufficient dimension reduction and model-free variable selection is elucidated in [21], where popular sufficient dimension reduction methods such as sliced inverse regression (SIR) [11], sliced average variance estimation [4], and directional regression [10] are used to construct corresponding model-free variable selection procedures.

To achieve dual model-free variable selection, we demonstrate that CCA can be viewed as a valid sufficient dimension reduction procedure under suitable conditions. There is an important difference between CCA and popular sufficient dimension reduction methods such as SIR: CCA maintains the symmetry between $\mathbf{x}$ and $\mathbf{y}$ while SIR does not. We follow Yu et al. [21] and develop CCA-based model-free variable selection procedures. Unlike the procedures proposed in Yu et al. [21] that select important variables among $\mathbf{x}$, the symmetry in CCA provides a unique opportunity to perform dual variable selection among both $\mathbf{x}$ and $\mathbf{y}$ simultaneously.

*Corresponding author.
Email address: ydong@temple.edu (Yuexiao Dong)

The rest of the paper is organized as follows. We review SIR-based trace test for variable selection in Section 2. The general framework for dual model-free variable selection is introduced in Section 3. CCA-based trace test for dual variable selection is developed in Section 4. Numerical studies are performed in Section 5 and we conclude the paper with some discussions in Section 6. All the proofs are relegated to the Appendix.

2. Review of SIR-based trace test

Let $\mathbf{x} = (X_1, \ldots, X_p)^\top$ and $\mathbf{y} = (Y_1, \ldots, Y_q)^\top$. Without loss of generality, assume $\mathrm{E}(\mathbf{x}) = \mathbf{0}$ and $\mathrm{E}(\mathbf{y}) = \mathbf{0}$. Denote $\mathbf{x}_{-k} = (X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_p)^\top$ for $k \in \{1, \ldots, p\}$. To test the importance of the kth predictor X_k, we may consider the following hypotheses

$$\mathcal{H}_0^{-k} : \mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{-k} \quad \text{vs.} \quad \mathcal{H}_a^{-k} : \mathbf{y} \not\perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{-k}, \tag{1}$$

where $\perp\!\!\!\perp$ means independence and $\not\perp\!\!\!\perp$ means no independence. The hypothesis $\mathcal{H}_0^{-k} : \mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{-k}$ above implies that X_k is not important for the prediction of $\mathbf{y}$ in the presence of all the other predictors. Hypotheses (1) are known as the marginal coordinate hypotheses [3]. Once we have a valid test for (1), sequential procedures such as forward selection, backward selection and stepwise regression can be designed in parallel to the classical procedures in linear regression. For example, Li et al. [13] consider backward selection through the marginal coordinate test proposed in [3], while forward selection and stepwise regression through the trace test are discussed in [21].

Since [3], different tests for (1) have been proposed in the literature. Most tests have the same flavor as the original marginal coordinate test in [3], such as [16, 19, 20]. Yu et al. [21] introduce a novel family of trace tests, which can be combined with various sufficient dimension reduction methods. In the following, we first review SIR as a sufficient dimension reduction method, and then we revisit the SIR-based trace test for (1).

Classical sufficient dimension reduction aims to find $\boldsymbol{\beta} \in \mathbb{R}^{p\times d}$ with the smallest possible column space such that $\mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \boldsymbol{\beta}^\top \mathbf{x}$. The corresponding column space is known as the central space for the regression of $\mathbf{y}$ on $\mathbf{x}$, and is denoted by $\mathcal{S}_{\mathbf{y}|\mathbf{x}}$. Let $\boldsymbol{\Sigma}_{\mathbf{x}} = \mathrm{var}(\mathbf{x})$ and let $\{J_1, \ldots, J_H\}$ denote a measurable partition of $\Theta_{\mathbf{y}}$, the sample space of $\mathbf{y}$. Under the linearity condition that $\mathrm{E}(\mathbf{x}|\boldsymbol{\beta}^\top \mathbf{x})$ is linear in $\boldsymbol{\beta}^\top \mathbf{x}$, we know from [11] that

$$\mathrm{E}(\mathbf{z}|\mathbf{y} \in J_h) \in \boldsymbol{\Sigma}_{\mathbf{x}}^{1/2} \mathcal{S}_{\mathbf{y}|\mathbf{x}}, \tag{2}$$

where $\mathbf{z} = \boldsymbol{\Sigma}_{\mathbf{x}}^{-1/2} \mathbf{x}$ is the standardized version of $\mathbf{x}$. Define $\mathbf{z}$-scale SIR kernel matrix as $\mathbf{M}^{\mathrm{SIR}} = \sum_{h=1}^{H} \pi_h \mathrm{E}(\mathbf{z}|\mathbf{y} \in J_h)\mathrm{E}^\top(\mathbf{z}|\mathbf{y} \in J_h)$, where $\pi_h = \Pr(\mathbf{y} \in J_h)$. From (2), we know

$$\mathrm{Span}(\mathbf{M}^{\mathrm{SIR}}) \subseteq \boldsymbol{\Sigma}_{\mathbf{x}}^{1/2} \mathcal{S}_{\mathbf{y}|\mathbf{x}}, \tag{3}$$

where Span denotes the column space.

Let $\boldsymbol{\Sigma}_{\mathbf{x}_{-k}} = \mathrm{var}(\mathbf{x}_{-k})$ and $\mathbf{z}_{-k} = \boldsymbol{\Sigma}_{\mathbf{x}_{-k}}^{-1/2} \mathbf{x}_{-k}$. Similar to $\mathbf{M}^{\mathrm{SIR}}$, we define $\mathbf{M}_{-k}^{\mathrm{SIR}} = \sum_{h=1}^{H} \pi_h \mathrm{E}(\mathbf{z}_{-k}|\mathbf{y} \in J_h)\mathrm{E}^\top(\mathbf{z}_{-k}|\mathbf{y} \in J_h)$. Yu et al. [21] consider

$$\delta_k^{\mathrm{SIR}} = \mathrm{tr}(\mathbf{M}^{\mathrm{SIR}}) - \mathrm{tr}(\mathbf{M}_{-k}^{\mathrm{SIR}}), \tag{4}$$

where tr denotes the trace. Yu et al. [21] provide the asymptotic distribution of $\hat{\delta}_k^{\mathrm{SIR}}$ under $\mathcal{H}_0^{-k}$, where $\hat{\delta}_k^{\mathrm{SIR}}$ is the sample version of δ_k^{SIR}. Because $\delta_k^{\mathrm{SIR}} = 0$ under $\mathcal{H}_0^{-k}$ in (1), $\mathcal{H}_0^{-k}$ is rejected if $\hat{\delta}_k^{\mathrm{SIR}}$ is larger than a threshold determined by its asymptotic distribution under null.

3. The principle of dual model-free variable selection

Denote $\mathcal{I}_{\mathbf{x}} = \{1, \ldots, p\}$ as the full index set for $\mathbf{x}$. Define the active set $\mathcal{A}$ for the regression of $\mathbf{y}$ on $\mathbf{x}$ as

$$\mathcal{A} = \{k \in \mathcal{I}_{\mathbf{x}} : \mathbf{y} \text{ depends on } \mathbf{x} \text{ through } X_k\}. \tag{5}$$

Similarly, let $\mathcal{I}_{\mathbf{y}} = \{1, \ldots, q\}$ denote the full index set for $\mathbf{y}$, and the active set $\mathcal{B}$ for the regression of $\mathbf{x}$ on $\mathbf{y}$ be defined as

$$\mathcal{B} = \{j \in \mathcal{I}_{\mathbf{y}} : \mathbf{x} \text{ depends on } \mathbf{y} \text{ through } Y_j\}. \tag{6}$$

Let $\mathbf{x}_{\mathcal{A}} = \{X_k : k \in \mathcal{A}\}$ and $\mathbf{y}_{\mathcal{B}} = \{Y_j : j \in \mathcal{B}\}$. We have the following result.

Proposition 1. *The following three conditions are equivalent, and all are implied from the definitions of $\mathcal{A}$ in* (5) *and $\mathcal{B}$ in* (6).

(i) $\mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{\mathcal{A}}$ *and* $\mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{y}_{\mathcal{B}}$*;*

(ii) $\mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{\mathcal{A}}$ *and* $\mathbf{y} \perp\!\!\!\perp \mathbf{x}_{\mathcal{A}} \mid \mathbf{y}_{\mathcal{B}}$*;*

(iii) $\mathbf{y}_{\mathcal{B}} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{\mathcal{A}}$ *and* $\mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{y}_{\mathcal{B}}$.

Let $\varnothing$ denote the empty set. It follows from Proposition 1 that $\mathcal{A} = \varnothing$ if and only if $\mathcal{B} = \varnothing$. We remark that Proposition 1 is parallel to Proposition 1 in [8], where the dual central spaces for sufficient dimension reduction are studied.

The goal of dual model-free variable selection is to identify $\mathcal{A}$ and $\mathcal{B}$ without assuming specific models between $\mathbf{x}$ and $\mathbf{y}$. Let $\mathbf{x}_{\mathcal{F}} = \{X_k : k \in \mathcal{F}\}$ and $\mathbf{y}_{\mathcal{G}} = \{Y_j : j \in \mathcal{G}\}$, where $\mathcal{F} \subseteq \mathcal{I}_{\mathbf{x}}$ is the working active set for $\mathbf{x}$ and $\mathcal{G} \subseteq \mathcal{I}_{\mathbf{y}}$ is the working active set for $\mathbf{y}$. Motivated from part (i) in Proposition 1 and the marginal coordinate hypotheses (1) in Section 2, we consider the following dual marginal coordinate hypotheses

$$\mathcal{H}_0^{\mathcal{F},[\mathcal{G}]} : \mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{\mathcal{F}} \text{ and } \mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{y}_{\mathcal{G}} \quad \text{vs.} \quad \mathcal{H}_a^{\mathcal{F},[\mathcal{G}]} : \mathbf{y} \not\perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{\mathcal{F}} \text{ or } \mathbf{y} \not\perp\!\!\!\perp \mathbf{x} \mid \mathbf{y}_{\mathcal{G}}. \tag{7}$$

If $\mathcal{H}_0^{\mathcal{F},[\mathcal{G}]}$ in (7) is true, then obviously we have $\mathcal{A} \subseteq \mathcal{F}$ and $\mathcal{B} \subseteq \mathcal{G}$. We can then recover $\mathcal{A}$ and $\mathcal{B}$ by looking for the combination of the smallest possible $\mathcal{F}$ and the smallest possible $\mathcal{G}$ such that $\mathcal{H}_0^{\mathcal{F},[\mathcal{G}]}$ is not rejected.

4. CCA-based trace tests and dual model-free variable selection

We have reviewed in Section 2 that SIR-based trace test can be used to test the marginal coordinate hypotheses (1). To test the dual marginal coordinate hypotheses (7), where the roles of $\mathbf{x}$ and $\mathbf{y}$ are symmetric, we need a dimension reduction method that maintains the symmetry between $\mathbf{x}$ and $\mathbf{y}$. In Section 4.1, we introduce CCA as a dual sufficient dimension reduction method. In Section 4.2, we study CCA-based trace tests for selecting variables among either $\mathbf{x}$ or $\mathbf{y}$. In Section 4.3, CCA-based test for the dual marginal coordinate hypotheses (7) is developed. In Section 4.4, we propose a sample level algorithm for dual model-free variable selection.

4.1. CCA for dual sufficient dimension reduction

Recall that $\mathbf{z} = \boldsymbol{\Sigma}_{\mathbf{x}}^{-1/2}\mathbf{x}$ is the standardized version of $\mathbf{x}$. Let $\mathbf{w} = \boldsymbol{\Sigma}_{\mathbf{y}}^{-1/2}\mathbf{y}$ be the standardized version of $\mathbf{y}$, where $\boldsymbol{\Sigma}_{\mathbf{y}} = \mathrm{var}(\mathbf{y})$. Define kernel matrices

$$\mathbf{M} = \mathrm{E}(\mathbf{z}\mathbf{w}^\top)\mathrm{E}(\mathbf{w}\mathbf{z}^\top) \quad \text{and} \quad \widetilde{\mathbf{M}} = \mathrm{E}(\mathbf{w}\mathbf{z}^\top)\mathrm{E}(\mathbf{z}\mathbf{w}^\top). \tag{8}$$

Given $\mathbf{x} \in \mathbb{R}^p$ and $\mathbf{y} \in \mathbb{R}^q$, the ℓth pair of canonical covariates (u_ℓ, v_ℓ) is defined as $u_\ell = \mathbf{a}_\ell^\top\mathbf{x}$ and $v_\ell = \mathbf{b}_\ell^\top\mathbf{y}$, such that $\mathrm{var}(u_\ell) = \mathrm{var}(v_\ell) = 1$ and $\mathrm{cov}(u_\ell, v_\ell)$ is maximized. For $\ell > 1$, u_ℓ and v_ℓ satisfy the additional constraints that $\mathrm{cov}(u_\ell, u_k) = 0$ and $\mathrm{cov}(v_\ell, v_k) = 0$ for all $k < \ell$. It is well-known that the $\mathbf{a}_\ell = \boldsymbol{\Sigma}_{\mathbf{x}}^{-1/2}\mathbf{c}_\ell$, where $\mathbf{c}_\ell$ is the eigenvector corresponding to the ℓth largest eigenvalue of $\mathbf{M}$. Similarly, $\mathbf{b}_\ell = \boldsymbol{\Sigma}_{\mathbf{y}}^{-1/2}\mathbf{d}_\ell$, where $\mathbf{d}_\ell$ is the ℓth eigenvector of $\widetilde{\mathbf{M}}$. Denote $\mathcal{S}_{\mathbf{y}|\mathbf{x}}$ and $\mathcal{S}_{\mathbf{x}|\mathbf{y}}$ as the dual central spaces for the regression of $\mathbf{y}$ on $\mathbf{x}$ and the regression of $\mathbf{x}$ on $\mathbf{y}$, respectively. The next result states that matrices $\mathbf{M}$ and $\widetilde{\mathbf{M}}$ are closely related to sufficient dimension reduction.

Proposition 2. *Suppose* $\mathrm{E}(\mathbf{x}) = \mathbf{0}$ *and* $\mathrm{E}(\mathbf{y}) = \mathbf{0}$. *Assume* $\boldsymbol{\beta}$ *is the basis for* $\mathcal{S}_{\mathbf{y}|\mathbf{x}}$ *and* $\boldsymbol{\eta}$ *is the basis for* $\mathcal{S}_{\mathbf{x}|\mathbf{y}}$.

(i) If $\mathrm{E}(\mathbf{x}|\boldsymbol{\beta}^\top\mathbf{x})$ *is linear in* $\boldsymbol{\beta}^\top\mathbf{x}$, *then* $\mathrm{Span}(\mathbf{M}) \subseteq \boldsymbol{\Sigma}_{\mathbf{x}}^{1/2}\mathcal{S}_{\mathbf{y}|\mathbf{x}}$;

(ii) If $\mathrm{E}(\mathbf{y}|\boldsymbol{\eta}^\top\mathbf{y})$ *is linear in* $\boldsymbol{\eta}^\top\mathbf{y}$, *then* $\mathrm{Span}(\widetilde{\mathbf{M}}) \subseteq \boldsymbol{\Sigma}_{\mathbf{y}}^{1/2}\mathcal{S}_{\mathbf{x}|\mathbf{y}}$.

The assumptions made in this proposition are common in the sufficient dimension reduction literature. Proposition 2 implies that the column space of $\boldsymbol{\Sigma}_{\mathbf{x}}^{-1/2}\mathbf{M}$ can recover the central space for the regression of $\mathbf{y}$ on $\mathbf{x}$, while the column space of $\boldsymbol{\Sigma}_{\mathbf{y}}^{-1/2}\widetilde{\mathbf{M}}$ can recover the central space for the regression of $\mathbf{x}$ on $\mathbf{y}$. It follows that $\mathbf{a}_\ell \in \mathcal{S}_{\mathbf{y}|\mathbf{x}}$ and $\mathbf{b}_\ell \in \mathcal{S}_{\mathbf{x}|\mathbf{y}}$. We remark that the conclusions in Proposition 2 bare close resemblance to (3) about the SIR-based kernel matrix $\mathbf{M}^{\mathrm{SIR}}$.

4.2. CCA-based trace tests for marginal coordinate hypotheses

We consider two sets of marginal coordinate hypotheses in this section, both of which are related to (7). The first set is

$$\mathcal{H}_0^{\mathcal{F}} : \mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{\mathcal{F}} \quad \text{vs.} \quad \mathcal{H}_a^{\mathcal{F}} : \mathbf{y} \not\perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{\mathcal{F}}. \tag{9}$$

Hypotheses (9) include (1) as a special case, as $\mathbf{x}_{\mathcal{F}}$ becomes $\mathbf{x}_{-k}$ when we take $\mathcal{F} = \{1, \ldots, k-1, k+1, \ldots, p\}$. The second set is

$$\mathcal{H}_0^{[\mathcal{G}]} : \mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{y}_{\mathcal{G}} \quad \text{vs.} \quad \mathcal{H}_a^{[\mathcal{G}]} : \mathbf{y} \not\perp\!\!\!\perp \mathbf{x} \mid \mathbf{y}_{\mathcal{G}}. \tag{10}$$

While hypotheses (9) can be used for selecting important variables among $\mathbf{x}$, hypotheses (10) are useful for selecting important variables among $\mathbf{y}$.

First we focus on the CCA-based trace test for (9). Let $\boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}}} = \mathrm{var}(\mathbf{x}_{\mathcal{F}})$ and $\mathbf{z}_{\mathcal{F}} = \boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}}}^{-1/2}\mathbf{x}_{\mathcal{F}}$. Motivated by the SIR-based trace test, we consider

$$\delta_{-\mathcal{F}} = \mathrm{tr}(\mathbf{M}) - \mathrm{tr}(\mathbf{M}_{\mathcal{F}}), \tag{11}$$

where $\mathbf{M}_{\mathcal{F}} = \mathrm{E}(\mathbf{z}_{\mathcal{F}}\mathbf{w}^\top)\mathrm{E}(\mathbf{w}\mathbf{z}_{\mathcal{F}}^\top)$. We remark that $\delta_{-\mathcal{F}}$ is constructed as the trace difference of two $\mathbf{z}$-scale CCA kernel matrices, which has the same flavor as δ_k^{SIR} in (4).

Let $\mathcal{F}^c$ be the complement of $\mathcal{F}$ in $\mathcal{I}_{\mathbf{x}}$ and denote $\boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}^c}} = \mathrm{var}(\mathbf{x}_{\mathcal{F}^c})$. Define $\boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}} = \boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}^c}} - \mathrm{E}(\mathbf{x}_{\mathcal{F}^c}\mathbf{x}_{\mathcal{F}}^\top)\boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}}}^{-1}\mathrm{E}(\mathbf{x}_{\mathcal{F}}\mathbf{x}_{\mathcal{F}^c}^\top)$ and $\boldsymbol{\gamma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}} = \mathbf{x}_{\mathcal{F}^c} - \mathrm{E}(\mathbf{x}_{\mathcal{F}^c}\mathbf{x}_{\mathcal{F}}^\top)\boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}}}^{-1}\mathbf{x}_{\mathcal{F}}$. Then we have

Proposition 3. *Suppose* $\mathrm{E}(\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}})$ *is a linear function of* $\mathbf{x}_{\mathcal{F}}$. *Then*

(i) $\delta_{-\mathcal{F}} = \mathrm{tr}\{\boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}^{-1}\mathrm{E}(\boldsymbol{\gamma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}\mathbf{y}^\top)\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\mathrm{E}(\mathbf{y}\boldsymbol{\gamma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}^\top)\}$.

(ii) $\delta_{-\mathcal{F}} = 0$ *under* $\mathcal{H}_0^{\mathcal{F}} : \mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{\mathcal{F}}$.

The assumption made in this proposition is common in the model-free variable selection literature, and it is satisfied if $\mathbf{x}$ is normal. The first part of Proposition 3 provides the explicit formula to calculate $\delta_{-\mathcal{F}}$. The second part states that if $\mathbf{x}_{\mathcal{F}^c}$ is unimportant for the prediction of $\mathbf{y}$ given $\mathbf{x}_{\mathcal{F}}$, then $\delta_{-\mathcal{F}}$ becomes zero. Yu et al. [21] have shown that $\delta_k^{\mathrm{SIR}} = 0$ if X_k is unimportant for the prediction of $\mathbf{y}$ given $\mathbf{x}_{-k}$. Our result here is more general as $\mathcal{F}^c$ can contain more than one variable. Denote $\hat{\delta}_{-\mathcal{F}}$ as the sample version of $\delta_{-\mathcal{F}}$. We reject $\mathcal{H}_0^{\mathcal{F}}$ if $\hat{\delta}_{-\mathcal{F}}$ is too large. The asymptotic distribution of $\hat{\delta}_{-\mathcal{F}}$ under $\mathcal{H}_0^{\mathcal{F}}$ is provided in Corollary 1 in the Appendix.

Next we introduce the CCA-based trace test for (10). Let $\boldsymbol{\Sigma}_{\mathbf{y}_{\mathcal{G}}} = \mathrm{var}(\mathbf{y}_{\mathcal{G}})$ and $\mathbf{w}_{\mathcal{G}} = \boldsymbol{\Sigma}_{\mathbf{y}_{\mathcal{G}}}^{-1/2}\mathbf{y}_{\mathcal{G}}$. Denote $\widetilde{\mathbf{M}}_{\mathcal{G}} = \mathrm{E}(\mathbf{w}_{\mathcal{G}}\mathbf{z}^\top)\mathrm{E}(\mathbf{z}\mathbf{w}_{\mathcal{G}}^\top)$ and consider

$$\delta^{-\mathcal{G}} = \mathrm{tr}(\widetilde{\mathbf{M}}) - \mathrm{tr}(\widetilde{\mathbf{M}}_{\mathcal{G}}). \tag{12}$$

Let $\mathcal{G}^c$ be the complement of $\mathcal{G}$ in $\mathcal{I}_{\mathbf{y}}$ and denote $\boldsymbol{\Sigma}_{\mathbf{y}_{\mathcal{G}^c}} = \mathrm{var}(\mathbf{y}_{\mathcal{G}^c})$. Define $\boldsymbol{\Sigma}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}} = \boldsymbol{\Sigma}_{\mathbf{y}_{\mathcal{G}^c}} - \mathrm{E}(\mathbf{y}_{\mathcal{G}^c}\mathbf{y}_{\mathcal{G}}^\top)\boldsymbol{\Sigma}_{\mathbf{y}_{\mathcal{G}}}^{-1}\mathrm{E}(\mathbf{y}_{\mathcal{G}}\mathbf{y}_{\mathcal{G}^c}^\top)$. Let $\boldsymbol{\gamma}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}} = \mathbf{y}_{\mathcal{G}^c} - \mathrm{E}(\mathbf{y}_{\mathcal{G}^c}\mathbf{y}_{\mathcal{G}}^\top)\boldsymbol{\Sigma}_{\mathbf{y}_{\mathcal{G}}}^{-1}\mathbf{y}_{\mathcal{G}}$. Parallel to Proposition 3, we have

Proposition 4. *Suppose* $\mathrm{E}(\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}})$ *is a linear function of* $\mathbf{y}_{\mathcal{G}}$. *Then*

(i) $\delta^{-\mathcal{G}} = \mathrm{tr}\{\mathbf{\Sigma}^{-1}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}\mathrm{E}(\boldsymbol{\gamma}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}\mathbf{x}^\top)\mathbf{\Sigma}^{-1}_{\mathbf{x}}\mathrm{E}(\mathbf{x}\boldsymbol{\gamma}^\top_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}})\}$.

(ii) $\delta^{-\mathcal{G}} = 0$ *under* $\mathcal{H}_0^{[\mathcal{G}]} : \mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{y}_{\mathcal{G}}$.

Let $\hat{\delta}^{-\mathcal{G}}$ be the sample version of $\delta^{-\mathcal{G}}$. We reject $\mathcal{H}_0^{[\mathcal{G}]}$ if $\hat{\delta}^{-\mathcal{G}}$ is too large.

The asymptotic distribution of $\hat{\delta}^{-\mathcal{G}}$ under $\mathcal{H}_0^{[\mathcal{G}]}$ is provided in Corollary 2 in the Appendix.

4.3. CCA-based trace test for dual marginal coordinate hypotheses

In this section, we develop a test for $\mathcal{H}_0^{\mathcal{F},[\mathcal{G}]} : \mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{\mathcal{F}}$ and $\mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{y}_{\mathcal{G}}$ versus the alternative $\mathcal{H}_a^{\mathcal{F},[\mathcal{G}]} : \mathbf{y} \not\perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{\mathcal{F}}$ or $\mathbf{y} \not\perp\!\!\!\perp \mathbf{x} \mid \mathbf{y}_{\mathcal{G}}$. From the definition of $\mathbf{M} = \mathrm{E}(\mathbf{z}\mathbf{w}^\top)\mathrm{E}(\mathbf{w}\mathbf{z}^\top)$ and $\widetilde{\mathbf{M}} = \mathrm{E}(\mathbf{w}\mathbf{z}^\top)\mathrm{E}(\mathbf{z}\mathbf{w}^\top)$ in (8), we have $\mathrm{tr}(\mathbf{M}) = \mathrm{tr}(\widetilde{\mathbf{M}})$. Recall $\mathbf{M}_{\mathcal{F}} = \mathrm{E}(\mathbf{z}_{\mathcal{F}}\mathbf{w}^\top)\mathrm{E}(\mathbf{w}\mathbf{z}^\top_{\mathcal{F}})$ and define $\mathbf{M}^{\mathcal{G}} = \mathrm{E}(\mathbf{z}\mathbf{w}^\top_{\mathcal{G}})\mathrm{E}(\mathbf{w}_{\mathcal{G}}\mathbf{z}^\top)$. It is easy to see that $\mathrm{tr}(\mathbf{M}^{\mathcal{G}}) = \mathrm{tr}(\widetilde{\mathbf{M}}_{\mathcal{G}})$. Hence $\delta^{-\mathcal{G}}$ in (12) becomes

$$\delta^{-\mathcal{G}} = \mathrm{tr}(\mathbf{M}) - \mathrm{tr}(\mathbf{M}^{\mathcal{G}}). \tag{13}$$

We have seen that $\delta_{-\mathcal{F}} = \mathrm{tr}(\mathbf{M}) - \mathrm{tr}(\mathbf{M}_{\mathcal{F}})$ in (11) can be used to test $\mathcal{H}_0^{\mathcal{F}} : \mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{\mathcal{F}}$, and $\delta^{-\mathcal{G}}$ in (13) can be used to test $\mathcal{H}_0^{[\mathcal{G}]} : \mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{y}_{\mathcal{G}}$. This motivates us to consider

$$\delta^{-\mathcal{G}}_{-\mathcal{F}} = \mathrm{tr}(\mathbf{M}) - \mathrm{tr}(\mathbf{M}^{\mathcal{G}}_{\mathcal{F}}), \tag{14}$$

where $\mathbf{M}^{\mathcal{G}}_{\mathcal{F}} = \mathrm{E}(\mathbf{z}_{\mathcal{F}}\mathbf{w}^\top_{\mathcal{G}})\mathrm{E}(\mathbf{w}_{\mathcal{G}}\mathbf{z}^\top_{\mathcal{F}})$. Note that $\delta^{-\mathcal{G}}_{-\mathcal{F}}$ in (14) include $\delta_{-\mathcal{F}}$ and $\delta^{-\mathcal{G}}$ as special cases. If we take $\mathcal{F} = \mathcal{I}_{\mathbf{x}}$, then $\delta^{-\mathcal{G}}_{-\mathcal{F}}$ becomes $\delta^{-\mathcal{G}}$. On the other hand, $\delta^{-\mathcal{G}}_{-\mathcal{F}}$ reduces to $\delta_{-\mathcal{F}}$ when we set $\mathcal{G} = \mathcal{I}_{\mathbf{y}}$.

The symmetry between $\mathbf{z}$ and $\mathbf{w}$ in the definition of $\mathbf{M}$ and $\widetilde{\mathbf{M}}$ allows $\mathrm{tr}(\mathbf{M})$ to simultaneously capture the regression information between $\mathbf{y}$ and $\mathbf{x}$ as well as the regression information between $\mathbf{x}$ and $\mathbf{y}$. This is a unique feature of the CCA-based trace test, as $\mathrm{tr}(\mathbf{M}^{\mathrm{SIR}})$ in (4) only captures the regression information between $\mathbf{y}$ and $\mathbf{x}$. Parallel to Proposition 3 and Proposition 4, we have

Proposition 5. *Suppose* $\mathrm{E}(\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}})$ *is a linear function of* $\mathbf{x}_{\mathcal{F}}$ *and* $\mathrm{E}(\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}})$ *is a linear function of* $\mathbf{y}_{\mathcal{G}}$. *Then*

(i) $\delta^{-\mathcal{G}}_{-\mathcal{F}} = \mathrm{tr}\{\mathbf{\Sigma}^{-1}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}\mathrm{E}(\boldsymbol{\gamma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}\mathbf{y}^\top)\mathbf{\Sigma}^{-1}_{\mathbf{y}}\mathrm{E}(\mathbf{y}\boldsymbol{\gamma}^\top_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}})\} + \mathrm{tr}\{\mathbf{\Sigma}^{-1}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}\mathrm{E}(\boldsymbol{\gamma}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}\mathbf{x}^\top_{\mathcal{F}})\mathbf{\Sigma}^{-1}_{\mathbf{x}_{\mathcal{F}}}\mathrm{E}(\mathbf{x}_{\mathcal{F}}\boldsymbol{\gamma}^\top_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}})\}$.

(ii) $\delta^{-\mathcal{G}}_{-\mathcal{F}} = 0$ *under* $\mathcal{H}_0^{\mathcal{F},[\mathcal{G}]} : \mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{\mathcal{F}}$ *and* $\mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{y}_{\mathcal{G}}$.

Let $\{(\mathbf{x}^{(1)}, \mathbf{y}^{(1)}), \ldots, (\mathbf{x}^{(n)}, \mathbf{y}^{(n)})\}$ be an iid sample. Let $\bar{\mathbf{x}} = n^{-1}\sum_{i=1}^n \mathbf{x}^{(i)}$, $\tilde{\mathbf{x}}^{(i)} = \mathbf{x}^{(i)} - \bar{\mathbf{x}}$, and $\hat{\mathbf{\Sigma}}_{\mathbf{x}} = n^{-1}\sum_{i=1}^n \tilde{\mathbf{x}}^{(i)}(\tilde{\mathbf{x}}^{(i)})^\top$. Similarly let $\bar{\mathbf{y}} = n^{-1}\sum_{i=1}^n \mathbf{y}^{(i)}$, $\tilde{\mathbf{y}}^{(i)} = \mathbf{y}^{(i)} - \bar{\mathbf{y}}$, $\hat{\mathbf{\Sigma}}_{\mathbf{y}} = n^{-1}\sum_{i=1}^n \tilde{\mathbf{y}}^{(i)}(\tilde{\mathbf{y}}^{(i)})^\top$, $\mathrm{E}_n(\mathbf{x}\mathbf{y}^\top) = n^{-1}\sum_{i=1}^n \tilde{\mathbf{x}}^{(i)}(\tilde{\mathbf{y}}^{(i)})^\top$, and $\mathrm{E}_n(\mathbf{y}\mathbf{x}^\top) = n^{-1}\sum_{i=1}^n \tilde{\mathbf{y}}^{(i)}(\tilde{\mathbf{x}}^{(i)})^\top$. Then $\hat{\mathbf{M}} = \hat{\mathbf{\Sigma}}_{\mathbf{x}}^{-1/2}\mathrm{E}_n(\mathbf{x}\mathbf{y}^\top)\hat{\mathbf{\Sigma}}_{\mathbf{y}}^{-1}\mathrm{E}_n(\mathbf{y}\mathbf{x}^\top)\hat{\mathbf{\Sigma}}_{\mathbf{x}}^{-1/2}$. Similarly one can calculate

$$\hat{\mathbf{M}}^{\mathcal{G}}_{\mathcal{F}} = \hat{\mathbf{\Sigma}}_{\mathbf{x}_{\mathcal{F}}}^{-1/2}\mathrm{E}_n(\mathbf{x}_{\mathcal{F}}\mathbf{y}^\top_{\mathcal{G}})\hat{\mathbf{\Sigma}}_{\mathbf{y}_{\mathcal{G}}}^{-1}\mathrm{E}_n(\mathbf{y}_{\mathcal{G}}\mathbf{x}^\top_{\mathcal{F}})\hat{\mathbf{\Sigma}}_{\mathbf{x}_{\mathcal{F}}}^{-1/2}.$$

Then the sample version of $\delta^{-\mathcal{G}}_{-\mathcal{F}}$ in (14) becomes

$$\hat{\delta}^{-\mathcal{G}}_{-\mathcal{F}} = \mathrm{tr}(\hat{\mathbf{M}}) - \mathrm{tr}(\hat{\mathbf{M}}^{\mathcal{G}}_{\mathcal{F}}).$$

We conclude this section with the asymptotic distribution of $\hat{\delta}^{-\mathcal{G}}_{-\mathcal{F}}$ under $\mathcal{H}_0^{\mathcal{F},[\mathcal{G}]}$. Assume $|\mathcal{F}| = p_1$ and $|\mathcal{G}| = q_1$, where $|\cdot|$ denotes cardinality.

Theorem 1. *Suppose* $\mathrm{E}(\mathbf{x}) = \mathbf{0}$, $\mathrm{E}(\mathbf{y}) = \mathbf{0}$, $\mathrm{E}(\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}})$ *is a linear function of* $\mathbf{x}_{\mathcal{F}}$ *and* $\mathrm{E}(\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}})$ *is a linear function of* $\mathbf{y}_{\mathcal{G}}$. *Then under* $\mathcal{H}_0^{\mathcal{F},[\mathcal{G}]}$,

$$n\hat{\delta}^{-\mathcal{G}}_{-\mathcal{F}} \rightsquigarrow \sum_{\ell=1}^{L} \tau_\ell \chi^2_\ell(1)$$

as $n \to \infty$, where $\rightsquigarrow$ means convergence in distribution, $L = pq - p_1 q_1$, $\chi^2_\ell(1)$ is independent chi-square with one degree of freedom for all $\ell \in \{1, \ldots, L\}$ and $\tau_1 \geq \cdots \geq \tau_L$ are the eigenvalues of $\mathbf{\Omega}$, and the exact form of $\mathbf{\Omega}$ is provided in the Appendix.

The asymptotic distribution in Theorem 1 needs to be estimated in practice. Specifically, let $\hat{\mathbf{\Omega}}$ be the sample estimators of $\mathbf{\Omega}$, and let $\hat{\tau}_1 \geq \cdots \geq \hat{\tau}_L$ be the eigenvalues of $\hat{\mathbf{\Omega}}$. Denote $\boldsymbol{\zeta} = (\hat{\tau}_1, \ldots, \hat{\tau}_L)^\top \in \mathbb{R}^L$ and let $\mathbf{\Xi} \in \mathbb{R}^{N \times L}$ consist of i.i.d. $\chi^2(1)$ realizations. Then the N elements of $\mathbf{\Xi}\boldsymbol{\zeta}$ become realizations of the approximate asymptotic distribution of $n\hat{\delta}^{-\mathcal{G}}_{-\mathcal{F}}$ under null. The proportion of these N elements greater than $n\hat{\delta}^{-\mathcal{G}}_{-\mathcal{F}}$ become the approximate p-value. For a given significance level α, we reject $\mathcal{H}_0^{\mathcal{F},[\mathcal{G}]}$ if this p-value is smaller than α. We use $N = 500$ in our numerical studies.

4.4. Algorithm for dual model-free variable selection

Let $\{(\mathbf{x}^{(1)}, \mathbf{y}^{(1)}), \ldots, (\mathbf{x}^{(n)}, \mathbf{y}^{(n)})\}$ be an iid sample of $\{\mathbf{x} \in \mathbb{R}^p, \mathbf{y} \in \mathbb{R}^q\}$. We devise a sample-level algorithm for dual model-free variable selection in this section. From the development in Section 3, we have seen that the active sets $\mathcal{A}$ and $\mathcal{B}$ can be recovered by the smallest possible $\mathcal{F}$ and the smallest possible $\mathcal{G}$ such that $\mathcal{H}_0^{\mathcal{F},[\mathcal{G}]}$ is not rejected. This motivates us to consider the following joint backward selection procedure.

1. Initial step. Set $\mathcal{F}^{(0)} = \{1, \ldots, p\}$ and $\mathcal{G}^{(0)} = \{1, \ldots, q\}$. Let α be the pre-specified significance level.
   - 1.1 For each $i \in \{1, \ldots, p\}$, denote $\mathcal{F}^{(0)}_{-i}$ as the index set where i is removed from $\mathcal{F}^{(0)}$, and let $\varrho_{i,(0)}$ be the approximate p-value from testing $\mathcal{H}_0^{\mathcal{F}^{(0)}_{-i},[\mathcal{G}^{(0)}]}$ against its alternative.
   - 1.2 For each $j \in \{1, \ldots, q\}$, denote $\mathcal{G}^{(0)}_{-j}$ as the index set where j is removed from $\mathcal{G}^{(0)}$, and let $\varrho_{p+j,(0)}$ be the approximate p-value from testing $\mathcal{H}_0^{\mathcal{F}^{(0)},[\mathcal{G}^{(0)}_{-j}]}$ against its alternative.
   - 1.3 Let $k_{(0)} = \underset{\iota \in \{1,\ldots,p+q\}}{\arg\max}\, \varrho_{\iota,(0)}$ and $\varrho_{(0)} = \underset{\iota \in \{1,\ldots,p+q\}}{\max} \varrho_{\iota,(0)}$. If $\varrho_{(0)} \geq \alpha$ and $k_{(0)} \leq p$, then set $\mathcal{F}^{(1)} = \mathcal{F}^{(0)}_{-k_{(0)}}$, $\mathcal{G}^{(1)} = \mathcal{G}^{(0)}$, and go to Step 2. If $\varrho_{(0)} \geq \alpha$ and $k_{(0)} > p$, then set $\mathcal{F}^{(1)} = \mathcal{F}^{(0)}$, $\mathcal{G}^{(1)} = \mathcal{G}^{(0)}_{-\{k_{(0)}-p\}}$, and go to Step 2. If $\varrho_{(0)} < \alpha$, then stop the algorithm and return $\hat{\mathcal{A}} = \mathcal{F}^{(0)}$, $\hat{\mathcal{B}} = \mathcal{G}^{(0)}$.
2. Iteration step. At the beginning of the ℓth iteration, let $\mathcal{F}^{(\ell)}$ and $\mathcal{G}^{(\ell)}$ be the working index sets. Assume $|\mathcal{F}^{(\ell)}| = p_{(\ell)}$ and $|\mathcal{G}^{(\ell)}| = q_{(\ell)}$.
   - 2.1 For each $i \in \{1, \ldots, p_{(\ell)}\}$, denote $\mathcal{F}^{(\ell)}_{-i}$ as the index set where the ith element of $\mathcal{F}^{(\ell)}$ is removed from $\mathcal{F}^{(\ell)}$, and let $\varrho_{i,(\ell)}$ be the approximate p-value from testing $\mathcal{H}_0^{\mathcal{F}^{(\ell)}_{-i},[\mathcal{G}^{(\ell)}]}$ against its alternative.
   - 2.2 For each $j \in \{1, \ldots, q_{(\ell)}\}$, denote $\mathcal{G}^{(\ell)}_{-j}$ as the index set where the jth element of $\mathcal{G}^{(\ell)}$ is removed from $\mathcal{G}^{(\ell)}$, and let $\varrho_{p_{(\ell)}+j,(\ell)}$ be the approximate p-value from testing $\mathcal{H}_0^{\mathcal{F}^{(\ell)},[\mathcal{G}^{(\ell)}_{-j}]}$ against its alternative.
   - 2.3 Let $k_{(\ell)} = \underset{\iota \in \{1,\ldots,p_{(\ell)}+q_{(\ell)}\}}{\arg\max}\, \varrho_{\iota,(\ell)}$ and $\varrho_{(\ell)} = \underset{\iota \in \{1,\ldots,p_{(\ell)}+q_{(\ell)}\}}{\max} \varrho_{\iota,(\ell)}$. If $\varrho_{(\ell)} \geq \alpha$ and $k_{(\ell)} \leq p_{(\ell)}$, then set $\mathcal{F}^{(\ell+1)} = \mathcal{F}^{(\ell)}_{-k_{(\ell)}}$, $\mathcal{G}^{(\ell+1)} = \mathcal{G}^{(\ell)}$, and repeat Step 2. If $\varrho_{(\ell)} \geq \alpha$ and $k_{(\ell)} > p_{(\ell)}$, then set $\mathcal{F}^{(\ell+1)} = \mathcal{F}^{(\ell)}$, $\mathcal{G}^{(\ell+1)} = \mathcal{G}^{(\ell)}_{-\{k_{(\ell)}-p_{(\ell)}\}}$, and repeat Step 2. If $\varrho_{(\ell)} < \alpha$, then stop the iteration and return $\hat{\mathcal{A}} = \mathcal{F}^{(\ell)}$, $\hat{\mathcal{B}} = \mathcal{G}^{(\ell)}$.

In the initial step, we first test $\mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{-i}$ against its alternative for each $i \in \{1, \ldots, p\}$. Then we test $\mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{y}_{-j}$ for each $j \in \{1, \ldots, q\}$. The corresponding p-values are denoted as $\varrho_{\iota,(0)}$ for $\iota \in \{1, \ldots, p+q\}$. The maximum p-value $\varrho_{(0)}$ is then compared to α. If $\varrho_{(0)}$ is smaller than α, then $\varrho_{\iota,(0)} < \alpha$ for any $\iota \in \{1, \ldots, p+q\}$. Thus we reject $\mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{-i}$ for any $i \in \{1, \ldots, p\}$, and we reject $\mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{y}_{-j}$ for any $j \in \{1, \ldots, q\}$. Hence we estimate the active sets by $\mathcal{F}^{(0)}$ and $\mathcal{G}^{(0)}$. In the case with $\varrho_{(0)} \geq \alpha$, the least significant element, which is indexed by $k_{(0)}$, can be removed from the active sets. For $k_{(0)} \leq p$, we update $\mathcal{F}$ by removing the least significant element, which corresponds to an element in $\mathbf{x}$. For $k_{(0)} > p$, the least significant element corresponds to an element in $\mathbf{y}$ and we only update $\mathcal{G}$.

In the ℓth iteration, we start from working index sets $\mathcal{F}^{(\ell)}$ and $\mathcal{G}^{(\ell)}$. Note that $\mathcal{H}_0^{\mathcal{F}^{(\ell)},[\mathcal{G}^{(\ell)}]}$ is not rejected from the last iteration, as we only go to the ℓth iteration if $\varrho_{(\ell-1)} \geq \alpha$. In another word, the ℓth iteration is needed only when

Table 1: Frequencies of rejecting $\mathcal{H}_0^{\mathcal{F},[\mathcal{G}]}$ based on 1000 repetitions.

	Model 1			Model 2		
	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
$\mathcal{F} = \{3,4\}, \mathcal{G} = \{1,2\}$	1	1	1	0.007	0.033	0.085
$\mathcal{F} = \{1,2\}, \mathcal{G} = \{3,4\}$	0.007	0.044	0.093	1	1	1

$\mathcal{F}^{(\ell-1)}$ or $\mathcal{G}^{(\ell-1)}$ is updated. Parallel to the initial step, after removing one element at a time from either $\mathcal{F}^{(\ell)}$ or $\mathcal{G}^{(\ell)}$, we test the dual marginal coordinate hypotheses (7) and get p-value $\varrho_{\iota,(\ell)}$ for $\iota \in \{1, \ldots, p_{(\ell)} + q_{(\ell)}\}$. The maximum p-value $\varrho_{(\ell)}$ is then compared to α. If $\varrho_{(\ell)} < \alpha$, then $\mathcal{F}^{(\ell)}$ and $\mathcal{G}^{(\ell)}$ can not be further reduced. We stop the iteration and estimate the active sets by $\mathcal{F}^{(\ell)}$ and $\mathcal{G}^{(\ell)}$. Otherwise we go to the next iteration, where either $\mathcal{F}^{(\ell)}$ or $\mathcal{G}^{(\ell)}$ is updated by removing the least significant element. Note that in each iteration, we are testing the conditional independence between $\mathbf{x}$ and $\mathbf{y}$, and our procedure asymptotically controls the type-I error rate at the significance level α.

5. Numerical results

We use synthetic data in Section 5.1, and a real data analysis is considered in Section 5.2.

5.1. Simulation studies

Let $\mathbf{x} = (X_1, \ldots, X_p)^\top$ be multivariate normal with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}_\mathbf{x} = (\sigma_{ij})$, where $\sigma_{ij} = \sigma^{|i-j|}$ for $1 \le i, j \le p$. Similarly, let $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_q)^\top$ be multivariate normal with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}_{\boldsymbol{\epsilon}} = (\theta_{ij})$, where $\theta_{ij} = \theta^{|i-j|}$ for all $i, j \in \{1, \ldots, q\}$. The error $\boldsymbol{\epsilon}$ is independent of $\mathbf{x}$. Then we generate $\mathbf{y} = (Y_1, \ldots, Y_q)^\top$ from the following two models:

$$\text{Model 1: } Y_1 = \epsilon_1, Y_2 = \epsilon_2, Y_3 = \epsilon_3, Y_4 = X_1 + X_2 + \epsilon_4;$$
$$\text{Model 2: } Y_1 = e^{0.5X_3} + \sin(X_4) + \epsilon_1, Y_2 = X_3^3 + X_4 + \epsilon_2, Y_3 = \epsilon_3, Y_4 = \epsilon_4.$$

In both models, we set $p = 5$ and $q = 4$. In Model 1, we set $\sigma = 0$ and $\theta = 0.5$. The active set for the regression of $\mathbf{y}$ on $\mathbf{x}$ is $\mathcal{A} = \{1, 2\}$. Due to the nonzero correlation among the ϵ's, we cannot determine $\mathcal{B}$ by evaluating the forward regression between $\mathbf{y}$ and $\mathbf{x}$. Instead we calculate $\mathrm{E}(\mathbf{x} \mid \mathbf{y}) = (0, 0, 0, 4Y_4/11 - 2Y_3/11, 4Y_4/11 - 2Y_3/11)^\top$, and thus the active set for the regression of $\mathbf{x}$ on $\mathbf{y}$ is $\mathcal{B} = \{3, 4\}$. More details are provided in the Appendix. In Model 2, we set $\sigma = 0.5$ and $\theta = 0$. We have $\mathcal{A} = \{3, 4\}$ and $\mathcal{B} = \{1, 2\}$ as the result of $\theta = 0$.

First we evaluate the performance of the CCA-based trace test for the dual marginal coordinate hypotheses (7). For user-specified $\mathcal{F}$ and $\mathcal{G}$, we test $\mathcal{H}_0^{\mathcal{F},[\mathcal{G}]} : \mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_\mathcal{F}$ and $\mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{y}_\mathcal{G}$ versus the alternative $\mathcal{H}_a^{\mathcal{F},[\mathcal{G}]} : \mathbf{y} \not\perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_\mathcal{F}$ or $\mathbf{y} \not\perp\!\!\!\perp \mathbf{x} \mid \mathbf{y}_\mathcal{G}$. Based on 1000 repetitions, the frequencies of $\mathcal{H}_0^{\mathcal{F},[\mathcal{G}]}$ being rejected are reported in Table 1. Fix sample size $n = 300$ and take $\alpha \in \{0.01, 0.05, 0.1\}$. We consider two combinations of $\mathcal{F}$ and $\mathcal{G}$. When $\mathcal{F} = \{3, 4\}$ and $\mathcal{G} = \{1, 2\}$, $\mathcal{H}_0^{\mathcal{F},[\mathcal{G}]}$ does not hold for Model 1. We see from Table 1 that $\mathcal{H}_0^{\mathcal{F},[\mathcal{G}]}$ is always rejected regardless of the α value. When $\mathcal{F} = \{1, 2\}$ and $\mathcal{G} = \{3, 4\}$, $\mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_\mathcal{F}$ and $\mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{y}_\mathcal{G}$ for Model 1. We see that the frequencies of rejecting $\mathcal{H}_0^{\mathcal{F},[\mathcal{G}]}$ are close to the corresponding nominal level. The results for Model 2 are reversed. The first combination of $\mathcal{F}$ and $\mathcal{G}$ now corresponds to the Type-I error, and the frequencies are close to the nominal level. The second combination of $\mathcal{F}$ and $\mathcal{G}$ corresponds to the estimated power, which is 1 for all α values.

Next we investigate the performance of the joint backward selection algorithm proposed in Section 4.4. For each $i \in \{1, \ldots, 1000\}$, denote the estimated active sets of the ith repetition as $\hat{\mathcal{A}}_i$ and $\hat{\mathcal{B}}_i$. We define the over-fitted frequency (OF), the correctly-fitted frequency (CF), and the under-fitted frequency (UF) as

$$OF = 1000^{-1} \sum_{i=1}^{1000} \left\{ \mathbf{1}(\mathcal{A} \subseteq \hat{\mathcal{A}}_i)\mathbf{1}(\mathcal{B} \subseteq \hat{\mathcal{B}}_i) - \mathbf{1}(\mathcal{A} = \hat{\mathcal{A}}_i)\mathbf{1}(\mathcal{B} = \hat{\mathcal{B}}_i) \right\}, \quad CF = 1000^{-1} \sum_{i=1}^{1000} \mathbf{1}(\mathcal{A} = \hat{\mathcal{A}}_i)\mathbf{1}(\mathcal{B} = \hat{\mathcal{B}}_i),$$

and $UF = 1 - CF - OF$, where $\mathbf{1}$ denotes the indicator function. The average model size is defined as $MS = 1000^{-1} \sum_{i=1}^{1000} (|\hat{\mathcal{A}}_i| + |\hat{\mathcal{B}}_i|)$. Based on 1000 repetitions, we report UF, CF, OF, MS together with the frequencies of each variable being selected.

Table 2: Variable selection results for Model 1 based on 1000 repetitions.

n	X_1	X_2	X_3	X_4	X_5	Y_1	Y_2	Y_3	Y_4	UF	CF	OF	MS
100	1	1	0.01	0.01	0.02	0.02	0.04	0.15	1	0.85	0.13	0.02	3.25
300	1	1	0.01	0.01	0.01	0.01	0.02	0.9	1	0.1	0.86	0.04	3.96
700	1	1	0.02	0.02	0.02	0.02	0.03	1	1	0	0.94	0.06	4.1

Table 3: Variable selection results for Model 2 based on 1000 repetitions.

α	X_1	X_2	X_3	X_4	X_5	Y_1	Y_2	Y_3	Y_4	UF	CF	OF	MS
0.01	0	0	1	0.99	0	0.99	0.99	0	0	0.03	0.96	0.01	3.97
0.05	0.01	0.01	1	1	0.02	1	1	0.02	0.02	0	0.95	0.05	4.08
0.1	0.02	0.03	1	1	0.03	1	1	0.05	0.03	0	0.9	0.1	4.17

The variable selection results for Model 1 is summarized in Table 2. We fix $\alpha = 0.05$ and take $n \in \{100, 300, 700\}$. We see that the variable selection performance improves as n increases. For $n = 300$ and $n = 700$, the unimportant variables X_3, X_4, X_5, Y_1 and Y_2 are selected with very low frequencies; the important variables X_1, X_2, Y_3 and Y_4 are selected with frequency 1 or frequency close to 1; and the average model size is close to 4. We also note that the frequency of correctly-fitted model becomes close to $1 - \alpha$ with $n = 700$.

Table 3 summarizes the variable selection results for Model 2. We fix $n = 300$ and consider $\alpha \in \{0.01, 0.05, 0.1\}$. Our backward algorithm works well at all nominal levels. The important variables X_3, X_4, Y_1 and Y_2 are selected with high frequencies, the unimportant variables X_1, X_2, X_5, Y_3 and Y_4 are selected with low frequencies, and the average model size is close to 4. As expected, $\alpha = 0.01$ leads to smaller models and $\alpha = 0.1$ tend to result in larger models. Again, the frequency of correctly-fitted model is close to $1 - \alpha$.

5.2. Real data analysis

Beta-carotene and retinol are well studied chemical compounds in the human plasma. Several studies suggest that low levels of both compounds in plasma are associated with increased risk of an array of diseases such as cancer, cardiovascular disease, and cataracts. To determine the role of dietary habits and other health related metrics in plasma concentrations of beta-carotene and retinol, [15] did a cross-sectional study with 12 personal characteristics and dietary metrics for 315 patients with nonmelanoma skin cancer. After removing three categorical variables, we consider $\mathbf{x} = (X_1, \ldots, X_9)^\top$. The response variables are $\mathbf{y} = (Y_1, Y_2)^\top$, where Y_1 is the plasma concentration of beta-carotene and Y_2 is the plasma concentration of retinol. After exploratory data analysis, we remove six observations with extreme values and get $n = 309$. We apply our proposed dual variable selection procedure from Section 4.4 with significance level $\alpha = 0.05$, and end up with $\hat{\mathcal{A}} = \{1, 2, 6, 8\}$ and $\hat{\mathcal{B}} = \{1, 2\}$. This suggests that to further study the multivariate associations between dietary habits and the plasma compound concentrations, we can focus only on six variables X_1, X_2, X_6, X_8, Y_1 and Y_2 instead of the original $\mathbf{x}$ and $\mathbf{y}$.

To demonstrate the effect of variable selection on canonical correlation analysis, we first calculate the first two pairs of canonical covariates (u_1, v_1) and (u_2, v_2) based on the original data, where $\mathbf{x} \in \mathbb{R}^9$ and $\mathbf{y} \in \mathbb{R}^2$. Then we calculate the first two pairs of canonical covariates $(\tilde{u}_1, \tilde{v}_1)$ and $(\tilde{u}_2, \tilde{v}_2)$ based on the reduced data, where $\mathbf{x}_{\hat{\mathcal{A}}} \in \mathbb{R}^4$ and $\mathbf{x}_{\hat{\mathcal{B}}} \in \mathbb{R}^2$. The plots of the canonical covariate from the original data versus the corresponding canonical covariate from the reduced data are provided in Figure 1. The scatterplots are close to the dotted 45 degree line, suggesting that the canonical covariates before and after the data reduction largely agree with each other. This implies that the reduced data keeps the canonical information from the original data.

6. Concluding remarks

In this paper, we propose the CCA-based trace test for the dual marginal coordinate hypotheses and study the asymptotic properties of the resulting test statistic. The validity of the asymptotic test is justified through simulation studies. Based on this novel test, we design a joint backward selection algorithm for dual model-free variable selection.

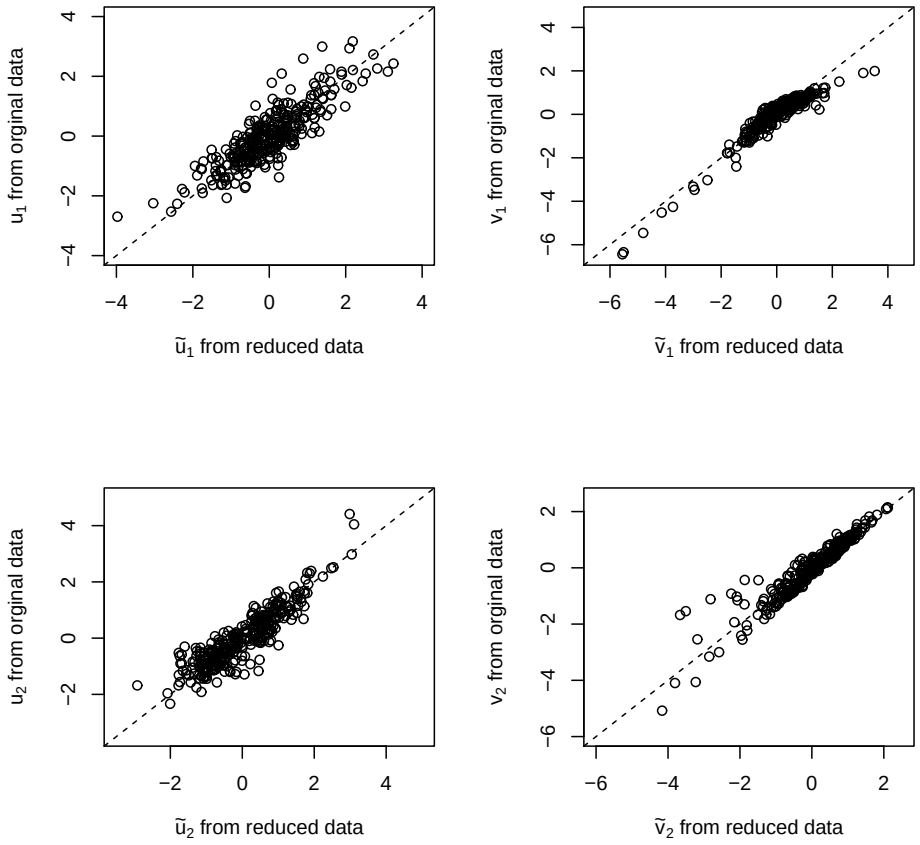

Figure 1. Scatterplots of the canonical covariates from the original data versus the canonical covariate from the reduced data.

The finite-sample performance of the proposed test and the variable selection algorithm are very promising. The dual variable selection and feature screening in the case of diverging p and q is worth future investigation.

Appendix

Proof of Proposition 1. The proof is similar to Proposition 1 in Iaci et al. [8], and is thus omitted. □

Proof of Proposition 2. For part (i), note that $\text{Span}(\mathbf{M}) = \text{Span}\{\text{E}(\mathbf{z}\mathbf{w}^\top)\}$. Plug in $\mathbf{z} = \boldsymbol{\Sigma}_\mathbf{x}^{-1/2}\mathbf{x}$ and $\mathbf{w} = \boldsymbol{\Sigma}_\mathbf{y}^{-1/2}\mathbf{y}$, and all we need to prove becomes

$$\text{Span}\{\boldsymbol{\Sigma}_\mathbf{x}^{-1}\text{E}(\mathbf{x}\mathbf{y}^\top)\} = \text{Span}\{\boldsymbol{\Sigma}_\mathbf{x}^{-1/2}\text{E}(\mathbf{z}\mathbf{w}^\top)\} \subseteq \mathcal{S}_{\mathbf{y}|\mathbf{x}} = \text{Span}(\boldsymbol{\beta}). \tag{A.1}$$

From the law of iterated expectations and the fact that $\mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \boldsymbol{\beta}^\top\mathbf{x}$, we have

$$\text{E}(\mathbf{x}\mathbf{y}^\top) = \text{E}\{\mathbf{x}\text{E}^\top(\mathbf{y}|\mathbf{x})\} = \text{E}\{\mathbf{x}\text{E}^\top(\mathbf{y}|\boldsymbol{\beta}^\top\mathbf{x})\}. \tag{A.2}$$

From the property of conditional expectation and the assumption that $\text{E}(\mathbf{x}|\boldsymbol{\beta}^\top\mathbf{x})$ is linear in $\boldsymbol{\beta}^\top\mathbf{x}$, we have

$$\text{E}\{\mathbf{x}\text{E}^\top(\mathbf{y}|\boldsymbol{\beta}^\top\mathbf{x})\} = \text{E}\{\text{E}(\mathbf{x}|\boldsymbol{\beta}^\top\mathbf{x})\mathbf{y}^\top\} = \boldsymbol{\Sigma}_\mathbf{x}\boldsymbol{\beta}(\boldsymbol{\beta}^\top\boldsymbol{\Sigma}_\mathbf{x}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^\top\text{E}(\mathbf{x}\mathbf{y}^\top). \tag{A.3}$$

(A.2) and (A.3) together lead to

$$\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathrm{E}(\mathbf{x}\mathbf{y}^\top) = \boldsymbol{\beta}(\boldsymbol{\beta}^\top\boldsymbol{\Sigma}_{\mathbf{x}}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^\top\mathrm{E}(\mathbf{x}\mathbf{y}^\top). \tag{A.4}$$

(A.1) follows from (A.4) and proof of part (i) is done. Proof of part (ii) is similar to the proof of part (i), and is thus omitted. □

Proof of Proposition 3. For part (i), assume $\mathbf{x}_{\mathcal{F}} \in \mathbb{R}^{p_1}$ and $\mathbf{x}_{\mathcal{F}^c} \in \mathbb{R}^{p_2}$ with $p_1 + p_2 = p$. Let $\mathbf{x} = (\mathbf{x}_{\mathcal{F}}^\top, \mathbf{x}_{\mathcal{F}^c}^\top)^\top$. Define $\mathbf{C}$ and $\mathbf{D}$ as

$$\mathbf{C} = \begin{pmatrix} \mathrm{I}_{p_1} & \mathbf{0} \\ -\mathrm{E}(\mathbf{x}_{\mathcal{F}^c}\mathbf{x}_{\mathcal{F}}^\top)\boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}}}^{-1} & \mathrm{I}_{p_2} \end{pmatrix} \quad \text{and} \quad \mathbf{D} = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}}} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}} \end{pmatrix}.$$

Then $\mathbf{C}\mathbf{x} = (\mathbf{x}_{\mathcal{F}}^\top, \boldsymbol{\gamma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}^\top)^\top$, $\mathbf{C}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{C}^\top = \mathbf{D}$ and $\boldsymbol{\Sigma}_{\mathbf{x}}^{-1} = \mathbf{C}^\top\mathbf{D}^{-1}\mathbf{C}$. It follows that

$$\begin{aligned}
\mathrm{tr}(\mathbf{M}) &= \mathrm{tr}\left\{\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathrm{E}(\mathbf{x}\mathbf{y}^\top)\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\mathrm{E}(\mathbf{y}\mathbf{x}^\top)\right\} = \mathrm{tr}\left\{\mathbf{C}^\top\mathbf{D}^{-1}\mathbf{C}\mathrm{E}(\mathbf{x}\mathbf{y}^\top)\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\mathrm{E}(\mathbf{y}\mathbf{x}^\top)\right\} \\
&= \mathrm{tr}\left\{\begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}}}^{-1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}^{-1} \end{pmatrix}\begin{pmatrix} \mathrm{E}(\mathbf{x}_{\mathcal{F}}\mathbf{y}^\top) \\ \mathrm{E}(\boldsymbol{\gamma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}\mathbf{y}^\top) \end{pmatrix}\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\left(\mathrm{E}(\mathbf{y}\mathbf{x}_{\mathcal{F}}^\top), \mathrm{E}(\mathbf{y}\boldsymbol{\gamma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}^\top)\right)\right\} \\
&= \mathrm{tr}\left\{\begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}}}^{-1}\mathrm{E}(\mathbf{x}_{\mathcal{F}}\mathbf{y}^\top)\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\mathrm{E}(\mathbf{y}\mathbf{x}_{\mathcal{F}}^\top) & \boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}}}^{-1}\mathrm{E}(\mathbf{x}_{\mathcal{F}}\mathbf{y}^\top)\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\mathrm{E}(\mathbf{y}\boldsymbol{\gamma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}^\top) \\ \boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}^{-1}\mathrm{E}(\boldsymbol{\gamma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}\mathbf{y}^\top)\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\mathrm{E}(\mathbf{y}\mathbf{x}_{\mathcal{F}}^\top) & \boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}^{-1}\mathrm{E}(\boldsymbol{\gamma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}\mathbf{y}^\top)\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\mathrm{E}(\mathbf{y}\boldsymbol{\gamma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}^\top) \end{pmatrix}\right\} \\
&= \mathrm{tr}\{\boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}}}^{-1}\mathrm{E}(\mathbf{x}_{\mathcal{F}}\mathbf{y}^\top)\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\mathrm{E}(\mathbf{y}\mathbf{x}_{\mathcal{F}}^\top)\} + \mathrm{tr}\{\boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}^{-1}\mathrm{E}(\boldsymbol{\gamma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}\mathbf{y}^\top)\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\mathrm{E}(\mathbf{y}\boldsymbol{\gamma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}^\top)\}.
\end{aligned}$$

Together with $\mathrm{tr}(\mathbf{M}_{\mathcal{F}}) = \mathrm{tr}\{\boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}}}^{-1}\mathrm{E}(\mathbf{x}_{\mathcal{F}}\mathbf{y}^\top)\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\mathrm{E}(\mathbf{y}\mathbf{x}_{\mathcal{F}}^\top)\}$, we get

$$\delta_{-\mathcal{F}} = \mathrm{tr}(\mathbf{M}) - \mathrm{tr}(\mathbf{M}_{\mathcal{F}}) = \mathrm{tr}\{\boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}^{-1}\mathrm{E}(\boldsymbol{\gamma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}\mathbf{y}^\top)\}\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\mathrm{E}(\mathbf{y}\boldsymbol{\gamma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}^\top)\}. \tag{A.5}$$

For part (ii), the assumption that $\mathrm{E}(\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}})$ is a linear function of $\mathbf{x}_{\mathcal{F}}$ implies $\mathrm{E}(\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}) = \mathrm{E}(\mathbf{x}_{\mathcal{F}^c}\mathbf{x}_{\mathcal{F}}^\top)\boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}}}^{-1}\mathbf{x}_{\mathcal{F}}$. It follows that

$$\mathrm{E}\{\mathrm{E}(\mathbf{x}_{\mathcal{F}^c}\mathbf{x}_{\mathcal{F}}^\top)\boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}}}^{-1}\mathbf{x}_{\mathcal{F}}\mathbf{y}^\top\} = \mathrm{E}\{\mathrm{E}(\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}})\mathbf{y}^\top\} = \mathrm{E}\{\mathbf{x}_{\mathcal{F}^c}\mathrm{E}^\top(\mathbf{y}|\mathbf{x}_{\mathcal{F}})\}. \tag{A.6}$$

Under $\mathcal{H}_0^{\mathcal{F}} : \mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{\mathcal{F}}$, we have $\mathrm{E}(\mathbf{y}|\mathbf{x}) = \mathrm{E}(\mathbf{y}|\mathbf{x}_{\mathcal{F}})$. Thus

$$\mathrm{E}\{\mathbf{x}_{\mathcal{F}^c}\mathrm{E}^\top(\mathbf{y}|\mathbf{x}_{\mathcal{F}})\} = \mathrm{E}\{\mathbf{x}_{\mathcal{F}^c}\mathrm{E}^\top(\mathbf{y}|\mathbf{x})\} = \mathrm{E}(\mathbf{x}_{\mathcal{F}^c}\mathbf{y}^\top). \tag{A.7}$$

The definition $\boldsymbol{\gamma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}} = \mathbf{x}_{\mathcal{F}^c} - \mathrm{E}(\mathbf{x}_{\mathcal{F}^c}\mathbf{x}_{\mathcal{F}}^\top)\boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}}}^{-1}\mathbf{x}_{\mathcal{F}}$ together with (A.6) and (A.7) leads to $\mathrm{E}(\boldsymbol{\gamma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}\mathbf{y}^\top) = \mathbf{0}$. It follows from part (i) that $\delta_{-\mathcal{F}} = 0$ under $\mathcal{H}_0^{\mathcal{F}}$. □

Proof of Proposition 4. The proof is similar to the proof of Proposition 3, and is thus omitted. □

Proof of Proposition 5. For part (i), assume $\mathbf{y}_{\mathcal{G}} \in \mathbb{R}^{q_1}$ and $\mathbf{y}_{\mathcal{G}^c} \in \mathbb{R}^{q_2}$ with $q_1 + q_2 = q$. Let $\mathbf{y} = (\mathbf{y}_{\mathcal{G}}^\top, \mathbf{y}_{\mathcal{G}^c}^\top)^\top$. Define $\mathbf{K}$ and $\mathbf{O}$ as

$$\mathbf{K} = \begin{pmatrix} \mathrm{I}_{q_1} & \mathbf{0} \\ -\mathrm{E}(\mathbf{y}_{\mathcal{G}^c}\mathbf{y}_{\mathcal{G}}^\top)\boldsymbol{\Sigma}_{\mathbf{y}_{\mathcal{G}}}^{-1} & \mathrm{I}_{q_2} \end{pmatrix} \text{ and } \mathbf{O} = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{y}_{\mathcal{G}}} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}} \end{pmatrix}.$$

Then $\mathbf{K}\mathbf{y} = (\mathbf{y}_{\mathcal{G}}^\top, \boldsymbol{\gamma}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}^\top)^\top$, $\mathbf{K}\boldsymbol{\Sigma}_{\mathbf{y}}\mathbf{K}^\top = \mathbf{O}$ and $\boldsymbol{\Sigma}_{\mathbf{y}}^{-1} = \mathbf{K}^\top\mathbf{O}^{-1}\mathbf{K}$. Thus

$$\begin{aligned}
\mathrm{tr}(\mathbf{M}_{\mathcal{F}}) &= \mathrm{tr}\left\{\boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}}}^{-1}\mathrm{E}(\mathbf{x}_{\mathcal{F}}\mathbf{y}^\top)\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\mathrm{E}(\mathbf{y}\mathbf{x}_{\mathcal{F}}^\top)\right\} = \mathrm{tr}\left\{\boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}}}^{-1}\mathrm{E}(\mathbf{x}\mathbf{y}^\top)\mathbf{K}^\top\mathbf{O}^{-1}\mathbf{K}\mathrm{E}(\mathbf{y}\mathbf{x}^\top)\right\} \\
&= \mathrm{tr}\left\{\boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}}}^{-1}\left(\mathrm{E}(\mathbf{x}_{\mathcal{F}}\mathbf{y}_{\mathcal{G}}^\top), \mathrm{E}(\mathbf{x}_{\mathcal{F}}\boldsymbol{\gamma}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}^\top)\right)\begin{pmatrix}\boldsymbol{\Sigma}_{\mathbf{y}_{\mathcal{G}}}^{-1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}^{-1}\end{pmatrix}\begin{pmatrix}\mathrm{E}(\mathbf{y}_{\mathcal{G}}\mathbf{x}_{\mathcal{F}}^\top) \\ \mathrm{E}(\boldsymbol{\gamma}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}\mathbf{x}_{\mathcal{F}}^\top)\end{pmatrix}\right\} \\
&= \mathrm{tr}\{\boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}}}^{-1}\mathrm{E}(\mathbf{x}_{\mathcal{F}}\mathbf{y}_{\mathcal{G}}^\top)\boldsymbol{\Sigma}_{\mathbf{y}_{\mathcal{G}}}^{-1}\mathrm{E}(\mathbf{y}_{\mathcal{G}}\mathbf{x}_{\mathcal{F}}^\top)\} + \mathrm{tr}\{\boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}}}^{-1}\mathrm{E}(\mathbf{x}_{\mathcal{F}}\boldsymbol{\gamma}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}^\top)\boldsymbol{\Sigma}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}^{-1}\mathrm{E}(\boldsymbol{\gamma}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}\mathbf{x}_{\mathcal{F}}^\top)\} \\
&= \mathrm{tr}(\mathbf{M}_{\mathcal{F}}^{\mathcal{G}}) + \mathrm{tr}\{\boldsymbol{\Sigma}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}^{-1}\mathrm{E}(\boldsymbol{\gamma}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}\mathbf{x}_{\mathcal{F}}^\top)\boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}}}^{-1}\mathrm{E}(\mathbf{x}_{\mathcal{F}}\boldsymbol{\gamma}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}^\top)\}.
\end{aligned}$$

Together with (A.5) from the proof of Proposition 3, we get the desired result in part (i).

For part (ii), we have seen that $\mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{\mathcal{F}}$ leads to $\mathrm{E}(\boldsymbol{\gamma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}\mathbf{y}^\top) = \mathbf{0}$ from the proof of Proposition 3. Following similar steps, we can show that $\mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{y}_{\mathcal{G}}$ leads to $\mathrm{E}(\boldsymbol{\gamma}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}\mathbf{x}_{\mathcal{F}}^\top) = \mathbf{0}$. It follows from part (i) that $\delta_{-\mathcal{F}}^{-\mathcal{G}} = 0$ under $\mathcal{H}_0^{\mathcal{F},[\mathcal{G}]} : \mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{\mathcal{F}}$ and $\mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{y}_{\mathcal{G}}$. □

We need Lemma 1 and Lemma 2 before we prove Theorem 1. Let $\bar{\mathbf{x}}_{\mathcal{F}} = n^{-1}\sum_{i=1}^n \mathbf{x}_{\mathcal{F}}^{(i)}$, $\tilde{\mathbf{x}}_{\mathcal{F}}^{(i)} = \mathbf{x}_{\mathcal{F}}^{(i)} - \bar{\mathbf{x}}_{\mathcal{F}}$, and $\hat{\boldsymbol{\Sigma}}_{\mathbf{x}_{\mathcal{F}}} = n^{-1}\sum_{i=1}^n \tilde{\mathbf{x}}_{\mathcal{F}}^{(i)}(\tilde{\mathbf{x}}_{\mathcal{F}}^{(i)})^\top$. Similarly we define $\tilde{\mathbf{y}}^{(i)}$ and $\tilde{\mathbf{x}}_{\mathcal{F}^c}^{(i)}$. Let $\mathrm{E}_n(\mathbf{x}_{\mathcal{F}}\mathbf{x}_{\mathcal{F}^c}^\top) = n^{-1}\sum_{i=1}^n \tilde{\mathbf{x}}_{\mathcal{F}}^{(i)}(\tilde{\mathbf{x}}_{\mathcal{F}^c}^{(i)})^\top$, $\hat{\boldsymbol{\gamma}}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}^{(i)} = \tilde{\mathbf{x}}_{\mathcal{F}^c}^{(i)} - \mathrm{E}_n(\mathbf{x}_{\mathcal{F}^c}\mathbf{x}_{\mathcal{F}}^\top)\hat{\boldsymbol{\Sigma}}_{\mathbf{x}_{\mathcal{F}}}^{-1}\tilde{\mathbf{x}}_{\mathcal{F}}^{(i)}$, and $\mathrm{E}_n(\mathbf{y}\hat{\boldsymbol{\gamma}}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}^\top) = n^{-1}\sum_{i=1}^n \tilde{\mathbf{y}}^{(i)}(\hat{\boldsymbol{\gamma}}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}^{(i)})^\top$. Define

$$\boldsymbol{\Pi}^{(i)} = \left\{\tilde{\mathbf{y}}^{(i)} - \mathrm{E}(\mathbf{y}\mathbf{x}_{\mathcal{F}}^\top)\boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}}}^{-1}\tilde{\mathbf{x}}_{\mathcal{F}}^{(i)}\right\}\left\{(\tilde{\mathbf{x}}_{\mathcal{F}^c}^{(i)})^\top - (\tilde{\mathbf{x}}_{\mathcal{F}}^{(i)})^\top\boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}}}^{-1}\mathrm{E}(\mathbf{x}_{\mathcal{F}}\mathbf{x}_{\mathcal{F}^c}^\top)\right\}$$

and we have the following result.

Lemma 1. *Suppose* $\mathrm{E}(\mathbf{x}) = \mathbf{0}$, $\mathrm{E}(\mathbf{y}) = \mathbf{0}$, *and* $\mathrm{E}(\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}})$ *is a linear function of* $\mathbf{x}_{\mathcal{F}}$. *If* $\mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{\mathcal{F}}$, *then*

$$\mathrm{E}_n(\mathbf{y}\hat{\boldsymbol{\gamma}}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}^\top) = \frac{1}{n}\sum_{i=1}^n \boldsymbol{\Pi}^{(i)} + O_p(n^{-1}),$$

where the first term on the right-hand side is of order $O_p(n^{-1/2})$.

Proof of Lemma 1. From the definition of $\hat{\boldsymbol{\gamma}}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}$ and $\boldsymbol{\gamma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}$, we have

$$\mathrm{E}_n(\mathbf{y}\hat{\boldsymbol{\gamma}}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}^\top) - \mathrm{E}(\mathbf{y}\boldsymbol{\gamma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}^\top) = \{\mathrm{E}_n(\mathbf{y}\mathbf{x}_{\mathcal{F}^c}^\top) - \mathrm{E}(\mathbf{y}\mathbf{x}_{\mathcal{F}^c}^\top)\} - [\mathrm{E}_n\{\mathbf{y}\mathbf{x}_{\mathcal{F}}^\top\hat{\boldsymbol{\Sigma}}_{\mathbf{x}_{\mathcal{F}}}^{-1}\mathrm{E}_n(\mathbf{x}_{\mathcal{F}}\mathbf{x}_{\mathcal{F}^c}^\top)\} - \mathrm{E}\{\mathbf{y}\mathbf{x}_{\mathcal{F}}^\top\boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}}}^{-1}\mathrm{E}(\mathbf{x}_{\mathcal{F}}\mathbf{x}_{\mathcal{F}^c}^\top)\}]. \tag{A.8}$$

Because $\mathrm{E}(\mathbf{x}) = \mathbf{0}$ and $\mathrm{E}(\mathbf{y}) = \mathbf{0}$, it can be shown that

$$\mathrm{E}_n(\mathbf{y}\mathbf{x}_{\mathcal{F}^c}^\top) - \mathrm{E}(\mathbf{y}\mathbf{x}_{\mathcal{F}^c}^\top) = \frac{1}{n}\sum_{i=1}^n\{\tilde{\mathbf{y}}^{(i)}(\tilde{\mathbf{x}}_{\mathcal{F}^c}^{(i)})^\top - \mathrm{E}(\mathbf{y}\mathbf{x}_{\mathcal{F}^c}^\top)\} + O_p(n^{-1}). \tag{A.9}$$

The asymptotic expansions of $\hat{\boldsymbol{\Sigma}}_{\mathbf{x}_{\mathcal{F}}}^{-1}$ and $\mathrm{E}_n(\mathbf{x}_{\mathcal{F}}\mathbf{x}_{\mathcal{F}^c}^\top)$ are, respectively,

$$\hat{\boldsymbol{\Sigma}}_{\mathbf{x}_{\mathcal{F}}}^{-1} - \boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}}}^{-1} = -\boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}}}^{-1}\left[\frac{1}{n}\sum_{i=1}^n\left\{\tilde{\mathbf{x}}_{\mathcal{F}}^{(i)}(\tilde{\mathbf{x}}_{\mathcal{F}}^{(i)})^\top - \boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}}}\right\}\right]\boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}}}^{-1} + O_p(n^{-1}) \text{ and} \tag{A.10}$$

$$\mathrm{E}_n(\mathbf{x}_{\mathcal{F}}\mathbf{x}_{\mathcal{F}^c}^\top) - \mathrm{E}(\mathbf{x}_{\mathcal{F}}\mathbf{x}_{\mathcal{F}^c}^\top) = \frac{1}{n}\sum_{i=1}^n\left\{\tilde{\mathbf{x}}_{\mathcal{F}}^{(i)}(\tilde{\mathbf{x}}_{\mathcal{F}^c}^{(i)})^\top - \mathrm{E}(\mathbf{x}_{\mathcal{F}}\mathbf{x}_{\mathcal{F}^c}^\top)\right\} + O_p(n^{-1}). \tag{A.11}$$

Eqs. (A.10) and (A.11) together lead to

$$\hat{\boldsymbol{\Sigma}}_{\mathbf{x}_\mathcal{F}}^{-1}\mathrm{E}_n(\mathbf{x}_\mathcal{F}\mathbf{x}_{\mathcal{F}^c}^\top) - \boldsymbol{\Sigma}_{\mathbf{x}_\mathcal{F}}^{-1}\mathrm{E}(\mathbf{x}_\mathcal{F}\mathbf{x}_{\mathcal{F}^c}^\top) = \frac{1}{n}\sum_{i=1}^n \boldsymbol{\Sigma}_{\mathbf{x}_\mathcal{F}}^{-1}\tilde{\mathbf{x}}_\mathcal{F}^{(i)}\left\{(\tilde{\mathbf{x}}_{\mathcal{F}^c}^{(i)})^\top - (\tilde{\mathbf{x}}_\mathcal{F}^{(i)})^\top\boldsymbol{\Sigma}_{\mathbf{x}_\mathcal{F}}^{-1}\mathrm{E}(\mathbf{x}_\mathcal{F}\mathbf{x}_{\mathcal{F}^c}^\top)\right\} + O_p(n^{-1}). \tag{A.12}$$

It follows from (A.12) that

$$\begin{aligned}\mathrm{E}_n\{\mathbf{y}\mathbf{x}_\mathcal{F}^\top\hat{\boldsymbol{\Sigma}}_{\mathbf{x}_\mathcal{F}}^{-1}\mathrm{E}_n(\mathbf{x}_\mathcal{F}\mathbf{x}_{\mathcal{F}^c}^\top)\} - \mathrm{E}\{\mathbf{y}\mathbf{x}_\mathcal{F}^\top\boldsymbol{\Sigma}_{\mathbf{x}_\mathcal{F}}^{-1}\mathrm{E}(\mathbf{x}_\mathcal{F}\mathbf{x}_{\mathcal{F}^c}^\top)\} &= \frac{1}{n}\sum_{i=1}^n \mathrm{E}(\mathbf{y}\mathbf{x}_\mathcal{F}^\top)\boldsymbol{\Sigma}_{\mathbf{x}_\mathcal{F}}^{-1}\tilde{\mathbf{x}}_\mathcal{F}^{(i)}\left\{(\tilde{\mathbf{x}}_{\mathcal{F}^c}^{(i)})^\top - (\tilde{\mathbf{x}}_\mathcal{F}^{(i)})^\top\boldsymbol{\Sigma}_{\mathbf{x}_\mathcal{F}}^{-1}\mathrm{E}(\mathbf{x}_\mathcal{F}\mathbf{x}_{\mathcal{F}^c}^\top)\right\} \\ &\quad + \frac{1}{n}\sum_{i=1}^n\{\tilde{\mathbf{y}}^{(i)}(\tilde{\mathbf{x}}_\mathcal{F}^{(i)})^\top - \mathrm{E}(\mathbf{y}\mathbf{x}_{\mathcal{F}^c}^\top)\}\boldsymbol{\Sigma}_{\mathbf{x}_\mathcal{F}}^{-1}\mathrm{E}(\mathbf{x}_\mathcal{F}\mathbf{x}_{\mathcal{F}^c}^\top) + O_p(n^{-1}).\end{aligned} \tag{A.13}$$

Eqs. (A.8), (A.9) and (A.13) together lead to

$$\begin{aligned}\mathrm{E}_n(\mathbf{y}\hat{\boldsymbol{\gamma}}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_\mathcal{F}}^\top) = \mathrm{E}_n(\mathbf{y}\mathbf{x}_{\mathcal{F}^c}^\top) - \mathrm{E}_n\{\mathbf{y}\mathbf{x}_\mathcal{F}^\top\hat{\boldsymbol{\Sigma}}_{\mathbf{x}_\mathcal{F}}^{-1}\mathrm{E}_n(\mathbf{x}_\mathcal{F}\mathbf{x}_{\mathcal{F}^c}^\top)\} &= \frac{1}{n}\sum_{i=1}^n\Big[\tilde{\mathbf{y}}^{(i)}(\tilde{\mathbf{x}}_{\mathcal{F}^c}^{(i)})^\top - \tilde{\mathbf{y}}^{(i)}(\tilde{\mathbf{x}}_\mathcal{F}^{(i)})^\top\boldsymbol{\Sigma}_{\mathbf{x}_\mathcal{F}}^{-1}\mathrm{E}(\mathbf{x}_\mathcal{F}\mathbf{x}_{\mathcal{F}^c}^\top) \\ &\quad -\mathrm{E}(\mathbf{y}\mathbf{x}_\mathcal{F}^\top)\boldsymbol{\Sigma}_{\mathbf{x}_\mathcal{F}}^{-1}\tilde{\mathbf{x}}_\mathcal{F}^{(i)}\left\{(\tilde{\mathbf{x}}_{\mathcal{F}^c}^{(i)})^\top - (\tilde{\mathbf{x}}_\mathcal{F}^{(i)})^\top\boldsymbol{\Sigma}_{\mathbf{x}_\mathcal{F}}^{-1}\mathrm{E}(\mathbf{x}_\mathcal{F}\mathbf{x}_{\mathcal{F}^c}^\top)\right\}\Big] + O_p(n^{-1}) \\ &= \frac{1}{n}\sum_{i=1}^n\left\{\tilde{\mathbf{y}}^{(i)} - \mathrm{E}(\mathbf{y}\mathbf{x}_\mathcal{F}^\top)\boldsymbol{\Sigma}_{\mathbf{x}_\mathcal{F}}^{-1}\tilde{\mathbf{x}}_\mathcal{F}^{(i)}\right\}\left\{(\tilde{\mathbf{x}}_{\mathcal{F}^c}^{(i)})^\top - (\tilde{\mathbf{x}}_\mathcal{F}^{(i)})^\top\boldsymbol{\Sigma}_{\mathbf{x}_\mathcal{F}}^{-1}\mathrm{E}(\mathbf{x}_\mathcal{F}\mathbf{x}_{\mathcal{F}^c}^\top)\right\} + O_p(n^{-1}),\end{aligned} \tag{A.14}$$

which is the desired result. □

Similarly, let $\mathrm{E}_n(\mathbf{y}_\mathcal{G}\mathbf{y}_{\mathcal{G}^c}^\top) = n^{-1}\sum_{i=1}^n\tilde{\mathbf{y}}_\mathcal{G}^{(i)}(\tilde{\mathbf{y}}_{\mathcal{G}^c}^{(i)})^\top$, $\hat{\boldsymbol{\gamma}}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_\mathcal{G}}^{(i)} = \tilde{\mathbf{y}}_{\mathcal{G}^c}^{(i)} - \mathrm{E}_n(\mathbf{y}_{\mathcal{G}^c}\mathbf{y}_\mathcal{G}^\top)\hat{\boldsymbol{\Sigma}}_{\mathbf{y}_\mathcal{G}}^{-1}\tilde{\mathbf{y}}_\mathcal{G}^{(i)}$, and $\mathrm{E}_n(\mathbf{x}_\mathcal{F}\hat{\boldsymbol{\gamma}}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_\mathcal{G}}^\top) = n^{-1}\sum_{i=1}^n\tilde{\mathbf{x}}_\mathcal{F}^{(i)}(\hat{\boldsymbol{\gamma}}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_\mathcal{G}}^{(i)})^\top$. Define

$$\boldsymbol{\Lambda}^{(i)} = \left\{\tilde{\mathbf{x}}_\mathcal{F}^{(i)} - \mathrm{E}(\mathbf{x}_\mathcal{F}\mathbf{y}_\mathcal{G}^\top)\boldsymbol{\Sigma}_{\mathbf{y}_\mathcal{G}}^{-1}\tilde{\mathbf{y}}_\mathcal{G}^{(i)}\right\}\left\{(\tilde{\mathbf{y}}_{\mathcal{G}^c}^{(i)})^\top - (\tilde{\mathbf{y}}_\mathcal{G}^{(i)})^\top\boldsymbol{\Sigma}_{\mathbf{y}_\mathcal{G}}^{-1}\mathrm{E}(\mathbf{y}_\mathcal{G}\mathbf{y}_{\mathcal{G}^c}^\top)\right\}$$

and we have

Lemma 2. *Suppose* $\mathrm{E}(\mathbf{x}) = \mathbf{0}$, $\mathrm{E}(\mathbf{y}) = \mathbf{0}$, *and* $\mathrm{E}(\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_\mathcal{G})$ *is a linear function of* $\mathbf{y}_\mathcal{G}$. *If* $\mathbf{y}\perp\!\!\!\perp\mathbf{x} \mid \mathbf{y}_\mathcal{G}$, *then*

$$\mathrm{E}_n(\mathbf{x}_\mathcal{F}\hat{\boldsymbol{\gamma}}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_\mathcal{G}}^\top) = \frac{1}{n}\sum_{i=1}^n\boldsymbol{\Lambda}^{(i)} + O_p(n^{-1}),$$

where the first term on the right-hand side is of order $O_p(n^{-1/2})$.

Proof of Lemma 2. The proof is similar to the proof of Lemma 1, and is thus omitted. □

Proof of Theorem 1. Recall that $|\mathcal{F}| = p_1$, $|\mathcal{F}^c| = p_2$, $|\mathcal{G}| = q_1$, $|\mathcal{G}^c| = q_2$, $p_1 + p_2 = p$ and $q_1 + q_2 = q$. Let $\boldsymbol{\phi}_1 = \mathrm{vec}\{\boldsymbol{\Sigma}_\mathbf{y}^{-1/2}\mathrm{E}(\mathbf{y}\boldsymbol{\gamma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_\mathcal{F}}^\top)\boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_\mathcal{F}}^{-1/2}\} \in \mathbb{R}^{qp_2}$, $\boldsymbol{\phi}_2 = \mathrm{vec}\{\boldsymbol{\Sigma}_{\mathbf{x}_\mathcal{F}}^{-1/2}\mathrm{E}(\mathbf{x}_\mathcal{F}\boldsymbol{\gamma}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_\mathcal{G}}^\top)\boldsymbol{\Sigma}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_\mathcal{G}}^{-1/2}\} \in \mathbb{R}^{p_1q_2}$, and $\boldsymbol{\psi} = (\boldsymbol{\phi}_1^\top, \boldsymbol{\phi}_2^\top)^\top$, where vec denotes vectorization. Then we have $\delta_{-\mathcal{F}}^{-\mathcal{G}} = \boldsymbol{\psi}^\top\boldsymbol{\psi}$. At the sample level, let $\hat{\boldsymbol{\psi}} = (\hat{\boldsymbol{\phi}}_1^\top, \hat{\boldsymbol{\phi}}_2^\top)^\top$, where $\hat{\boldsymbol{\phi}}_1 = \mathrm{vec}\{\hat{\boldsymbol{\Sigma}}_\mathbf{y}^{-1/2}\mathrm{E}_n(\mathbf{y}\hat{\boldsymbol{\gamma}}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_\mathcal{F}}^\top)\hat{\boldsymbol{\Sigma}}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_\mathcal{F}}^{-1/2}\}$ and $\hat{\boldsymbol{\phi}}_2 = \mathrm{vec}\{\hat{\boldsymbol{\Sigma}}_{\mathbf{x}_\mathcal{F}}^{-1/2}\mathrm{E}_n(\mathbf{x}_\mathcal{F}\hat{\boldsymbol{\gamma}}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_\mathcal{G}}^\top)\hat{\boldsymbol{\Sigma}}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_\mathcal{G}}^{-1/2}\}$. Then we have

$$\hat{\delta}_{-\mathcal{F}}^{-\mathcal{G}} = \hat{\boldsymbol{\psi}}^\top\hat{\boldsymbol{\psi}}. \tag{A.15}$$

Under $\mathcal{H}_0^{\mathcal{F},[\mathcal{G}]}$, we have $\mathbf{y}\perp\!\!\!\perp\mathbf{x} \mid \mathbf{x}_\mathcal{F}$. It follows that $\mathrm{E}(\mathbf{y}\boldsymbol{\gamma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_\mathcal{F}}^\top) = \mathbf{0}$ and $\boldsymbol{\phi}_1 = \mathbf{0}$. Together with Lemma 1, we have

$$\hat{\boldsymbol{\phi}}_1 = \frac{1}{n}\sum_{i=1}^n\mathrm{vec}\{\boldsymbol{\Sigma}_\mathbf{y}^{-1/2}\boldsymbol{\Pi}^{(i)}\boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_\mathcal{F}}^{-1/2}\} + O_p(n^{-1}), \tag{A.16}$$

where the first term on the right-hand side is of order $O_p(n^{-1/2})$. Similarly, we have $\mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{y}_{\mathcal{G}}$ under $\mathcal{H}_0^{\mathcal{F},[\mathcal{G}]}$. It follows that $\mathrm{E}(\mathbf{x}_{\mathcal{F}} \boldsymbol{\gamma}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}^{\top}) = \mathbf{0}$ and $\boldsymbol{\phi}_2 = \mathbf{0}$. Together with Lemma 2, we have

$$\hat{\boldsymbol{\phi}}_2 = \frac{1}{n} \sum_{i=1}^{n} \mathrm{vec}\{\boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}}}^{-1/2} \boldsymbol{\Lambda}^{(i)} \boldsymbol{\Sigma}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}^{-1/2}\} + O_p(n^{-1}), \tag{A.17}$$

where the first term on the right-hand side is of order $O_p(n^{-1/2})$. It follows from (A.16) and (A.17) that

$$\hat{\boldsymbol{\psi}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\vartheta}^{(i)} + O_p(n^{-1}), \tag{A.18}$$

where

$$\boldsymbol{\vartheta}^{(i)} = \left\{ \mathrm{vec}^{\top}(\boldsymbol{\Sigma}_{\mathbf{y}}^{-1/2} \boldsymbol{\Pi}^{(i)} \boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}^{-1/2}), \mathrm{vec}^{\top}(\boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}}}^{-1/2} \boldsymbol{\Lambda}^{(i)} \boldsymbol{\Sigma}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}^{-1/2}) \right\}^{\top} \in \mathbb{R}^{L}$$

with $\mathrm{E}(\boldsymbol{\vartheta}^{(i)}) = \mathbf{0}$ and $L = qp_2 + p_1q_2 = pq - p_1q_1$. As a result of (A.18), we have

$$\sqrt{n}\, \hat{\boldsymbol{\psi}} \rightsquigarrow N(\mathbf{0}, \boldsymbol{\Omega}) \tag{A.19}$$

as $n \to \infty$, where $\boldsymbol{\Omega} = \mathrm{E}\{\boldsymbol{\vartheta}^{(i)} (\boldsymbol{\vartheta}^{(i)})^{\top}\}$. Eqs. (A.19) and (A.15) lead to the desired result. □

As a special case, $\delta_{-\mathcal{F}}^{-\mathcal{G}}$ reduces to $\delta_{-\mathcal{F}}$ when we set $\mathcal{G} = \mathcal{I}_{\mathbf{y}}$. Then $\delta_{-\mathcal{F}} = \boldsymbol{\phi}_1^{\top} \boldsymbol{\phi}_1$ and $\hat{\delta}_{-\mathcal{F}} = \hat{\boldsymbol{\phi}}_1^{\top} \hat{\boldsymbol{\phi}}_1$. It follows from (A.16) that $\hat{\boldsymbol{\phi}}_1 = n^{-1} \sum_{i=1}^{n} \boldsymbol{\vartheta}_1^{(i)} + O_p(n^{-1})$, where $\boldsymbol{\vartheta}_1^{(i)} = \mathrm{vec}\{\boldsymbol{\Sigma}_{\mathbf{y}}^{-1/2} \boldsymbol{\Pi}^{(i)} \boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}^{-1/2}\} \in \mathbb{R}^{p_2 q}$. Thus $\sqrt{n}\, \hat{\boldsymbol{\phi}}_1 \rightsquigarrow \mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma})$, where $\boldsymbol{\Gamma} = \mathrm{E}\{\boldsymbol{\vartheta}_1^{(i)} (\boldsymbol{\vartheta}_1^{(i)})^{\top}\}$. The asymptotic distribution of $\hat{\delta}_{-\mathcal{F}}$ is summarized in the next result.

Corollary 1. *Suppose* $\mathrm{E}(\mathbf{x}) = \mathbf{0}$, $\mathrm{E}(\mathbf{y}) = \mathbf{0}$, *and* $\mathrm{E}(\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}})$ *is a linear function of* $\mathbf{x}_{\mathcal{F}}$. *Then under* $\mathcal{H}_0^{\mathcal{F}} : \mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{\mathcal{F}}$,

$$n\hat{\delta}_{-\mathcal{F}} \rightsquigarrow \sum_{\ell=1}^{p_2 q} \rho_\ell \chi_\ell^2(1)$$

as $n \to \infty$, *where* $\chi_\ell^2(1)$ *is independent chi-square with one degree of freedom for* $\ell \in \{1, \ldots, p_2 q\}$, *and* $\rho_1 \geq \cdots \geq \rho_{p_2 q}$ *are the eigenvalues of* $\boldsymbol{\Gamma}$.

Similarly, $\delta_{-\mathcal{F}}^{-\mathcal{G}}$ becomes $\delta^{-\mathcal{G}}$ when we set $\mathcal{F} = \mathcal{I}_{\mathbf{x}}$. Note that $\delta^{-\mathcal{G}} = \boldsymbol{\phi}_3^{\top} \boldsymbol{\phi}_3$ with $\boldsymbol{\phi}_3 = \mathrm{vec}\{\boldsymbol{\Sigma}_{\mathbf{x}}^{-1/2} \mathrm{E}(\mathbf{x} \boldsymbol{\gamma}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}^{\top}) \boldsymbol{\Sigma}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}^{-1/2}\} \in \mathbb{R}^{pq_2}$, and $\hat{\delta}^{-\mathcal{G}} = \hat{\boldsymbol{\phi}}_3^{\top} \hat{\boldsymbol{\phi}}_3$ with $\hat{\boldsymbol{\phi}}_3 = \mathrm{vec}\{\hat{\boldsymbol{\Sigma}}_{\mathbf{x}}^{-1/2} \mathrm{E}_n(\mathbf{x} \hat{\boldsymbol{\gamma}}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}^{\top}) \hat{\boldsymbol{\Sigma}}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}^{-1/2}\}$. Similar to (A.17), it can be shown that $\hat{\boldsymbol{\phi}}_3 = n^{-1} \sum_{i=1}^{n} \boldsymbol{\vartheta}_3^{(i)} + O_p(n^{-1})$, where $\boldsymbol{\vartheta}_3^{(i)} = \mathrm{vec}\{\boldsymbol{\Sigma}_{\mathbf{x}}^{-1/2} \boldsymbol{\Lambda}^{(i)} \boldsymbol{\Sigma}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}^{-1/2}\}$. Thus $\sqrt{n}\, \hat{\boldsymbol{\phi}}_3 \rightsquigarrow \mathcal{N}(\mathbf{0}, \boldsymbol{\Upsilon})$, where $\boldsymbol{\Upsilon} = \mathrm{E}\{\boldsymbol{\vartheta}_3^{(i)} (\boldsymbol{\vartheta}_3^{(i)})^{\top}\}$. The asymptotic distribution of $\hat{\delta}^{-\mathcal{G}}$ is summarized in the next result.

Corollary 2. *Suppose* $\mathrm{E}(\mathbf{x}) = \mathbf{0}$, $\mathrm{E}(\mathbf{y}) = \mathbf{0}$, *and* $\mathrm{E}(\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}})$ *is a linear function of* $\mathbf{y}_{\mathcal{G}}$. *Then under* $\mathcal{H}_0^{[\mathcal{G}]} : \mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{y}_{\mathcal{G}}$,

$$n\hat{\delta}^{-\mathcal{G}} \rightsquigarrow \sum_{\ell=1}^{pq_2} \omega_\ell \chi_\ell^2(1)$$

as $n \to \infty$, *where* $\chi_\ell^2(1)$ *is independent chi-square with one degree of freedom for* $\ell \in \{1, \ldots, pq_2\}$, *and* $\omega_1 \geq \cdots \geq \omega_{pq_2}$ *are the eigenvalues of* $\boldsymbol{\Upsilon}$.

Derivation of $\mathcal{B}$ *for Model 1.* Let

$$\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = \begin{bmatrix} 1 & .5 & .25 & .125 \\ .5 & 1 & .5 & .25 \\ .25 & .5 & 1 & .5 \\ .125 & .25 & .5 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{H} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \end{bmatrix}.$$

Because $\mathbf{y} = \mathbf{H}\mathbf{x} + \boldsymbol{\epsilon}$, we have

$$\boldsymbol{\Sigma}_{\mathbf{y}} = \mathbf{H}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{H}^{\top} + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = \begin{bmatrix} 1 & .5 & .25 & .125 \\ .5 & 1 & .5 & .25 \\ .25 & .5 & 1 & .5 \\ .125 & .25 & .5 & 3 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma}_{\mathbf{y}}^{-1} = \begin{bmatrix} 4/3 & -2/3 & 0 & 0 \\ -2/3 & 5/3 & -2/3 & 0 \\ 0 & -2/3 & 47/33 & -2/11 \\ 0 & 0 & -2/11 & 4/11 \end{bmatrix}.$$

It follows that

$$\mathrm{E}(\mathbf{x}|\mathbf{y}) = \boldsymbol{\Sigma}_{\mathbf{xy}}\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\mathbf{y} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 4/3 & -2/3 & 0 & 0 \\ -2/3 & 5/3 & -2/3 & 0 \\ 0 & -2/3 & 47/33 & -2/11 \\ 0 & 0 & -2/11 & 4/11 \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{bmatrix} = \frac{2}{11} \begin{bmatrix} 2Y_4 - Y_3 \\ 2Y_4 - Y_3 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

Thus we have $\mathcal{B} = \{3, 4\}$ for Model 1. □

References

[1] E. Candès, T. Tao, The Dantzig selector: Statistical estimation when p is much larger than n, Ann. Statist. 35 (2007) 2313–2351.
[2] R.D. Cook, Regression Graphics: Ideas for Studying Regressions through Graphics, Wiley, New York, 1998.
[3] R.D. Cook, Testing predictor contributions in sufficient dimension reduction, Ann. Statist. 32 (2004) 1062–1092.
[4] R.D. Cook, S. Weisberg, Discussion of "sliced inverse regression for dimension reduction" by K.C. Li, J. Amer. Statist. Assoc. 86 (2004) 328–332.
[5] J. Fan, R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, J. Amer. Statist. Assoc. 96 (2001) 1348–1360.
[6] H. Hotelling, Relations between two sets of variables. Biometrika 58 (1936) 433–451.
[7] J. Huang, J.L. Horowitz, F. Wei, Variable selection in nonparametric additive models, Ann. Statist. 38 (2010) 2282–2313.
[8] R. Iaci, X. Yin, L.X. Zhu, The dual central subspaces in dimension reduction, J. Multivariate Anal. 145 (2016) 178–189.
[9] B. Jiang, J. Liu, Variable selection for general index models via sliced inverse regression, Ann. Statist. 42 (2014) 1751–1786.
[10] B. Li, S. Wang, On directional regression for dimension reduction, J. Amer. Statist. Assoc. 102 (2007) 997–1008.
[11] K.C. Li, Sliced inverse regression for dimension reduction (with discussion), J. Amer. Statist. Assoc. 86 (1991) 316–342.
[12] L. Li, Sparse sufficient dimension reduction, Biometrika 94 (2007) 603–613.
[13] L. Li, R.D. Cook, C.J. Nachtsheim, Model-free variable selection, J. R. Stat. Soc. Ser. B 67 (2005) 285–299.
[14] J.Y. Liu, R. Li, R. Wu, Feature selection for varying coefficient models with ultrahigh dimensional covariates, J. Amer. Statist. Assoc. 109 (2014) 266–274.
[15] D.W. Nierenberg, T.A. Stukel, J.A. Baron, B.J. Dain, E.R. Greenberg, S.C.P.S. Group, Determinants of plasma levels of beta-carotene and retinol, Amer. J. Epidemiol. 130 (1989) 511–521.
[16] Y. Shao, R.D. Cook, S. Weisberg, Mariginal tests with sliced average variance estimation, Biometrika 94 (2007) 285–296.
[17] R.J. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Ser. B 58 (1996) 267–288.
[18] L. Wang, L. Xue, A. Qu, H. Liang, Estimation and model selection in generalized additive partial linear models for correlated data with diverging number of covariates, Ann. Statist. 42 (2014) 592–624.
[19] Z. Yu, Y. Dong. Model-free coordinate test and variable selection via directional regression, Statist. Sinica 26 (2016) 1159–1174.
[20] Z. Yu, Y. Dong, Y. Fang, Marginal coordinate tests for central mean subspace with principal hessian directions, Chinese J. Appl. Probab. Statist. 26 (2010) 544–552.
[21] Z. Yu, Y. Dong, L.X. Zhu, Trace pursuit: A general framework for model-free variable selection, J. Amer. Statist. Assoc. 111 (2016) 813–821.
[22] H. Zou, The adaptive lasso and its oracle properties, J. Amer. Statist. Assoc. 101 (2006) 1418–1429.