Taylor & Francis
Taylor & Francis Group

Journal of Nonparametric Statistics

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

# Sufficient dimension reduction via distance covariance with multivariate responses

Xianyan Chen, Qingcong Yuan & Xiangrong Yin

**To cite this article:** Xianyan Chen, Qingcong Yuan & Xiangrong Yin (2019) Sufficient dimension reduction via distance covariance with multivariate responses, Journal of Nonparametric Statistics, 31:2, 268-288, DOI: 10.1080/10485252.2018.1562065

**To link to this article:** https://doi.org/10.1080/10485252.2018.1562065

Published online: 28 Dec 2018.

JOURNAL OF NONPARAMETRIC STATISTICS
2019, VOL. 31, NO. 2, 268–288
https://doi.org/10.1080/10485252.2018.1562065

Taylor & Francis
Taylor & Francis Group



# Sufficient dimension reduction via distance covariance with multivariate responses

Xianyan Chen[a], Qingcong Yuan[b] and Xiangrong Yin[c]

[a]Department of Statistics, University of Georgia, Athens, GA, USA; [b]Department of Statistics, Miami University, Oxford, OH, USA; [c]Department of Statistics, University of Kentucky, Lexington, KY, USA

**ABSTRACT**

In this article, we propose a new method for sufficient dimension reduction when both response and predictor are vectors. The new method, using distance covariance, keeps the model-free advantage, and can fully recover the central subspace even when many predictors are discrete. We then extend this method to the dual central subspace, including a special case of canonical correlation analysis. We illustrated estimators through extensive simulations and real datasets, and compared to some existing methods, showing that our estimators are competitive and robust.

**ARTICLE HISTORY**
Received 23 June 2018
Accepted 18 December 2018

**KEYWORDS**
Central subspace; distance covariance; dual central subspace; projective resampling; sufficient dimension reduction

## 1. Introduction

Suppose $Y$ is a response (scalar or vector) and $\mathbf{X}$ is a $p \times 1$ predictor vector. Sufficient dimension reduction (SDR; Li 1991; Cook 1994, 1996) is a methodology for reducing the dimension of predictors without loss of regression information. The ultimate goal of sufficient dimension reduction is to search $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}$, where $\boldsymbol{\beta}$ is a $p \times d$ matrix, $d < p$, such that $Y$ depends on $\mathbf{X}$ only through $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}$. That is: $Y \perp\!\!\!\perp \mathbf{X} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}$, where $\perp\!\!\!\perp$ means independence. The column space of $\boldsymbol{\beta}$, denoted by $\mathcal{S}(\boldsymbol{\beta})$, forms a dimension reduction subspace (Li 1991; Cook 1996). The intersection of all such subspaces, if itself is a dimension reduction subspace, is called the central subspace (CS; Cook 1996), and is denoted by $\mathcal{S}_{Y|\mathbf{X}}$. The dimension of $\mathcal{S}_{Y|\mathbf{X}}$, denoted by $dim(\mathcal{S}_{Y|\mathbf{X}}) = d$, is called the structural dimension. Under mild conditions (Cook 1996; Yin, Li, and Cook 2008), the CS exists and is unique. We assume CS exists throughout this article.

Many methods have been proposed in this area. These include the inverse approaches: SIR (Li 1991), SAVE (Cook and Weisberg 1991), IR (Cook and Ni 2005), DR (Li and Wang 2007); forward approaches: Hristache, Juditsky, Polzehl, and Spokoiny (2001), MAVE (Xia, Tong, Li, and Zhu 2002) and SR (Wang and Xia 2008); correlation approaches: CANCOR (Fung, He, Liu, and Shi 2002), Kullback–Leibler (KL)-distance (Yin and Cook 2005) and Fourier transform (Zhu and Zeng 2006; Zeng and Zhu 2010). However, these methods require either the linearity condition or constant covariance condition, or the predictors to be multivariate normal, continuous and the link function to be smooth. Recently, Sheng

**CONTACT** Xiangrong Yin yinxiangrong@uky.edu Department of Statistics, University of Kentucky, Lexington, KY 40536, USA

and Yin (2013, 2016) developed a novel method using distance covariance (DCOV; Székely, Rizzo, and Bakirov 2007; Székely and Rizzo 2009) for SDR. The method does not require linearity condition or constant covariance condition, or any particular distribution on $\mathbf{X}$, $\mathbf{X} \mid Y$ or $Y \mid \mathbf{X}$. These advantages enable the method to work effectively under a variety of $\mathbf{X}$: $\mathbf{X}$ could be normal, non-normal, continuous, or discrete.

Various dimension reduction concepts can be extended to a multivariate response by replacing random scalar $Y$ with random vector $\mathbf{Y}$. Generally, there are three approaches to extend dimension reduction objects. The first approach is to slice the multidimensional $\mathbf{Y}$ into hypercubes. However, this method faces 'curse of dimensionality' since the number of observations within each hypercube decreases sharply as when the dimension of response increases. The second approach is to target the central mean subspace (Cook and Setodji 2003) or the central moment subspace (Yin and Bura 2006). The third approach is to estimate the marginal dimension reduction spaces, and then pool these estimates to recover the central subspace (Saracco 2005). However, the latter two methods are not guaranteed to fully recover the dimension reduction space. The projective resampling method (Li, Wen, and Zhu 2008) solves these problems by projecting the multivariate response along $m$ randomly sampled directions, where $m$ is a pre-selected integer, to obtain $m$ scalar-valued responses, then use any dimension reduction method for a univariate response to get a subspace. Averaging these $m$ subspaces, we can estimate the CS. It is shown that this method can fully recover the CS.

Canonical correlation analysis (CCA) extracts a pairwise linear relationship between two random vectors. Kettenring (1971) extended CCA to multiple sets, by maximising a generalised measure of correlation between the random vectors. Burg and Leeuw (1983) first proposed a method termed nonlinear canonical correlation analysis using an alternating least squares algorithm. Yin (2004) used KL information to find linear and nonlinear relationships between two sets of random vectors. Yin and Sriram (2008), Iaci, Yin, Sriram, and Klingenberg (2008) and Iaci, Sriram, and Yin (2010) extended this idea to independent groups and multiple sets of random vectors. However, all of these CCA methods require that the number of coefficient vectors from both sets that provide the dimension reduction be equal. Iaci, Yin, and Zhu (2015) introduced the dual central subspaces (DCS), which is to provide a dimension reduction of both vectors without requiring the dimensions of the reduction to be equal, with the idea that the true associations between the random vectors may not be equal.

In this article, based on the advantages of DCOV, we develop several methods (combining projective resampling and sequential search) to implement dimension reduction for a multivariate response. Among them, one approach is to average the $m$ subspaces to get the CS. The other is to sum $m$ distance covariance functions and then obtain the CS. We also introduce a novel idea of $k$ nearest neighbours kNN procedure to estimate the dimension of the CS. We extend the two DCOV methods to canonical analysis as canonical distance covariance analysis (CDCA) and to estimate DCS, and use the bootstrap method to estimate the dimension of DCS. Through a number of simulation studies, we demonstrate the better performance of the proposed methods.

The rest of the article is organised as follows: in Section 2, we describe our method in details, including DCOV, projective resampling approach, DCS, methods to estimate the dimensions of CS and DCS. In Section 3, we conduct simulation comparisons between our estimators and others in a variety of models; and in Section 4, we summarise our work.

## 2. Methodology

### 2.1. Distance covariance (DCOV) for sufficient dimension reduction

DCOV is introduced by Székely et al. (2007) as a new measurement of multivariate dependence. Let $\boldsymbol{Z}_1 \in \mathbb{R}^p$ and $\boldsymbol{Z}_2 \in \mathbb{R}^q$ be random variables, where $p$ and $q$ are positive integers. Let $\mathcal{V}(\boldsymbol{Z}_1, \boldsymbol{Z}_2)$ be the distance covariance between $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$. The squared distance covariance can be defined as the weighted $L_2$ norm of the distance between the joint characteristic function of the random variables and the product of their marginal characteristic functions:

$$\mathcal{V}^2(\boldsymbol{Z}_1, \boldsymbol{Z}_2) = \int_{\mathbb{R}^{p+q}} |f_{\boldsymbol{Z}_1,\boldsymbol{Z}_2}(t,s) - f_{\boldsymbol{Z}_1}(t) f_{\boldsymbol{Z}_2}(s)|^2 w(t,s)\, \mathrm{d}t\, \mathrm{d}s$$

where $f_{\boldsymbol{Z}_1}, f_{\boldsymbol{Z}_2}$, and $f_{\boldsymbol{Z}_1,\boldsymbol{Z}_2}$ are the characteristic functions of $\boldsymbol{Z}_1, \boldsymbol{Z}_2$, and $(\boldsymbol{Z}_1, \boldsymbol{Z}_2)$, respectively. The weight function $w(t,s) = (c_p c_q |s|_p^{1+p} |t|_q^{1+q})^{-1}$, where $c_q$ and $c_q$ are positive constants, and $|s|_p^{1+p}$ is the $1+p$ power of the Euclidean norm of $s$ in $\mathbb{R}^p$. Székely and Rizzo (2009) developed an equivalent form of DCOV:

$$\begin{aligned}\mathcal{V}^2(\boldsymbol{Z}_1, \boldsymbol{Z}_2) = {} & E|\boldsymbol{Z}_1 - \boldsymbol{Z}_1'|\, |\boldsymbol{Z}_2 - \boldsymbol{Z}_2'| + E|\boldsymbol{Z}_1 - \boldsymbol{Z}_1'| E|\boldsymbol{Z}_2 - \boldsymbol{Z}_2'| \\ & - E|\boldsymbol{Z}_1 - \boldsymbol{Z}_1'| |\boldsymbol{Z}_2 - \boldsymbol{Z}_2''| - E|\boldsymbol{Z}_1 - \boldsymbol{Z}_1''| E|\boldsymbol{Z}_2 - \boldsymbol{Z}_2'|,\end{aligned}$$

where $(\boldsymbol{Z}_1, \boldsymbol{Z}_2)$, $(\boldsymbol{Z}_1', \boldsymbol{Z}_2')$, $(\boldsymbol{Z}_1'', \boldsymbol{Z}_2'')$ are *i.i.d.* copies. In this form, DCOV requires $E|\boldsymbol{Z}_1| < \infty$ and $E|\boldsymbol{Z}_2| < \infty$ so that DCOV is finite (Székely et al. 2007).

DCOV equals to 0 if and only if two random vectors are independent (Székely et al. 2007). Based on this property, Sheng and Yin (2013, 2016) proposed DCOV as an SDR tool. Suppose $\boldsymbol{\beta}$ is a $p \times d$ matrix, where $1 \le d \le q$. Under $E|\mathbf{X}| < \infty$ and $E|Y| < \infty$ (Székely et al. 2007), the solution to the following optimisation problem will yield a basis of the CS:

$$\max_{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{\Sigma}_X \boldsymbol{\beta} = \mathbf{I}_d} \mathcal{V}^2(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}, Y). \tag{1}$$

Throughout the article we assume $E|\mathbf{X}| < \infty$ and $E|Y| < \infty$. The constraint $\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{\Sigma}_X \boldsymbol{\beta} = \mathbf{I}_d$ in the optimisation problem guarantees the solution of $\boldsymbol{\beta}$ in the same scale and the optimisation solver does not diverge.

### 2.2. DCOV for multivariate response

The method developed by Sheng and Yin (2013, 2016) is for a scalar response. We now extend their approach and results to a multivariate response, say, $\boldsymbol{Y}$, a $q \times 1$ random vector. To facilitate our discussion, let $\boldsymbol{B}$ be a $p \times d$ matrix and let $\mathcal{S}(\boldsymbol{B})$ be the subspace of $\mathbb{R}^p$ spanned by the columns of $\boldsymbol{B}$. Let $\boldsymbol{\Sigma}_X$ be the covariance matrix of $\mathbf{X}$, which is assumed to be nonsingular. Let $\mathbf{P}_{\boldsymbol{B}(\boldsymbol{\Sigma}_X)}$ denote the orthogonal projection onto $\mathcal{S}(\boldsymbol{B})$ with respect to the inner product $\langle \boldsymbol{a}, \boldsymbol{b} \rangle = \boldsymbol{a}^{\mathrm{T}} \boldsymbol{\Sigma} \boldsymbol{b}$. That is, $\mathbf{P}_{\boldsymbol{B}(\boldsymbol{\Sigma}_X)} = \boldsymbol{B}(\boldsymbol{B}^{\mathrm{T}} \boldsymbol{\Sigma}_X \boldsymbol{B})^{-1} \boldsymbol{B}^{\mathrm{T}} \boldsymbol{\Sigma}_X$. Let $\mathbf{Q}_{\boldsymbol{B}(\boldsymbol{\Sigma}_X)} = \mathbf{I} - \mathbf{P}_{\boldsymbol{B}(\boldsymbol{\Sigma}_X)}$, where $\mathbf{I}$ is the identity matrix. Following the previous section, then a basis of the CS can be obtained by solving (1) with $Y$ replaced by $\boldsymbol{Y}$, and obtain the following.

**Proposition 2.1:** *Let $\boldsymbol{\eta}$ be a basis of the CS with dimension $d$, $\boldsymbol{\beta}$ be a $p \times d_0$ matrix, $d_0 \le d$, $\dim(\mathcal{S}(\boldsymbol{\beta})) = d_0$, $\boldsymbol{\eta}^{\top} \boldsymbol{\Sigma}_X \boldsymbol{\eta} = \boldsymbol{I}_d$, and $\boldsymbol{\beta}^{\top} \boldsymbol{\Sigma}_X \boldsymbol{\beta} = \boldsymbol{I}_{d_0}$. Assume $\mathcal{S}(\boldsymbol{\beta}) \subseteq \mathcal{S}(\boldsymbol{\eta})$, then $\mathcal{V}^2(\boldsymbol{\beta}^{\top} \mathbf{X}, \boldsymbol{Y}) \le \mathcal{V}^2(\boldsymbol{\eta}^{\top} \mathbf{X}, \boldsymbol{Y})$. The equality holds if and only if $\mathcal{S}(\boldsymbol{\beta}) = \mathcal{S}(\boldsymbol{\eta})$.*

**Proposition 2.2:** *Let* $\boldsymbol{\eta}$ *be a basis of the CS with dimension* $d$, $\boldsymbol{\beta}$ *be a* $p \times d_0$ *matrix,* $\boldsymbol{\eta}^\top \boldsymbol{\Sigma}_X \boldsymbol{\eta} = \boldsymbol{I}_d$, *and* $\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_X \boldsymbol{\beta} = \boldsymbol{I}_{d_0}$. *Here* $d_0$ *could be bigger, less than or equal to* $d$. *Suppose* $\mathbf{P}_{\boldsymbol{B}(\boldsymbol{\Sigma}_X)}\boldsymbol{X} \perp\!\!\!\perp \mathbf{Q}_{\boldsymbol{B}(\boldsymbol{\Sigma}_X)}\boldsymbol{X}$, *and* $\mathcal{S}(\boldsymbol{\beta}) \nsubseteq \mathcal{S}(\boldsymbol{\eta})$, *then* $\mathcal{V}^2(\boldsymbol{\beta}^\top \boldsymbol{X}, \boldsymbol{Y}) < \mathcal{V}^2(\boldsymbol{\eta}^\top \boldsymbol{X}, \boldsymbol{Y})$.

Proposition 2.1 suggests that if $\mathcal{S}(\boldsymbol{\beta})$ is a subspace of $\mathcal{S}(\boldsymbol{\eta})$, then the squared distance covariance between $\boldsymbol{\beta}^\top \boldsymbol{X}$ and $\boldsymbol{Y}$ is always less than or equal to that between $\boldsymbol{\eta}^\top \boldsymbol{X}$ and $\boldsymbol{Y}$. The equation holds if and only if $\mathcal{S}(\boldsymbol{\beta}) = \mathcal{S}(\boldsymbol{\eta})$. Proposition 2.2 suggests that if $\mathcal{S}(\boldsymbol{\beta})$ is not a subspace of $\mathcal{S}(\boldsymbol{\eta})$, then under a mild condition, the DCOV between $\boldsymbol{\beta}^\top \boldsymbol{X}$ and $\boldsymbol{Y}$ is always less than the DCOV between $\boldsymbol{\eta}^\top \boldsymbol{X}$ and $\boldsymbol{Y}$. These two propositions together indicate that by maximising $\mathcal{V}^2(\boldsymbol{\beta}^\top \boldsymbol{X}, \boldsymbol{Y})$ with a constraint of $\boldsymbol{\beta}$ can always identify the CS. Following Székely et al. (2007), a sample version for a multivariate response can be defined as $\mathcal{V}^2(\boldsymbol{\beta}^\top \boldsymbol{X}, \boldsymbol{Y}) = (1/n^2) \sum_{k,l=1}^n A_{kl}(\boldsymbol{\beta}) B_{kl}$, where, for $k, l = 1, \ldots, n$,

$$A_{kl}(\boldsymbol{\beta}) = a_{kl}(\boldsymbol{\beta}) - \bar{a}_{k.}(\boldsymbol{\beta}) - \bar{a}_{.l}(\boldsymbol{\beta}) + \bar{a}_{..}(\boldsymbol{\beta})$$

$$a_{kl}(\boldsymbol{\beta}) = |\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_k - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_l|, \bar{a}_{k.}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{l=1}^{n} a_{kl}(\boldsymbol{\beta}),$$

$$\bar{a}_{.l}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{k=1}^{n} a_{kl}(\boldsymbol{\beta}), \bar{a}_{..}(\boldsymbol{\beta}) = \frac{1}{n^2}\sum_{k,l=1}^{n} a_{kl}(\boldsymbol{\beta}).$$

Similarly, define $b_{kl} = |\boldsymbol{Y}_k - \boldsymbol{Y}_l|$ and $B_{kl} = b_{kl} - \bar{b}_{k.} - \bar{b}_{.l} + \bar{b}_{..}$, where $|\cdot|$ is the Euclidean norm in the respective dimension. Replacing $\boldsymbol{\Sigma}_X$ with its sample version $\hat{\boldsymbol{\Sigma}}_X$, the estimated basis matrix of the CS is

$$\boldsymbol{\eta}_n = \arg \max_{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_X\boldsymbol{\beta}=\mathbf{I}_d} \mathcal{V}_n^2(\boldsymbol{\beta}^\top \boldsymbol{X}, \boldsymbol{Y}). \tag{2}$$

Using a Sequential Quadratic Programming (SQP) method, we can solve the nonlinear optimisation problem in Equation (2).

### 2.3. DCOV with projective resampling

Projective resampling (Li et al. 2008) is an SDR method for multivariate responses. Let $\boldsymbol{t}$ be a generic vector in $\mathbb{R}^q$. It is established on the statement: $\mathbf{Y} \perp\!\!\!\perp \mathbf{X} \,|\, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}$ if and only if $\boldsymbol{t}^{\mathrm{T}}\mathbf{Y} \perp\!\!\!\perp \mathbf{X} \,|\, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}$ for all $t \in \mathbb{R}^q$. That is

$$\mathcal{S}_{\mathbf{Y}|\mathbf{X}} = \mathrm{Span}\{\mathcal{S}_{\boldsymbol{t}^{\mathrm{T}}\mathbf{Y}\,|\,\mathbf{X}}, \boldsymbol{t} \in \mathbb{R}^q\}.$$

In this way, the multivariate response problem is reduced to the many univariate response problem. Thus, all SDR methods developed for the univariate response can be employed to the multivariate response by estimating $\mathcal{S}_{\boldsymbol{t}^{\mathrm{T}}\mathbf{Y}\,|,pmbX}$ for all $\boldsymbol{t} \in \mathbb{R}^q$. However, it is impossible to conduct dimension reduction for all $\boldsymbol{t} \in \mathbb{R}^q$. Hilafu and Yin (2013) discuss the size of $\mathbf{t}$ as:

(i) if the structural dimension is $d$, there exist $d$ $\boldsymbol{t}_i$'s such that $\mathcal{S}_{\mathbf{Y}\,|\,\mathbf{X}} = \mathrm{Span}\{\mathcal{S}_{\mathbf{t}^{\mathrm{T}}\mathbf{Y}\,|\,\mathbf{X}}\}$;
(ii) if the size of $\boldsymbol{t}$ is large enough, the subspace will be recovered through those univariate CSs.

Li et al. (2008) proposed projective resampling SIR, SAVE, and DR. In addition to the multivariate DCOV (denote it as DCOV0) that is described in Section 2.2, we apply projective resampling to univariate DCOV. Suppose the sample size of random direction $\boldsymbol{t}$ is $m$. With different approaches to combine all generated univariate $\boldsymbol{t}^{\mathrm{T}}\mathbf{Y}$, we develop DCOV1 and DCOV2 methods:

**DCOV 1:** For each of the $m$ combinations of $\mathbf{Y}, \boldsymbol{t}_i{}^{\mathrm{T}}\mathbf{Y},\ i = 1, \ldots, m$, solve the optimisation problem to get

$$\hat{\boldsymbol{\beta}}_i = \arg \max_{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}_X\boldsymbol{\beta}=\mathbf{I}_d} \mathcal{V}^2(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{t}_i{}^{\mathrm{T}}\mathbf{Y}).$$

Then an estimated basis of CS can be the first $d$ eigenvectors of

$$\frac{1}{m}\sum_{i=1}^{m} \hat{\boldsymbol{\beta}}_i \hat{\boldsymbol{\beta}}_i^{\mathrm{T}}.$$

**DCOV 2:** Sum the squared distance covariance for each $\boldsymbol{t}_i{}^{\mathrm{T}}\mathbf{Y}$ as the new objective function, and then solve the optimisation problem

$$\hat{\boldsymbol{\beta}} = \arg \max_{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}_X\boldsymbol{\beta}=\mathbf{I}_d} \sum_{i=1}^{m} \mathcal{V}^2(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{t}_i{}^{\mathrm{T}}\mathbf{Y}).$$

DCOV1 is similar to the outer product gradient (OPG) type. We get a basis for each univariate $\boldsymbol{t}_i{}^{\mathrm{T}}\mathbf{Y}$, $\hat{\boldsymbol{\beta}}_i$, for $i = 1, \ldots, m$, and then apply singular value decomposition (SVD) to $(1/m)\sum_{i=1}^{m} \hat{\boldsymbol{\beta}}_i \hat{\boldsymbol{\beta}}_i^{\mathrm{T}}$ to obtain the estimated $\hat{\boldsymbol{\beta}}$. While DCOV2 is similar to a MAVE type, we sum $\mathcal{V}^2(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{t}^{\mathrm{T}}\mathbf{Y})$ first and get the estimated $\hat{\boldsymbol{\beta}}$. In the simulation section, results of both methods are given for comparison.

Note that $\boldsymbol{t}_i$, $i = 1, \ldots, m$, is a random direction that the multivariate response is projected onto, and $m$ is the total number of random directions. Typically, they can be generated by using multivariate normal with unit length (Li et al. 2008). For DCOV1, $m$ estimates of $\hat{\boldsymbol{B}}_i$ are obtained for each random direction $\boldsymbol{t}_i$, $i = 1, \ldots, m$. The estimate $\hat{\boldsymbol{B}}$ is calculated by singular value decomposition of the sum of $\hat{\boldsymbol{B}}_i{}^{\top}\hat{\boldsymbol{B}}_i$. Note that by invariance law, we can equivalently work on a standardised predictor $\boldsymbol{Z}$-scale. As such, we first standardise $\boldsymbol{X}$- to $\boldsymbol{Z}$-scale. After obtaining the estimate under the $\boldsymbol{Z}$-scale, we transform the estimate back to the $\boldsymbol{X}$-scale, $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\Sigma}}_X^{-1/2}\hat{\boldsymbol{\beta}}_Z$. This scheme seems to work well in our simulations and real data studies. An alternative procedure is to use a successive one-at-a-time search similar to that of Yin et al. (2008).

Sheng and Yin (2016) showed in their paper that the estimator of univariate DCOV, $\boldsymbol{\eta}_n = \arg\max_{\boldsymbol{\beta}^{\top}\hat{\boldsymbol{\Sigma}}_X\boldsymbol{\beta}=\boldsymbol{I}} \mathcal{V}_n^2(\boldsymbol{\beta}^{\top}\boldsymbol{X}, Y)$, is consistent and asymptotically normal. Here for DCOV1, by the consistency proposition for the univariate response in the work of Sheng and Yin (2016), we have $\boldsymbol{\eta}_n^i \xrightarrow{p} \boldsymbol{\eta}\boldsymbol{Q}$, where $\boldsymbol{Q}$ is a rotation matrix, for each $\boldsymbol{t}_i^{\top}\boldsymbol{Y}$, with $i = 1, \ldots m$. We combine all these $\boldsymbol{\eta}_n^i$ and use SVD to obtain the estimator in DCOV1, thus it also has the consistency property, that is, $\boldsymbol{\eta}_n \xrightarrow{p} \boldsymbol{\eta}\boldsymbol{Q}$. The asymptotically normal property can be shown in the same way. For each univariate response $\boldsymbol{t}_i^{\top}\boldsymbol{Y}$, $i = 1, \ldots m$, by the normality property, $\sqrt{n}[vec(\boldsymbol{\eta}_n^i) - vec(\boldsymbol{\eta}\boldsymbol{Q})] \xrightarrow{\mathfrak{D}} N(0, V(\boldsymbol{\eta}_Q))$, then when adding these

estimators, the final estimator by DCOV1 has asymptotic normality as that in Sheng and Yin (2016). Consider $m=1$ in DCOV2, the estimator has $\sqrt{n}$-consistency and asymptotic normality, when increasing $m$, that is, adding squared distance covariance, the estimator also has $\sqrt{n}$-consistency and asymptotic normality, but with tedious calculations based on Sheng and Yin (2016).

### 2.4. Estimating the DCS via distance covariance

Consider two sets of random vectors, $\mathbf{X}$ is $p \times 1$ and $\mathbf{Y}$ is $q \times 1$, exchange the role of $\mathbf{X}$ and $\mathbf{Y}$, if $\boldsymbol{\alpha}$ is a $q \times s$ matrix, $s < q$, $\mathbf{X}$ depends on $\mathbf{Y}$ only through $\boldsymbol{\alpha}^\top \mathbf{Y}$. That is,

$$\mathbf{Y} \perp\!\!\!\perp \mathbf{X} \,|\, \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{Y}.$$

The column space of $\boldsymbol{\alpha}$ is called $\mathcal{S}(\boldsymbol{\alpha})$, and the intersection of such subspaces is defined as the central subspace of $\mathbf{X}$ given $\mathbf{Y}$, denoted by $\mathcal{S}_{X\,|\,\mathbf{Y}}$. The reduction subspace $\boldsymbol{\beta}, \boldsymbol{\alpha}$ is called DCS by Iaci et al. (2015) as the combination of $\mathcal{S}_{\mathbf{Y}\,|\,\mathbf{X}}$ and $\mathcal{S}_{\mathbf{X}\,|\,\mathbf{Y}}$.

The proposition below suggests ways to recover the DCS.

**Proposition 2.3 (Iaci et al. 2015):** *Let $\boldsymbol{B}$ and $\boldsymbol{A}$ be the base for $\mathcal{S}_{\mathbf{Y}\,|\,\mathbf{X}}$ and $\mathcal{S}_{\mathbf{X}\,|\,\mathbf{Y}}$, respectively. The following conditions are equivalent:*

$$\text{(i)} \quad \mathbf{Y} \perp\!\!\!\perp \mathbf{X} \,|\, \boldsymbol{B}^\top\mathbf{X} \quad \text{and} \quad \mathbf{Y} \perp\!\!\!\perp \mathbf{X} \,|\, \boldsymbol{A}^\top\mathbf{Y},$$

$$\text{(ii)} \quad \mathbf{Y} \perp\!\!\!\perp \mathbf{X} \,|\, \boldsymbol{B}^\top\mathbf{X} \quad \text{and} \quad \mathbf{Y} \perp\!\!\!\perp \boldsymbol{B}^\top\mathbf{X} \,|\, \boldsymbol{A}^\top\mathbf{Y},$$

$$\text{(iii)} \quad \boldsymbol{B}^\top\mathbf{Y} \perp\!\!\!\perp \mathbf{X} \,|\, \boldsymbol{B}^\top\mathbf{X} \quad \text{and} \quad \mathbf{Y} \perp\!\!\!\perp \mathbf{X} \,|\, \boldsymbol{A}^\top\mathbf{Y}.$$

Proposition 2.3 suggests that we can first reduce the dimension of $\mathbf{X}$ by treating $\mathbf{Y}$ as response and then reduce the dimension of $\mathbf{Y}$ by treating $\mathbf{X}$ or $\boldsymbol{B}^\top\mathbf{X}$ as response.

Assume the dimensions $d_x$ and $d_y$ are known. Let $(\boldsymbol{x}_i, \boldsymbol{y}_i)$, $i = 1, \ldots, n$ be the random sample from $(\boldsymbol{X}, \boldsymbol{Y})$. The estimates of the matrices that form the bases of the DCS, $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{B}}$ can be obtained by finding the maximum of squared distance covariance:

$$(\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}}) = \arg \max_{\substack{\boldsymbol{A}^\top \hat{\boldsymbol{\Sigma}}_X \boldsymbol{A} = I_{d_x} \\ \boldsymbol{B}^\top \hat{\boldsymbol{\Sigma}}_Y \boldsymbol{B} = I_{d_y}}} \mathcal{V}^2(\boldsymbol{B}^\top \boldsymbol{x}, \boldsymbol{A}^\top \boldsymbol{y})$$

The two constraints $\boldsymbol{A}^\top \hat{\boldsymbol{\Sigma}}_X \boldsymbol{A} = \mathbf{I}_{d_x}$, and $\boldsymbol{B}^\top \hat{\boldsymbol{\Sigma}}_Y \boldsymbol{B} = \mathbf{I}_{d_y}$ guarantee the estimated directions has unit length and is orthogonal to each other. Here, $\hat{\boldsymbol{\Sigma}}_X$ and $\hat{\boldsymbol{\Sigma}}_Y$ are the sample covariance matrixes for $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively.

Since there are too many parameters when we estimate $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{B}}$ simultaneously, we propose two approaches to estimate $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{B}}$, separately, with a difference of the estimation of $\boldsymbol{A}$ depends on $\hat{\boldsymbol{B}}$ or not. The procedure of these two approaches are described as Approach 1 and Approach 2, with the multivariate response in the squared distance covariance (DCOV0) as the objective function in the optimisation problem.

**Approach 1:** Estimate $\boldsymbol{B}$ considering $\boldsymbol{Y}$ as a response, and estimate $\boldsymbol{A}$ considering $\boldsymbol{X}$ as a response. This means, we can calculate $\hat{\boldsymbol{B}} = \arg\max_{\boldsymbol{B}^\top \hat{\boldsymbol{\Sigma}}_X \boldsymbol{B} = \mathbf{I}_{d_x}} \mathcal{V}_n^2(\boldsymbol{B}^\top \mathbf{x}, \boldsymbol{y})$ and $\hat{\boldsymbol{A}} =$

**Table 1.** Methods for DCS.

| Method | Estimate $\boldsymbol{B}$ | Estimate $\boldsymbol{A}$ |
|---|---|---|
| Approach1 DCOV0 | $\max_{\boldsymbol{B}^\top \hat{\boldsymbol{\Sigma}}_X \boldsymbol{B}=\boldsymbol{I}_{d_x}} \mathcal{V}_n^2(\boldsymbol{B}^\top \mathbf{X}, \boldsymbol{Y})$ | $\max_{\boldsymbol{A}^\top \hat{\boldsymbol{\Sigma}}_Y \boldsymbol{A}=\boldsymbol{I}_{d_y}} \mathcal{V}_n^2(\mathbf{X}, \boldsymbol{A}^\top \boldsymbol{Y})$ |
| Approach2 DCOV0 | $\max_{\boldsymbol{B}^\top \hat{\boldsymbol{\Sigma}}_X \boldsymbol{B}=\boldsymbol{I}_{d_x}} \mathcal{V}_n^2(\boldsymbol{B}^\top \mathbf{X}, \boldsymbol{Y})$ | $\max_{\boldsymbol{A}^\top \hat{\boldsymbol{\Sigma}}_Y \boldsymbol{A}=\boldsymbol{I}_{d_y}} \mathcal{V}_n^2(\hat{\boldsymbol{B}}^\top \mathbf{X}, \boldsymbol{A}^\top \boldsymbol{Y})$ |
| Approach1 DCOV1[a] | $\max_{\boldsymbol{B}^\top \hat{\boldsymbol{\Sigma}}_X \boldsymbol{B}=\boldsymbol{I}_{d_x}} \mathcal{V}^2(\boldsymbol{B}^\top \mathbf{X}, \boldsymbol{t}_i^\top \mathbf{Y})$ | $\max_{\boldsymbol{A}^\top \hat{\boldsymbol{\Sigma}}_Y \boldsymbol{A}=\boldsymbol{I}_{d_y}} \mathcal{V}^2(\boldsymbol{t}_i^\top \mathbf{X}, \boldsymbol{A}^\top \mathbf{Y})$ |
| Approach2 DCOV1[a] | $\max_{\boldsymbol{B}^\top \hat{\boldsymbol{\Sigma}}_X \boldsymbol{B}=\boldsymbol{I}_{d_x}} \mathcal{V}^2(\boldsymbol{B}^\top \mathbf{X}, \boldsymbol{t}_i^\top \mathbf{Y})$ | $\max_{\boldsymbol{A}^\top \hat{\boldsymbol{\Sigma}}_Y \boldsymbol{A}=\boldsymbol{I}_{d_y}} \mathcal{V}^2(\boldsymbol{t}_i^\top \hat{\boldsymbol{B}}^\top \mathbf{X}, \boldsymbol{A}^\top \mathbf{Y})$ |
| Approach1 DCOV2 | $\max_{\boldsymbol{B}^\top \hat{\boldsymbol{\Sigma}}_X \boldsymbol{B}=\boldsymbol{I}_{d_x}} \sum_{i=1}^{m} \mathcal{V}^2(\boldsymbol{B}^\top \mathbf{X}, \boldsymbol{t}_i^\top \mathbf{Y})$ | $\max_{\boldsymbol{A}^\top \hat{\boldsymbol{\Sigma}}_Y \boldsymbol{A}=\boldsymbol{I}_{d_y}} \sum_{i=1}^{m} \mathcal{V}^2(\boldsymbol{t}_i^\top \mathbf{X}, \boldsymbol{A}^\top \mathbf{Y})$ |
| Approach2 DCOV2 | $\max_{\boldsymbol{B}^\top \hat{\boldsymbol{\Sigma}}_X \boldsymbol{B}=\boldsymbol{I}_{d_x}} \sum_{i=1}^{m} \mathcal{V}^2(\boldsymbol{B}^\top \mathbf{X}, \boldsymbol{t}_i^\top \mathbf{Y})$ | $\max_{\boldsymbol{A}^\top \hat{\boldsymbol{\Sigma}}_Y \boldsymbol{A}=\boldsymbol{I}_{d_y}} \sum_{i=1}^{m} \mathcal{V}^2(\boldsymbol{t}_i^\top \hat{\boldsymbol{B}}^\top \mathbf{X}, \boldsymbol{A}^\top \mathbf{Y})$ |

[a] Refer to the introduction above in this section for detailed calculation.

$\arg\max_{\boldsymbol{A}^\top \hat{\boldsymbol{\Sigma}}_Y \boldsymbol{A}=\mathbf{I}_{d_y}} \mathcal{V}_n^2(\mathbf{x}, \boldsymbol{A}^\top \boldsymbol{y})$ at the same time, since the two steps do not depend on each other.

**Approach 2:** Estimate $\boldsymbol{B}$ considering $\boldsymbol{Y}$ as a response, and then estimate $\boldsymbol{A}$ considering $\boldsymbol{B}^\top \boldsymbol{X}$ as a response. That is, after calculating

$$\hat{\boldsymbol{B}} = \arg \max_{\boldsymbol{B}^\top \hat{\boldsymbol{\Sigma}}_X \boldsymbol{B}=I_{d_x}} \mathcal{V}_n^2(\boldsymbol{B}^\top \mathbf{x}, \boldsymbol{y}),$$

obtain $\boldsymbol{A}$ with the projection $\hat{\boldsymbol{B}}^\top \boldsymbol{x}$,

$$\hat{\boldsymbol{A}} = \arg \max_{\boldsymbol{A}^\top \hat{\boldsymbol{\Sigma}}_Y \boldsymbol{A}=I_{d_y}} \mathcal{V}_n^2(\hat{\boldsymbol{B}}^\top \mathbf{x}, \boldsymbol{A}^\top \boldsymbol{y});$$

We call the above two approaches 'Approach1 DCOV0' and 'Approach2 DCOV0', respectively, since DCOV0 is used in the procedure. When using DCOV derivatives with projective resampling on the multivariate response, we can develop methods 'Approach 1 DCOV1', 'Approach 2 DCOV1', 'Approach 1 DCOV2' and 'Approach 2 DCOV2', whose optimisation problems are summarised in Table 1. Based on Sheng and Yin (2016), it can be easily shown that the estimator in Table 1 are consistent and asymptotically normal. Canonical analysis, as a special case of DCS, is termed as CDCA for this setting, where it requires $d_x = d_y$, and the calculation is through pairwise, not matrix optimisation. That is, like in CCA, we search one pair of vectors, and after this, we search another pair of vectors in the respective orthogonal spaces.

## 2.5. Estimating dimension

### 2.5.1. Estimating d for multivariate response

In practice, $d$, the dimension of CS is unknown and must be inferred from data. A few methods have been proposed in the literature, for example, a sequential test based on a chi-squared statistic proposed by Li (1991, 1992), a permutation based test by Cook and Yin (2001), and a bootstrap procedure initialled by Ye and Weiss (2003), followed by Zhu and Zeng (2006), and Sheng and Yin (2016). In this article, we introduce a *kNN* procedure for the purpose of choosing $d$, following the idea of the kNN method (Wang, Yin, and Critchley 2015).

Given $\{(\mathbf{X}_i, \mathbf{Y}_i)\},\ 1 \le i \le n$, $d$ can be evaluated by the following kNN procedure:

(1) for each point in $\{(\mathbf{X}_i, \mathbf{Y}_i),\ 1 \le i \le n\}$, obtain the $k$ nearest neighbours of sample point $i$ using Euclidean distance $|\mathbf{X}_i - \mathbf{X}_j|$, where $1 \le j \le n$. The $k$ nearest neighbours of sample point $i$ is denoted as $\{(\mathbf{X}_j^{(i)}, \mathbf{Y}_j^{(i)}),\ 1 \le j \le k\}$;
(2) for each sample point $i$, apply any dimension reduction method to its $k$ nearest neighbours $\{(\mathbf{X}_j^{(i)}, \mathbf{Y}_j^{(i)}),\ 1 \le j \le k\}$, and estimate $\hat{\boldsymbol{\beta}}_i$. Setting the dimension of $\hat{\boldsymbol{\beta}}_i$ as 1 or 2 is usually good enough;
(3) after all $\hat{\boldsymbol{\beta}}_i,\ 1 \le i \le n$ are obtained, get the eigenvalues of $\sum_{i=1}^n \hat{\boldsymbol{\beta}}_i \hat{\boldsymbol{\beta}}_i^{\mathrm{T}}$, denote as $\lambda_1, \lambda_2, \ldots, \lambda_p$;
(4) calculate the ratio $r_i = \lambda_i/\lambda_{i+1},\ 1 \le i \le p-1$. Choose $d$ as where the largest $r_i$ happens in the sequence.

In the last step, this maximal eigenvalue ratio criterion was suggested by Luo, Wang, and Tsai (2009) and was also used by Li and Yin (2009).

### 2.5.2. Estimating dimension of DCS

In practice, the dimension of the DCS $(d_x, d_y)$, is unknown and needs to be estimated. We adopt the idea of Iaci et al. (2015) to estimate the dimensions of the DCS. Suppose $\boldsymbol{B}_{d_x}$ and $\boldsymbol{A}_{d_y}$ are the true bases for $\mathcal{S}_{\mathbf{Y}\,|\,\mathbf{X}}$ and $\mathcal{S}_{\mathbf{X}\,|\,\mathbf{Y}}$, respectively. Let $\mathcal{S}_{\boldsymbol{B}_k}$ and $\mathcal{S}_{\boldsymbol{A}_l}$ be subspace for a fixed pair of dimensions $k$ and $l$. Calculate the estimated dual subspace on the original data, denoted by $\mathcal{S}_{\hat{\boldsymbol{B}}_k}$ and $\mathcal{S}_{\hat{\boldsymbol{A}}_l}$. Then calculate the bootstrap estimated dual subspaces $\mathcal{S}_{\hat{\boldsymbol{B}}_k^b}$ and $\mathcal{S}_{\hat{\boldsymbol{A}}_l^b}$. If $k = d_x$, and $l = d_y$, the variabilities of $\mathcal{S}_{\hat{\boldsymbol{B}}_k^b}$ and $\mathcal{S}_{\hat{\boldsymbol{A}}_l^b}$, respectively, from $\mathcal{S}_{\hat{\boldsymbol{B}}_k}$ and $\mathcal{S}_{\hat{\boldsymbol{A}}_l}$ are expected to be small, i.e. $\mathcal{S}_{\hat{\boldsymbol{B}}_k^b}$ and $\mathcal{S}_{\boldsymbol{B}_k}$ estimate the central subspace $\mathcal{S}_{\mathbf{Y}\,|\,\mathbf{X}}$, and $\mathcal{S}_{\hat{\boldsymbol{A}}_l^b}$ and $\mathcal{S}_{\boldsymbol{A}_l}$ estimate central subspace of $\mathcal{S}_{\mathbf{X}\,|\,\mathbf{Y}}$. $\Delta_m(\hat{\mathcal{S}}_1, \mathcal{S}_2)$, defined in the next section, is used to measure the distance between the $\mathcal{S}_{\hat{\boldsymbol{B}}_k^b}$ and $\mathcal{S}_{\boldsymbol{B}_k}$, and $\mathcal{S}_{\hat{\boldsymbol{A}}_l^b}$ and $\mathcal{S}_{\boldsymbol{A}_l}$. Given $\{(\mathbf{X}_i, \mathbf{Y}_i)\},\ 1 \le i \le n$, the following procedure can be used to estimate the dimensions of the DCS:

(1) fix $(k, l)$, calculate the $\mathcal{S}_{\hat{\boldsymbol{B}}_k}$ and $\mathcal{S}_{\hat{\boldsymbol{A}}_l}$ based on the original data;
(2) from $\{(\mathbf{X}_i, \mathbf{Y}_i)\},\ 1 \le i \le n$, generate $N$ bootstrap samples each with size $n$, denote by $\{(\mathbf{X}_i^{(j)}, \mathbf{Y}_i^{(j)})\}$ for $1 \le j \le N$;
(3) for each bootstrap sample $\{(\mathbf{X}_i^{(j)}, \mathbf{Y}_i^{(j)})\}$ for $1 \le j \le N$, calculate the bootstrap subspace $\mathcal{S}_{\hat{\boldsymbol{B}}_k^{b(j)}}$ and $\mathcal{S}_{\hat{\boldsymbol{A}}_l^{b(j)}}$, for $1 \le j \le N$;
(4) calculate the distance $\Delta_m(\mathcal{S}_{\hat{\boldsymbol{B}}_k}, \mathcal{S}_{\hat{\boldsymbol{B}}_k^{b(j)}})$, and $\Delta_m(\mathcal{S}_{\hat{\boldsymbol{A}}_l}, \mathcal{S}_{\hat{\boldsymbol{A}}_l^{b(j)}})$, for $1 \le j \le N$
(5) calculate the average $\Delta_{m,k,l} = [\Delta_m(\mathcal{S}_{\hat{\boldsymbol{B}}_k}, \mathcal{S}_{\hat{\boldsymbol{B}}_k^{b(j)}}) + \Delta_m(\mathcal{S}_{\hat{\boldsymbol{A}}_l}, \mathcal{S}_{\hat{\boldsymbol{A}}_l^{b(j)}})]/2$ for the estimation of the variability of the dual subspace. Find a pair of $(k, l)$ that the smallest value of average $\bar{\Delta}_{m,k,l}$ with smallest standard deviation occurs.

## 3. Numerical studies

In this section, we assess the proposed methods through simulation and real data study. In the simulations, we compare the performance of our methods DCOV0, DCOV1 and DCOV2 with some well-established SDR methods: PRSIR (Li et al. 2008), PRSAVE (Li et al. 2008), and RMAVE-$\mathfrak{F}_C$ (Yin and Li 2011). We choose these three methods because

SIR and SAVE are the most well-known methods in SDR, and RMAVE-$\mathfrak{F}_C$ is the most efficient method for multivariate SDR. We include the results in a sequential way (Yin et al. 2008) of DCOV1 and DCOV2, and denote them as DCOV1-seq and DCOV2-seq, which is to calculate the first single direction, and calculate the second direction in the orthogonal subspace of the first direction and so on. In the simulations for DCS, we compare the performance of six methods for CDCA and DCS.

Two measures of accuracies are used in the simulation study.

(1) Distance between two projection matrices: $\Delta_m(\mathcal{S}_1, \mathcal{S}_2) = \|\mathbf{P}_{\mathcal{S}_1} - \mathbf{P}_{\mathcal{S}_2}\|$ (Li, Zha, and Chiaromonte 2005), where $\|\cdot\|$ is the maximum singular value of a matrix, $\mathcal{S}_1$, $\mathcal{S}_2$ are two subspaces with the same dimensions, and $\mathbf{P}_{\mathcal{S}_1}$ and $\mathbf{P}_{\mathcal{S}_2}$ are the orthogonal projections onto the subspace $\mathcal{S}_1$ and $\mathcal{S}_2$, respectively. The smaller the $\Delta_m$ is, the closer the two subspaces.

(2) The squared vector correlation coefficient: $\rho^2(\hat{\boldsymbol{D}}) = |\boldsymbol{D}^\top \hat{\boldsymbol{D}} \hat{\boldsymbol{D}}^\top \boldsymbol{D}| = \prod_i^p \lambda_i$ (Hotelling 1936), where $\boldsymbol{D}$ and $\hat{\boldsymbol{D}}$ are the true and estimated bases, and $\lambda_i$ are the eigenvalues of $\boldsymbol{D}^\top \hat{\boldsymbol{D}} \hat{\boldsymbol{D}}^\top \boldsymbol{D}$, and $0 \leq \rho(\hat{\boldsymbol{D}}) \leq 1$. The statistic $\rho(\hat{\boldsymbol{D}})$ is a measure of correlation between two subspaces. The larger the $\rho(\hat{\boldsymbol{D}})$ is, the better the estimate is.

Distance between two projection matrices is evaluated for all models, and the squared vector correlation coefficient is used for CDCA (Model 4) and DCS (Model 5).

The R package Nlcoptim (Chen and Yin 2018) is used to solve the above nonlinear optimisation problem. This package implements an SQP method to solve nonlinear optimisation problems with nonlinear objective and nonlinear constraint function. The initial value for the optimisation problem can be generated randomly, but it is not efficient when the dimension of $\boldsymbol{X}$ is large, since we need variation on each parameter. Thus, we suggest to use the SIR and SAVE estimates and choose the one which gives the larger squared distance covariance as the initial values. Codes are available upon request to the first author.

### 3.1. Simulations

Here we simulate five models. The first three models are for the multivariate response, with the relationships of $\boldsymbol{Y}$ and $\boldsymbol{X}$ linear (Model 1), quadratic (Model 2), and other nonlinear (Model 3). We use these models to demonstrate that our methods perform well for linear and nonlinear relationships between two random vectors. The forth model is for CDCA, where $d_x = d_y = 1$ and the relationship between this pair of vectors is quadratic. This model is used to confirm that CDCA outperforms CCA at nonlinear setting. And the last model is for DCS, where $d_x \neq d_y$, and we use it to demonstrate that our methods work well in finding dual central subspace.

For the first three models, 100 replicates of the data are generated. The comparison is made for three sample size $n = 100$, 200 and 400. For PRSIR and PRSAVE, we use $m = 200$ random directions; for RMAVE-$\mathfrak{F}_C$, we take $m = 100$ random directions; and for DCOV1 and DCOV2, we use $m = 50$ after plotting the number of random directions and accuracy via the first three models in this section under design part (1): standard multivariate normal distribution (not reported here). We consider four different designs on predictors for each model to examine if the model assumption can go beyond normal distribution. The four

designs are: part (1), standard multivariate normal predictors, $\boldsymbol{X} \sim N(\mathbf{0}, \boldsymbol{I})$; part (2), non-normal but continuous predictors; part (3), discrete predictors, and part (4), multivariate normal predictors with covariance, $\boldsymbol{X} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\rho)$. The following three models come from Li et al. (2008)[Respective their examples of 4.1, 4.5, and 4.4].

**Model 1:** Let $p = 6, q = 4, \boldsymbol{X} \sim N(\mathbf{0}, \mathbf{I}_6)$. Generate $\boldsymbol{Y}$ as

$$\begin{aligned} Y_1 &= \boldsymbol{\beta}_1^{\mathrm{T}} \boldsymbol{X} + \epsilon_1, \\ Y_2 &= \boldsymbol{\beta}_2^{\mathrm{T}} \boldsymbol{X} + \epsilon_2, \\ Y_3 &= \epsilon_3, \\ Y_4 &= \epsilon_4. \end{aligned}$$

where $\boldsymbol{\beta}_1 = (1, 0, 0, 0, 0, 0)^{\mathrm{T}}$, $\boldsymbol{\beta}_2 = (0, 2, 1, 0, 0, 0)^{\mathrm{T}}$, and $\boldsymbol{\epsilon} \sim N_4(\mathbf{0}, \boldsymbol{\Sigma})$ with

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & -.5 & \mathbf{0} \\ -.5 & 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_2 \end{bmatrix}$$

In this model, $d = 2$. In part (1), $\boldsymbol{X}$ follows the standard normal distribution; in part (2), $X_i \sim \mathit{Unif}(-\sqrt{3}, \sqrt{3})$, for $i = 1, \ldots, 6$; in part (3), $X_i \sim \mathit{Poisson}(1)$, for $i = 1, \ldots, 3$, and $X_i \sim N(0, 1)$, for $i = 4, \ldots, 6$, and in part (4), $\boldsymbol{X} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\rho)$, where $\boldsymbol{\Sigma}_\rho = (\sigma_{ij} = (\rho^{|i-j|}))$ and $\rho = 0.5$. For this model, Table 2 gives the mean and standard deviation of the estimation accuracy ($\Delta_m$) based on $N = 100$ simulated samples for each combination of eight methods and three sample sizes.

PRSIR performs relatively well, because the response is a linear function of the predictors. DCOV0 performs the best for the standard normal design and second best for the other three designs (very close to the top one: DCOV2-seq). RMAVE-$\mathfrak{F}_C$ performs relatively well. DCOV1 and DCOV2 are not better than DCOV0 in all cases; this may be due to the fact that the objective functions in DCOV1 and DCOV2 in the optimisation problem are much more complicated than DCOV0. As the sample size increases, the error decreases substantially for all methods but DCOV1-seq, reflecting the fact that they are consistent, while DCOV1-seq may not be stable.

**Model 2:** Let $p = 6, q = 4, \boldsymbol{X} \sim N(\mathbf{0}, \mathbf{I}_6)$. Generate $\boldsymbol{Y}$ as

$$\begin{aligned} Y_1 &= 1 + (\boldsymbol{\beta}_1^{\mathrm{T}} \boldsymbol{X})^2 + \epsilon_1, \\ Y_2 &= \boldsymbol{\beta}_2^{\mathrm{T}} \boldsymbol{X} + \epsilon_2, \\ Y_3 &= \epsilon_3, \\ Y_4 &= \epsilon_4, \end{aligned}$$

where $\boldsymbol{\beta}_1 = (1, 0, 0, 0, 0, 0)^{\mathrm{T}}$, $\boldsymbol{\beta}_2 = (0, 2, 1, 0, 0, 0)^{\mathrm{T}}$, and $\boldsymbol{\epsilon} \sim N_4(\mathbf{0}, \boldsymbol{\Sigma})$ with

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & -.5 & \mathbf{0} \\ -.5 & 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_2 \end{bmatrix}.$$

**Table 2.** Comparison based on Model 1.

| | | Part (1) | | Part (2) | | Part (3) | | Part (4) | |
|---|---|---|---|---|---|---|---|---|---|
| $n$ | Method | $\bar{\Delta}_m$ | $SE_{\Delta_m}$ | $\bar{\Delta}_m$ | $SE_{\Delta_m}$ | $\bar{\Delta}_m$ | $SE_{\Delta_m}$ | $\bar{\Delta}_m$ | $SE_{\Delta_m}$ |
| 100 | PRSIR | 0.4001 | 0.1773 | 0.5552 | 0.1458 | 0.3918 | 0.1889 | 0.4746 | 0.1929 |
| | PRSAVE | 0.9097 | 0.0977 | 0.9398 | 0.0950 | 0.9273 | 0.1125 | 0.9057 | 0.1245 |
| | RMAVE-$\mathfrak{F}_C$ | 0.4144 | 0.1283 | 0.4387 | 0.1492 | 0.5454 | 0.2420 | 0.6214 | 0.2703 |
| | DCOV0 | 0.2730 | 0.0923 | 0.2609 | 0.0940 | 0.2373 | 0.1051 | 0.4044 | 0.1410 |
| | DCOV1 | 0.5831 | 0.2267 | 0.7695 | 0.1743 | 0.4993 | 0.2611 | 0.7137 | 0.2036 |
| | DCOV2 | 0.4083 | 0.2136 | 0.4696 | 0.2448 | 0.3729 | 0.2372 | 0.5342 | 0.2312 |
| | DCOV1-seq | 0.4956 | 0.2088 | 0.4709 | 0.2131 | 0.5840 | 0.2802 | 0.6624 | 0.2079 |
| | DCOV2-seq | 0.3536 | 0.1934 | 0.2287 | 0.0855 | 0.2040 | 0.1737 | 0.4454 | 0.1750 |
| 200 | PRSIR | 0.3372 | 0.1520 | 0.4938 | 0.1540 | 0.3313 | 0.1767 | 0.4149 | 0.2007 |
| | PRSAVE | 0.7482 | 0.2174 | 0.8891 | 0.1587 | 0.8237 | 0.1985 | 0.8564 | 0.1476 |
| | RMAVE-$\mathfrak{F}_C$ | 0.2595 | 0.0817 | 0.3378 | 0.1441 | 0.3263 | 0.0828 | 0.4138 | 0.2136 |
| | DCOV0 | 0.1943 | 0.0624 | 0.1825 | 0.0583 | 0.1453 | 0.0545 | 0.3196 | 0.0798 |
| | DCOV1 | 0.4802 | 0.2437 | 0.4232 | 0.2395 | 0.5873 | 0.2922 | 0.6587 | 0.2304 |
| | DCOV2 | 0.2745 | 0.2102 | 0.3816 | 0.2349 | 0.3129 | 0.2373 | 0.3435 | 0.2031 |
| | DCOV1-seq | 0.5028 | 0.2294 | 0.4312 | 0.2248 | 0.6700 | 0.2893 | 0.6264 | 0.2320 |
| | DCOV2-seq | 0.3699 | 0.1669 | 0.1621 | 0.0972 | 0.1190 | 0.1272 | 0.2803 | 0.1004 |
| 400 | PRSIR | 0.2985 | 0.1603 | 0.4314 | 0.1136 | 0.3025 | 0.1572 | 0.4190 | 0.2131 |
| | PRSAVE | 0.6288 | 0.2392 | 0.8773 | 0.1594 | 0.8047 | 0.2153 | 0.7562 | 0.2003 |
| | RMAVE-$\mathfrak{F}_C$ | 0.1963 | 0.0602 | 0.1797 | 0.0771 | 0.2507 | 0.0918 | 0.2764 | 0.1826 |
| | DCOV0 | 0.1412 | 0.0417 | 0.1308 | 0.0399 | 0.0945 | 0.0366 | 0.2615 | 0.0580 |
| | DCOV1 | 0.4470 | 0.2576 | 0.5143 | 0.2588 | 0.6652 | 0.3203 | 0.5840 | 0.2283 |
| | DCOV2 | 0.2114 | 0.1860 | 0.3167 | 0.2165 | 0.2806 | 0.2151 | 0.3682 | 0.2221 |
| | DCOV1-seq | 0.5251 | 0.2163 | 0.5220 | 0.2777 | 0.7447 | 0.2814 | 0.5501 | 0.2475 |
| | DCOV2-seq | 0.3449 | 0.1886 | 0.1298 | 0.0437 | 0.0766 | 0.1033 | 0.2312 | 0.1370 |

This model is the same as Model 1 except that $Y_1$ is a quadratic form of $\boldsymbol{\beta}_1^{\mathrm{T}}\boldsymbol{X}$. Again, $d=2$. Table 3 reports the results for Model 2. We can see that except the discrete design, PRSAVE performs better than PRSIR, because $Y_1$ has a quadratic function of $\boldsymbol{\beta}_1^{\mathrm{T}}\boldsymbol{X}$. However, they are not stable across the designs, neither do DCOV1 and DCOV2. DCOV1-seq and DCOV2-seq perform well except for the non-normal design. For large samples, DCOV0 performs the best on the discrete design and second best on the standard normal and non-normal designs, and RMAVE-$\mathfrak{F}_C$ performs the best on the non-normal design and the second best on the correlated normal design.

**Model 3:** Let $p=6$, $q=5$, $\boldsymbol{X}\sim N(\mathbf{0},\mathbf{I}_6)$. Generate $\boldsymbol{Y}$ as

$$
\begin{aligned}
Y_1 &= X_2 + \frac{3x_2}{.5+(X_1+1.5)^2} + \epsilon_1,\\
Y_2 &= X_1 + \mathrm{e}^{.5X_2} + \epsilon_2,\\
Y_3 &= X_1 + X_2 + \epsilon_3,\\
Y_4 &= \epsilon_4,\\
Y_5 &= \epsilon_5.
\end{aligned}
$$

where $\boldsymbol{\epsilon}\sim N_5(\mathbf{0},\boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}=\mathrm{diag}(\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2)$

$$
\boldsymbol{\Sigma}_1=\begin{bmatrix}1 & -.5\\ -.5 & 0.5\end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma}_2=\begin{bmatrix}1/2 & 0 & 0\\ 0 & 1/3 & 0\\ 0 & 0 & 1/4\end{bmatrix}.
$$

**Table 3.** Comparison based on Model 2.

| | | Part (1) | | Part (2) | | Part (3) | | Part (4) | |
|---|---|---|---|---|---|---|---|---|---|
| $n$ | Method | $\bar{\Delta}_m$ | $SE_{\Delta_m}$ | $\bar{\Delta}_m$ | $SE_{\Delta_m}$ | $\bar{\Delta}_m$ | $SE_{\Delta_m}$ | $\bar{\Delta}_m$ | $SE_{\Delta_m}$ |
| 100 | PRSIR | 0.8851 | 0.1563 | 0.8503 | 0.1597 | 0.2685 | 0.0815 | 0.8067 | 0.2014 |
| | PRSAVE | 0.8235 | 0.1902 | 0.6010 | 0.2377 | 0.8765 | 0.1632 | 0.8227 | 0.1627 |
| | RMAVE-$\mathfrak{F}_C$ | 0.4486 | 0.2715 | 0.6451 | 0.5345 | 0.3092 | 0.0720 | 0.9359 | 0.3878 |
| | DCOV0 | 0.5547 | 0.2951 | 0.7795 | 0.2317 | 0.1819 | 0.0609 | 0.6292 | 0.246 |
| | DCOV1 | 0.5316 | 0.2543 | 0.8683 | 0.1528 | 0.2665 | 0.0809 | 0.5339 | 0.2428 |
| | DCOV2 | 0.5200 | 0.2926 | 0.8695 | 0.1508 | 0.2931 | 0.1511 | 0.7509 | 0.2529 |
| | DCOV1-seq | 0.3328 | 0.1425 | 0.6704 | 0.2221 | 0.2491 | 0.1105 | 0.4249 | 0.1336 |
| | DCOV2-seq | 0.4946 | 0.2937 | 0.6486 | 0.3352 | 0.1640 | 0.0759 | 0.5531 | 0.2599 |
| 200 | PRSIR | 0.8750 | 0.1484 | 0.8087 | 0.1835 | 0.2024 | 0.0648 | 0.7280 | 0.2283 |
| | PRSAVE | 0.3405 | 0.1450 | 0.5542 | 0.1826 | 0.5266 | 0.2175 | 0.5030 | 0.2049 |
| | RMAVE-$\mathfrak{F}_C$ | 0.1963 | 0.0644 | 0.2928 | 0.2912 | 0.2409 | 0.0862 | 0.5149 | 0.3685 |
| | DCOV0 | 0.3527 | 0.2700 | 0.5980 | 0.3017 | 0.1155 | 0.0378 | 0.3800 | 0.1496 |
| | DCOV1 | 0.3612 | 0.2015 | 0.7736 | 0.2233 | 0.1964 | 0.0727 | 0.3866 | 0.2039 |
| | DCOV2 | 0.4011 | 0.3173 | 0.8793 | 0.1641 | 0.2765 | 0.1857 | 0.6440 | 0.2814 |
| | DCOV1-seq | 0.2090 | 0.0697 | 0.4272 | 0.2010 | 0.1952 | 0.0904 | 0.3249 | 0.0716 |
| | DCOV2-seq | 0.2381 | 0.2025 | 0.4869 | 0.3907 | 0.1199 | 0.0651 | 0.3984 | 0.2731 |
| 400 | PRSIR | 0.6731 | 0.2335 | 0.7350 | 0.2212 | 0.1750 | 0.0579 | 0.3497 | 0.1672 |
| | PRSAVE | 0.2262 | 0.0763 | 0.5288 | 0.1589 | 0.4727 | 0.2266 | 0.2869 | 0.1148 |
| | RMAVE-$\mathfrak{F}_C$ | 0.2042 | 0.1482 | 0.1303 | 0.0353 | 0.1845 | 0.0577 | 0.2187 | 0.1515 |
| | DCOV0 | 0.1443 | 0.0466 | 0.3870 | 0.3038 | 0.0733 | 0.0250 | 0.2998 | 0.0718 |
| | DCOV1 | 0.2059 | 0.1547 | 0.6723 | 0.2815 | 0.1514 | 0.0464 | 0.3168 | 0.1170 |
| | DCOV2 | 0.2983 | 0.3333 | 0.8060 | 0.2490 | 0.2489 | 0.1803 | 0.4693 | 0.2773 |
| | DCOV1-seq | 0.1398 | 0.0460 | 0.6594 | 0.2375 | 0.1623 | 0.0712 | 0.2855 | 0.0616 |
| | DCOV2-seq | 0.1712 | 0.1800 | 0.5090 | 0.4282 | 0.1014 | 0.0759 | 0.2132 | 0.1082 |

Again, $d=2$. Table 4 reports the results for Model 3. The four designs are same as in Models 1 and 2. All methods perform well with high accuracies. The majority of the performances for DCOV0 stays on top one or two for all four designs and three sample sizes. RMAVE-$\mathfrak{F}_C$ performs well for the normal design, but not as good as other DCOV0 methods on the non-normal design and discrete design. The errors decrease rapidly when sample size increases for all methods, which means that all estimates are consistent.

To summarise our simulation studies, we conclude that DCOV0 provides the most stable estimator across models with different designs, standard normal, non-normal, discrete and correlated normal predictors, especially when sample size is large.

To illustrate the *kNN* method for estimating $d$, we use the three models under design part (1), set $(n,p) = (400,6)$, and $k=20$. The ratios of $\lambda$ for Models 1, 2, and 3 are summarised in Table 5. It indicates that maximum ratios happen at the dimension of the CS, $d=2$, as we expected.

Now we simulate examples to conduct CDCA. For each model setting, 100 replicates of the data are generated. The comparison is made for three sample size $n = 100, 200$, and 300.

**Model 4:** Let $p=6,q=4, \boldsymbol{X} \sim N(\mathbf{0}, \mathbf{I}_6)$. The four-dimensional response random vector $\boldsymbol{Y}$ is generated as

$$
\begin{aligned}
Y_1 &= 1 + (\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X})^2 + \epsilon_1,\\
Y_2 &= \epsilon_2,\\
Y_3 &= \epsilon_3,\\
Y_4 &= \epsilon_4.
\end{aligned}
$$

**Table 4.** Comparison based on Model 3.

| | | Part (1) | | Part (2) | | Part (3) | | Part (4) | |
|---|---|---|---|---|---|---|---|---|---|
| $n$ | Method | $\bar{\Delta}_m$ | $SE_{\Delta_m}$ | $\bar{\Delta}_m$ | $SE_{\Delta_m}$ | $\bar{\Delta}_m$ | $SE_{\Delta_m}$ | $\bar{\Delta}_m$ | $SE_{\Delta_m}$ |
| 100 | PRSIR | 0.2716 | 0.1048 | 0.7269 | 0.1860 | 0.3071 | 0.1393 | 0.3407 | 0.1271 |
| | PRSAVE | 0.8732 | 0.1547 | 0.5679 | 0.2320 | 0.8142 | 0.1995 | 0.8667 | 0.15555 |
| | RMAVE-$\mathfrak{F}_C$ | 0.2386 | 0.0827 | 0.3131 | 0.1367 | 0.4173 | 0.1538 | 0.5722 | 0.3146 |
| | DCOV0 | 0.2477 | 0.0877 | 0.2002 | 0.0791 | 0.1556 | 0.0902 | 0.4107 | 0.1287 |
| | DCOV1 | 0.2809 | 0.1155 | 0.2012 | 0.0603 | 0.2436 | 0.1124 | 0.5135 | 0.2008 |
| | DCOV2 | 0.2718 | 0.2008 | 0.4010 | 0.2062 | 0.2633 | 0.1983 | 0.5666 | 0.2664 |
| | DCOV1-seq | 0.3032 | 0.0877 | 0.2341 | 0.0758 | 0.2715 | 0.0995 | 0.4711 | 0.1947 |
| | DCOV2-seq | 0.2503 | 0.1242 | 0.1876 | 0.0921 | 0.1049 | 0.0866 | 0.4173 | 0.2092 |
| 200 | PRSIR | 0.2250 | 0.0743 | 0.6074 | 0.1818 | 0.2618 | 0.0903 | 0.2539 | 0.1064 |
| | PRSAVE | 0.5462 | 0.2452 | 0.2944 | 0.1279 | 0.3823 | 0.2069 | 0.7569 | 0.2010 |
| | RMAVE-$\mathfrak{F}_C$ | 0.1700 | 0.0610 | 0.2179 | 0.0885 | 0.2595 | 0.0612 | 0.3373 | 0.1772 |
| | DCOV0 | 0.1569 | 0.0519 | 0.1283 | 0.0414 | 0.0987 | 0.0677 | 0.3369 | 0.0861 |
| | DCOV1 | 0.2170 | 0.1252 | 0.1493 | 0.0440 | 0.1579 | 0.0498 | 0.3643 | 0.1096 |
| | DCOV2 | 0.1573 | 0.1021 | 0.3787 | 0.2134 | 0.2377 | 0.1789 | 0.4399 | 0.2823 |
| | DCOV1-seq | 0.2338 | 0.0800 | 0.1597 | 0.0499 | 0.1986 | 0.0730 | 0.4344 | 0.1635 |
| | DCOV2-seq | 0.1430 | 0.0488 | 0.1331 | 0.0640 | 0.0698 | 0.0686 | 0.2160 | 0.0538 |
| 400 | PRSIR | 0.1660 | 0.0592 | 0.2826 | 0.1235 | 0.2229 | 0.0905 | 0.1981 | 0.0736 |
| | PRSAVE | 0.2855 | 0.1578 | 0.2405 | 0.0876 | 0.1991 | 0.0719 | 0.5068 | 0.2457 |
| | RMAVE-$\mathfrak{F}_C$ | 0.1071 | 0.0367 | 0.1302 | 0.0424 | 0.2069 | 0.0702 | 0.1866 | 0.0845 |
| | DCOV0 | 0.1071 | 0.0319 | 0.0870 | 0.0231 | 0.0633 | 0.0175 | 0.2849 | 0.0715 |
| | DCOV1 | 0.1597 | 0.1061 | 0.1273 | 0.1560 | 0.1189 | 0.0357 | 0.3321 | 0.1178 |
| | DCOV2 | 0.1223 | 0.1579 | 0.3539 | 0.2556 | 0.2235 | 0.1786 | 0.4472 | 0.2823 |
| | DCOV1-seq | 0.1650 | 0.0623 | 0.1182 | 0.0305 | 0.1260 | 0.0414 | 0.3246 | 0.0677 |
| | DCOV2-seq | 0.1114 | 0.0467 | 0.1147 | 0.0822 | 0.0542 | 0.0556 | 0.1461 | 0.0538 |

**Table 5.** Ratio of eigenvalues for Models 1, 2, and 3.

| Model | $r_1 = \lambda_1/\lambda_2$ | $r_2 = \lambda_2/\lambda_3$ | $r_3 = \lambda_3/\lambda_4$ | $r_4 = \lambda_4/\lambda_5$ | $r_5 = \lambda_5/\lambda_6$ |
|---|---|---|---|---|---|
| 1 | 1.0356 | 1.4020[a] | 1.0317 | 1.0651 | 1.0593 |
| 2 | 1.1668 | 1.3368[a] | 1.0551 | 1.0254 | 1.0053 |
| 3 | 1.8969 | 2.1869[a] | 1.0869 | 1.0123 | 1.0381 |

[a] Where the largest ratio occurs.

where $\boldsymbol{\beta} = (0, 2, 1, 0, 0, 0)^{\mathrm{T}}$, and $\boldsymbol{\epsilon} \sim N_4(\mathbf{0}, \boldsymbol{\Sigma})$ with

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & -.5 & \mathbf{0} \\ -.5 & 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_2 \end{bmatrix}.$$

For this model, $d_x = d_y = 1$. Therefore, we use $\boldsymbol{a}$ and $\boldsymbol{b}$ to denote $\mathbf{A}$ and $\mathbf{B}$, and use $\hat{\boldsymbol{a}}$ and $\hat{\boldsymbol{b}}$ to denote $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$. The results for Model 4 is shown in Table 6. All DCOV methods perform better than traditional CCA with lower $\bar{\Delta}_m(\hat{\boldsymbol{b}})$ and $\bar{\Delta}_m(\hat{\boldsymbol{a}})$, and higher $\bar{\rho}(\hat{\boldsymbol{b}})$ and $\bar{\rho}(\hat{\boldsymbol{a}})$, this is because the relationship in this model is quadratic, and CCA cannot capture this nonlinear pattern. The performance of recovering the directions gets better when the sample size increases, indicating consistency. Overall, Approach 2 outperforms Approach 1, which suggests that recovering $\boldsymbol{a}$ based on the distance covariance of $\boldsymbol{b}^{\top}\boldsymbol{X}$ is better than based on that of $\boldsymbol{X}$. When the relationship between $\boldsymbol{X}$ and $\boldsymbol{Y}$ is linear, CCA performs as well as the DCOV methods under both estimation approaches (not reported here).

We then investigate the performance of our methods for DCS using an example that is similar to the examples provided by Wang et al. (2015). The relationships between two

**Table 6.** Comparison based on Model 4.

| $n$ | Order | DCOV | $\bar{\Delta}_m(\hat{\boldsymbol{a}})$ | $SE_{\Delta_m}(\hat{\boldsymbol{a}})$ | $\bar{\rho}(\hat{\boldsymbol{a}})$ | $SE_{\rho}(\hat{\boldsymbol{a}})$ | $\bar{\Delta}_m(\hat{\boldsymbol{b}})$ | $SE_{\Delta_m}(\hat{\boldsymbol{b}})$ | $\bar{\rho}(\hat{\boldsymbol{b}})$ | $SE_{\rho}(\hat{\boldsymbol{b}})$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 100 | Approach1 | DCOV0 | 0.4209 | 0.2876 | 0.7409 | 0.3041 | 0.4663 | 0.2214 | 0.7339 | 0.2374 |
| | | DCOV1 | 0.2337 | 0.0976 | 0.9359 | 0.0579 | 0.3055 | 0.1470 | 0.8852 | 0.1146 |
| | | DCOV2 | 0.1413 | 0.1067 | 0.9687 | 0.0964 | 0.6944 | 0.2120 | 0.4732 | 0.2736 |
| | Approach2 | DCOV0 | 0.4209 | 0.2876 | 0.7409 | 0.3041 | 0.4757 | 0.2461 | 0.7136 | 0.2628 |
| | | DCOV1 | 0.2337 | 0.0976 | 0.9359 | 0.0579 | 0.1689 | 0.0726 | 0.9662 | 0.0272 |
| | | DCOV2 | 0.1413 | 0.1067 | 0.9687 | 0.0964 | 0.3087 | 0.1013 | 0.8944 | 0.0731 |
| | | CCA | 0.8160 | 0.1927 | 0.2974 | 0.2778 | 0.9738 | 0.0675 | 0.0472 | 0.1077 |
| 200 | Approach1 | DCOV0 | 0.1818 | 0.0919 | 0.9585 | 0.0435 | 0.2929 | 0.1494 | 0.8920 | 0.1117 |
| | | DCOV1 | 0.1645 | 0.0575 | 0.9696 | 0.0200 | 0.2568 | 0.1316 | 0.9169 | 0.0929 |
| | | DCOV2 | 0.1066 | 0.1175 | 0.9749 | 0.1056 | 0.5374 | 0.2199 | 0.6632 | 0.2508 |
| | Approach2 | DCOV0 | 0.1818 | 0.0919 | 0.9585 | 0.0435 | 0.3261 | 0.1756 | 0.8630 | 0.1532 |
| | | DCOV1 | 0.1645 | 0.0575 | 0.9696 | 0.0200 | 0.1366 | 0.0528 | 0.9785 | 0.0168 |
| | | DCOV2 | 0.1066 | 0.1175 | 0.9749 | 0.1056 | 0.2727 | 0.0966 | 0.9163 | 0.0961 |
| | | CCA | 0.7878 | 0.1962 | 0.3413 | 0.2845 | 0.9773 | 0.0337 | 0.0437 | 0.0637 |
| 300 | Approach1 | DCOV0 | 0.1393 | 0.0506 | 0.9780 | 0.0159 | 0.2529 | 0.1107 | 0.9238 | 0.0652 |
| | | DCOV1 | 0.1382 | 0.0548 | 0.9779 | 0.0188 | 0.1945 | 0.0841 | 0.9551 | 0.0400 |
| | | DCOV2 | 0.0882 | 0.0871 | 0.9846 | 0.0509 | 0.4032 | 0.1831 | 0.8041 | 0.1847 |
| | Approach2 | DCOV0 | 0.1393 | 0.0506 | 0.9780 | 0.0159 | 0.2543 | 0.1139 | 0.9224 | 0.0686 |
| | | DCOV1 | 0.1382 | 0.0548 | 0.9779 | 0.0188 | 0.1144 | 0.0406 | 0.9852 | 0.0099 |
| | | DCOV2 | 0.0882 | 0.0871 | 0.9846 | 0.0509 | 0.2558 | 0.0569 | 0.9313 | 0.0291 |
| | | CCA | 0.7521 | 0.2184 | 0.3872 | 0.3066 | 0.9726 | 0.0608 | 0.0503 | 0.0956 |

random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ are linear as well as nonlinear. For each example setting, 100 replicates of the data are generated. The comparison is made for three sample sizes $n = 100, 200, 300$. For the two projective resampling-based methods DCOV1 and DCOV2, we choose the number of random directions $m = 50$, and transfer the multivariate response to 50 univariate responses. For DCOV0, we treat the response as it is – multivariate form. The following models are considered:

**Model 5:** Let $p = 5, q = 4, \boldsymbol{X} \sim N(\boldsymbol{0}, \mathbf{I}_5)$. The four-dimensional response random vector $\boldsymbol{Y}$ is generated as

$$
\begin{aligned}
Y_1 &= 4\cos(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}) + 0.3\epsilon_1,\\
Y_2 &= \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X} + 0.5\epsilon_2,\\
Y_3 &= \epsilon_3,\\
Y_4 &= \epsilon_4.
\end{aligned}
$$

where $\boldsymbol{\beta} = (1, 1, 0, 0, 0)^{\mathrm{T}}$ and $\boldsymbol{\epsilon} \sim N_4(\boldsymbol{0}, I_4)$. In this model, $\boldsymbol{A}_1 = (1, 0, 0, 0)^{\mathrm{T}}$, $\boldsymbol{A}_2 = (0, 1, 0, 0)^{\mathrm{T}}$, $\boldsymbol{A} = (\boldsymbol{A}_1, \boldsymbol{A}_2)$, and $\boldsymbol{B} = \boldsymbol{\beta}$.

Table 7 shows the results of six methods to recover DCS for Model 1. All the six methods perform well. For each method, $\bar{\Delta}_m$ decreases and $\bar{\rho}$ increases with the sample size increase, indicating consistency. Overall, Approach 2 outperforms Approach 1, suggesting that recovering $\boldsymbol{A}$ based on the distance covariance of $\boldsymbol{B}^{\top}\boldsymbol{X}$ is better then based on that of $\boldsymbol{X}$. A dataset of size $n = 300$ is selected to illustrate the bootstrap method, with 100 bootstrap iterations. Table 8 shows the results for different $(k, l)$. We are looking for the smallest mean and least variability. We can see that $(k, l) = (1, 1)$ produces the smallest mean but $(k, l) = (1, 2)$ has a very close mean! In such a case, the variability plays an important role:

**Table 7.** Comparison based on Model 5.

| $n$ | Order | DCOV | $\bar{\Delta}_m(\hat{\boldsymbol{A}})$ | $SE_{\Delta_m}(\hat{\boldsymbol{A}})$ | $\bar{\rho}(\hat{\boldsymbol{A}})$ | $SE_{\rho}(\hat{\boldsymbol{A}})$ | $\bar{\Delta}_m(\hat{\boldsymbol{B}})$ | $SE_{\Delta_m}(\hat{\boldsymbol{B}})$ | $\bar{\rho}(\hat{\boldsymbol{B}})$ | $SE_{\rho}(\hat{\boldsymbol{B}})$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 100 | Approach1 | DCOV0 | 0.1166 | 0.0410 | 0.9847 | 0.0107 | 0.2962 | 0.1943 | 0.8714 | 0.1833 |
| | | DCOV1 | 0.1154 | 0.0443 | 0.9847 | 0.0115 | 0.2880 | 0.1786 | 0.8799 | 0.1656 |
| | | DCOV2 | 0.1262 | 0.1130 | 0.9714 | 0.0635 | 0.4409 | 0.2238 | 0.7521 | 0.2317 |
| | Approach2 | DCOV0 | 0.1166 | 0.0410 | 0.9847 | 0.0107 | 0.3106 | 0.1778 | 0.8687 | 0.1578 |
| | | DCOV1 | 0.1154 | 0.0443 | 0.9847 | 0.0115 | 0.0938 | 0.0410 | 0.9881 | 0.0099 |
| | | DCOV2 | 0.1262 | 0.1130 | 0.9714 | 0.0635 | 0.5510 | 0.2429 | 0.6181 | 0.2833 |
| 200 | Approach1 | DCOV0 | 0.0790 | 0.0299 | 0.9928 | 0.0052 | 0.2069 | 0.1248 | 0.9394 | 0.0824 |
| | | DCOV1 | 0.0817 | 0.0286 | 0.9925 | 0.0049 | 0.2161 | 0.1252 | 0.9353 | 0.0806 |
| | | DCOV2 | 0.0696 | 0.0863 | 0.9877 | 0.0714 | 0.4069 | 0.2284 | 0.7800 | 0.2422 |
| | Approach2 | DCOV0 | 0.0790 | 0.0299 | 0.9928 | 0.0052 | 0.2009 | 0.1117 | 0.9448 | 0.0617 |
| | | DCOV1 | 0.0817 | 0.0286 | 0.9925 | 0.0049 | 0.0651 | 0.0225 | 0.9947 | 0.0035 |
| | | DCOV2 | 0.0696 | 0.0863 | 0.9877 | 0.0714 | 0.2946 | 0.2366 | 0.8466 | 0.2112 |
| 300 | Approach1 | DCOV0 | 0.0657 | 0.0271 | 0.9949 | 0.0037 | 0.1443 | 0.0784 | 0.9716 | 0.0370 |
| | | DCOV1 | 0.0684 | 0.0234 | 0.9947 | 0.0036 | 0.1605 | 0.0919 | 0.9633 | 0.0431 |
| | | DCOV2 | 0.0621 | 0.0615 | 0.9923 | 0.0339 | 0.3702 | 0.2236 | 0.8122 | 0.2109 |
| | Approach2 | DCOV0 | 0.0657 | 0.0271 | 0.9949 | 0.0037 | 0.1451 | 0.0767 | 0.9716 | 0.0364 |
| | | DCOV1 | 0.0684 | 0.0234 | 0.9947 | 0.0036 | 0.0540 | 0.0208 | 0.9963 | 0.0027 |
| | | DCOV2 | 0.0621 | 0.0615 | 0.9923 | 0.0339 | 0.2018 | 0.1846 | 0.9189 | 0.1363 |

**Table 8.** Bootstrap distance measure for Models 5.

| $k$ | $l$ | $\bar{\Delta}_{m,k,l}$ | $SE_{\Delta_{m,k,l}}$ |
|---|---|---|---|
| 1 | 1 | 0.1489 | 0.1409 |
| 1 | 2 | 0.1604 | 0.0175 |
| 1 | 3 | 0.4215 | 0.1260 |
| 2 | 1 | 0.4402 | 0.1037 |
| 2 | 2 | 0.4745 | 0.1170 |
| 2 | 3 | 0.7224 | 0.1985 |
| 3 | 1 | 0.4240 | 0.0942 |
| 3 | 2 | 0.4970 | 0.1214 |
| 3 | 3 | 0.7746 | 0.1828 |

our ad hoc experience suggests that one should choose the dimension with the least variability! Note that $(k, l) = (1, 2)$ has the least variability (0.0175 vs. 0.1409 of $(k, l) = (1, 1)$). So we will choose $(k, l) = (1, 2)$, which agrees with the true dimension of DCS.

## 3.2. Application

In this section, we analyse the Minneapolis elementary schools data set (Cook 1998, p. 216) and the LA pollution data set (Shumway, Azari, and Pawitan 1988), to illustrate the DCOV methods and DCS approaches, respectively.

### 3.2.1. Minneapolis elementary schools data

These data were used to explore the relationship between students' performance and characteristics of school. It has 63 observations (schools) and 13 variables. The response is a four-dimensional multivariate response, which is described as

- 4BELOW: percentage of 4th graders scoring BELOW average on a standard 4th grade vocabulary test in 1972.

**Table 9.** Ratio of eigenvalues **r** for Minneapolis elementary schools data.

| | $r_1$ | $r_2$ | $r_3$ | $r_4$ | $r_5$ | $r_6$ | $r_7$ |
|---|---|---|---|---|---|---|---|
| Ratio-original | 1.812453 | 1.323173 | 1.384807 | 1.508781 | 1.525245 | 1.202717 | 1.074232 |
| Ratio-sqrt transformation | 5.917090 | 1.149957 | 1.117052 | 1.308639 | 1.153622 | 1.229887 | 1.236647 |

- 4ABOVE: percentage of 4th graders scoring ABOVE average on a standard 4th grade vocabulary test in 1972.
- 6BELOW: percentage of 6th graders scoring BELOW average on a standard 6th grade comprehension test in 1972.
- 6ABOVE: percentage of 6th graders scoring ABOVE average on a standard 6th grade comprehension test in 1972.

And the explanatory variables are

- BP: percentage of children in the school living with Both Parents
- AFDC: percentage of children receiving Aid to Families with Dependent Children
- Poverty: percentage of persons in the school area who are above the federal poverty levels
- HSchl: percentage of adults in the school area who have completed high school
- Attend: average percentage of children in attendance during the year
- Mobility: percentage of children who started in a school, but did not finish there
- PT-ratio: pupil–teacher ratio
- Minority: percentage minority children in the area.

These data were analysed by Yin and Bura (2006) to demonstrate their moment-based SDR method for the multivariate response. In order to satisfy the two assumptions of their method, they used square-root transform on the response as well as the explanatory variables of percentages. The DCOV method does not require the assumption of distribution, so we can perform the dimension reduction on the original data. But in order to compare our result to the work of Yin and Bura (2006), we use the transformed data, where all the response variable and percentages are square-root transformed. The kNN method described in Section 2.5 results in Table 9. The maximum ratios for both cases happen at one. Thus we conclude $d = 1$, which also agrees with the analysis of Yin and Bura (2006). Table 10 shows the estimated directions at the original scale and the transformed scale. AFDC and HSchl contribute most to the estimated direction for the original scale, and $\sqrt{AFDC}$ and $\sqrt{HSchl}$ contribute most to the estimated direction for the transformed scale.

### 3.2.2. LA pollution data

These data are obtained from Shumway et al. (1988) and were used to explore the effects of temperature and pollution on daily mortality in Los Angeles (LA). The data set has 508 observations and 11 variables (daily records from 1970 to 1979). These 11 variables include 3 mortality measures (total mortality, respiratory mortality and cardiovascular mortality) which counted all deaths of LA area, 2 weather measures (temperature and relative humidity), and 6 pollution measures include carbon monoxide, sulfur dioxide,

**Table 10.** Estimated direction by DCOV1 on original data and transformed data.

| Variables | $\hat{\beta}_1$ (original) | $\hat{\beta}_1$ (transformed) |
|---|---|---|
| BP | −0.1444 | −0.0631 |
| AFDC | 0.6569 | −0.7443 |
| Poverty | −0.0820 | 0.0741 |
| HSchl | −0.6608 | 0.5218 |
| Attend | −0.0385 | 0.0376 |
| Mobility | 0.2546 | −0.1502 |
| PT-ratio | −0.0535 | 0.1948 |
| Minority | 0.1873 | −0.3198 |

**Table 11.** Bootstrap distance measure for LA pollution data.

| $k$ | $l$ | $\bar{\Delta}_{m,k,l}$ | $SE_{\Delta_{m,k,l}}$ |
|---|---|---|---|
| 1 | 1 | 0.1381 | 0.0257 |
| 1 | 2 | 0.5430 | 0.1073 |
| 2 | 1 | 0.3398 | 0.1224 |
| 2 | 2 | 0.4486 | 0.1342 |
| 3 | 1 | 0.4954 | 0.0883 |
| 3 | 2 | 0.7428 | 0.2067 |

**Table 12.** Estimated direction of the multivariate response and predictors by Approch2 DCOV0 for LA pollution data.

| Variables | $\hat{\boldsymbol{A}}$ | Variables | $\hat{\boldsymbol{B}}$ |
|---|---|---|---|
| Total mortality | 0.5875 | Temperature | 0.6585 |
| Respiratory mortality | 0.0075 | Relative humidity | 0.2534 |
| Cardiovascular mortality | 0.8095 | Carbon monoxide | −0.4162 |
| | | Hydrocarbons | −0.3917 |
| | | Ozone | 0.1255 |
| | | Particulates | −0.3998 |

nitrogen dioxide, hydrocarbons, ozone, and particulates. The data are also discussed by Iaci et al. (2010, 2015).

We apply our method to the data to identify the DCS with the mortality variables as the multivariate response, and two weather measures and four pollution measures as predictors. Note that sulfur dioxide, nitrogen dioxide are excluded since they are highly correlated with other predictors. Thus the multivariate response vector is $\boldsymbol{Y} = (Y_1, Y_2, Y_3)^\top$, where $Y_1$ = total mortality, $Y_2$ = respiratory mortality and $Y_3$ = cardiovascular mortality; the predictor vector $\boldsymbol{X} = (X_1, \ldots, X_6)^\top$, where $X_1$ = temperature, $X_2$ = relative humidity, $X_3$ = carbon monoxide, $X_4$ = hydrocarbons, $X_5$ = ozone, $X_6$ = particulates.

Table 11 shows results from the bootstrap method of Section 2.5 which estimates the dimension of the DCS to be $(k, l) = (1, 1)$. Table 12 shows the estimated directions of the multivariate response and predictors. The loadings for the multivariate response indicate that $Y_1$ and $Y_3$ contribute the most to the estimated direction, while $Y_2$ does not contribute to $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$. For the estimated direction corresponding to the subspace $\mathcal{S}_{\mathbf{X}|\mathbf{Y}}$, $X_1$ contribute most positive to the estimated direction, $X_3$, $X_4$ and $X_6$ contribute equally negative to the estimated direction. The plot of the estimated directions of the multivariate response and

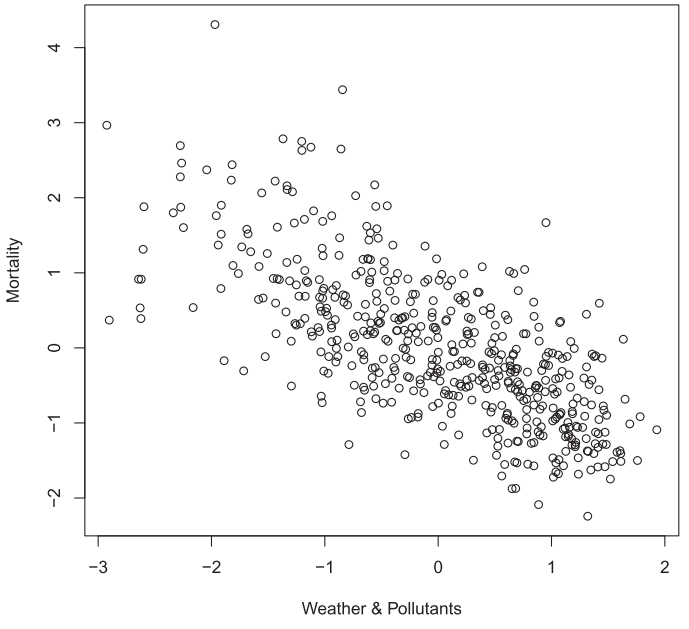


**Figure 1.** Relationship of direction of multivariate response and direction of predictors.

the predictors is in Figure 1, indicating a linear relationship between these two estimated directions, which agrees with Iaci et al. (2015).

## 4. Discussion

In this article, we develop DCOV0 for SDR with a multivariate response. We present DCOV1 and DCOV2 by using projective resampling idea. DCOV0 performs well on different models by stably achieving highest accuracy. DCOV1 and DCOV2 perform relatively well, better than projection resampling with SIR and SAVE. In addition, we introduced a *kNN* method for estimating $d$, showing that this approach is quite useful in estimating the dimension under different models. We extend the DCOV method to CDCA. Comparing to the traditional CCA, our methods can capture nonlinear relationship. We recovered DCS using DCOV. The results show that all DCOV methods estimate the central dual subspaces with high accuracy by our simulation.

## Acknowledgments

The authors would like to thank the Editor, an Associate Editor and two referees for their valuable comments and suggestions, which lead to a greatly improved paper.

## Disclosure statement

No potential conflict of interest was reported by the authors.

## Funding

Yin's research was supported in part by National Science Foundation (Directorate for Computer and Information Science and Engineering) Grant CIF-1813330.

## References

Burg, E., and Leeuw, J. (1983), 'Non-linear Canonical Correlation', *British Journal of Mathematical and Statistical Psychology*, 36, 54–80.

Chen, X., and Yin, X. (2018), 'NlcOptim: An R Package for Nonlinear Constrained Optimization Program', *Journal of Statistical Software*.

Cook, R.D. (1994), 'On the Interpretation of Regression Plots', *Journal of the American Statistical Association*, 89, 177–189.

Cook, R.D. (1996), 'Graphics for Regressions With a Binary Response', *Journal of the American Statistical Association*, 91, 983–992.

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

Cook, R.D., and Ni, L. (2005), 'Sufficient Dimension Reduction via Inverse Regression: A Minimum Discrepancy Approach', *Journal of the American Statistical Association*, 100, 410–428.

Cook, R.D., and Setodji, C.M. (2003), 'A Model-free Test for Reduced Rank in Multivariate Regression', *Journal of the American Statistical Association*, 98, 340–351.

Cook, R., and Weisberg, S. (1991), 'Discussion of a Paper by KC Li', *Journal of the American Statistical Association*, 86, 328–32.

Cook, R.D. and Yin, X. (2001), 'Dimension Reduction and Visualization in Discriminant Analysis (with Discussion)', *Australian and New Zealand Journal of Statistics*, 43, 147–199.

Fung, W.K., He, X., Liu, L., and Shi, P. (2002), 'Dimension Reduction Based on Canonical Correlation', *Statistica Sinica*, 12, 1093–1113.

Hilafu, H., and Yin, X. (2013), 'Sufficient Dimension Reduction in Multivariate Regressions with Categorical Predictors', *Computational Statistics & Data Analysis*, 63, 139–147.

Hotelling, H. (1936), 'Relations Between Two Sets of Variates', *Biometrika*, 28, 321–377.

Hristache, M., Juditsky, A., Polzehl, J., and Spokoiny, V. (2001), 'Structure Adaptive Approach for Dimension Reduction', *The Annals of Statistics*, 29, 1537–1566.

Iaci, R., Sriram, T., and Yin, X. (2010), 'Multivariate Association and Dimension Reduction: A Generalization of Canonical Correlation Analysis', *Biometrics*, 66, 1107–1118.

Iaci, R., Yin, X., Sriram, T., and Klingenberg, C.P. (2008), 'An Informational Measure of Association and Dimension Reduction for Multiple Sets and Groups with Applications in Morphometric Analysis', *Journal of the American Statistical Association*, 103, 1166–1176.

Iaci, R., Yin, X., and Zhu, L. (2015), 'The Dual Central Subspaces in Dimension Reduction', *Journal of Multivariate Analysis*, 145, 178–189.

Kettenring, J.R. (1971), 'Canonical Analysis of Several Sets of Variables', *Biometrika*, 58, 433–451.

Li, K. (1991), 'Sliced Inverse Regression for Dimension Reduction', *Journal of the American Statistical Association*, 86, 316–327.

Li, K. (1992), 'On Principal Hessian Directions for Data Visualization and Dimension Reduction: Another Application of Stein's Lemma', *Journal of the American Statistical Association*, 87, 1025–1039.

Li, B., and Wang, S. (2007), 'On Directional Regression for Dimension Reduction', *Journal of the American Statistical Association*, 102, 997–1008.

Li, B., Wen, S., and Zhu, L. (2008), 'On a Projective Resampling Method for Dimension Reduction with Multivariate Responses', *Journal of the American Statistical Association*, 103, 1177–1186.

Li, L., and Yin, X. (2009), 'Longitudinal Data Analysis Using Sufficient Dimension Reduction Method', *Computational Statistics and Data Analysis*, 53, 4106–4115.

Li, B., Zha, H., and Chiaromonte, F. (2005), 'Contour Regression: A General Approach to Dimension Reduction', *Annals of Statistics*, 33, 1580–1616.

Luo, R., Wang, H., and Tsai, C.L. (2009), 'Contour Projected Dimension Reduction', *The Annals of Statistics*, 37, 3743–3778.
Saracco, J. (2005), 'Asymptotics for Pooled Marginal Slicing Estimator Based on $SIR_\alpha$ Approach', *Journal of Multivariate Analysis*, 96, 117–135.
Sheng, W., and Yin, X. (2013), 'Direction Estimation in Single-index Models via Distance Covariance', *Journal of Multivariate Analysis*, 122, 148–161.
Sheng, W., and Yin, X. (2016), 'Sufficient Dimension Reduction via Distance Covariance', *Journal of Computational and Graphical Statistics*, 25, 91–104.
Shumway, R., Azari, A., and Pawitan, Y. (1988), 'Modeling Mortality Fluctuations in Los Angeles as Functions of Pollution and Weather Effects', *Environmental Research*, 45, 224–241.
Székely, G.J., and Rizzo, M.L. (2009), 'Brownian Distance Covariance', *The Annals of Applied Statistics*, 3, 1236–1265.
Székely, G.J., Rizzo, M.L., and Bakirov, N.K. (2007), 'Measuring and Testing Dependence by Correlation of Distances', *The Annals of Statistics*, 35, 2769–2794.
Wang, H., and Xia, Y. (2008), 'Sliced Regression for Dimension Reduction', *Journal of the American Statistical Association*, 103, 811–821.
Wang, Q., Yin, X., and Critchley, F. (2015), 'Dimension Reduction Based on the Hellinger Integral', *Biometrika*, 102, 95–106.
Xia, Y., Tong, H., Li, W., and Zhu, L. (2002), 'An Adaptive Estimation of Dimension Reduction Space', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64, 363–410.
Ye, Z., and Weiss, R.E. (2003), 'Using the Bootstrap to Select one of a New Class of Dimension Reduction Methods', *Journal of the American Statistical Association*, 98, 968–979.
Yin, X. (2004), 'Canonical Correlation Analysis Based on Information Theory', *Journal of Multivariate Analysis*, 91, 161–176.
Yin, X., and Bura, E. (2006), 'Moment-based Dimension Reduction for Multivariate Response Regression', *Journal of Statistical Planning and Inference*, 136, 3675–3688.
Yin, X., and Cook, R.D. (2005), 'Direction Estimation in Single-index Regressions', *Biometrika*, 92, 371–384.
Yin, X., and Li, B. (2011), 'Sufficient Dimension Reduction Based on an Ensemble of Minimum Average Variance Estimators', *The Annals of Statistics*, 39, 3392–3416.
Yin, X., Li, B., and Cook, R.D. (2008), 'Successive Direction Extraction for Estimating the Central Subspace in a Multiple-index Regression', *Journal of Multivariate Analysis*, 99, 1733–1757.
Yin, X., and Sriram, T. (2008), 'Common Canonical Variates for Independent Groups Using Information Theory', *Statistica Sinica*, 18, 335–353.
Zeng, P., and Zhu, Y. (2010), 'An Integral Transform Method for Estimating the Central Mean and Central Subspaces', *Journal of Multivariate Analysis*, 101, 271–290.
Zhu, Y., and Zeng, P. (2006), 'Fourier Methods for Estimating the Central Subspace and the Central Mean Subspace in Regression', *Journal of the American Statistical Association*, 101, 1638–1651.

## Appendix

**Lemma A.1:** *Suppose* $\boldsymbol{\eta}$ *is a basis of the CS. Let* $(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$ *be any partition of* $\boldsymbol{\eta}$, *where* $\boldsymbol{\eta}^\top \boldsymbol{\Sigma}_X \boldsymbol{\eta} = \boldsymbol{I}_d$. *We have* $\mathcal{V}^2(\boldsymbol{\eta}_i^\top \boldsymbol{X}, \boldsymbol{Y}) < \mathcal{V}^2(\boldsymbol{\eta}^\top \boldsymbol{X}, \boldsymbol{Y})$, $i = 1, 2$.

***Proof:*** Let $\tilde{\boldsymbol{X}}_1 = \boldsymbol{\eta}_1^\top \boldsymbol{X}$, $\tilde{\boldsymbol{X}}_2 = \boldsymbol{\eta}_2^\top \boldsymbol{X}$, $F(a,b) = \mathcal{V}^2\left(\begin{pmatrix} a\tilde{\boldsymbol{X}}_1 \\ b\tilde{\boldsymbol{X}}_2 \end{pmatrix}, \boldsymbol{Y}\right)$, $a \in R$, and $b \in R$,and $G_1(a,b) = \partial F(a,b)/\partial a$, $G_2(a,b) = \partial F(a,b)/\partial b$. A simple calculation shows that $aG_1(a,b) + bG_2(a,b) = F(a,b)$.

If $(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2) \in \mathcal{S}(\boldsymbol{\eta})$, then $F(0,1), F(1,0) > 0$.

Claim, if $0 \le \lambda < 1$, then $F(1,\lambda) < F(1,1)$, and $F(\lambda,1) < F(1,1)$.

If not, then there exist a $0 \le \lambda_0 < 1$ such that $F(1,\lambda_0) \ge F(1,1)$ or $F(\lambda_0,1) \ge F(1,1)$.

Without loss of generality, we assume there exist a $0 \le \lambda_0 < 1$ such that $F(1,\lambda_0) \ge F(1,1)$.

However, $F(1,\lambda) = \lambda F(\frac{1}{\lambda}, 1)$, and as $\lambda \to \infty$, $F(1/\lambda, 1) \to F(0,1) > 0$. Thus $F(1,\lambda) \to \infty$, as $\lambda \to \infty$. That means, there exists a $\lambda_1 \in (\lambda_0, \infty)$ such that $F(1,\lambda_1)$ achieves a minimum in $(\lambda_0, \infty)$.

Hence, $G_2(1,\lambda_1)=0$. Note that function $F(a,b)$ is a 'ray' function, i.e. $F(ca,cb)=cF(a,b)$. Thus using the fact that $F(1,\lambda)=\lambda F(1/\lambda,1)$, we can have $G_1(1/\lambda,1)=0$. And it is easy to calculate that $G_1(1,\lambda_1)=G_1(1/\lambda,1)=0$.

However, $0=1G_1(1,\lambda_1)+\lambda_1 G_2(1,\lambda_1)=F(1,\lambda_1)$. $F(1,\lambda_1)=0$ means that $\begin{pmatrix} a\tilde{\boldsymbol{X}}_1 \\ b\tilde{\boldsymbol{X}}_2 \end{pmatrix} \perp\!\!\!\perp \boldsymbol{Y}$, which conflicts with the assumption. ■

***Proof of Proposition*** 2.1**.:** Since $\mathcal{S}(\boldsymbol{\beta})\subseteq\mathcal{S}(\boldsymbol{\eta})=\mathcal{S}_{\boldsymbol{Y}|\boldsymbol{X}}$, $d_1\le d$, there exists a matrix $\boldsymbol{A}$, which satisfies $\boldsymbol{\beta}=\boldsymbol{\eta}\boldsymbol{A}$. Thus, $\mathcal{V}^2(\boldsymbol{\beta}^\top\boldsymbol{X},\boldsymbol{Y})=\mathcal{V}^2(\boldsymbol{A}^\top\boldsymbol{\eta}^\top\boldsymbol{X},\boldsymbol{Y})$.

Suppose the single value decomposition of $\boldsymbol{A}$ is $\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$, where $\boldsymbol{U}$ is a $d\times d$ orthogonal matrix, $\boldsymbol{V}$ is a $d_1\times d_1$ orthogonal matrix, and $\boldsymbol{D}$ is a $d\times d_1$ diagonal matrix with non-negative numbers on the diagonal. It is easy to prove that all non-negative values on the diagonal of $\boldsymbol{D}$ are 1. According to Székely and Rizzo (2009), Theorem 3, (ii),

$$\mathcal{V}^2(\boldsymbol{\beta}^\top\boldsymbol{X},\boldsymbol{Y})=\mathcal{V}^2(\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^\top\boldsymbol{\eta}^\top\boldsymbol{X},\boldsymbol{Y})=\mathcal{V}^2(\boldsymbol{D}\boldsymbol{U}^\top\boldsymbol{\eta}^\top\boldsymbol{X},\boldsymbol{Y}).$$

Let $\boldsymbol{U}^\top\boldsymbol{\eta}^\top\boldsymbol{X}=(\tilde{\boldsymbol{X}}_1,\ldots,\tilde{\boldsymbol{X}}_d)^\top$. Since all non-negative values on the diagonal of $\boldsymbol{D}$ are 1, and $\boldsymbol{D}^\top\boldsymbol{U}^\top\boldsymbol{\eta}^\top\boldsymbol{X}=(\tilde{\boldsymbol{X}}_1,\ldots,\tilde{\boldsymbol{X}}_{d_1})^\top$, by Lemma .1, we get

$$\mathcal{V}^2(\boldsymbol{D}\boldsymbol{U}^\top\boldsymbol{\eta}^\top\boldsymbol{X},\boldsymbol{Y})\le\mathcal{V}^2(\boldsymbol{U}^\top\boldsymbol{\eta}^\top\boldsymbol{X},\boldsymbol{Y}).$$

The equality holds if and only if $d=d_1$. According to Székely and Rizzo (2009), Theorem 3, (ii)

$$\mathcal{V}^2(\boldsymbol{U}^\top\boldsymbol{\eta}^\top\boldsymbol{X},\boldsymbol{Y})=\mathcal{V}^2(\boldsymbol{\eta}^\top\boldsymbol{X},\boldsymbol{Y}).$$

Thus

$$\mathcal{V}^2(\boldsymbol{\beta}^\top\boldsymbol{X},\boldsymbol{Y})\le\mathcal{V}^2(\boldsymbol{\eta}^\top\boldsymbol{X},\boldsymbol{Y}),$$

and equality holds if and only if $\mathcal{S}(\boldsymbol{\beta})=\mathcal{S}(\boldsymbol{\eta})$. ■

***Proof of Proposition*** 2.2**.:** For the $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ in Proposition 2.2, there exists a rotation matrix $\boldsymbol{Q}$ such that $\boldsymbol{\beta}\boldsymbol{Q}=(\boldsymbol{\eta}_a,\boldsymbol{\eta}_b)$, and $\mathcal{S}(\boldsymbol{\eta}_a)\subseteq\mathcal{S}(\boldsymbol{\eta})$, and $\mathcal{S}(\boldsymbol{\eta}_b)\subseteq\mathcal{S}(\boldsymbol{\eta})^\perp$, where $\mathcal{S}(\boldsymbol{\eta})^\perp$ is the orthogonal space of $\mathcal{S}(\boldsymbol{\eta})$.

Since $\boldsymbol{Y}\perp\!\!\!\perp\boldsymbol{\eta}_b^\top\boldsymbol{X}\,|\,\boldsymbol{\eta}^\top\boldsymbol{X}$ and $\mathbf{P}^\top_{\boldsymbol{B}(\boldsymbol{\Sigma}_X)}\boldsymbol{X}\perp\!\!\!\perp\mathbf{Q}^\top_{\boldsymbol{B}(\boldsymbol{\Sigma}_X)}\boldsymbol{X}$, therefore

$$\begin{pmatrix}\boldsymbol{Y}\\ \boldsymbol{\eta}^\top\boldsymbol{X}\end{pmatrix}\perp\!\!\!\perp\boldsymbol{\eta}_b^\top\boldsymbol{X}.$$

According to Proposition 4.3 (Cook 1998),

$$\begin{pmatrix}\boldsymbol{Y}\\ \boldsymbol{\eta}_a^\top\boldsymbol{X}\end{pmatrix}\perp\!\!\!\perp\boldsymbol{\eta}_b^\top\boldsymbol{X}.$$

Let $\boldsymbol{W}_1=\begin{pmatrix}\boldsymbol{\eta}_a^\top\boldsymbol{X}\\ 0\end{pmatrix}$, $\boldsymbol{V}_1=\boldsymbol{Y}$, $\boldsymbol{W}_2=\begin{pmatrix}0\\ \boldsymbol{\eta}_a^\top\boldsymbol{X}\end{pmatrix}$, and $\boldsymbol{V}_2=0$, then $(\boldsymbol{W}_1,\boldsymbol{V}_1)\perp\!\!\!\perp(\boldsymbol{W}_2,\boldsymbol{V}_2)$. According to Székely and Rizzo (2009), Theorem 3, (iii),

$$\mathcal{V}^2(\boldsymbol{W}_1+\boldsymbol{W}2,\boldsymbol{V}_1+\boldsymbol{V}2)<\mathcal{V}^2(\boldsymbol{W}_1,\boldsymbol{V}_1)+\mathcal{V}^2(\boldsymbol{W}2,\boldsymbol{V}2),$$

that is

$$\mathcal{V}^2(\boldsymbol{Q}^\top\boldsymbol{\beta}^\top\boldsymbol{X},\boldsymbol{Y})=\mathcal{V}^2(\boldsymbol{\beta}^\top\boldsymbol{X},\boldsymbol{Y})<\mathcal{V}^2(\boldsymbol{\eta}_a^\top\boldsymbol{X},\boldsymbol{Y})\le\mathcal{V}^2(\boldsymbol{\eta}^\top\boldsymbol{X},\boldsymbol{Y}).$$

■