Nonlinear Constrained Optimization in R and Its Application for Sufficient Dimension Reduction and Variable Selection

by

Xianyan Chen

(Under the Direction of Xiangrong Yin)

Abstract

In this dissertation, we develop an R package, NlcOptim, to solve the optimization problems with nonlinear objective functions and nonlinear constraints. This package can be used to solve problems in sufficient dimension reduction and variable selection because of its capability to accept the input parameters as a constrained matrix. We propose a framework for dimension reduction problems via Distance Covariance (DCOV) where both the response and the predictor are vectors. In this framework, distance covariance method is employed to estimate the central subspace effectively, and we also propose two different methods based on projective resampling technique to transfer multivariate response to univariate response. This approach keeps the model-free advantage, and can fully recover the central subspace even when many predictors are discrete. We then extend DCOV methods to canonical analysis, termed as Canonical Distance Covariance Analysis (CDCA), where we explore the relationships between two multivariate sets of variables. In addition, we extend DCOV to estimate the dual central subspace (DCS), which is to find the basis that span the subspace of $\boldsymbol{Y}$ as well as the basis that span the subspace of $\boldsymbol{X}$. At last, we develop a new concept, termed the Dual Variable Selection (DVS), to propose a method for simultaneously selecting subsets for each of the two random vectors, by employing DCOV method combined with LASSO penalty.

Index words: Central subspace, Distance Covariance, Dual Central subspace, Dual Variable Selection, Nonlinear constraints, Nonlinear optimization, Projective Resampling, Sequential quadratic programming, Sufficient Dimension Reduction

Nonlinear Constrained Optimization in R and Its Application for Sufficient Dimension Reduction and Variable Selection

by

Xianyan Chen

B.S., Hunan University, 2004

M.S., Hunan University, 2007

M.S., The George Washington University, 2011

A Dissertation Submitted to the Graduate Faculty
of The University of Georgia in Partial Fulfillment
of the
Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2016

Nonlinear Constrained Optimization in R and Its Application for Sufficient Dimension Reduction and Variable Selection

by

Xianyan Chen

Dissertation

Major Professor:	Xiangrong Yin
Committee:	Lynne Billard Pengsheng Ji William Mccormick Wenxuan Zhong

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
August 2016

Acknowledgments

First, I would like to express my deep gratitude to my advisor, Professor Xiangrong Yin, for his guidance, support and encouragement. I have been benefited so much from his teaching and extensive discussions with him. I have been impressed by his extreme diligence and his desire to understand every aspects in the field. His enthusiasm and confidence will always encourage me in my future professional life. Dear Dr. Yin, thanks for everything I have learned from you!

I would like to take this opportunity to thank Dr. Lynne Billard, Dr. Pengsheng Ji, Dr. William Mccormick, and Dr. Wenxuan Zhong for serving on my dissertation committee. Their encouragement, thoughtful comments, and invaluable advices are always appreciated.

I would also like to thank the Department of Statistics of UGA for bringing me into this wonderful department and give me the direct financial support. I also want to thank Dr. Reeves and the Statistical Consulting Center for giving me the opportunity to work as a graduate consultant. I will miss the time I spent with Julie, Daphney, Melissa, Mollie and Kirsten. I thank them for the kind help I received from them.

Too many people to mention individually have assisted me in so many ways during my studies at UGA. They all have my sincere gratitude. In particular, I would like to thank Yuanwen, Wenhui, Wenbo, Xinyi, Jing, Soyeon, Guannan, Zhen, Xijue, Zhengbo, Debin, Fei, Xiaoxiao, Yiwen, Xin, Rui, Xinlian, Lina, Hejiao, Xiaodong, all currently, or previously, of UGA.

I would like to thank my parents, without whom I would never have been able to achieve so much. I cordially thank my parents for supporting me remotely, even though they have so seldom opportunity to see me. Thanks for your great understanding.

My final, and most heartfelt, thanks must go to my husband Xianqiao and my son Hong-Yi. Their support and companionship have made my graduate study a pleasure.

Contents

[1] Chen, X. and Yin, X. Submitted to *[Journal of Statistical Software]*, 10/22/2015

[2] Chen, X. and Yin, X. *To be Submitted to [Journal of Computational and Graphical Statistics].*

[3] Chen, X. and Yin, X. *To be Submitted to [Computational Statistics & Data Analysis].*

[4] Chen, X. and Yin, X. *To be Submitted to [Technometrics].*

List of Figures

List of Tables

Chapter 1

Introduction

In recent years, we have witnessed the explosive growth of interest in the research of "Big Data". As technology advances, data now stream from daily life: phones, credit cards, televisions, and computers; from the infrastructure of cities; from sensor-equipped buildings, trains, buses, planes, bridges, and factories. Meanwhile, data flow so fast that the total accumulation or storage in the past two years-a zettabyte-dwarfs the prior record of human civilization. How to effectively process these dynamic large data sets so as to enable discovery and innovations is becoming critical and enjoying the great popularity in the statistical field.

Sufficient Dimension Reduction (SDR) can be brought into this picture. The essence of SDR is to extract information about statistical dependence of a response on predictors from the data without loss of any regression information. Ever since the establishment of the ground-breaking framework and theoretical foundation two decades ago (Li 1991; Cook 1994; Cook 1996; Cook 1998), SDR methods have been widely applied in many scientific disciplines. For example, they are often used as an intermediate step in data analysis and model fitting that efficiently downsizes the data from high or even ultra-high dimension to a relatively low one in order to avoid the effects of the curse of dimensionality, a situation where most classic regression methods fail.

Most existing SDR methods have shown their merits in coefficient estimation. However, they often involve kernel smoothing techniques and/or tend to put strong assumptions on link functions and/or underlying distributions of predictors. For instance, inverse approaches such as well-known sliced inverse regression (SIR; Li 1991) and Sliced Average Variance Estimates (SAVE; Cook and Weisberg 1991) require linearity and/or constant variance conditions; Forward approaches such as rMAVE (Xia et al., 2002) require smoothing techniques; joint approaches such as informational methods of Yin and Cook (2005) also require smoothing techniques while Fourier method by Zhu and Zeng (2006) requires strong distribution of predictors. Recently, Ma and Zhu (2012) developed a seminal paper using semi-parametric methods, but it again requires kernel smoothing. In addition, when predictors are categorical, these approaches may perform badly, due to their estimating algorithms or strong conditions.

More recently, Sheng and Yin (2013) and Sheng and Yin (2015) developed a novel method using distance covariance (DCOV; Székely, Rizzo, Bakirov, et al. 2007; Székely, Rizzo, et al. 2009) for sufficient dimension reduction. The method does not require linearity condition or constant covariance condition, neither does it any particular distribution on $\mathbf{X}$, $\mathbf{X}|Y$ or $Y|\mathbf{X}$. These advantages enable the method to work effectively under a variety of $\mathbf{X}$: $\mathbf{X}$ could be normal, non-normal but continuous, or discrete. In this dissertation, based on the advantages of DCOV, we apply DCOV to deal with multivariate response problems, and develop penalized procedure for variable selection.

Many SDR methods and other optimization problems involve a constrained matrix. However, there seems a lack of optimization algorithm in R, comparing with

Matlab, the commercial software. In Chapter 2, we present an R package "NlcOptim" to solve optimization problem with nonlinear objective function and nonlinear constraints. Using DCOV method for dimension reduction involves solving nonlinear optimization problems, but the existing R packages dealing with these problems do not fulfill the expectation. They are either inefficient or do not accept matrix input indeed, and most methods involving nonlinear constraint optimization use non-free software Matlab. NlcOptim utilizes gradient-based algorithms to tackle a general optimization problem with nonlinear constraints and nonlinear objective functions. In particular, it also accepts the input constrained parameters as a matrix.

In Chapter 3, we extend DCOV to sufficient dimension reduction with multiple-index models and multivariate responses. We also present two DCOV methods using projective resampling on multivariate response to convert the SDR with multivariate response to univariate response. One is to average the m subspaces to get the central subspace, where m is a pre-selected integer. The other is to sum m distance covariance functions and then obtain the central subspace. We also introduce an kNN procedure to estimate the dimension of the central subspace. Theoretical properties for DCOV on multivariate response such as asymptotic results are established based on the work of Sheng and Yin (2015). Our developed R package in Chapter 2 is used for the algorithms.

In Chapter 4, we extend DCOV method to canonical distance covariance analysis, where we explore the relationships between two multivariate sets of variables. Comparing to the traditional CCA, our methods can capture linear relationship as well as nonlinear relationship. We also use DCOV for recovering dual central subspaces. Two approaches based on distance covariance have been proposed. A bootstrap procedure is used to identify the dimension of dual central subspace. Asymptotic theory for the developed methods may be established, following the development of Sheng and Yin

(2013) and Sheng and Yin (2015). Again algorithms are developed using R package in Chapter 2.

In Chapter 5, we develop a new concept, termed the Dual Variable Selection (DVS), to propose a method for simultaneously selecting subsets for each of the two random vectors, by employing Distance Covariance (DCOV) method combined with LASSO (Tibshirani, 1996) penalty. This method is a model-free approach and does not need nonparametric smoothing. Algorithms, the performance of the proposed methods, and their theoretical studies are under investigation as our future work.

Bibliography

[1] R. D. Cook. “Graphics for regressions with a binary response”. In: *Journal of the American Statistical Association* 91.435 (1996), pp. 983–992.

[2] R. D. Cook. “On the interpretation of regression plots”. In: *Journal of the American Statistical Association* 89.425 (1994), pp. 177–189.

[3] R. D. Cook. *Regression graphics: Ideas for studying regression through graphics.* Wiley, 1998.

[4] R.D. Cook and S. Weisberg. “Discussion of a paper by KC Li”. In: *Journal of the American Statistical Association* 86 (1991), pp. 328–32.

[5] K.C. Li. “Sliced inverse regression for dimension reduction”. In: *Journal of the American Statistical Association* 86.414 (1991), pp. 316–327.

[6] Y. Ma and L. Zhu. “A semiparametric approach to dimension reduction”. In: *Journal of the American Statistical Association* 107.497 (2012), pp. 168–179.

[7] W. Sheng and X. Yin. “Direction estimation in single-index models via distance covariance”. In: *Journal of Multivariate Analysis* 122 (2013), pp. 148–161.

[8] W. Sheng and X. Yin. “Sufficient dimension reduction via distance covariance”. In: *Journal of Computational and Graphical Statistics* just-accepted (2015).

[9] G. J. Székely, M. L. Rizzo, N. K. Bakirov, et al. "Measuring and testing dependence by correlation of distances". In: *The Annals of Statistics* 35.6 (2007), pp. 2769–2794.

[10] G. J. Székely, M. L. Rizzo, et al. "Brownian distance covariance". In: *The annals of applied statistics* 3.4 (2009), pp. 1236–1265.

[11] R. Tibshirani. "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1996), pp. 267–288.

[12] Y. Xia et al. "An adaptive estimation of dimension reduction space". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64.3 (2002), pp. 363–410.

[13] Y. Zhu and P. Zeng. "Fourier methods for estimating the central subspace and the central mean subspace in regression". In: *Journal of the American Statistical Association* 101.476 (2006), pp. 1638–1651.

Chapter 2

NlcOptim: An R Package for Nonlinear Constrained Optimization Program [1]

[1] Chen, X. and Yin, X. Submitted to *[Journal of Statistical Software]*, 10/22/2015

Abstract

In this chapter, we present an R package, NlcOptim, to solve the optimization problem with nonlinear objective function and nonlinear constraints by using gradient-based algorithms. In particular, it accepts the input parameters as a constrained matrix, thus leveraging existing R packages where only vector parameters are acceptable.

Key Words: Nonlinear constraints; Nonlinear optimization; Quasi-Newton approximation; Sequential quadratic programming.

2.1 Introduction

R (Team, 2015) provides various packages and functions for different optimizations. Some packages can solve linear or quadratic optimization problems, for example, **liprog** (Henningsen, 2010), **quadprog** (Turlach and Weingessel, 2013), **limSolve** (Soetaert, Meersche, and Oevelen, 2009), and **rcdd** (Geyer and Meeden, 2015) etc. Others support unconstrained or box-constrained optimization problems with nonlinear objective functions, for example, **nlm()** function in **stat** package, **optimx** (Nash and Varadhan, 2011), **dfoptim** (Varadhan and Borchers, 2011), **subplex** (King and King, 2014), and **trustOptim** (Braun, 2014) etc. In addition, some packages can deal with linear equality and inequality constraints, such as **constrOptim()** function in stat package, and package **BB** (Varadhan and Gilbert, 2009).

There are limited packages that can solve an optimization problem with nonlinear constraints and nonlinear objective functions. Among the few, packages **alabama** (Varadhan and Grothendieck, 2015) and **Rsolnp** (Ghalanos and Theussl, 2014) aim to solve a general nonlinear optimization with both nonlinear constraints and objective function based on the augmented Lagrange multiplier method. Packages employing the differential evolution method to solve this kind of nonlinear problem are for example, **DEoptim** (Mullen et al., 2011) and **DEoptimR** (Conceicao and Maechler, 2015). This differential evolution method is useful when the objective function is difficult to differentiate, but inefficient for a problem with a smooth objective function.

There are plenty of chances in statistics that we are dealing with optimization problems with parameters in the form of matrix. For example, in the field of sufficient dimension reduction (Li, 1991; Cook, 1994; Cook, 1996), many methods have an optimization with a constraint $\boldsymbol{\beta}^\top\boldsymbol{\beta} = \boldsymbol{I}_d$, where $\boldsymbol{\beta}$ is a $q \times d$ matrix, $\boldsymbol{I}$ is an identity matrix and d is the dimension of a subspace. In this situation, we cannot use the aforementioned packages, since they all cannot deal with parameters in a constrained matrix form. Even if we transform the parameter matrix into a vector, in our experiences packages such as **alabama** and **Rsolnp** raise the error message because of redundant constraints while package **DeoptimR** experiences a convergence problem.

In this paper, we present an R package, called **NlcOptim**, to overcome the drawbacks that other packages have experienced. Especially our package can be efficiently used for nonlinear constraints. The newly-developed package utilizes gradient-based algorithms to tackle a general optimization problem with nonlinear constraints and nonlinear objective functions. In particular, it accepts the input parameters as a constrained matrix. The rest of this article is organized as follows. In Section 2.2, we give a brief description of the algorithms that guide the programming of **NlcOptim**. In Section 2.3, we show an example of how to use **NlcOptim** function. Section 2.4 presents four general nonlinear optimization examples and three sufficient dimension reduction examples. We compare the results for these examples with the solutions from **fmincon** function in MATLAB, **Rsolnp** (Ghalanos and Theussl, 2014), and **DEoptimR** (Conceicao and Maechler, 2015). Finally, a short discussion is presented in Section 2.5.

2.2 Theoretical background

2.2.1 SQP framework

A general constrained optimization problem can be written as:

$$\begin{aligned} \min_{x \in \Re^n} \quad & f(x), \\ s.t. \quad & c_i(x) = 0 \quad i \in \mathcal{E}, \\ & c_i(x) \leq 0 \quad i \in \mathcal{I}, \end{aligned}$$

where the objective f and the constraint function c_i are all smooth, real-valued functions on a subset of $\Re^n$; while $\mathcal{E}$ and $\mathcal{I}$ are two finite sets of equality constraints and inequality constraints, respectively.

The Lagrangian function for this special problem can be expressed as $\mathcal{L}(x, \lambda) = f(x) + \boldsymbol{\lambda}^\top \mathbf{c}(x)$, where $\boldsymbol{\lambda}$ is a tuning parameter vector while $\boldsymbol{c}(x)$ is a vector of constrained functions, consisting of $c_i(x)$. With the Lagrangian objective function, the nonlinear problem converts to a linearly constrained optimization problem. In the kth iteration, with $\mathcal{L}_k$ and f_k, the problem is simplified to solve:

$$\begin{aligned} \min_{\boldsymbol{p} \in \Re^n} \quad & \tfrac{1}{2}\boldsymbol{p}^\top \nabla^2_{xx}\mathcal{L}_k \boldsymbol{p} + \nabla f_k^\top \boldsymbol{p}, \\ s.t. \quad & \nabla c_i(x_k)^\top \boldsymbol{p} + c_i(x_k) = 0 \quad i \in \mathcal{E}, \\ & \nabla c_i(x_k)^\top \boldsymbol{p} + c_i(x_k) \leq 0 \quad i \in \mathcal{I}, \end{aligned}$$

where ∇ denotes the gradient, and ∇^2_{xx} denotes the Hessian matrix; while $\boldsymbol{p}$ is the parameter vector that we are interested in. Let $\boldsymbol{p}_k$ be the the optimal solution of the above linearly constrained problem. The new iterate is obtained by $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + a_k \boldsymbol{p}_k$, where $a_k \in (0, 1]$ is a suitable step length parameter determined by a line search procedure in a merit function (Section 2.3). This linearly constrained problem can be solved using any quadratic programming (QP) algorithm. The whole procedure is called sequential quadratic programming (SQP; Powell 1978). In each iterate of SQP, we first compute the gradient of $c_i(x)$ and the Hessian matrix $\nabla^2_{xx}\mathcal{L}_k$, then use any algorithms for quadratic programming to solve the linearly constrained problem, and at last update $\boldsymbol{x}_k$. The Hessian of the Lagrangian $\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda})$ is replaced by a quasi-Newton approximation method BFGS (Broyden, 1970).

2.2.2 Update Hessian of the Lagrangian

Let $H_k := \nabla^2_{xx}\mathcal{L}_k$ be the Hessian of the Lagrange $\mathcal{L}(x, \lambda)$ at the kth iteration step. The update for $\mathbf{H}_k$ from iterate k to iterate $k+1$ makes use of the vectors below:

$$\begin{aligned}
&\mathbf{s}_k = \mathbf{x}_{k+1} - \mathbf{x}_k,\\
&\mathbf{y}_k = \nabla\mathcal{L}(\mathbf{x}_{k+1}, \lambda_{k+1}) - \nabla\mathcal{L}(\mathbf{x}_k, \lambda_{k+1}),\\
&\mathbf{r}_k = \theta_k\mathbf{y}_k + (1-\theta_k)\mathbf{H}_k\mathbf{s}_k,
\end{aligned}$$

where the scalar θ_k is defined as

$$\theta_k = \begin{cases} 1 & \text{if } \mathbf{s}_k^\top\mathbf{y}_k \geq 0.2\mathbf{s}_k^\top\mathbf{H}_k\mathbf{s}_k, \\ \frac{0.8\mathbf{s}_k^\top\mathbf{H}_k\mathbf{s}_k}{\mathbf{s}_k^\top\mathbf{H}_k\mathbf{s}_k - \mathbf{s}_k^\top\mathbf{y}_k} & \text{if } \mathbf{s}_k^\top\mathbf{y}_k < 0.2\mathbf{s}_k^\top\mathbf{H}_k\mathbf{s}_k. \end{cases}$$

Then update $\mathbf{H}_k$ as

$$\mathbf{H}_{k+1} = \mathbf{H}_k - \frac{\mathbf{H}_k\mathbf{s}_k\mathbf{s}_k^\top\mathbf{H}_k}{\mathbf{s}_k^\top\mathbf{H}_k\mathbf{s}_k} + \frac{\mathbf{r}_k\mathbf{r}_k^\top}{\mathbf{s}_k^\top\mathbf{r}_k}.$$

Standard BFGS (Broyden, 1970) updates $\mathbf{y}_k$ and $\mathbf{s}_k$. However, the method requires $\mathbf{s}_k^\top\mathbf{y}_k > 0$ so as to make sure that the Hessian matrix is positive definite. Using $\mathbf{r}_k$ instead of $\mathbf{y}_k$ in the above formula meets this requirement, since when $\theta_k \neq 1$, $\mathbf{s}_k^\top\mathbf{r}_k = 0.2\mathbf{s}_k^\top\mathbf{H}_k\mathbf{s}_k > 0$.

2.2.3 Merit functions

SQP methods often use a merit function to determine the step length parameter a_k in a line search where the merit function decreases sufficiently. A variety of merit functions have been used in SQP methods. In our implementation, we choose an L_1-penalty merit function (Powell, 1978; Han, 1977), detailed as follows:

$$\phi(\boldsymbol{x}) = f(\boldsymbol{x}) + \sum_{i \in \mathcal{E}} v_i |c_i(\boldsymbol{x})| + \sum_{i \in \mathcal{I}} v_i |\min(0, c_i(\boldsymbol{x}))|,$$

where v_i is recommended as $v_i = \max(|\lambda_i|, \frac{v_i+|\lambda_i|}{2})$, $|\cdot|$ denotes absolute value.

2.3 Using the package

The main function in this package is named as **NlcOptim()**. Before starting to call this function, we need to prepare the objective function and constraint functions. The objective function `objfun` should be with one argument in form of a vector, and should return as a scalar. The constraint function `confun` should be one argument in form of a vector, and should return a `ceq` vector and a `c` vector as nonlinear equality constraints and inequality constraints, respectively. Set the vector to `NULL` in the constraint function if there is no such nonlinear constraints.

NlcOptim can have thirteen arguments–`X, objfun, confun, A, B, Aeq, Beq, lb, ub, tolX, tolObj, tolCon, nFunmax, Itmax`. `tolX`, `tolObj`, and `tolCon` are the tolerances in X, the objective function, and the constraint function, and with their respective default values 1e-5, 1e-6 and 1e-6. `nFunmax` and `Itmax` are the respective maximum numbers of parameters updated and maximum iteration steps, with the default values 1,000,000 and 4,000, respectively. If the optimization problem has no more constraints other than the nonlinear constraint written in `confun`, we can call the function as: `NlcOptim(X0, objfun=objfun, confun=confun)`, where `X0` is the initial value. If the problem has linear constraints, we should write the constraint as $A * X \leq B$, and $Aeq * X = Beq$, then put `A, B, Aeq, Beq` in the argument when calling the function `NlcOptim(X0, objfun=objfun, confun=confun, A, B, Aeq, Beq)`. If the parameters have lower bounds and up bounds, one should add them as `lb` and `ub`. The outputs of this function are `p`, as the optimum solution; `fval`, as the value of the objective function at the optimal point; lambda, as the Lagrangian multiplier; grad and hessian, as the gradient and hessian of the objective function at the optimal point, respectively.

In what follows, we take problem A1 in section 2.4.1 as the example to better describe the definitions used in this function. First, we write the objective and constraint functions as:

```
R> obj = function(x){
+   return(exp(x[1] * x[2] * x[3] * x[4] * x[5]))
+ }
R> con = function(x){
+   f = NULL
+   f = rbind(f,x[1] ^ 2 + x[2] ^ 2 + x[3] ^ 2 + x[4] ^ 2 + x[5] ^ 2 - 10)
+   f = rbind(f,x[2] * x[3] - 5 * x[4] * x[5])
+   f = rbind(f,x[1] ^ 3 + x[2] ^ 3 + 1)
+   return(list(ceq = f, c = NULL))
+ }
```

Then we choose an initial value and call the **NlcOptim** function, respectively as

```
R> x0 = c(-2, 2, 2, -1, -1)
```

```
R> NlcOptim(x0, objfun = obj, confun = con)
```

The output of the solution looks like:

```
  $p
           [,1]
[1,] -1.7171435
[2,]  1.5957096
[3,]  1.8272459
[4,] -0.7636431
[5,] -0.7636431
$fval
[1] 0.05394985
$lambda
$lambda$lower
     [,1]
[1,]    0
[2,]    0
[3,]    0
[4,]    0
[5,]    0
$lambda$upper
     [,1]
[1,]    0
[2,]    0
[3,]    0
```

```
[4,]    0
[5,]    0
$lambda$eqnonlin
[1]  0.040162737  0.037957774 -0.005222606
$grad
            [,1]
[1,]  0.09173207
[2,] -0.09871287
[3,] -0.08620505
[4,]  0.20627103
[5,]  0.20627103
$hessian
           [,1]        [,2]        [,3]        [,4]        [,5]
[1,]  0.6490970  0.12265571 -0.14577731  0.21968845  0.21968842
[2,]  0.1226557  0.61355598  0.21559060 -0.04869998 -0.04870000
[3,] -0.1457773  0.21559060  0.35647651 -0.06607590 -0.06607589
[4,]  0.2196885 -0.04869998 -0.06607590  1.45587101  0.45587101
[5,]  0.2196884 -0.04870000 -0.06607589  0.45587101  1.45587102
```

2.4 Examples

2.4.1 Comparisons of optimizations

In this section, we compare methods, **Rsolnp**, **NlcOptim** and **DEoptimR** in R, and **fmincon** in MATLAB by the following four problems:

$$
\begin{aligned}
(A1) \quad & \min \quad e^{(x_1x_2x_3x_4x_5)},\\
& \text{s.t.} \quad x_1^2+x_2^2+x_3^2+x_4^2+x_5^2=10,\\
& \qquad x_2x_3-5x_4x_5=0,\\
& \qquad x_1^3+x_2^3+1=0.\\
(B1) \quad & \min \quad (x_1-1)^2+(x_1-x_2)^2+(x_2-x_3)^3+(x_3-x_4)^4+(x_4-x_5)^4,\\
& \text{s.t.} \quad x_1+x_2^2+x_3^3=2+3\sqrt{2},\\
& \qquad x_2-x_3^2+x_4=-2+2\sqrt{2},\\
& \qquad x_1x_5=2.\\
(C1) \quad & \min \quad (1-x_1)^2+(x_2-x_1^2)^2,\\
& \text{s.t.} \quad x_1^2+x_2^2-1.5\le 0.\\
(D1) \quad & \min \quad x_1^2+x_2^2,\\
& \text{s.t.} \quad -x_1-x_2+1\le 0,\\
& \qquad -x_1^2-x_2^2+1\le 0,\\
& \qquad -9x_1^2-x_2^2+9\le 0,\\
& \qquad -x_1^2+x_2\le 0,\\
& \qquad x_1-x_2^2\le 0.
\end{aligned}
$$

The initial value chosen for (A1) is $(-2,2,2,-1,-1)^\top$, for (B1) is $(1,1,1,1,1)^\top$, for (C1) is $(-1.9, 2)^\top$, and for (D1) is $(3, 1)^\top$. Besides the above constraints, in the application of **DEoptimR**, a range (-1,1) for the parameters is considered to shorten the computing time.

Results from different packages are close to each other. **Rsolnp** generates an error message in inverting Hessian matrix when solving problem D1. **DEoptimR** gets a slightly different solution from others in problem A1, which may be due to the computing tolerance. However, it is worthwhile to point out that the value of the objective is -2.900 at the minimum point given by **DEoptimR**, slightly greater than -2.9192, the minimum objective value given by the other three methods.

Table 2.1: Solutions for A1, B1, C1 and D1 by different methods.

Problem	Method	Solution
A1	MATLAB	$(-1.7171\ 1.5957\ 1.8272\ -0.7636\ -0.7636)^\top$
	Rsolnp	$(-1.7171\ 1.5957\ 1.8272\ -0.7636\ -0.7636)^\top$
	NlcOptim	$(-1.7171\ 1.5957\ 1.8272\ -0.7636\ -0.7636)^\top$
	DEoptimR	$(-1.6921\ 1.5667\ -1.8686\ -0.8386\ 0.6981)^\top$
B1	MATLAB	$(1.1168\ 1.2206\ 1.5377\ 1.9724\ 1.7907)^\top$
	Rsolnp	$(1.1166\ 1.2204\ 1.5377\ 1.9727\ 1.7910)^\top$
	NlcOptim	$(1.1168\ 1.2206\ 1.5377\ 1.9724\ 1.7907)^\top$
	DEoptimR	$(1.1166\ 1.2204\ 1.5377\ 1.9726\ 1.7910)^\top$
C1	MATLAB	$(0.9167\ 0.8122)^\top$
	Rsolnp	$(0.9167\ 0.8122)^\top$
	NlcOptim	$(0.9167\ 0.8122)^\top$
	DEoptimR	$(0.9167\ 0.8121)^\top$
D1	MATLAB	$(1.0000\ 1.0000)^\top$
	Rsolnp *	...
	NlcOptim	$(1.0000\ 1.0000)^\top$
	DEoptimR	$(1.0000\ 1.0000)^\top$

* Rsolnp generates error.

In order to demonstrate the novel capability of this newly-developed function, we carry out three models in sufficient dimension reduction by comparing with **Rsolnp**, **NlcOptim**, and **DEoptimR** in R, and **fmincon** function in MATLAB. The sufficient dimension reduction method is DCOV of Sheng and Yin, 2013; Sheng and Yin, 2015. We use their MATLAB code, while we code the DCOV method in R with the three optimization approaches. Let $\boldsymbol{\beta}_1 = (1, 1, 1, 1, 1, 0, 0, 0, 0, 0)^\top$, $\boldsymbol{\beta}_2 = (1, 0, 0, 0, 0, 0)^\top$, and $\boldsymbol{\beta}_3 = (0, 1, 0, 0, 0, 0)^\top$. The constraint is $\boldsymbol{\beta}^\top \boldsymbol{\beta} = \boldsymbol{I}_d$, where d is the dimension of central subspace (Li, 1991; Cook, 1994; Cook, 1996). Sample sizes n=100, 200, and 400 are tested for each model. The accuracies are measured by the distance between the two subspaces (Li, Zha, and Chiaromonte, 2005), $\Delta_m(\boldsymbol{B}, \hat{\boldsymbol{B}}) = ||\boldsymbol{P}_B - \boldsymbol{P}_{\hat{B}}||$, where $\boldsymbol{P}_B$ and $\boldsymbol{P}_{\hat{B}}$ are the orthogonal projections onto the subspaces spanned by the columns of $\boldsymbol{B}$ and $\hat{\boldsymbol{B}}$, respectively, while $|| \cdot ||$ is the maximum singular value of a matrix. Thus the smaller Δ_m is, the better the estimate is. For each sample size n, the results of the mean $\bar{\Delta}_m$, standard error (SE), and mean computing time, and its standard error over 100 replicates are reported. Set the error $\epsilon \sim N(0, 1)$. The three models are:

$$\begin{aligned}
(A2) \quad & Y = \boldsymbol{\beta}_1^\top \boldsymbol{X} + \epsilon, \\
(B2) \quad & Y = (\boldsymbol{\beta}_1^\top \boldsymbol{X})^2 + \epsilon, \\
(C2) \quad & Y = \boldsymbol{\beta}_2^\top \boldsymbol{X} + (\boldsymbol{\beta}_3^\top \boldsymbol{X})^2 + 0.1\epsilon.
\end{aligned}$$

Tables 2.2, 2.3 and 2.4 show the respective results for Models A2, B2, and C2. In general, it can be seen that the accuracy increases and the computing time increases when n gets larger. MATLAB performs best in terms of computing time, with smallest mean time and SE(time) in all four packages, while **NlcOptim** is the best among all three R packages. For example, the average computing time for Model A2 with $n = 100$ is 4.75 sec by **NlcOptim**, but 78.13 sec by **Rsolnp**, 845.26 sec by **DEoptimR** with 20,000 iterations, and 3,324.40 sec with 100,000 iterations. The solution from **DEoptimR** does not converge with 100,000 iterations. Because of the computing time, we do not consider **DEoptimR** in our Model C2.

With respect to the accuracy, **NlcOptim** also outperforms other R packages, sometimes even better than that of MATLAB. For example, for Models A2 and B2, **NlcOptim** gives slightly better accuracy than that of MATLAB, but similar to that of **Rsolnp**. **DEoptimR**, even with 100,000 iterations, still yields much lower accuracy than other three methods did for Models A2 and B2.

NlcOptim and **fmincon** in MATLAB can solve the optimization problems with a constrained matrix of parameters, while **Rsolnp** does not have this option. Even we vactorize the matrix, **Rsolnp** still fails.

Table 2.2: Average estimation accuracy ($\bar{\Delta}_m$) and its standard error (SE_{δ_m}), average computing time ($t\bar{im}e$) and SE_{time} for Model A2.

n	Method	Δ_m	SE_{δ_m}	$t\bar{im}e(sec)$	$SE_{time}(sec)$
100	MATLAB	0.2536	0.0648	0.60	0.07
	Rsolnp	0.2197	0.0534	78.13	17.39
	NlcOptim	0.2240	0.0571	4.75	1.68
	DEoptimR1 $*$	0.4074	0.0803	845.26	27.73
	DEoptimR2 $*$	0.4051	0.0917	3324.40	69.31
200	MATLAB	0.1755	0.0432	0.84	0.55
	Rsolnp	0.1553	0.0332	207.70	40.32
	NlcOptim	0.1521	0.0401	7.64	2.98
	DEoptimR1 $*$	0.3650	0.0761	1569.30	37.40
	DEoptimR2 $*$	0.3444	0.0774	4265.07	115.04
400	MATLAB	0.1281	0.0301	1.42	0.18
	Rsolnp	0.1019	0.0234	630.12	112.91
	NlcOptim	0.1032	0.0222	28.64	8.38
	DEoptimR1 $*$	0.3495	0.0767	4382.22	231.96
	DEoptimR2 $*$	0.3404	0.0785	8069.70	296.81

* **DEoptimR**1: iteration=20,000, not converged; **DEoptimR**2: iteration=100,000, not converged.

Table 2.3: Average estimation accuracy ($\bar{\Delta}_m$) and its standard error (SE_{δ_m}), average computing time ($t\bar{im}e$) and SE_{time} for Model B2.

n	Method	Δ_m	SE_{δ_m}	$t\bar{im}e(sec)$	$SE_{time}(sec)$
100	MATLAB	0.1900	0.1791	0.45	0.07
	Rsolnp	0.1737	0.1372	95.55	72.76
	NlcOptim	0.1535	0.0422	2.35	1.38
	DEoptimR1 $*$	0.4381	0.1342	823.84	12.36
	DEoptimR2 $*$	0.4045	0.1081	3313.58	74.49
200	MATLAB	0.1062	0.0281	0.79	0.13
	Rsolnp	0.0978	0.0239	193.06	43.06
	NlcOptim	0.0972	0.0246	21.73	10.89
	DEoptimR1 $*$	0.3703	0.0772	1531.65	57.78
	DEoptimR2 $*$	0.3732	0.0960	4261.85	117.78
400	MATLAB	0.0719	0.0192	2.04	0.28
	Rsolnp	0.0648	0.0145	504.16	90.63
	NlcOptim	0.0619	0.0134	63.64	18.43
	DEoptimR1 $*$	0.3433	0.0835	4440.58	244.88
	DEoptimR2 $*$	0.3348	0.0711	8081.10	391.61

* **DEoptimR**1: iteration=20,000, not converged; **DEoptimR**2: iteration=100,000, not converged.

2.5 Discussion

In this paper, we have developed a new R package, **NlcOptim**, for nonlinear objective and nonlinear constrained optimization. Our comparisons show that it outperforms the existing two R packages dealing with the same problem, with respect to computing time and accuracy of the solutions. Although **NlcOptim** is a little bit more time-consuming than **fmincon** function in MATLAB, our package is free. We can further improve the computing speed of our package. Furthermore, we only programme this package based on one optimization approach, there are other optimization methods in literature which may be used as well for our package.

NlcOptim can accept the constrained arguments in form of a vector as well as a matrix. Especially, in sufficient dimension reduction, we often have an orthogonal constraint $\boldsymbol{\beta}^\top\boldsymbol{\beta} = I_d$, where $\boldsymbol{\beta}$ is a $q \times d$ matrix. This makes it very essential to develop a R package to accept a constrained matrix argument in order to implement sufficient dimension reduction techniques.

It is necessary to mention that **NlcOptim** finds local minima. But if the problem is convex optimization – minimizing a convex function over a set of convex constraints,

Table 2.4: Average estimation accuracy ($\bar{\Delta}_m$) and its standard error (SE_{δ_m}), average computing time ($\bar{time}$) and SE_{time} for Model C2.

n	Method	$\bar{\Delta}_m$	SE_{δ_m}	$\bar{time}(sec)$	$SE_{time}(sec)$
100	MATLAB	0.2026	0.1027	0.72	0.20
	Rsolnp	...	...	...	...
	NlcOptim $*$	0.2967	0.2415	8.33	3.69
200	MATLAB	0.1211	0.0354	1.32	0.33
	Rsolnp $*$	...	...	...	...
	NlcOptim	0.1378	0.0420	10.69	4.04
400	MATLAB	0.0820	0.0253	3.02	0.47
	Rsolnp $*$	...	...	...	...
	NlcOptim	0.0981	0.0300	62.73	13.60

* **Rsolnp** cannot deal with matrix input.

the local minima is also a global minima. Thus for a non-convex optimization, choosing a good starting point is very important. We encourage more tries with different starting points when using **NlcOptim** for a non-convex optimization problem. For the sufficient dimension reduction examples in our paper, the solutions from existing dimension reduction methods such as SIR (Li, 1991) and SAVE (Cook and Weisberg, 1991) are used as the initial value.

Bibliography

[1] M. Braun. "**trustOptim**: An R Package for trust region optimization with sparse hessians". In: *Journal of Statistical Software* 60.4 (2014), pp. 1–16.

[2] C. G. Broyden. "The convergence of a class of double-rank minimization algorithms. 2. the new algorithm". In: *IMA Journal of Applied Mathematics* 6.3 (1970), pp. 222–231.

[3] Eduardo L.T. Conceicao and M. Maechler. **DEoptimR***: Differential evolution optimization in pure R*. 2015. URL: `https://cran.r-project.org/web/packages/DEoptimR`.

[4] R. D. Cook. "Graphics for regressions with a binary response". In: *Journal of the American Statistical Association* 91.435 (1996), pp. 983–992.

[5] R. D. Cook. "On the interpretation of regression plots". In: *Journal of the American Statistical Association* 89.425 (1994), pp. 177–189.

[6] R.D. Cook and S. Weisberg. "Discussion of a paper by KC Li". In: *Journal of the American Statistical Association* 86 (1991), pp. 328–32.

[7] C.J. Geyer and G.D. Meeden. *R package* **rcdd** *(C double description for R)*. R package version 3.0.0. 2015. URL: `https://cran.r-project.org/web/packages/rcdd`.

[8] A. Ghalanos and S. Theussl. **Rsolnp**: *General non-linear optimization using augmented lagrange multiplier method.* R package version 1.15. 2014. URL: `https://cran.r-project.org/web/packages/Rsolnp`.

[9] S.P. Han. "A globally convergent method for nonlinear programming". In: *Journal of optimization theory and applications* 22.3 (1977), pp. 297–309.

[10] A. Henningsen. **linprog**: *linear programming.* Optimization. R package version 0.9-0. 2010. URL: `http://linprog.r-forge.r-project.org/`.

[11] A. A. King and M. A. A. King. **subplex**: *Unconstrained optimization using the subplex algorithm.* R package version 2.5.1. 2014. URL: `https://cran.r-project.org/web/packages/subplex`.

[12] B. Li, H. Zha, and F. Chiaromonte. "Contour Regression: A general approach to dimension reduction". In: *Annals of statistics* (2005), pp. 1580–1616.

[13] K.C. Li. "Sliced inverse regression for dimension reduction". In: *Journal of the American Statistical Association* 86.414 (1991), pp. 316–327.

[14] K. Mullen et al. "**DEoptim**: An R Package for global optimization by differential evolution". In: *Journal of Statistical Software* 40.6 (2011), pp. 1–26.

[15] J. C. Nash and R. Varadhan. "Unifying optimization algorithms to aid software system users: **optimx** for R". In: *Journal of Statistical Software* 43.9 (2011), pp. 1–14.

[16] M.J.D. Powell. "A fast algorithm for nonlinearly constrained optimization calculations". In: *Numerical analysis.* Springer-Verlag, 1978, pp. 144–157.

[17] W. Sheng and X. Yin. "Direction estimation in single-index models via distance covariance". In: *Journal of Multivariate Analysis* 122 (2013), pp. 148–161.

[18] W. Sheng and X. Yin. "Sufficient dimension reduction via distance covariance". In: *Journal of Computational and Graphical Statistics* just-accepted (2015).

[19] K. Soetaert, K. Van den Meersche, and D. van Oevelen. **limSolve***: solving linear inverse models*. R package version 1.5.1. 2009. URL: `https://cran.r-project.org/web/packages/limSolve`.

[20] R Core Team. *R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2015*. 2015.

[21] B. A. Turlach and A. Weingessel. **quadprog***: Functions to solve quadratic programming problems*. R package version 2.15.0. 2013. URL: `https://cran.r-project.org/web/packages/quadprog`.

[22] R. Varadhan and H.W. Borchers. **dfoptim***: Derivative-free optimization*. R package version 2.10.1. 2011. URL: `https://cran.r-project.org/web/packages/dfoptim`.

[23] R. Varadhan and P. Gilbert. "**BB**: An R package for solving a large system of nonlinear equations and for optimizing a high-dimensional nonlinear objective function". In: *Journal of Statistical Software* 32.4 (2009), pp. 1–26.

[24] R. Varadhan and G. Grothendieck. **alabama***: Constrained nonlinear optimization*. R package version 2.10.1. 2015. URL: `https://cran.r-project.org/web/packages/alabama`.

Chapter 3

Sufficient Dimension Reduction via Distance Covariance with Multivariate Responses[1]

[1] Chen, X. and Yin, X. *To be Submitted to [Journal of Computational and Graphical Statistics].*

Abstract

In this article, we propose a new method for dimension reduction problems where both response and predictors are vectors. Distance covariance (DCOV) method, which finds the maximum of the dependency between response and reduced direction, is used to estimate the central subspace effectively. Projective resampling method, which converts multivariate responses to univariate response, is used to combine with distance covariance method to conduct dimension reduction. This approach keeps the model-free advantage, and can fully recover the central subspace even when many predictors are categorical or discrete. Based on DCOV method, we defined three optimization problems and present three estimators. All measures are illustrated through extensive simulations and data sets, and compared with some existing methods. The comparison suggests that our method is competitive and robust.

Key Words: Central subspace; Distance Covariance; Projective Resampling; Sufficient Dimension Reduction.

3.1 Introduction

Suppose Y is a scalar response and $\mathbf{X}$ is a $p \times 1$ predictor vector. Sufficient dimension reduction (SDR; Li 1991; Cook 1994; Cook 1996) is a methodology for reducing the dimension of predictors while preserving the regression relation with response. Many methods have been proposed to estimate $\mathcal{S}_{Y|\mathbf{X}}$ or part of it. 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 et al. (2001), MAVE (Xia et al., 2002) and SR (Wang and Xia, 2008); Correlation approaches: CANCOR (Fung et al., 2002), KL-distance (Yin and Cook, 2005) and (Yin and Cook, 2005), Fourier transform (Zeng and Zhu, 2010) and (Zhu and Zeng, 2006), and Reproducing Kernel Hilbert Space type (Fukumizu, Bach, and Jordan, 2004). However, these methods either need the linearity condition or constant covariance condition, or require the predictors to be multivariate normal or at least continuous and the link function to be smooth. More recently, Sheng and Yin (2013) and Sheng and Yin (2015) developed a novel method using distance covariance for sufficient dimension reduction (DCOV; Székely, Rizzo, Bakirov, et al. (2007); Székely, Rizzo, et al. (2009)). The method does not require linearity condition or constant covariance condition, or any particular distribution on $\mathbf{X}$, $\mathbf{X}|Y$ or $Y|\mathbf{X}$. These advantages enable the method to work effectively under a variety of $\mathbf{X}$: $\mathbf{X}$ could be normal, non-normal but continuous, or discrete or categorical.

Various dimension reduction concepts can be extended to 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. However, the latter two methods are not guaranteed to fully recover the dimension reduction space. 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, and then use any dimension reduction method to get a subspace. Averaging these m subspaces, we can estimate the central subspace. It is shown that this method can fully recover the central subspace.

In this article, we extend DCOV to dimension reduction for multivariate response. Also based on projective resampling method, we propose two estimates combining it with DCOV. One is to average the m subspaces to get the central subspace. And the other is to sum m distance covariance functions and then obtain the central subspace. We also introduce a kNN procedure to estimate the dimension of the central subspace. Through a number of simulation experiments, most of which are based on published models, we demonstrate the superb performance of DCOV and projective resampling method.

The rest of the article is organized as follows: In section 3.2, we describe our method in details, including DCOV, projective resampling method, and kNN procedure. In section 3.3, we conduct simulation comparisons between the our estimators and others in a large variety of models; and in section 3.4, we summarize our work.

3.2 Methodology

3.2.1 Sufficient dimension reduction

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 column vectors 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 on to $\mathcal{S}(\boldsymbol{B})$ with respect to the inner product $<\boldsymbol{a},\boldsymbol{b}>=\boldsymbol{a}^T\boldsymbol{\Sigma}\boldsymbol{b}$. That is, $\mathbf{P}_{\boldsymbol{B}(\boldsymbol{\Sigma}_X)} = \boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{\Sigma}_X\boldsymbol{B})^{-1}\boldsymbol{B}^T\boldsymbol{\Sigma}_X$. Let $\mathbf{Q}_{\boldsymbol{B}(\boldsymbol{\Sigma}_X)} = \mathbf{I} - \mathbf{P}_{\boldsymbol{B}(\boldsymbol{\Sigma}_X)}$, where $\mathbf{I}$ is the identity matrix.

The ultimate goal of sufficient dimension reduction is to search a number of linear combinations of $\mathbf{X}$, say $\boldsymbol{\beta}^T\mathbf{X}$, where $\boldsymbol{\beta}$ is a $p \times d$ matrix, $d < p$, such that Y depends on $\mathbf{X}$ only through $\boldsymbol{\beta}^T\mathbf{X}$. That is:

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

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 (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 central subspace is well-defined and unique. We assume central subspace exists throughout this article.

3.2.2 Distance covariance as a sufficient dimension reduction tool

DCOV is introduced by Székely, Rizzo, Bakirov, 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:

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

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, c_q are constants, and is chosen to be positive. An equivalent form of the squared DCOV is given by Székely, Rizzo, et al. (2009) as

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

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

DCOV equals to 0 if and only if two random vectors are independent (Székely, Rizzo, Bakirov, et al., 2007). Based on this property, Sheng and Yin (2013) and Sheng and Yin (2015) proposed DCOV as a sufficient dimension reduction tool. Suppose $\boldsymbol{\beta}$ is a $p \times d$ matrix, where $1 \leq d \leq q$. The solution to the following optimization problem will yield a basis of the central subspace.

$$\max_{\boldsymbol{\beta}^T \boldsymbol{\Sigma}_X \boldsymbol{\beta} = \mathbf{I}_d} \mathcal{V}^2(\boldsymbol{\beta}^T \mathbf{X}, Y) \tag{3.1}$$

under $E|\mathbf{X}| < \infty$ and $E|Y| < \infty$ (Székely, Rizzo, Bakirov, et al., 2007). In this article we assume $E|\mathbf{X}| < \infty$ and $E|Y| < \infty$. The constraint $\boldsymbol{\beta}^T \boldsymbol{\Sigma}_X \boldsymbol{\beta} = \mathbf{I}_d$ in the optimization problem guarantees the solution of $\boldsymbol{\beta}$ in the same scale and the optimization solver does not diverge.

3.2.3 DCOV for multivariate response

Sheng and Yin (2013) and Sheng and Yin (2015) developed the DCOV method for the case that the response is a scalar. In this article, we extend DCOV method to multivariate response. Suppose $\boldsymbol{X}$ is $p \times 1$ random vector, $\boldsymbol{Y}$ is $q \times 1$ random vector, and $\boldsymbol{\beta}$ is a $p \times d$ matrix, where $1 \leq d \leq p$. A basis of the central subspace can be obtained by solving the following optimization problem.

$$\max_{\boldsymbol{\beta}^T\boldsymbol{\Sigma}_X\boldsymbol{\beta}=\mathbf{I}_d} \mathcal{V}^2(\boldsymbol{\beta}^T\boldsymbol{X},\boldsymbol{Y}). \tag{3.2}$$

under $E|\boldsymbol{X}| < \infty$ and $E|\boldsymbol{Y}| < \infty$ (Székely, Rizzo, Bakirov, et al., 2007). Sheng and Yin (2013) and Sheng and Yin (2015) demonstrated that under some mild conditions, the solution to (1) always spans the central subspace. We generalize their propositions to multivariate response cases for (2).

Proposition 1 *Let $\boldsymbol{\eta}$ be a basis of the central subspace with dimension d, $\boldsymbol{\beta}$ be a $p \times d_0$ matrix, $d_0 \leq d$, $dim(\mathcal{S}(\beta)) = d_0$, $\boldsymbol{\eta}^\top\boldsymbol{\Sigma}_X\boldsymbol{\eta} = \boldsymbol{I}_d$, and $\boldsymbol{\beta}^\top\boldsymbol{\Sigma}_X\boldsymbol{\beta} = \boldsymbol{I}_{d_0}$. Assume $\mathcal{S}(\boldsymbol{\beta}) \subseteq \mathcal{S}(\boldsymbol{\eta})$, then $\mathcal{V}^2(\boldsymbol{\beta}^\top\boldsymbol{X},\boldsymbol{Y}) \leq \mathcal{V}^2(\boldsymbol{\eta}^\top\boldsymbol{X},\boldsymbol{Y})$. The equality holds if and only if $\mathcal{S}(\boldsymbol{\beta}) = \mathcal{S}(\boldsymbol{\eta})$.*

Proposition 2 *Let $\boldsymbol{\eta}$ be a basis of the central subspace 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 or equal to d. Suppose $\mathbf{P}^\top_{\boldsymbol{B}(\boldsymbol{\Sigma}_X)}\boldsymbol{X} \perp\!\!\!\perp \mathbf{Q}^\top_{\boldsymbol{B}(\boldsymbol{\Sigma}_X)}\boldsymbol{X}$, 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 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 equals 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 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 maximizing $\mathcal{V}^2(\boldsymbol{\beta}^\top\boldsymbol{X},\boldsymbol{Y})$ with a constraint of $\boldsymbol{\beta}$ can always identify the central subspace.

Based on the sample version of squared distance covariance $\mathcal{V}^2_n(\boldsymbol{\beta}^T\mathbf{X}, Y)$ proposed by Székely, Rizzo, Bakirov, et al. (2007), a sample version for multivariate response can be defined as

$$\mathcal{V}^2(\boldsymbol{\beta}^\top \boldsymbol{X}, \boldsymbol{Y}) = \frac{1}{n^2} \sum_{k,l=1}^{n} A_{kl}(\boldsymbol{\beta}) B_{kl}, \tag{3.3}$$

where, for $k, l = 1, \cdots, n$,

$$\begin{aligned} A_{kl}(\boldsymbol{\beta}) &= a_{kl}(\boldsymbol{\beta}) - \bar{a}_{k.}(\boldsymbol{\beta}) - \bar{a}_{.l}(\boldsymbol{\beta}) + \bar{a}_{..}(\boldsymbol{\beta}) \\ a_{kl}(\boldsymbol{\beta}) &= |\boldsymbol{\beta}^T \mathbf{X}_k - \boldsymbol{\beta}^T \mathbf{X}_l|, \bar{a}_{k.}(\boldsymbol{\beta}) = \frac{1}{n} \sum_{l=1}^{n} a_{kl}(\boldsymbol{\beta}), \\ \bar{a}_{.l}(\boldsymbol{\beta}) &= \frac{1}{n} \sum_{k=1}^{n} a_{kl}(\boldsymbol{\beta}), \bar{a}_{..}(\boldsymbol{\beta}) = \frac{1}{n^2} \sum_{k,l=1}^{n} a_{kl}(\boldsymbol{\beta}). \end{aligned}$$

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 central subspace is

$$\boldsymbol{\eta}_n = \arg \max_{\boldsymbol{\beta}^T \hat{\boldsymbol{\Sigma}}_X \boldsymbol{\beta} = \mathbf{I}_d} \mathcal{V}_n^2(\boldsymbol{\beta}^\top \boldsymbol{X}, \boldsymbol{Y}). \tag{3.4}$$

Using Sequential Quadratic Programming method for R introduced in Chen and Yin (2015), we can solve this nonlinear optimization problem. The asymptotic properties of $\boldsymbol{\eta}_n$ can be derived in the same way as in the study of Sheng and Yin (2015), which is presented in section 3.2.5.

3.2.4 DCOV with projective resampling

Existing dimension reduction methods for univariate response can be extend to multivariate response by combining with projective resampling method (Li, Wen, and Zhu, 2008). Let $\boldsymbol{t}$ be a generic vector in $\mathbb{R}^q$. The projective was established on the statement: $\mathbf{Y} \perp\!\!\!\perp \mathbf{X}|\boldsymbol{\beta}^T\mathbf{X}$ if and only if $\boldsymbol{t}^T\mathbf{Y} \perp\!\!\!\perp \mathbf{X}|\boldsymbol{\beta}^T\mathbf{X}$ for all $t \in \mathbb{R}^q$. That is

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

In this way, the multivariate response problem is reduced to the many univariate response problem. Thus, all sufficient dimension reduction methods developed for univariate response can be employed to multivariate response by estimating $\mathcal{S}_{\boldsymbol{t}^T\mathbf{Y}|X}$ 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 $\mathcal{S}_{\mathbf{Y}|X} = Span\{\mathcal{S}_{\mathbf{t}^T\mathbf{Y}|X}\}$;

(ii) If the size of $\boldsymbol{t}$ is large enough, the subspace will be recovered through those univariate central subspaces. In practice, we may take the size of $\boldsymbol{t}$ as large as the computer allowed.

Li, Wen, and Zhu (2008) proposed projective resampling SIR, SAVE, and DR. In this article, beside the multivariate DCOV (DCOV0) in section 3.2.3, 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 $\boldsymbol{t}^T\mathbf{Y}$, we develop DCOV1 and DCOV2 similar to the idea of outer product gradient (OPG) and rMAVE (Xia et al., 2002):

DCOV 1 For each of the m combinations of $\mathbf{Y}$, ${\boldsymbol{t}_i}^T\mathbf{Y}, i = 1, ..., m$, solve the optimization problem to get

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

Then an estimated basis of central subspace can be the first d eigenvectors of

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

DCOV 2 Instead of obtaining a basis for each $\boldsymbol{t}_i{}^T\mathbf{Y}$, sum the squared distance covariance for each $\boldsymbol{t}_i{}^T\mathbf{Y}$ as the new objective function, and then solve the optimization problem

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

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

The R package Nlcoptim (Chen and Yin, 2015) is used to solve the above nonlinear optimization problem. This package implements Sequential Quadratic Programming (SQP) method to solve nonlinear optimization problems with nonlinear objective and nonlinear constraint functions. The initial value for the optimization problem can be generated randomly, but it is not efficient when the dimension of $\boldsymbol{X}$ is not small, since we need variation on each parameter. While in this article, we obtain the initial value by comparing the estimates by SIR and SAVE, and choose the one which gives the larger squared distance covariance.

Note that by invariance law, we can equivalently work on standardized predictor Z-scale, then transform back to X-scale. We first standardize X into Z-scale, where $\boldsymbol{Z} = \boldsymbol{\Sigma}_X^{-\frac{1}{2}}(\boldsymbol{X} - E(\boldsymbol{X}))$, such that $\boldsymbol{\Sigma}_Z = \boldsymbol{I}$, the $p \times p$ identity matrix. Then the direction of in X-scale would be $\boldsymbol{\eta} = \boldsymbol{\Sigma}_X^{-\frac{1}{2}}\boldsymbol{\eta}_Z$, where $\boldsymbol{\eta}_Z$ is the direction in Z-scale. Then the optimization problem is transformed to Z-scale, where the constraint becomes

$\boldsymbol{\beta}_Z^\top\boldsymbol{\beta}_Z = \mathbf{I}_d$. After obtaining the estimate under Z scale, we transform the estimate back into X scale, $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\Sigma}}_X^{-\frac{1}{2}}\hat{\boldsymbol{\beta}}_Z$. This scheme works well in our simulations and real data analysis. An alternative procedure is to use a successive one-at-a-time search similar to that of Yin, Li and Cook (2008).

3.2.5 Asymptotic properties

Sheng and Yin (2015) 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. Based on their results, we develop the asymptotic properties of the estimator of multivariate DCOV (DCOV0), $\boldsymbol{\eta}_n = \arg\max_{\boldsymbol{\beta}^\top\hat{\boldsymbol{\Sigma}}_X\boldsymbol{\beta}=\boldsymbol{I}} \mathcal{V}_n^2(\boldsymbol{\beta}^\top\boldsymbol{X}, \boldsymbol{Y})$.

Proposition 3 *Assume η is a basis matrix of the central subspace $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ and $\boldsymbol{\eta}^T\boldsymbol{\Sigma}_X\boldsymbol{\eta} = \mathbf{I}_d$. Suppose the support of $\mathbf{X} \in \mathbb{R}^p$, say S, is a compact set, $\mathrm{E}|Y| < \infty$, and $\mathbf{P}^T_{\boldsymbol{\eta}(\boldsymbol{\Sigma}_X)}\boldsymbol{X} \perp\!\!\!\perp \mathbf{Q}^\top_{\boldsymbol{\eta}(\boldsymbol{\Sigma}_{\boldsymbol{X}})}\boldsymbol{X}$. Let $\boldsymbol{\eta}_n = \arg\max_{\boldsymbol{\beta}^T\hat{\boldsymbol{\Sigma}}_{\mathbf{X}}\boldsymbol{\beta}=I_d} \mathcal{V}_n^2(\boldsymbol{\beta}^\top X, \boldsymbol{Y})$, then $\boldsymbol{\eta}_n$ is a consistent estimator of a basis of $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$, that is, there exists a rotation matrix $\boldsymbol{Q}$: $\boldsymbol{Q}^\top\boldsymbol{Q} = \boldsymbol{I}_d$, such that $\boldsymbol{\eta}_n \xrightarrow{p} \boldsymbol{\eta}\boldsymbol{Q}$.*

Proposition 4 *Assume η is a basis matrix of the central subspace $\mathcal{S}_{\boldsymbol{Y}|\boldsymbol{X}}$ and $\boldsymbol{\eta}^T\boldsymbol{\Sigma}_X\boldsymbol{\eta} = \mathbf{I}_d$. Suppose the support of $\mathbf{X} \in \mathbb{R}^p$, S, be a compact set, $\mathrm{E}|Y| < \infty$ and $\mathbf{P}^T_{\boldsymbol{\eta}(\boldsymbol{\Sigma}_X)}\boldsymbol{X} \perp\!\!\!\perp \mathbf{Q}^\top_{\boldsymbol{\eta}(\boldsymbol{\Sigma}_{\boldsymbol{X}})}\boldsymbol{X}$. Let $\boldsymbol{\eta}_n = \arg\max_{\boldsymbol{\beta}^T\hat{\Sigma}_{\mathbf{X}}\boldsymbol{\beta}=1} \mathcal{V}_n^2(\boldsymbol{\beta}^\top\boldsymbol{X}, \boldsymbol{Y})$, then under some regularity conditions, there exists a rotation matrix $\boldsymbol{Q}$: $\boldsymbol{Q}^\top\boldsymbol{Q} = \boldsymbol{I}_d$, such that $\sqrt{n}[vec(\boldsymbol{\eta}_n) - vec(\boldsymbol{\eta}\boldsymbol{Q})] \xrightarrow{\mathfrak{D}} N(0, V(\boldsymbol{\eta}_Q))$, where $V(\boldsymbol{\eta}_Q)$ is covariance matrix.*

If the support of $\boldsymbol{X}$ is compact, $E|\boldsymbol{Y}| < \infty$ and $\mathbf{P}^T_{\boldsymbol{\eta}(\boldsymbol{\Sigma}_X)}\boldsymbol{X} \perp\!\!\!\perp \mathbf{Q}^\top_{\boldsymbol{\eta}(\boldsymbol{\Sigma}_{\boldsymbol{X}})}\mathbf{X}$, then $\boldsymbol{\eta}_n$ is a consistent estimator of basis of $\mathcal{S}_{Y|\mathbf{X}}$. Sheng and Yin (2015) also discussed that the support of $\boldsymbol{X}$ does not necessarily to be compact, this assumption is set just to simplify the proof. They proved the $\sqrt{n}$−consistency and asymptotic normality of the estimator $\boldsymbol{\eta}_n$.The proofs of Proposition 3 and 4 are straightforward from Sheng and Yin (2015), replacing the scalar Y by a multivariate $\boldsymbol{Y}$. We omit the proofs.

For DCOV1, by the consistency proposition for univariate response in the work of Sheng and Yin (2015), $\boldsymbol{\eta}_n^i \xrightarrow{p} \boldsymbol{\eta Q}$ for each $\boldsymbol{t}_i^\top \boldsymbol{Y}$, with $i = 1, ...m$. Then combining all these $\boldsymbol{\eta}_n^i$ to obtain the estimator in DCOV1 should also have these consistency property, that is $\boldsymbol{\eta}_n \xrightarrow{p} \boldsymbol{\eta Q}$. The asymptotically normal property can be shown in the same way. For each univariate response $\boldsymbol{t}_i^\top \boldsymbol{Y}$, $i = 1, ...m$, by the normality property, $\sqrt{n}[vec(\boldsymbol{\eta}_n^i) - vec(\boldsymbol{\eta Q})] \xrightarrow{\mathfrak{D}} N(0, V(\boldsymbol{\eta}_Q))$, then when adding these estimators, the final estimator by DCOV1 should also has asymptotic normality as in Sheng and Yin (2015). 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 should also have $\sqrt{n}$-consistency and asymptotic normality, but with tedious calculations based on Sheng and Yin (2015).

3.2.6 Estimating d

In practice, the dimension of the central subspace $d = dim(\mathcal{S}_{Y|\mathbf{X}})$ is unknown and must be inferred from data. A few methods have been proposed in literature, such as a sequential test based on a chi-squared statistic proposed by Li (1991) and Li (1992), a permutation based test by Cook and Yin (2001), and a bootstrap procedure initialed by Ye and Weiss (2011), followed by Zhu and Zeng (2006), and Sheng and Yin (2015). In this article we introduce a kNN procedure for the purpose of choosing d, following the idea of k nearest neighbors (kNN) method (Wang, Yin, and Critchley, 2015).

Given $\{(\mathbf{X}_i, Y_i)\}, 1 \leq i \leq n$, the structural dimension can be evaluated by the following kNN procedure:

1. For each point in $\{(\mathbf{X}_i, Y_i), 1 \leq i \leq n\}$, obtain the k nearest neighbors of sample point i using Euclidean distance $|\mathbf{X}_i - \mathbf{X}_j|$,where $1 \leq j \leq n$. The k nearest neighbors of sample point i is denoted as $\{(\mathbf{X}_j^{(i)}, Y_j^{(i)}),\ 1 \leq j \leq k\}$;

2. For each sample point i, apply any dimension reduction method to its k nearest neighbors $\{(\mathbf{X}_j^{(i)}, Y_j^{(i)}),\ 1 \leq j \leq k\}$, we can estimate $\hat{\boldsymbol{\beta}}_i$. Setting the dimension of $\hat{\boldsymbol{\beta}}_i$ as 1 or 2 is good enough usually ;

3. After all $\hat{\boldsymbol{\beta}}_i,\ 1 \leq i \leq n$ are obtained, we can get the eigenvalues of $\sum\limits_{i=1}^{n} \hat{\boldsymbol{\beta}}_i \hat{\boldsymbol{\beta}}_i^T$, denote as $\lambda_1, \lambda_2, ..., \lambda_p$;

4. Calculate the ratio $r_i = \lambda_i / \lambda_{i+1},\ 1 \leq i \leq p-1$. Choose d as where the largest r_i happens in the sequence.

3.3 Numerical studies

In this section, we assess the efficiency of the proposed DCOV methods through the simulation and application to a real data. In the simulations, we compare the performance of our methods DCOV0, DCOV1 and DCOV2 with some well-established dimension reduction methods, that is, the PRSIR (Li, Wen, and Zhu, 2008), PRSAVE (Li, Wen, and Zhu, 2008), and RMAVE-$\mathfrak{F}_C$ (Yin and Li, 2011). We choose these three methods because SIR and SAVE are the most common methods in the area of dimension reduction, and RMAVE-$\mathfrak{F}_C$ is the most efficient method for multivariate dimension reduction since it is a nonparametric way to estimate the basis of the subspace. We include the results from sequential way of DCOV1 and DCOV2 as well, which is to calculate the first single direction, and calculate the second direction in the orthogonal subspace of the first direction and so on.

The accuracies are measured by the distance between the two subspaces (Li, Zha, and Chiaromonte, 2005):

$$\Delta_m(\hat{\mathcal{S}}_1, \mathcal{S}_2) = ||\mathbf{P}_{\hat{\mathcal{S}}_1} - \mathbf{P}_{\mathcal{S}_1}||$$

where $||\cdot||$ is the maximum singular value of a matrix, $\mathcal{S}_1$, and $\mathcal{S}_2$ are two same dimensional subspace, and $\mathbf{P}_{\hat{\mathcal{S}}_1}$ and $\mathbf{P}_{\mathcal{S}_2}$ are the orthogonal projections onto the subspace $\hat{\mathcal{S}}_1$ and $\mathcal{S}_2$, respectively. The smaller the Δ_m is, the better the estimate is.

3.3.1 Simulations

Here we simulate several models. For each model setting, 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 study relationship between the number of random directions and accuracy, and choose $m = 50$. We also consider three different designs on predictors for each model to examine that if the model assumption can be extended to distributions other than normal. The three different designs are: part (I), standard multivariate normal predictors, $\boldsymbol{X} \sim N(0, \boldsymbol{I})$; part (II), non-normal but continues predictors; and part (III), discrete predictors. The following three models come from the paper by Li, Wen, and Zhu (2008).

Model 1 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:

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

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

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

In part (I), $\boldsymbol{X}$ follows the standard normal distribution; in part (II), $X_i \sim Unif(-\sqrt{3}, \sqrt{3})$, for $i = 1, ..., 6$; and in part (III), $X_i \sim Poisson(1)$, for $i = 1, ..., 3$, and $X_i \sim N(0, 1)$, for $i = 4, ..., 6$. For this model, Table 3.1 gives the mean and standard deviation of the estimation accuracy (Δ_m) based on $N = 100$ simulated samples for each combination of eight models and three sample sizes.

PRSIR performs relatively well, because the response is a linear function of the predictors. DCOV0 performs the best for all three sample sizes and three different designs. RMAVE-$\mathfrak{F}_C$ also performs well, but worse in all cases than DCOV0, especially for discrete setting. DCOV1 and DCOV2 is not better than DCOV0 in all cases, which may due to the fact that the objective functions in DCOV1 and DCOV2 in the optimization problem is much more complicated than DCOV0, and the algorithm is stuck in the local optimum. The error decreases substantially for all methods but sequential DCOV1 as the sample size increases, reflecting the fact that they are consistent. Figure 3.1 shows the relationship between accuracy and different number of random directions from simulation of sample size 100 for method DCOV1. The figure indicates that a choice of $m = 50$ is reasonable.

Table 3.1: Comparison based on Model 1

Part (1)				Part (2)				Part (3)			
n	Method	$\bar{\Delta}_m$	SE_{Δ_m}	n	Method	$\bar{\Delta}_m$	SE_{Δ_m}	n	Method	$\bar{\Delta}_m$	SE_{Δ_m}
100	PRSIR	0.4001	0.1773	100	PRSIR	0.5552	0.1458	100	PRSIR	0.3918	0.1889
	PRSAVE	0.9097	0.0977		PRSAVE	0.9398	0.0950		PRSAVE	0.9273	0.1125
	RMAVE-$\mathfrak{F}_C$	0.4144	0.1283		RMAVE-$\mathfrak{F}_C$	0.4387	0.1492		RMAVE-$\mathfrak{F}_C$	0.5454	0.2420
	DCOV0	0.2730	0.0923		DCOV0	0.2609	0.0940		DCOV0	0.2373	0.1051
	DCOV1	0.5831	0.2267		DCOV1	0.7695	0.1743		DCOV1	0.4993	0.2611
	DCOV2	0.4083	0.2136		DCOV2	0.4696	0.2448		DCOV2	0.3729	0.2372
	DCOV1-seq	0.4956	0.2088		DCOV1-seq	0.4709	0.2131		DCOV1-seq	0.5840	0.2802
	DCOV2-seq	0.3536	0.1934		DCOV2-seq	0.2287	0.0855		DCOV2-seq	0.2040	0.1737
200	PRSIR	0.3372	0.1520	200	PRSIR	0.4938	0.1540	200	PRSIR	0.3313	0.1767
	PRSAVE	0.7482	0.2174		PRSAVE	0.8891	0.1587		PRSAVE	0.8237	0.1985
	RMAVE-$\mathfrak{F}_C$	0.2595	0.0817		RMAVE-$\mathfrak{F}_C$	0.3378	0.1441		RMAVE-$\mathfrak{F}_C$	0.3263	0.0828
	DCOV0	0.1943	0.0624		DCOV0	0.1825	0.0583		DCOV0	0.1453	0.0545
	DCOV1	0.4802	0.2437		DCOV1	0.4232	0.2395		DCOV1	0.5873	0.2922
	DCOV2	0.2745	0.2102		DCOV2	0.3816	0.2349		DCOV2	0.3129	0.2373
	DCOV1-seq	0.5028	0.2294		DCOV1-seq	0.4312	0.2248		DCOV1-seq	0.6700	0.2893
	DCOV2-seq	0.3699	0.1669		DCOV2-seq	0.1621	0.0972		DCOV2-seq	0.1190	0.1272
400	PRSIR	0.2985	0.1603	400	PRSIR	0.4314	0.1136	400	PRSIR	0.3025	0.1572
	PRSAVE	0.6288	0.2392		PRSAVE	0.8773	0.1594		PRSAVE	0.8047	0.2153
	RMAVE-$\mathfrak{F}_C$	0.1963	0.0602		RMAVE-$\mathfrak{F}_C$	0.1797	0.0771		RMAVE-$\mathfrak{F}_C$	0.2507	0.0918
	DCOV0	0.1412	0.0417		DCOV0	0.1308	0.0399		DCOV0	0.0945	0.0366
	DCOV1	0.4470	0.2576		DCOV1	0.5143	0.2588		DCOV1	0.6652	0.3203
	DCOV2	0.2114	0.1860		DCOV2	0.3167	0.2165		DCOV2	0.2806	0.2151
	DCOV1-seq	0.5251	0.2163		DCOV1-seq	0.5220	0.2777		DCOV1-seq	0.7447	0.2814
	DCOV2-seq	0.3449	0.1886		DCOV2-seq	0.1298	0.0437		DCOV2-seq	0.0766	0.1033

Model 2 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:

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

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

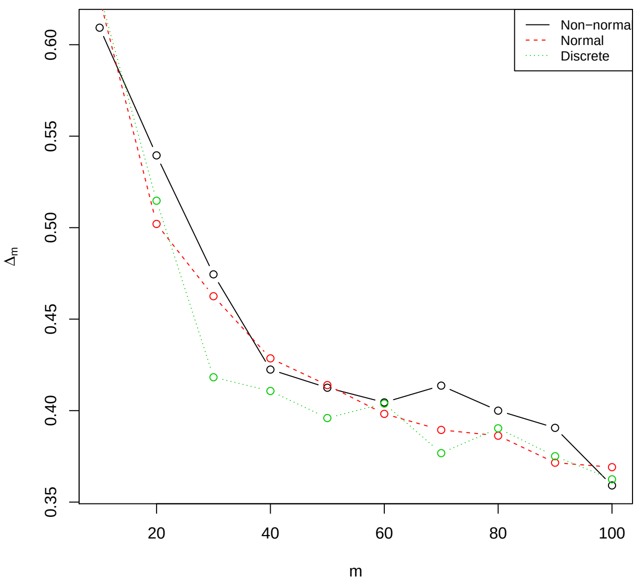


Figure 3.1: Relationship between the number of random directions m and accuracy $\bar{\Delta}_m$ for Model 1 using DCOV1.

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

This model is the same as Model 1 except that Y_1 is a quadratic form of $\boldsymbol{\beta}_1^T\boldsymbol{X}$. Table 3.2 reports the results for Model 2. We can see that PRSAVE performs better than PRSIR for normal design, because Y_1 has a quadratic function of $\boldsymbol{\beta}_1^T\boldsymbol{X}$. DCOV0 outperforms other methods for normal design and discrete design, and is the second best for non-normal design. Its derivatives, DCOV1 and DCOV2, both perform better than PRSIR and PRSAVE. Sequential DCOV1 and DCOV2 performs well on normal design and discrete design, but it seems that they are not consistent for non-normal design, while the accuracy decrease with sample size from 200 to 400. RMAVE-$\mathfrak{F}_C$ performs the best on non-normal design, but not as well as DCOV0, and sequential DCOV1 or DCOV2 on normal design and discrete design. The accuracy increases when the sample size increases for almost all these methods, indicating consistent estimates. Figure 3.2 suggests that number of random direction $m = 50$ is reasonable for DCOV1.

Table 3.2: Comparison based on Model 2

Part (1)				Part (2)				Part (3)			
n	Method	Δ_m	SE_{Δ_m}	n	Method	Δ_m	SE_{Δ_m}	n	Method	Δ_m	SE_{Δ_m}
100	PRSIR	0.8851	0.1563	100	PRSIR	0.8503	0.1597	100	PRSIR	0.2685	0.0815
	PRSAVE	0.8235	0.1902		PRSAVE	0.6010	0.2377		PRSAVE	0.8765	0.1632
	RMAVE-$\mathfrak{F}_C$	0.4486	0.2715		RMAVE-$\mathfrak{F}_C$	0.6451	0.5345		RMAVE-$\mathfrak{F}_C$	0.3092	0.0720
	DCOV0	0.5547	0.2951		DCOV0	0.7795	0.2317		DCOV0	0.1819	0.0609
	DCOV1	0.5316	0.2543		DCOV1	0.8683	0.1528		DCOV1	0.2665	0.0809
	DCOV2	0.5200	0.2926		DCOV2	0.8695	0.1508		DCOV2	0.2931	0.1511
	DCOV1-seq	0.3328	0.1425		DCOV1-seq	0.6704	0.2221		DCOV1-seq	0.2491	0.1105
	DCOV2-seq	0.4946	0.2937		DCOV2-seq	0.6486	0.3352		DCOV2-seq	0.1640	0.0759
200	PRSIR	0.8750	0.1484	200	PRSIR	0.8087	0.1835	200	PRSIR	0.2024	0.0648
	PRSAVE	0.3405	0.1450		PRSAVE	0.5542	0.1826		PRSAVE	0.5266	0.2175
	RMAVE-$\mathfrak{F}_C$	0.1963	0.0644		RMAVE-$\mathfrak{F}_C$	0.2928	0.2912		RMAVE-$\mathfrak{F}_C$	0.2409	0.0862
	DCOV0	0.3527	0.2700		DCOV0	0.5980	0.3017		DCOV0	0.1155	0.0378
	DCOV1	0.3612	0.2015		DCOV1	0.7736	0.2233		DCOV1	0.1964	0.0727
	DCOV2	0.4011	0.3173		DCOV2	0.8793	0.1641		DCOV2	0.2765	0.1857
	DCOV1-seq	0.2090	0.0697		DCOV1-seq	0.4272	0.2010		DCOV1-seq	0.1952	0.0904
	DCOV2-seq	0.2381	0.2025		DCOV2-seq	0.4869	0.3907		DCOV2-seq	0.1199	0.0651
400	PRSIR	0.6731	0.2335	400	PRSIR	0.7350	0.2212	400	PRSIR	0.1750	0.0579
	PRSAVE	0.2262	0.0763		PRSAVE	0.5288	0.1589		PRSAVE	0.4727	0.2266
	RMAVE-$\mathfrak{F}_C$	0.2042	0.1482		RMAVE-$\mathfrak{F}_C$	0.1303	0.0353		RMAVE-$\mathfrak{F}_C$	0.1845	0.0577
	DCOV0	0.1443	0.0466		DCOV0	0.3870	0.3038		DCOV0	0.0733	0.0250
	DCOV1	0.2059	0.1547		DCOV1	0.6723	0.2815		DCOV1	0.1514	0.0464
	DCOV2	0.2983	0.3333		DCOV2	0.8060	0.2490		DCOV2	0.2489	0.1803
	DCOV1-seq	0.1398	0.0460		DCOV1-seq	0.6594	0.2375		DCOV1-seq	0.1623	0.0712
	DCOV2-seq	0.1712	0.1800		DCOV2-seq	0.5090	0.4282		DCOV2-seq	0.1014	0.0759

Model 3 Let $p = 6, q = 5$, $\boldsymbol{X} \sim N(\mathbf{0}, \mathbf{I}_6)$. We increase the dimension of $\boldsymbol{Y}$ to five,

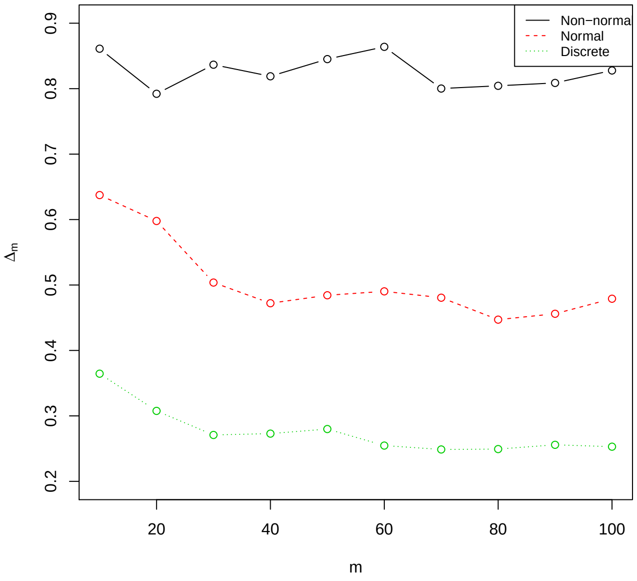


Figure 3.2: Relationship between the number of random directions m and accuracy $\bar{\Delta}_m$ for Model 2 using DCOV1.

which is generated as:

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

where $\boldsymbol{\epsilon} \sim N_4(\mathbf{0}, \boldsymbol{\Delta})$ with $\boldsymbol{\Delta} = diag(\boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2)$

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

Table 3.3 reports the results for Model 3. All methods perform well with high accuracies. DCOV0 outperforms all other methods with smallest $\bar{\Delta}_m$, and smallest SE_{Δ_m} for all three designs and three sample sizes. RMAVE-$\mathfrak{F}_C$ performs well for normal design, but not as well as other DCOV methods on non-normal design and discrete design. The errors decrease rapidly when sample size increases for all methods, which means that all estimates are consistent. Using DCOV1 again, Figure 3.3 shows that $m = 50$ is reasonable for Model 3.

Table 3.3: Comparison based on Model 3

Part (1)				Part (2)				Part (3)			
n	Method	$\bar{\Delta}_m$	SE_{Δ_m}	n	Method	$\bar{\Delta}_m$	SE_{Δ_m}	n	Method	$\bar{\Delta}_m$	SE_{Δ_m}
100	PRSIR	0.2716	0.1048	100	PRSIR	0.7269	0.1860	100	PRSIR	0.3071	0.1393
	PRSAVE	0.8732	0.1547		PRSAVE	0.5679	0.2320		PRSAVE	0.8142	0.1995
	RMAVE-$\mathfrak{F}_C$	0.2386	0.0827		RMAVE-$\mathfrak{F}_C$	0.3131	0.1367		RMAVE-$\mathfrak{F}_C$	0.4173	0.1538
	DCOV0	0.2477	0.0877		DCOV0	0.2002	0.0791		DCOV0	0.1556	0.0902
	DCOV1	0.2809	0.1155		DCOV1	0.2012	0.0603		DCOV1	0.2436	0.1124
	DCOV2	0.2718	0.2008		DCOV2	0.4010	0.2062		DCOV2	0.2633	0.1983
	DCOV1-seq	0.3032	0.0877		DCOV1-seq	0.2341	0.0758		DCOV1-seq	0.2715	0.0995
	DCOV2-seq	0.2503	0.1242		DCOV2-seq	0.1876	0.0921		DCOV2-seq	0.1049	0.0866
200	PRSIR	0.2250	0.0743	200	PRSIR	0.6074	0.1818	200	PRSIR	0.2618	0.0903
	PRSAVE	0.5462	0.2452		PRSAVE	0.2944	0.1279		PRSAVE	0.3823	0.2069
	RMAVE-$\mathfrak{F}_C$	0.1700	0.0610		RMAVE-$\mathfrak{F}_C$	0.2179	0.0885		RMAVE-$\mathfrak{F}_C$	0.2595	0.0612
	DCOV0	0.1569	0.0519		DCOV0	0.1283	0.0414		DCOV0	0.0987	0.0677
	DCOV1	0.2170	0.1252		DCOV1	0.1493	0.0440		DCOV1	0.1579	0.0498
	DCOV2	0.1573	0.1021		DCOV2	0.3787	0.2134		DCOV2	0.2377	0.1789
	DCOV1-seq	0.2338	0.0800		DCOV1-seq	0.1597	0.0499		DCOV1-seq	0.1986	0.0730
	DCOV2-seq	0.1430	0.0488		DCOV2-seq	0.1331	0.0640		DCOV2-seq	0.0698	0.0686
400	PRSIR	0.1660	0.0592	400	PRSIR	0.2826	0.1235	400	PRSIR	0.2229	0.0905
	PRSAVE	0.2855	0.1578		PRSAVE	0.2405	0.0876		PRSAVE	0.1991	0.0719
	RMAVE-$\mathfrak{F}_C$	0.1071	0.0367		RMAVE-$\mathfrak{F}_C$	0.1302	0.0424		RMAVE-$\mathfrak{F}_C$	0.2069	0.0702
	DCOV0	0.1071	0.0319		DCOV0	0.0870	0.0231		DCOV0	0.0633	0.0175
	DCOV1	0.1597	0.1061		DCOV1	0.1273	0.1560		DCOV1	0.1189	0.0357
	DCOV2	0.1223	0.1579		DCOV2	0.3539	0.2556		DCOV2	0.2235	0.1786
	DCOV1-seq	0.1650	0.0623		DCOV1-seq	0.1182	0.0305		DCOV1-seq	0.1260	0.0414
	DCOV2-seq	0.1114	0.0467		DCOV2-seq	0.1147	0.0822		DCOV2-seq	0.0542	0.0556

The estimation of the dimension (d) of the central subspace for the three different models is conducted to illustrate how the kNN method works. For each model, we

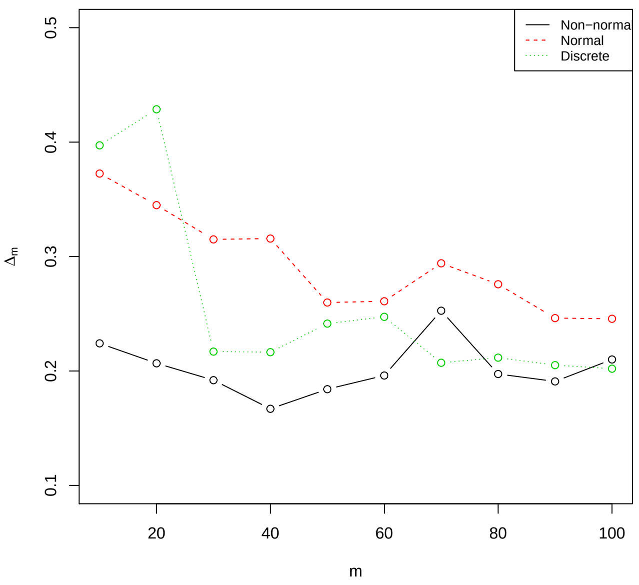


Figure 3.3: Relationship between the number of random directions m and accuracy $\bar{\Delta}_m$ for Model 3 using DCOV1.

simulation examples with (n, p) = (400, 6). The ratios of λ for Models 1, 2, and 3 are summarized in the Table 3.4. One can see from the table that maximum ratios happen at the dimension of the central subspace, which indicates that the kNN method correctly estimates the dimension under different models.

Table 3.4: Ratio of eigenvalues **r** 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*	1.0317	1.0651	1.0593
2	1.1668	1.3368*	1.0551	1.0254	1.0053
3	1.8969	2.1869*	1.0869	1.0123	1.0381

3.3.2 Application

In this section, we analyze the Minneapolis Elementary Schools data set, which is obtained from Cook (1998, p.216), and is used to explore the relationship between students' performance and characteristics of school. The data set has 63 observations (schools) and 13 variables. The response is a four dimension multivariate variable, which is described as:

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

- 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: percent of children in the school living with Both Parents

- AFDC: percent 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: percent of adults in the school area who have completed high school

- Enrol: number of children enrolled in the 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: percent minority children in the area.

This data is discussed by Yin and Bura (2006) to demonstrate their moment based dimension reduction method for multivariate response. In order to satisfy the two assumptions of their method, they square-root transformed the response as well as the explanatory variables. DCOV method does not require the assumption of distribution, so we perform the dimension reduction on the original data. But in order to compare our result with the work of Yin and Bura (2006), we also conduct the dimension reduction on the square-root transformed data. The kNN method described in section 3.2.6 results in Table 3.5. The maximum ratios for both cases are the first one, suggesting that the estimated dimension is one. This also agrees with the analysis of Yin and Bura (2006).

Table 3.5: Ratio of eigenvalues **r** for Minneapolis Elementary Schools data

	r_1	r_2	r_3	r_4	r_5	r_6	r_7	r_8
Ratio-original	4.439447*	2.093754	1.092490	1.336794	1.162351	1.213189	1.126388	1.097821
Ratio-sqrt tranformation	6.793484*	1.087571	1.234982	1.164279	1.403584	1.150167	1.102947	1.095085

Table 3.6 shows the estimated direction at the original scale. AFDC and HSchl contribute most to the estimated direction, which makes sense.

Table 3.7 shows the estimated direction at the original scale. Similar to the results of the original data, the coefficients of $\sqrt{AFDC}$ and $\sqrt{HSchl}$ have the highest absolute values, indicating they contribute most to the estimated direction.

Table 3.6: Estimated direction by DCOV1 on original data.

Variables	$\hat{\beta}_1$
BP	-0.2531
AFDC	0.5762
Poverty	-0.2060
HSchl	-0.4777
Enrol	0.37029
Attend	-0.0362
Mobility	0.3007
PT-ratio	-0.0270
Minority	0.3217

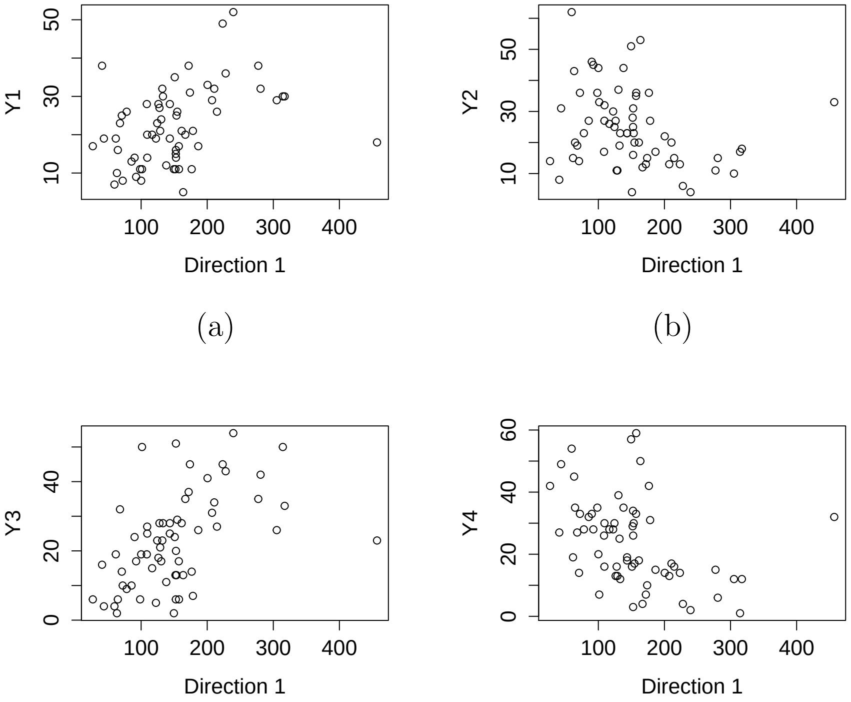
Table 3.7: Estimated direction by DCOV1 on square-root scale.

Variables	$\hat{\beta}_1$
BP	-0.0502
AFDC	-0.6843
Poverty	0.0627
HSchl	0.5334
Enrol	-0.2467
Attend	0.0388
Mobility	-0.3060
PT-ratio	0.0770
Minority	-0.2807

The scatterplot of the four response variables vs. the estimated direction for the original scale can be found in Figure 3.4, and the square-root in Figure 3.5. From both figures, it seems that using square-root scale may be better to establish models for each response. Linear model seems good enough for each response.

3.4 Discussion

In this article, we extend DCOV methods to sufficient dimension reduction with multivariate response. We also present two DCOV methods (DCOV1 and DCOV2) using projective resampling on multivariate response to convert the SDR with multivariate response into univariate response. DCOV0 performs well on different models with



(a)

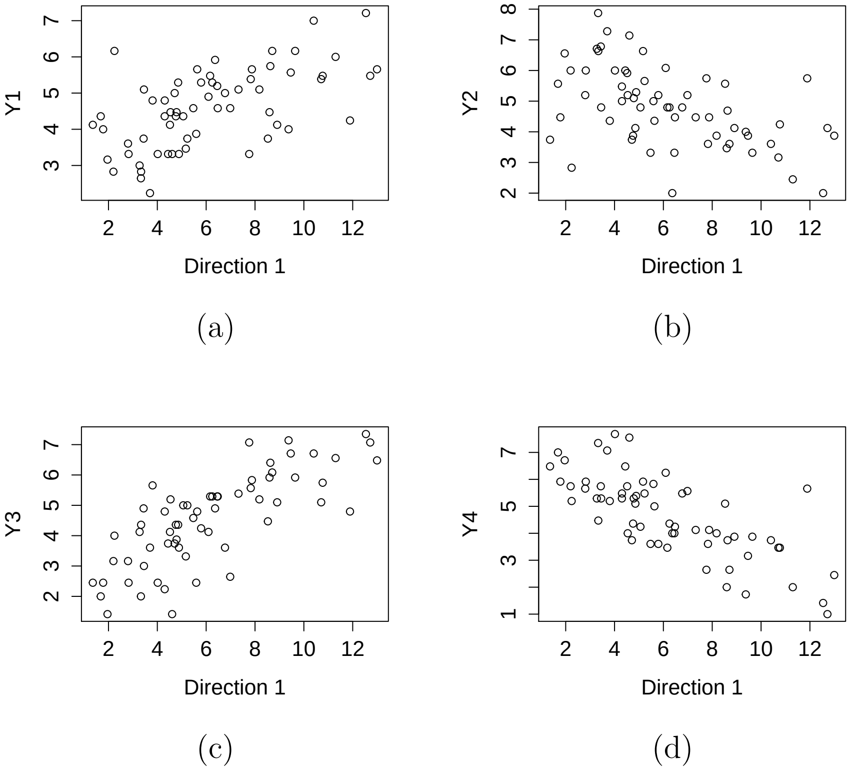
(b)

(c)

(d)

Figure 3.4: Scatterplot of four responses and the estimated direction under original scale of predictors for Minneapolis Elementary Schools data.

highest accuracy. DCOV1 and DCOV2 perform relatively well, better than projection resampling with SIR and SAVE. We also introduced a kNN method for estimating d. Results show that this method correctly chooses the dimension under different models. Theoretical properties for DCOV on multivariate response such as asymptotic results are established based on the work of Sheng and Yin (2015). Along the line, we apply our method to a real Minneapolis Elementary Schools data. The result agrees with previous study as well.



(a)

(b)

(c)

(d)

Figure 3.5: Scatterplot of four responses and the estimated direction in square-root scale for Minneapolis Elementary Schools data.

Bibliography

[1] X. Chen and X. Yin. "NlcOptim: an R package for nonlinear constrained optimization program". In: *Journal of Statistical Software* (2015), submitted.

[2] R. D. Cook. "Graphics for regressions with a binary response". In: *Journal of the American Statistical Association* 91.435 (1996), pp. 983–992.

[3] R. D. Cook. "On the interpretation of regression plots". In: *Journal of the American Statistical Association* 89.425 (1994), pp. 177–189.

[4] R. D. Cook. *Regression graphics: Ideas for studying regression through graphics*. Wiley, 1998.

[5] R. D. Cook and L. Ni. "Sufficient dimension reduction via inverse regression: a minimum discrepancy approach". In: *Journal of the American Statistical Association* 100 (2005), pp. 410–428.

[6] R. D. Cook and C. M. Setodji. "A model-free test for reduced rank in multivariate regression". In: *Journal of the American Statistical Association* 98.462 (2003), pp. 340–351.

[7] R.D. Cook and S. Weisberg. "Discussion of a paper by KC Li". In: *Journal of the American Statistical Association* 86 (1991), pp. 328–32.

[8] K. Fukumizu, F. R. Bach, and M. I. Jordan. “Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces”. In: *The Journal of Machine Learning Research* 5 (2004), pp. 73–99.

[9] W. K. Fung et al. “Dimension reduction based on canonical correlation”. In: *Statistica Sinica* (2002), pp. 1093–1113.

[10] H. Hilafu and X. Yin. “Sufficient dimension reduction in multivariate regressions with categorical predictors”. In: *Computational Statistics & Data Analysis* 63 (2013), pp. 139–147.

[11] M. Hristache et al. “Structure adaptive approach for dimension reduction”. In: *The Annals of Statistics* 29.6 (2001), pp. 1537–1566.

[12] B. Li, S. Wen, and L. Zhu. “On a projective resampling method for dimension reduction with multivariate responses”. In: *Journal of the American Statistical Association* 103.483 (2008), pp. 1177–1186.

[13] B. Li, H. Zha, and F. Chiaromonte. “Contour Regression: A general approach to dimension reduction”. In: *Annals of statistics* (2005), pp. 1580–1616.

[14] K.C. Li. “On principal hessian directions for data visualization and dimension reduction: another application of stein’s lemma”. In: *Journal of the American Statistical Association* 87.420 (1992), pp. 1025–1039.

[15] K.C. Li. “Sliced inverse regression for dimension reduction”. In: *Journal of the American Statistical Association* 86.414 (1991), pp. 316–327.

[16] W. Sheng and X. Yin. “Direction estimation in single-index models via distance covariance”. In: *Journal of Multivariate Analysis* 122 (2013), pp. 148–161.

[17] W. Sheng and X. Yin. “Sufficient dimension reduction via distance covariance”. In: *Journal of Computational and Graphical Statistics* just-accepted (2015).

[18] G. J. Székely, M. L. Rizzo, N. K. Bakirov, et al. “Measuring and testing dependence by correlation of distances”. In: *The Annals of Statistics* 35.6 (2007), pp. 2769–2794.

[19] G. J. Székely, M. L. Rizzo, et al. “Brownian distance covariance”. In: *The annals of applied statistics* 3.4 (2009), pp. 1236–1265.

[20] H. Wang and Y. Xia. “Sliced regression for dimension reduction”. In: *Journal of the American Statistical Association* 103.482 (2008), pp. 811–821.

[21] Q. Wang, X. Yin, and F. Critchley. “Dimension reduction based on the hellinger integral”. In: *Biometrika* 102.1 (2015), pp. 95–106.

[22] Y. Xia et al. “An adaptive estimation of dimension reduction space”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64.3 (2002), pp. 363–410.

[23] Z. Ye and R. E. Weiss. “Using the bootstrap to select one of a new class of dimension reduction methods”. In: *Journal of the American Statistical Association* (2011).

[24] X. Yin and E. Bura. “Moment-based dimension reduction for multivariate response regression”. In: *Journal of Statistical Planning and Inference* 136.10 (2006), pp. 3675–3688.

[25] X. Yin and R. D. Cook. “Direction estimation in single-index regressions”. In: *Biometrika* 92.2 (2005), pp. 371–384.

[26] X. Yin and B. Li. “Sufficient dimension reduction based on an ensemble of minimum average variance estimators”. In: *The Annals of Statistics* (2011), pp. 3392–3416.

[27] X. Yin, B. Li, and R. D. Cook. “Successive direction extraction for estimating the central subspace in a multiple-index regression”. In: *Journal of Multivariate Analysis* 99.8 (2008), pp. 1733–1757.

[28] P. Zeng and Y. Zhu. “An integral transform method for estimating the central mean and central subspaces”. In: *Journal of Multivariate Analysis* 101.1 (2010), pp. 271–290.

[29] Y. Zhu and P. Zeng. “Fourier methods for estimating the central subspace and the central mean subspace in regression”. In: *Journal of the American Statistical Association* 101.476 (2006), pp. 1638–1651.

Appendix

Lemma 1 *Suppose* $\boldsymbol{\eta}$ *is a basis of the central subspace. 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 $\tilde{\boldsymbol{X}}_1 = \boldsymbol{\eta}_1^\top \boldsymbol{X}$, $\tilde{\boldsymbol{X}}_2 = \boldsymbol{\eta}_2^\top \boldsymbol{X}$, $F(a,b) = \mathcal{V}^2(\begin{pmatrix} a\tilde{\boldsymbol{X}}_1 \\ b\tilde{\boldsymbol{X}}_2 \end{pmatrix}, \boldsymbol{Y})$, $a \in R$, 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 $0 \leq \lambda < 1$, then $F(1,\lambda) < F(1,1)$, and $F(\lambda,1) < F(1,1)$.

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

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

However, $F(1, \lambda) = \lambda F(\frac{1}{\lambda}, 1)$, and as $\lambda \to \infty$, $F(\frac{1}{\lambda}, 1) \to F(0, 1) > 0$. 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 (λ_0, ∞). 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 $F(1, \lambda) = \lambda F(\frac{1}{\lambda}, 1)$, we can have $G_1(\frac{1}{\lambda}, 1) = 0$. And it is easy to calculate that $G_1(1, \lambda_1) = G_1(\frac{1}{\lambda}, 1) = 0$.

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 1.

Since $\mathcal{S}(\boldsymbol{\beta}) \subseteq \mathcal{S}(\boldsymbol{\eta}) = \mathcal{S}_{Y|X}$, $d_1 \leq d$, 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, Rizzo, et al. (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

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

The equality holds if and only if $d = d_1$. According to Székely, Rizzo, et al. (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

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

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

Proof of Proposition 2.

For the $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ in Proposition 2, there exists a rotation matrix $\boldsymbol{Q}$ such that $\boldsymbol{\beta}\boldsymbol{Q} = (\eta_a, \eta_b)$, 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 $\mathbf{P}_{\boldsymbol{B}(\boldsymbol{\Sigma}_X)}^\top \boldsymbol{X} \perp\!\!\!\perp \mathbf{Q}_{\boldsymbol{B}(\boldsymbol{\Sigma}_X)}^\top \boldsymbol{X}$, 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, Rizzo, et al. (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

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

Chapter 4

Canonical Analysis and Dual Central Subspace Estimation via Distance Covariance[1]

[1] Chen, X. and Yin, X. *To be Submitted to [Computational Statistics & Data Analysis].*

Abstract

In this article, we extend Distance Covariance (DCOV) methods to Canonical Correlation Analysis (CCA), termed as Canonical Distance Covariance Analysis (CDCA), where we explore the relationship between two multivariate sets of variables. Compared with traditional CCA, CDCA captures both nonlinear and linear relationship. In addition, we extend DCOV to estimate the dual central subspace (DCS)–that is to find the basis that span the subspace of **Y** as well as the basis that span the subspace of **X**–by adding another constraint in the optimization problem. This approach keeps the model-free advantage, and its performance is investigated through multiple simulation examples and a real data analysis.

Key Words: Distance Covariance; Dual Central subspace; Projective Resampling; Sufficient Dimension Reduction.

4.1 Introduction

Suppose there are two sets of variables: **Y** a $q \times 1$ vector and **X** a $p \times 1$ vector. If one set, say **X**, is known as the predictor, then sufficient dimension reduction (SDR; Li 1991; Cook 1994; Cook 1996) can be used as a methodology for reducing the dimension of **X** while preserving the regression relation with response **Y**. However, there are cases that the role of predictor and response is not important, but the relationship between them is interesting. Methods for dimension reduction in multivariate association investigate this kind of problem. Canonical Correlation Analysis (CCA), introduced by Hotelling (1936), is a standard method in multivariate analysis to extract pairwise linear relationship between two random vectors, by maximizing their correlation. Kettenring (1971) extended CCA to multiple sets, by maximizing a generalized 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 Kullback-Leibler (KL) information to find linear and nonlinear relationships between two sets of random vectors. Yin and Sriram (2008), Iaci et al. (2008) and Iaci, Sriram, and Yin (2010) extended this idea to independent groups and multiple sets of random vectors.

Importantly, all of these CCA methods require that the number of coefficient vectors from both sets that provide the dimension reduction be equal. While this restriction simplifies the problem, if the number of coefficient vectors that recover the true associations between the random vectors is not equal then this could result in a critical loss of information. Therefore, methods that allow the number of coefficient vectors to be different and thus, provide a sufficient dimension reduction, are crucial in multivariate analysis. Iaci, Yin, and Zhu (2015) introduced the Dual Central Subspaces (DCS), which provides a dimension reduction of both vectors without requiring the dimensions of the reduction to be equal.

Some traditional methods in dimension reduction can be extended for estimating the Dual Central Subspaces. Since sliced inverse regression (SIR; Li 1991) method, many statistical studies have focused on dimension reduction in a regression setting, for example, Sliced Average Variance Estimate (SAVE; Cook and Weisberg 1991), Principal Hessian Directions (PHD; Li 1992), Minimum Average Variance Estimate (MAVE; Xia et al. 2002). All of these methods consider only a univariate response and thus, dimension reduction is performed only on the predictor variables. A few methods have been developed in a multivariate regression setting, but the dimension reduction is focused only on the predictors; see for example Cook and Setodji (2003), Yin and Bura (2006) and Li, Wen, and Zhu (2008). Methods for sufficient dimension reduction, especially with a multivariate response, such as Zhu, Zhu, and Wen

(2010) and Setodji and Cook (2004), could also be considered to develop a method to identify the DCS. More recently, Cook, Li, and Chiaromonte (2010) developed an envelope model for multivariate linear regression that not only reduces the dimension of the predictors, but also eliminates the noninformative responses in order to obtain a more efficient estimator. While their method and those of others, such as Su and Cook (2011), Su and Cook (2012), and Su and Cook (2013), have made significant advances in this area, the focus of these techniques are only on the regression mean function for a specified regression model. The proposed method of Li et al. (2003) in order to achieve a dimension reduction in a multivariate response regression setting could be considered for the identification of the DCS, however the linearity conditions and the exhaustive nature of recovering all the directions using SIR based method are viewed to be somewhat restrictive. Iaci, Yin, and Zhu (2015) considered a higher-order information measure based on the Kullback-Leibler (KL) divergence, which is able to detect both linear and nonlinear relationships that exist between random vectors.

More recently, Sheng and Yin (2013) and Sheng and Yin (2015) developed a novel method using distance covariance for sufficient dimension reduction (Székely, Rizzo, Bakirov, et al., 2007; Székely, Rizzo, et al., 2009). The method does not require linearity condition or constant covariance condition, or any particular distribution on $\mathbf{X}$, $\mathbf{X}|Y$ or $Y|\mathbf{X}$. These advantages enable the method to work effectively under a variety of $\mathbf{X}$: $\mathbf{X}$ could be normal, non-normal but continuous, or discrete or categorical. Together with projective resampling method, DCOV is generalized to multivariate response setting. In this article, we extend DCOV to canonical analysis as Canonical Distance Covariance Analysi (CDCA) and to estimate DCS.

The article is organized as follows: In Section 4.2.2 we introduce the procedure of the CDCA, which extends distance covariance method to extract pairwise relation between two sets. The extension of DCOV to the identification of the DCS and computational aspects of our approach are described in Sections 4.2.3 and 4.2.4. Estimation of diemnsion is in Section 4.2.5. Simulation studies and the analysis of the Los Angeles County dataset that was initially investigated in Shumway, Azari, and Pawitan (1988) to gain further insight into the associations that exist between mortality and environmental conditions using our method are in Section 4.3. A short discussion for our methods is in Section 4.4.

4.2 Methodology

4.2.1 Distance covariance as a sufficient dimension reduction tool

DCOV is introduced by Székely, Rizzo, Bakirov, et al. (2007) as a new measure of multivariate dependence. Let $\boldsymbol{Z}_1 \in \mathbb{R}^p, \boldsymbol{Z}_2 \in \mathbb{R}^q$ 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:

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

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, c_q are constants, is chosen as positive. An equivalent form of squared DCOV is given by Székely, Rizzo, et al. (2009) under finite assumption of $E|\boldsymbol{Z}_1|$ and $E|\boldsymbol{Z}_2|$ as

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

where $(\boldsymbol{Z}_1,\boldsymbol{Z}_2)$,$(\boldsymbol{Z}_1',\boldsymbol{Z}_2')$,$(\boldsymbol{Z}_1'',\boldsymbol{Z}_2'')$ are *i.i.d.* distributed.

DCOV equals to 0 if and only if two random vectors are independent (Székely, Rizzo, Bakirov, et al., 2007). Based on this property, Sheng and Yin (2013) and Sheng and Yin (2015) proposed DCOV as a sufficient dimension reduction tool. Suppose $\boldsymbol{\beta}$ is a $p \times d$ matrix, where $1 \leq d \leq q$. The solution to the following optimization problem will yield a basis of the central subspace.

$$\max_{\boldsymbol{\beta}^T\boldsymbol{\Sigma}_X\boldsymbol{\beta}=\mathbf{I}_d} \mathcal{V}^2(\boldsymbol{\beta}^T\mathbf{X},Y) \tag{4.1}$$

under $E|\mathbf{X}| < \infty$ and $E|Y| < \infty$. So throughout the article we assume $E|\mathbf{X}| < \infty$ and $E|Y| < \infty$. The constraint in the optimization problem guarantees the solution of $\boldsymbol{\beta}$ in the same scale and the optimization solver does not diverge.

Sheng and Yin (2013) and Sheng and Yin (2015) developed the DCOV method for the case that the response is a scalar. Chen and Yin (2016) extended DCOV method to multivariate response. Suppose $\boldsymbol{X}$ is $p\times 1$ random vector, $\boldsymbol{Y}$ is $q \times 1$ random vector, and $\boldsymbol{\beta}$ is a $p \times d$ matrix, where $1 \leq d \leq q$. A basis of the central subspace can be obtained by solving the following optimization problem.

$$\max_{\boldsymbol{\beta}^T\boldsymbol{\Sigma}_X\boldsymbol{\beta}=\mathbf{I}_d} \mathcal{V}^2(\boldsymbol{\beta}^T\boldsymbol{X},\boldsymbol{Y}). \tag{4.2}$$

under $E|\boldsymbol{X}| < \infty$ and $E|\boldsymbol{Y}| < \infty$. Sheng and Yin (2013) and Sheng and Yin (2015) demonstrated that under some mild conditions, the solution to (1) always spans the central subspace. Chen and Yin (2016) generalized the conclusion to multivariate response cases for (2).

Based on the sample version of squared distance covariance $\mathcal{V}_n^2(\boldsymbol{\beta}^T\mathbf{X}, \mathbf{Y})$ proposed by Székely, Rizzo, Bakirov, et al. (2007), a sample version for multivariate response can be defined as

$$\mathcal{V}^2(\boldsymbol{\beta}^\top\boldsymbol{X}, \boldsymbol{Y}) = \frac{1}{n^2}\sum_{k,l=1}^{n} A_{kl}(\boldsymbol{\beta})B_{kl}, \tag{4.3}$$

where, for $k, l = 1, \cdots, n$,

$$\begin{gathered}
A_{kl}(\boldsymbol{\beta}) = a_{kl}(\boldsymbol{\beta}) - \bar{a}_{k.}(\boldsymbol{\beta}) - \bar{a}_{.l}(\boldsymbol{\beta}) + \bar{a}_{..}(\boldsymbol{\beta}) \\
a_{kl}(\boldsymbol{\beta}) = |\boldsymbol{\beta}^T\mathbf{X}_k - \boldsymbol{\beta}^T\mathbf{X}_l|, \bar{a}_{k.}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{l=1}^{n} a_{kl}(\boldsymbol{\beta}), \\
\bar{a}_{.l}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{k=1}^{n} a_{kl}(\boldsymbol{\beta}), \bar{a}_{..}(\boldsymbol{\beta}) = \frac{1}{n^2}\sum_{k,l=1}^{n} a_{kl}(\boldsymbol{\beta}).
\end{gathered}$$

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 central subspace is

$$\boldsymbol{\eta}_n = \arg\max_{\boldsymbol{\beta}^T\hat{\boldsymbol{\Sigma}}_X\boldsymbol{\beta}=\mathbf{I}_d} \mathcal{V}_n^2(\boldsymbol{\beta}^\top\boldsymbol{X}, \boldsymbol{Y}). \tag{4.4}$$

We can solve this nonlinear optimization problem using Sequential Quadratic Programming method.

4.2.2 Canonical distance covariance analysis (CDCA)

Canonical correlation analysis (CCA) is the most well-known multivariate methods that explore the amount of linear relationship between two sets of variables, $\boldsymbol{X}$ and

$\boldsymbol{Y}$. Suppose that $\boldsymbol{X}$ is a $p \times 1$ random vector, $\boldsymbol{Y}$ is a $q \times 1$ random vector, and $\boldsymbol{X}$ and $\boldsymbol{Y}$ have zero mean and covariance matrix $\boldsymbol{\Sigma_{XY}}$. Consider two linear combination $\boldsymbol{a}^T\boldsymbol{X}$ and $\boldsymbol{b}^T\boldsymbol{Y}$, the goal of CCA is to find the vectors of $\boldsymbol{a}$ and $\boldsymbol{b}$ that maximize $\rho(\boldsymbol{a}, \boldsymbol{b})$, where $\rho(\boldsymbol{a}, \boldsymbol{b})$ is defined as

$$\rho(\boldsymbol{a}, \boldsymbol{b}) = \frac{\boldsymbol{a}^T\boldsymbol{\Sigma_{XY}}\boldsymbol{b}}{(\boldsymbol{a}^T\boldsymbol{\Sigma_X}\boldsymbol{a}\boldsymbol{b}^T\boldsymbol{\Sigma_Y}\boldsymbol{b})^{1/2}}.$$

Equivalently, CCA is to solve the following optimization problem for $\boldsymbol{a}$ and $\boldsymbol{b}$:

$$\max_{\substack{\boldsymbol{a}^T\boldsymbol{\Sigma}_X\boldsymbol{a}=1 \\ \boldsymbol{b}^T\boldsymbol{\Sigma}_Y\boldsymbol{b}=1}} \boldsymbol{a}^T\boldsymbol{\Sigma_{XY}}\boldsymbol{b}$$

However, CCA fails when the relationship is nonlinear. In order to solve this problem, Yin (2004) developed a technique based on an information theory which enables a generalized CCA to capture nonlinear relationship. Recently, Iaci and Sriram (2013) applied this method using beta-divergence and power divergence, and Mandal and Cichocki (2013) generalized their method.

We propose a new method which uses distance covariance to capture the nonlinear relationship between two sets of variables. The problem can be defined as to find the vectors of $\boldsymbol{a}$ and $\boldsymbol{b}$ that maximize the squared distance covariance

$$\max_{\substack{\boldsymbol{b}^\top\boldsymbol{\Sigma}_X\boldsymbol{b}=1 \\ \boldsymbol{a}^\top\boldsymbol{\Sigma}_Y\boldsymbol{a}=1}} \mathcal{V}^2(\boldsymbol{b}^\top\boldsymbol{X}, \boldsymbol{a}^\top\boldsymbol{Y}).$$

While estimating the coefficient vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ simultaneously, that is $<\hat{\boldsymbol{a}}, \hat{\boldsymbol{b}}>=$ $\arg\max_{\substack{\boldsymbol{b}^\top\boldsymbol{\Sigma}_X\boldsymbol{b}=1 \\ \boldsymbol{a}^\top\boldsymbol{\Sigma}_Y\boldsymbol{a}=1}} \mathcal{V}^2(\boldsymbol{b}^\top\boldsymbol{X}, \boldsymbol{a}^\top\boldsymbol{Y})$ may result in solving nonlinear optimization problem with many parameters, we propose two different approaches dividing this procedure into steps that estimate $\boldsymbol{a}$ and $\boldsymbol{b}$ separately:

Approach 1 Estimate the coefficient vectors $\boldsymbol{b}$ and $\boldsymbol{a}$, i.e. $\hat{\boldsymbol{b}} = \arg\max_{\boldsymbol{b}^\top \hat{\boldsymbol{\Sigma}}_X \boldsymbol{b}=1} \mathcal{V}_n^2(\boldsymbol{b}^\top \mathbf{X}, \boldsymbol{Y})$ and $\hat{\boldsymbol{a}} = \arg\max_{\boldsymbol{a}^\top \hat{\boldsymbol{\Sigma}}_Y \boldsymbol{a}=1} \mathcal{V}_n^2(\mathbf{X}, \boldsymbol{a}^\top \boldsymbol{Y})$ at the same time. This means that we obtain the estimate of $\hat{\boldsymbol{a}}$ using $\boldsymbol{X}$, not $\boldsymbol{b}^\top \boldsymbol{X}$, in the squared distance covariance.

Approach 2 After Calculating

$$\hat{\boldsymbol{b}} = \arg \max_{\boldsymbol{b}^\top \hat{\boldsymbol{\Sigma}}_X \boldsymbol{b}=1} \mathcal{V}_n^2(\boldsymbol{b}^\top \mathbf{X}, \boldsymbol{Y}),$$

obtain $\boldsymbol{a}$ with the projection $\boldsymbol{b}^\top \boldsymbol{X}$,

$$\hat{\boldsymbol{a}} = \arg \max_{\boldsymbol{a}^\top \hat{\boldsymbol{\Sigma}}_Y \boldsymbol{a}=1} \mathcal{V}_n^2(\hat{\boldsymbol{b}}^\top \mathbf{X}, \boldsymbol{a}^\top \boldsymbol{Y});$$

The main difference between these two approaches is whether to use $\hat{\boldsymbol{b}}$ to calculate $\hat{\boldsymbol{a}}$ or not. In both two approaches described above, DCOV method with multivariate response is used in the optimization problem to obtain estimates of $\boldsymbol{a}$ and $\boldsymbol{b}$, we call them method "Approach 1 DCOV0" and "Approach 2 DCOV0" respectively. Chen and Yin (2016) introduced two derivatives of DCOV method with projective resampling on the multivariate response, DCOV1 and DCOV2. When combining with the procedure of Approach 1 and Approach 2, we can have methods "Approach 1 DCOV1","Approach 2 DCOV1","Approach 1 DCOV2", and "Approach 2 DCOV2". The optimization problem of these methods are summarized briefly in Table 4.1, where $\boldsymbol{t}_i$, $i = 1, ..., m$ is a random direction that the multivariate response is projected onto, and m is the total number of random directions. Once the first pair of the directions is estimated, the search for the second pair of the directions is the same way as before but in the space that is orthogonal to the first direction. We compare six models and classical CCA in the simulation section.

Table 4.1: Methods for CDCA

method	estimate $\boldsymbol{b}$	estimate $\boldsymbol{a}$
Approach1 DCOV0	$\max\limits_{\boldsymbol{b}^\top \hat{\boldsymbol{\Sigma}}_X \boldsymbol{b}=1} \mathcal{V}_n^2(\boldsymbol{b}^\top \mathbf{X}, \boldsymbol{Y})$	$\max\limits_{\boldsymbol{a}^\top \hat{\boldsymbol{\Sigma}}_Y \boldsymbol{a}=1} \mathcal{V}_n^2(\mathbf{X}, \boldsymbol{a}^\top \boldsymbol{Y})$
Approach2 DCOV0	$\max\limits_{\boldsymbol{b}^\top \hat{\boldsymbol{\Sigma}}_X \boldsymbol{b}=1} \mathcal{V}_n^2(\boldsymbol{b}^\top \mathbf{X}, \boldsymbol{Y})$	$\max\limits_{\boldsymbol{a}^\top \hat{\boldsymbol{\Sigma}}_Y \boldsymbol{a}=1} \mathcal{V}_n^2(\hat{\boldsymbol{b}}^\top \mathbf{X}, \boldsymbol{a}^\top \boldsymbol{Y})$
Approach1 DCOV1*	$\max\limits_{\boldsymbol{b}^\top \hat{\boldsymbol{\Sigma}}_X \boldsymbol{b}=1} \mathcal{V}^2(\boldsymbol{b}^\top \mathbf{X}, \boldsymbol{t}_i{}^\top \mathbf{Y})$	$\max\limits_{\boldsymbol{a}^\top \hat{\boldsymbol{\Sigma}}_Y \boldsymbol{a}=1} \mathcal{V}^2(\boldsymbol{t}_i^\top \mathbf{X}, \boldsymbol{a}^\top \mathbf{Y})$
Approach2 DCOV1*	$\max\limits_{\boldsymbol{b}^\top \hat{\boldsymbol{\Sigma}}_X \boldsymbol{b}=1} \mathcal{V}^2(\boldsymbol{b}^\top \mathbf{X}, \boldsymbol{t}_i{}^\top \mathbf{Y})$	$\max\limits_{\boldsymbol{a}^\top \hat{\boldsymbol{\Sigma}}_Y \boldsymbol{a}=1} \mathcal{V}^2(\hat{\boldsymbol{b}}^\top \mathbf{X}, \boldsymbol{a}^\top \mathbf{Y})$
Approach1 DCOV2	$\max\limits_{\boldsymbol{b}^\top \hat{\boldsymbol{\Sigma}}_X \boldsymbol{b}=1} \sum\limits_{i=1}^{m} \mathcal{V}^2(\boldsymbol{b}^\top \mathbf{X}, \boldsymbol{t}_i{}^\top \mathbf{Y})$	$\max\limits_{\boldsymbol{a}^\top \hat{\boldsymbol{\Sigma}}_Y \boldsymbol{a}=1} \sum\limits_{i=1}^{m} \mathcal{V}^2(\boldsymbol{t}_i^\top \mathbf{X}, \boldsymbol{a}^\top \mathbf{Y})$
Approach2 DCOV2	$\max\limits_{\boldsymbol{b}^\top \hat{\boldsymbol{\Sigma}}_X \boldsymbol{b}=1} \sum\limits_{i=1}^{m} \mathcal{V}^2(\boldsymbol{b}^\top \mathbf{X}, \boldsymbol{t}_i{}^\top \mathbf{Y})$	$\max\limits_{\boldsymbol{a}^\top \hat{\boldsymbol{\Sigma}}_Y \boldsymbol{a}=1} \mathcal{V}^2(\hat{\boldsymbol{b}}^\top \mathbf{X}, \boldsymbol{a}^\top \mathbf{Y})$

* Approach1 DCOV1 and Approach 2 DCOV1: For each $\boldsymbol{t}_i$, we obtain estimates, $\hat{\boldsymbol{b}}_i$ and $\hat{\boldsymbol{a}}_i$, the final estimates $\hat{\boldsymbol{b}}$ and $\hat{\boldsymbol{a}}$ can be obtained by singular value decomposition of the $\sum_{i=1}^m \hat{\boldsymbol{b}}_i \hat{\boldsymbol{b}}_i^\top$, $\sum_{i=1}^m \hat{\boldsymbol{a}}_i \hat{\boldsymbol{a}}_i^\top$, respectively.

4.2.3 The Dual Central Subspaces

Consider two sets of random vectors, $\mathbf{X}$ is $p \times 1$ and $\mathbf{Y}$ is $q \times 1$. The ultimate goal of sufficient dimension reduction is to search a number of linear combinations of $\mathbf{X}$, say $\boldsymbol{\beta}^\top \mathbf{X}$, where $\boldsymbol{\beta}$ is a $p \times d$ matrix, $d < p$, such that $\mathbf{Y}$ depends on $\mathbf{X}$ only through $\boldsymbol{\beta}^\top \mathbf{X}$. That is,

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

The column space of $\boldsymbol{\beta}$ is called $\mathcal{S}(\boldsymbol{\beta})$, and the intersection of such subspaces is defined as the central subspace, denoted by $\mathcal{S}_{Y|\mathbf{X}}$. By exchange the role of $\mathbf{X}$ and $\mathbf{Y}$, we can define the CS of $\mathbf{Y}$ as $\mathcal{S}_{\mathbf{X}|\mathbf{Y}}$, which plays an important role in reducing the dimension of $\mathbf{Y}$ sufficiently.

In the sense of treating $\mathbf{X}$ and $\mathbf{Y}$ equally and reducing the dimension of both variables, Iaci, Yin, and Zhu (2015) defined the Dual Central Subspaces (DCS) as the combination of $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ and $\mathcal{S}_{\mathbf{X}|\mathbf{Y}}$. Different from the linear combination in usual canonical approach, the dimensions of the central subspaces $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ and $\mathcal{S}_{\mathbf{X}|\mathbf{Y}}$, d_x and d_y can differ in DCS. Iaci, Yin, and Zhu (2015) also demonstrated the difference between DCS and CCA and its extensions with an example in their paper. If the true dimension of central subspace differs, that is $d_x \neq d_y$, making the reduced dimensions equal will underestimate the CS with higher dimension or overestimate the CS with lower dimension. This calls for the necessity of recovering DCS while studying the multivariate association.

The proposition below suggests ways to recover the DCS.

Proposition 5 *(Iaci, Yin, and Zhu, 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:

$$\begin{aligned}
(i) \quad & \mathbf{Y} \perp\!\!\!\perp \mathbf{X} | \boldsymbol{B}^\top \mathbf{X} \quad \textit{and} \quad \mathbf{Y} \perp\!\!\!\perp \mathbf{X} | \boldsymbol{A}^\top \mathbf{Y}, \\
(ii) \quad & \mathbf{Y} \perp\!\!\!\perp \mathbf{X} | \boldsymbol{B}^\top \mathbf{X} \quad \textit{and} \quad \mathbf{Y} \perp\!\!\!\perp \boldsymbol{B}^\top \mathbf{X} | \boldsymbol{A}^\top \mathbf{Y}, \\
(iii) \quad & \boldsymbol{A}^\top \mathbf{Y} \perp\!\!\!\perp \mathbf{X} | \boldsymbol{B}^\top \mathbf{X} \quad \textit{and} \quad \mathbf{Y} \perp\!\!\!\perp \mathbf{X} | \boldsymbol{A}^\top \mathbf{Y}.
\end{aligned}$$

Proposition 5 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.

4.2.4 Estimating DCS via distance covariance

Assume the dimensions d_x and d_y are known. Let $(\boldsymbol{x}_i, \boldsymbol{y}_i), i = 1, ..., n$ 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:

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

The two constraints $\boldsymbol{B}^\top\hat{\boldsymbol{\Sigma}}_X\boldsymbol{B} = I_{d_x}$, and $\boldsymbol{A}^\top\hat{\boldsymbol{\Sigma}}_Y\boldsymbol{A} = I_{d_y}$ guarantee the estimated directions have unit length and are 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.

As the discussion in section 4.2.2, 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 difference of the estimation of $\boldsymbol{A}$ depends on $\hat{\boldsymbol{B}}$ or not. The procedure of these two approaches are described as Approach 3 and Approach 4 below, with multivariate response in the squared distance covariance (DCOV0) as the objective function in optimization problem.

Approach 3 Estimate $\hat{\boldsymbol{B}}$ with $\boldsymbol{Y}$ as response, and $\hat{\boldsymbol{A}}$ with $\boldsymbol{X}$ as response. This means, we can calculate $\hat{\boldsymbol{B}} = \arg\max_{\boldsymbol{B}^\top\hat{\boldsymbol{\Sigma}}_X\boldsymbol{B}=I_{d_x}} \mathcal{V}_n^2(\boldsymbol{B}^\top\mathbf{x}, \boldsymbol{y})$ and $\hat{\boldsymbol{A}} = \arg\max_{\boldsymbol{A}^\top\hat{\boldsymbol{\Sigma}}_Y\boldsymbol{A}=I_{d_y}} \mathcal{V}_n^2(\mathbf{x}, \boldsymbol{A}^\top\boldsymbol{y})$ at the same time, since the two steps do not depend on each other;

Approach 4 Estimate $\hat{\boldsymbol{B}}$ with $\boldsymbol{Y}$ as response, and then $\hat{\boldsymbol{A}}$ with $\boldsymbol{B}^\top\boldsymbol{X}$ as response. That is, after calculating

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

we 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 "Approach3 DCOV0" and "Approach4 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 3 DCOV1","Approach 4 DCOV1","Approach 3 DCOV2" and "Approach 4 DCOV2", whose optimization problems are summarized in Table 4.2. Note that $\boldsymbol{t}_i$,

$i = 1, ..., m$, is a random direction that the multivariate response is projected onto, and m is the total number of random directions. For DCOV1, m estimates of $\hat{\boldsymbol{B}}_i$ is obtained for each random direction $\boldsymbol{t}_i$, $i = 1, ..., m$. The estimate $\hat{\boldsymbol{B}}$ is calculated by singular value decomposition of the sum of $\hat{\boldsymbol{B}}_i \hat{\boldsymbol{B}}_i^\top$.

Table 4.2: Methods for DCS

method	estimate $\boldsymbol{B}$	estimate $\boldsymbol{A}$
Approach3 DCOV0	$\max\limits_{\boldsymbol{B}^\top \hat{\boldsymbol{\Sigma}}_X \boldsymbol{B} = \boldsymbol{I}_{d_x}} \mathcal{V}_n^2(\boldsymbol{B}^\top \mathbf{X}, \boldsymbol{Y})$	$\max\limits_{\boldsymbol{A}^\top \hat{\boldsymbol{\Sigma}}_Y \boldsymbol{A} = \boldsymbol{I}_{d_y}} \mathcal{V}_n^2(\mathbf{X}, \boldsymbol{A}^\top \boldsymbol{Y})$
Approach4 DCOV0	$\max\limits_{\boldsymbol{B}^\top \hat{\boldsymbol{\Sigma}}_X \boldsymbol{B} = 1} \mathcal{V}_n^2(\boldsymbol{B}^\top \mathbf{X}, \boldsymbol{Y})$	$\max\limits_{\boldsymbol{A}^\top \hat{\boldsymbol{\Sigma}}_Y \boldsymbol{A} = \boldsymbol{I}_{d_y}} \mathcal{V}_n^2(\hat{\boldsymbol{A}}^\top \mathbf{X}, \boldsymbol{A}^\top \boldsymbol{Y})$
Approach3 DCOV1*∗	$\max\limits_{\boldsymbol{B}^\top \hat{\boldsymbol{\Sigma}}_X \boldsymbol{B} = \boldsymbol{I}_{d_x}} \mathcal{V}^2(\boldsymbol{B}^\top \mathbf{X}, \boldsymbol{t}_i{}^\top \mathbf{Y})$	$\max\limits_{\boldsymbol{A}^\top \hat{\boldsymbol{\Sigma}}_Y \boldsymbol{A} = \boldsymbol{I}_{d_y}} \mathcal{V}^2(\boldsymbol{t}_i^\top \mathbf{X}, \boldsymbol{A}^\top \mathbf{Y})$
Approach4 DCOV1*∗	$\max\limits_{\boldsymbol{B}^\top \hat{\boldsymbol{\Sigma}}_X \boldsymbol{B} = \boldsymbol{I}_{d_x}} \mathcal{V}^2(\boldsymbol{B}^\top \mathbf{X}, \boldsymbol{t}_i{}^\top \mathbf{Y})$	$\max\limits_{\boldsymbol{A}^\top \hat{\boldsymbol{\Sigma}}_Y \boldsymbol{A} = \boldsymbol{I}_{d_y}} \mathcal{V}^2(\boldsymbol{t}_i^\top \hat{\boldsymbol{B}}^\top \mathbf{X}, \boldsymbol{A}^\top \mathbf{Y})$
Approach3 DCOV2	$\max\limits_{\boldsymbol{B}^\top \hat{\boldsymbol{\Sigma}}_X \boldsymbol{B} = \boldsymbol{I}_{d_x}} \sum\limits_{i=1}^{m} \mathcal{V}^2(\boldsymbol{B}^\top \mathbf{X}, \boldsymbol{t}_i{}^\top \mathbf{Y})$	$\max\limits_{\boldsymbol{A}^\top \hat{\boldsymbol{\Sigma}}_Y \boldsymbol{A} = \boldsymbol{I}_{d_y}} \sum\limits_{i=1}^{m} \mathcal{V}^2(\boldsymbol{t}_i^\top \mathbf{X}, \boldsymbol{A}^\top \mathbf{Y})$
Approach4 DCOV2	$\max\limits_{\boldsymbol{B}^\top \hat{\boldsymbol{\Sigma}}_X \boldsymbol{B} = \boldsymbol{I}_{d_x}} \sum\limits_{i=1}^{m} \mathcal{V}^2(\boldsymbol{B}^\top \mathbf{X}, \boldsymbol{t}_i{}^\top \mathbf{Y})$	$\max\limits_{\boldsymbol{A}^\top \hat{\boldsymbol{\Sigma}}_Y \boldsymbol{A} = \boldsymbol{I}_{d_y}} \sum\limits_{i=1}^{m} \mathcal{V}^2(\boldsymbol{t}_i^\top \hat{\boldsymbol{B}}^\top \mathbf{X}, \boldsymbol{A}^\top \mathbf{Y})$

* Approach1 DCOV1 and Approach 2 DCOV1: For each $\boldsymbol{t}_i$, we obtain estimates, $\hat{\boldsymbol{B}}_i$ and $\hat{\boldsymbol{A}}_i$, the final estimates $\hat{\boldsymbol{B}}$ and $\hat{\boldsymbol{A}}$ can be obtained by singular value decomposition of the $\sum_{i=1}^{m} \hat{\boldsymbol{B}}_i \hat{\boldsymbol{B}}_i^\top$, $\sum_{i=1}^{m} \hat{\boldsymbol{A}}_i \hat{\boldsymbol{a}}_i^\top$, respectively.

4.2.5 Estimating the dimensions of the DCS

In practice, the dimension of the dual central subspaces $< d_x, d_y >$, is unknown and needs to be estimated. There are multiple ways that are developed to estimate the central subspace. Li (1991) and Li (1992) proposed a sequential test based on a chi-squared statistic. Cook and Yin (2001) proposed the permutation test to determine the structural dimension. Ye and Weiss (2011), Zhu and Zeng (2006), and Sheng and Yin (2015) used bootstrap method to estimate d. Iaci, Yin, and Zhu (2015) modified bootstrap method to estimate the dimensions of the DCS using squared vector correlation coefficient. In this section, we adopt this idea by using the accuracy measured between the two subspaces. Two measures of accuracy are used in the simulation study:

1. Distance between the projection matrix (Li, Zha, and Chiaromonte, 2005):

$$\Delta_m(\hat{\mathcal{S}}_1, \mathcal{S}_2) = ||\mathbf{P}_{\hat{\mathcal{S}}_1} - \mathbf{P}_{\mathcal{S}_1}||$$

   where $||\cdot||$ is the maximum singular value of a matrix, $\mathcal{S}_1$,$\mathcal{S}_2$ are two same dimensional subspace, and $\mathbf{P}_{\hat{\mathcal{S}}_1}$ and $\mathbf{P}_{\mathcal{S}_2}$ are the orthogonal projections onto the subspace $\hat{\mathcal{S}}_1$ and $\mathcal{S}_2$ respectively. The smaller the Δ_m is, the better the estimate is.

2. Hotelling's squared vector correlation coefficient:

$$\rho^2(\hat{\boldsymbol{A}}) = |\boldsymbol{A}^\top \hat{\boldsymbol{A}} \hat{\boldsymbol{A}}^\top \boldsymbol{A}| = \prod_i^p \lambda_i,$$

   where λ_i are the eigenvalues of $\boldsymbol{A}^\top \hat{\boldsymbol{A}} \hat{\boldsymbol{A}}^\top \boldsymbol{A}$, and $0 \leq \rho(\hat{\boldsymbol{A}}) \leq 1$. The larger the $\rho(\hat{\boldsymbol{A}})$ is, the better the estimate is.

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, denote $\mathcal{S}_{\hat{\boldsymbol{B}}_k}$ 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 $\mathcal{S}_{\hat{\boldsymbol{A}}_l^b}$ 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 $\mathcal{S}_{\mathbf{X}|\mathbf{Y}}$. Use accuracy $\Delta_m(\hat{\mathcal{S}}_1, \mathcal{S}_2)$ 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 $\{(\mathbf{X}_i, \mathbf{Y}_i)\}, 1 \leq i \leq n$, the following procedure can be used to estimate the dimensions of the dual central subspaces:

1. For the fixed dimensions of (k,l), calculate the $\mathcal{S}_{\hat{\boldsymbol{B}}_k}$ and $\mathcal{S}_{\hat{\boldsymbol{A}}_l}$ based on the original data;

2. Randomly resample from $\{(\mathbf{X}_i,\mathbf{Y}_i)\}, 1 \le i \le n$, with replacement to 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.

4.3 Numerical studies

In this section, we assess the efficiency of the proposed methods through simulations and application to a real data. In the simulations, we compare the performance of six models for CDCA and DCS that described in Sections 4.2.2 and 4.2.4. For CDCA, classic canonical correlation analysis is also conducted to compare with the performance of our methods. Two measures of accuracy in Section 4.2.5 are presented to describe the the distance between the two subspaces.

4.3.1 Simulations

Here we simulate several models. For each model setting, 100 replicates of the data are generated. The comparison is made for three sample size $n = 100, 200$, and 300. For DCOV1 and DCOV2, we choose the number of random directions $m = 50$.

Canonical distance covariance analysis

In this section, we simulate examples to conduct canonical distance covariance analysis.

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:

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

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

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

The results for Model 4 is shown in Table 4.3. 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 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 $\boldsymbol{X}$.

Model 5 Now consider a linear relationship. We use the same model as Model 4 except that $Y_1 = \boldsymbol{\beta}_1^T \boldsymbol{X} + \epsilon_1$.

Table 4.4 compares the results of different approach of recovering central subspace for Model 5, as well as the results from CCA. CCA performs well in this case since the relationship between $\boldsymbol{X}$ and $\boldsymbol{Y}$ is linear. The DCOV methods perform well under both estimation approaches. The performance of recovering the directions gets better when sample size increases, 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 $\boldsymbol{X}$.

Dual Central Subspace

We investigate the performance of our methods for DCS by studying two examples. These two examples are similar to examples provided by Wang, Yin, and Critchley (2015). There are linear and nonlinear relationships between two sets of random

Table 4.3: Comparison based on Model 4

n	Order	DCOV	$\bar{\Delta}_m(\hat{\boldsymbol{b}})$	$SE_{\Delta_m}(\hat{\boldsymbol{b}})$	$\bar{\rho}(\hat{\boldsymbol{b}})$	$SE_{\rho}(\hat{\boldsymbol{b}})$	$\bar{\Delta}_m(\hat{\boldsymbol{a}})$	$SE_{\Delta_m}(\hat{\boldsymbol{a}})$	$\bar{\rho}(\hat{\boldsymbol{a}})$	$SE_{\rho}(\hat{\boldsymbol{a}})$
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

vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$. For each example setting, 100 replicates of the data are generated. The comparison is made for three sample size $n = 100, 200, 300$. For the two projective resampling based method 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 6 Let $p = 5, q = 4$, $\boldsymbol{X} \sim N(\mathbf{0}, \mathbf{I}_5)$. The four dimensional response random

Table 4.4: Comparison based on Model 5

n	Order	DCOV	$\bar{\Delta}_m(\hat{\boldsymbol{b}})$	$SE_{\Delta_m}(\hat{\boldsymbol{b}})$	$\bar{\rho}(\hat{\boldsymbol{b}})$	$SE_{\rho}(\hat{\boldsymbol{b}})$	$\bar{\Delta}_m(\hat{\boldsymbol{a}})$	$SE_{\Delta_m}(\hat{\boldsymbol{a}})$	$\bar{\rho}(\hat{\boldsymbol{a}})$	$SE_{\rho}(\hat{\boldsymbol{a}})$
100	Approach1	DCOV0	0.1691	0.0640	0.9673	0.0247	0.1934	0.0779	0.9565	0.0327
		DCOV1	0.1680	0.0643	0.9676	0.0242	0.1902	0.0877	0.9561	0.0415
		DCOV2	0.1447	0.0756	0.9733	0.0315	0.2577	0.0766	0.9277	0.0420
	Approach2	DCOV0	0.1691	0.0640	0.9673	0.0247	0.2024	0.0897	0.9510	0.0390
		DCOV1	0.1680	0.0643	0.9676	0.0242	0.1686	0.0639	0.9675	0.0240
		DCOV2	0.1447	0.0756	0.9733	0.0315	0.2441	0.0744	0.9348	0.0431
		CCA	0.0889	0.0296	0.9912	0.0057	0.4750	0.0727	0.7691	0.0667
200	Approach1	DCOV0	0.1152	0.0473	0.9844	0.0125	0.1613	0.0618	0.9701	0.0227
		DCOV1	0.1160	0.0431	0.9846	0.0108	0.1484	0.0527	0.9752	0.0160
		DCOV2	0.1003	0.0504	0.9874	0.0157	0.2450	0.0583	0.9365	0.0285
	Approach2	DCOV0	0.1152	0.0473	0.9844	0.0125	0.1608	0.0511	0.9715	0.0179
		DCOV1	0.1160	0.0431	0.9846	0.0108	0.1348	0.0462	0.9796	0.0132
		DCOV2	0.1003	0.0504	0.9874	0.0157	0.2420	0.0445	0.9394	0.0217
		CCA	0.0579	0.0160	0.9964	0.0019	0.4536	0.0539	0.7914	0.0487
300	Approach1	DCOV0	0.0862	0.0344	0.9913	0.0069	0.1377	0.0408	0.9793	0.0115
		DCOV1	0.0979	0.0361	0.9891	0.0077	0.1451	0.0428	0.9770	0.0131
		DCOV2	0.0892	0.0577	0.9887	0.0253	0.2505	0.0464	0.9350	0.0229
	Approach2	DCOV0	0.0862	0.0344	0.9913	0.0069	0.1373	0.0421	0.9793	0.0121
		DCOV1	0.0979	0.0361	0.9891	0.0077	0.1301	0.0357	0.9817	0.0093
		DCOV2	0.0892	0.0577	0.9887	0.0253	0.2490	0.0351	0.9367	0.0169
		CCA	0.0414	0.0135	0.9981	0.0012	0.4517	0.0343	0.7948	0.0309

vector $\boldsymbol{Y}$ is generated as:

$$\begin{aligned}
Y_1 &= 4\cos(\boldsymbol{B}_1^T\boldsymbol{X}) + 0.3\epsilon_1, \\
Y_2 &= \boldsymbol{B}_1^T\boldsymbol{X} + 0.5\epsilon_2, \\
Y_3 &= \epsilon_3, \\
Y_4 &= \epsilon_4.
\end{aligned}$$

where $\boldsymbol{B}_1 = (1,1,0,0,0)$, $\boldsymbol{A}_1 = (1,0,0,0)$, $\boldsymbol{A}_2 = (0,1,0,0)$ and $\boldsymbol{\epsilon} \sim N_4(\mathbf{0}, I_4)$.

Table 4.5 shows the results of six methods to recover dual central subspaces for Model 6. All the six methods perform well. For each method, $\bar{\Delta}_m$ decreases and $\bar{\rho}$ increases with sample size increases, 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 4.6 shows the results for different (k,l). We are looking for the smallest mean and least variability. We can see that $(k,l) = (1,1)$ and $(k,l) = (1,2)$ has the similar smallest mean, but $(k,l) = (1,2)$ has the least variability (0.0175 vs. 0.1409). So we will choose $(k,l) = (1,2)$, which agrees with the true dimension of dual central subspaces.

Model 7 Let $p = 5, q = 4$, $\boldsymbol{X} \sim N(\mathbf{0}, \mathbf{I}_5)$. The four dimensional response random vector $\boldsymbol{Y}$ is generated as:

$$\begin{aligned}
Y_1 &= 4\cos(\boldsymbol{B}_1^T\boldsymbol{X}) + 0.3\epsilon_1, \\
Y_2 &= \boldsymbol{B}_1^T\boldsymbol{X} + 0.5\epsilon_2, \\
Y_3 &= Y_4 + X_5 + 0.6\epsilon_3, \\
Y_4 &= \epsilon_4.
\end{aligned}$$

where $\boldsymbol{B}_1 = (1,0,1,0,0)$,$\boldsymbol{B}_2 = (0,0,0,0,1)$, $\boldsymbol{A}_1 = (1,0,0,0)$, $\boldsymbol{A}_2 = (0,1,0,0)$, $\boldsymbol{A}_3 = (0,0,1,-1)$ and $\boldsymbol{\epsilon} \sim N_4(\mathbf{0}, \boldsymbol{\Delta})$ with

$$\boldsymbol{\Delta} = \begin{bmatrix} 2 & -1 & \mathbf{0} \\ -1 & 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_2 \end{bmatrix}$$

The results of six methods to recover dual central subspaces for Model 7 is shown in Table 4.7. 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 $\boldsymbol{X}$. Table 4.8 shows the results for different (k,l). The smallest mean and least variability happen at $(k,l) = (2,3)$, which identify the dimension of dual central subspaces correctly.

Table 4.5: Comparison based on Model 6

n	Order	DCOV	$\bar{\Delta}_m(\hat{\boldsymbol{B}})$	$SE_{\Delta_m}(\hat{\boldsymbol{B}})$	$\bar{\rho}(\hat{\boldsymbol{B}})$	$SE_\rho(\hat{\boldsymbol{B}})$	$\bar{\Delta}_m(\hat{\boldsymbol{A}})$	$SE_{\Delta_m}(\hat{\boldsymbol{A}})$	$\bar{\rho}(\hat{\boldsymbol{A}})$	$SE_\rho(\hat{\boldsymbol{A}})$
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 4.6: Bootstrap distance measure for Model 6

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

4.3.2 LA pollution data

In this section, we analyze the LA pollution data set, which is obtained from Shumway, Azari, and Pawitan (1988), and was 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

Table 4.7: Comparison based on Model 7

n	Order	DCOV	$\bar{\Delta}_m(\hat{\boldsymbol{B}})$	$SE_{\Delta_m}(\hat{\boldsymbol{B}})$	$\bar{\rho}(\hat{\boldsymbol{B}})$	$SE_\rho(\hat{\boldsymbol{B}})$	$\bar{\Delta}_m(\hat{\boldsymbol{A}})$	$SE_{\Delta_m}(\hat{\boldsymbol{A}})$	$\bar{\rho}(\hat{\boldsymbol{A}})$	$SE_\rho(\hat{\boldsymbol{A}})$
100	Simultaneous	DCOV0	0.2383	0.0798	0.9310	0.0444	0.4909	0.1745	0.7288	0.1914
		DCOV1	0.3446	0.1822	0.8431	0.1756	0.3583	0.1584	0.8467	0.1471
		DCOV2	0.3107	0.1642	0.8672	0.1541	0.3839	0.2380	0.7964	0.2570
	Sequential	DCOV0	0.2383	0.0798	0.9310	0.0444	0.4802	0.1581	0.7446	0.1654
		DCOV1	0.3446	0.1822	0.8431	0.1756	0.2859	0.1336	0.9005	0.0958
		DCOV2	0.3107	0.1642	0.8672	0.1541	0.3169	0.2073	0.8569	0.1729
200	Simultaneous	DCOV0	0.1606	0.0594	0.9679	0.0212	0.4074	0.0948	0.8251	0.0805
		DCOV1	0.2216	0.0862	0.9392	0.0431	0.2743	0.1550	0.9009	0.1374
		DCOV2	0.2613	0.1739	0.8955	0.1613	0.4192	0.2900	0.7409	0.3151
	Sequential	DCOV0	0.1606	0.0594	0.9679	0.0212	0.4055	0.0942	0.8267	0.0794
		DCOV1	0.2216	0.0862	0.9392	0.0431	0.2437	0.1012	0.9304	0.0513
		DCOV2	0.2613	0.1739	0.8955	0.1613	0.2369	0.1638	0.9172	0.1435
300	Simultaneous	DCOV0	0.1339	0.0526	0.9774	0.0166	0.4192	0.1119	0.8118	0.1238
		DCOV1	0.1627	0.0761	0.9647	0.0283	0.2546	0.1063	0.9239	0.0580
		DCOV2	0.1909	0.1448	0.9387	0.1181	0.3961	0.2942	0.7573	0.3043
	Sequential	DCOV0	0.1339	0.0526	0.9774	0.0166	0.4112	0.0989	0.8211	0.1048
		DCOV1	0.1627	0.0761	0.9647	0.0283	0.2342	0.0758	0.9394	0.0362
		DCOV2	0.1909	0.1448	0.9387	0.1181	0.1607	0.1176	0.9604	0.0751

Table 4.8: Bootstrap distance measure for Model 7

k	l	$\Delta_{m,k,l}$	$SE_{\Delta_{m,k,l}}$
1	1	0.6473	0.1462
1	2	0.6647	0.1421
1	3	0.5537	0.1562
2	1	0.2046	0.0964
2	2	0.1332	0.0410
2	3	0.1090	0.0322
3	1	0.5217	0.1791
3	2	0.4263	0.1557
3	3	0.3992	0.1548

three mortality measures (total mortality, respiratory mortality and cardiovascular mortality) which countes all deaths of LA area, two weather measures (temperature and relative humidity), and six pollution measures including carbon monoxide, sulfur dioxide, nitrogen dioxide, hydrocarbons, ozone, particulates. The data is also discussed by Iaci, Sriram, and Yin (2010) and Iaci, Yin, and Zhu (2015).

We apply our method to the data in order 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 $\boldsymbol{X} = (X_1, ..., X_3)^\top$, where X_1 = temperature, X_2 = relative humidity, X_3 = carbon monoxide, X_4 = hydrocarbons, X_5 = ozone, and X_6 = particulates.

Table 4.9 shows results form the bootstrap method of Section 4.2.5 which estimates the dimension of the DCS to be $(k, l) = (1, 1)$. Table 4.10 shows the estimated direction of the multivariate response, and Table 4.11 is the estimated direction of the predictors. The loadings for the multivariate response indicate that Y_1 = total mortality and Y_2 = respiratory mortality contribute the most to the estimated direction, while Y_2 = respiratory mortality 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 = temperature contribute most positive to the estimated direction, and X_3 = carbon monoxide, X_4 = hydrocarbons and X_6 = particulates contribute equally negative to the estimated direction. The plot of the estimated directions of the multivariate response and the predictors is in Figure 4.1, indicating linear relationship between these two estimated directions, which agrees with Iaci, Yin, and Zhu (2015) very well. Projecting the data to the two estimated directions, that is, let $v = \hat{\boldsymbol{B}}_1^\top \boldsymbol{x}$ and $w = \hat{\boldsymbol{A}}_1^\top \boldsymbol{y}$, and fitting a linear regression model, we get $\hat{w} = -0.7033v$.

Table 4.9: Bootstrap distance measure for LA pollution data

k	l	$\bar{\Delta}_{m,k,l}$	$SE_{\bar{\Delta}_{m,k,l}}$
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 4.10: Estimated direction of the multivariate response by Approch2 DCOV0 for LA pollution data

Variables	$\hat{\boldsymbol{A}}$
total mortality	0.5875
respiratory mortality	0.0075
cardiovascular mortality	0.8095

4.4 Discussion

In this article, we extend DCOV method to canonical distance covariance analysis, where we explore the relationships between two multivariate sets of variables. The

Table 4.11: Estimated direction of the predictor by Approch2 DCOV0 for LA pollution data

Variables	$\hat{\boldsymbol{B}}$
temperature	0.6585
relative humidity	0.2534
carbon monoxide	-0.4162
hydrocarbons	-0.3917
ozone	0.1255
particulates	-0.3998

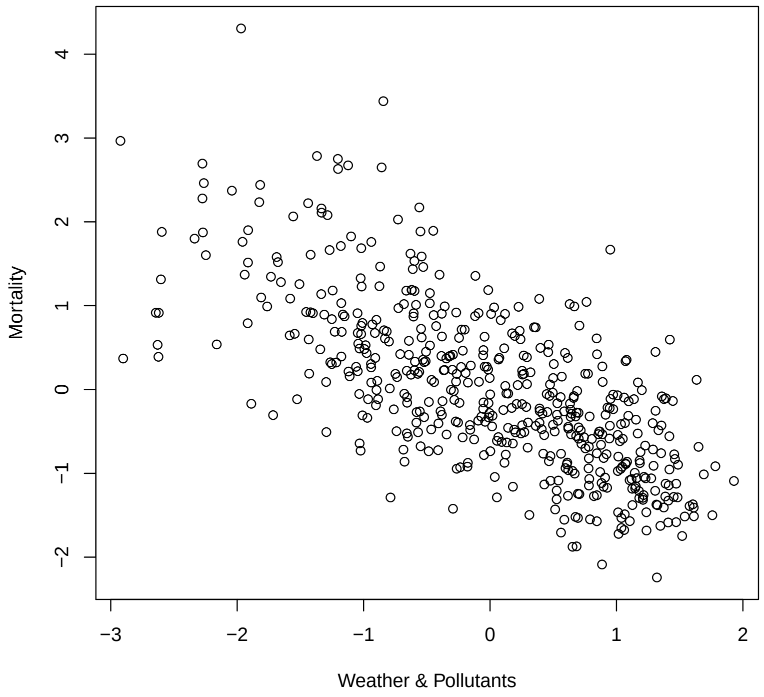


Figure 4.1: Relationship between direction of multivariate response and direction of predictors.

results show that comparing with the traditional CCA, our methods can capture linear relationship as well as nonlinear relationship. We also use DCOV for recovering dual central subspaces. The results show that all DCOV methods estimate the central subspaces of dual form with high accuracy. Two approaches for estimating $\boldsymbol{A}$, –based on $\boldsymbol{X}$ or $\boldsymbol{B}^\top \boldsymbol{X}$ – have been proposed. The results indicate that the latter approach recovers $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ with higher accuracy. A bootstrap procedure is used to identify the dimension of dual central subspace. Simulation examples show that this approach can find the true $< d_x, d_y >$ correctly. Asymptotic theory for the developed methods may be established, following the development of Sheng and Yin (2013) and Sheng and Yin (2015). Although the calculations can be tedious, the idea of the proofs is very straightforward based on Sheng and Yin (2013) and Sheng and Yin (2015).

Bibliography

[1] E. Burg and J. Leeuw. “Non-linear canonical correlation”. In: *British journal of mathematical and statistical psychology* 36.1 (1983), pp. 54–80.

[2] X. Chen and X. Yin. “Sufficient dimension reduction via distance covariance with multivariate responses”. In: (2016).

[3] R. D. Cook. “Graphics for regressions with a binary response”. In: *Journal of the American Statistical Association* 91.435 (1996), pp. 983–992.

[4] R. D. Cook. “On the interpretation of regression plots”. In: *Journal of the American Statistical Association* 89.425 (1994), pp. 177–189.

[5] R. D. Cook, B. Li, and F. Chiaromonte. “Envelope models for parsimonious and efficient multivariate linear regression”. In: *Statistica Sinica* (2010), pp. 927–960.

[6] R. D. Cook and C. M. Setodji. “A model-free test for reduced rank in multivariate regression”. In: *Journal of the American Statistical Association* 98.462 (2003), pp. 340–351.

[7] R.D. Cook and S. Weisberg. “Discussion of a paper by KC Li”. In: *Journal of the American Statistical Association* 86 (1991), pp. 328–32.

[8] H. Hotelling. “Relations between two sets of variants”. In: *Biometrika* 28 (1936), pp. 321–377.

[9] R. Iaci and T.N. Sriram. “Robust multivariate association and dimension reduction using density divergences”. In: *Journal of Multivariate Analysis* 117 (2013), pp. 281–295.

[10] R. Iaci, T.N. Sriram, and X. Yin. “Multivariate association and dimension reduction: A generalization of canonical correlation analysis”. In: *Biometrics* 66.4 (2010), pp. 1107–1118.

[11] R. Iaci, X. Yin, and L. Zhu. “The dual central subspaces in dimension reduction”. In: *Journal of Multivariate Analysis* (2015).

[12] R. Iaci et al. “An informational measure of association and dimension reduction for multiple sets and groups with applications in morphometric analysis”. In: *Journal of the American Statistical Association* 103.483 (2008), pp. 1166–1176.

[13] J. R. Kettenring. “Canonical analysis of several sets of variables”. In: *Biometrika* 58 (1971), pp. 433–451.

[14] B. Li, S. Wen, and L. Zhu. “On a projective resampling method for dimension reduction with multivariate responses”. In: *Journal of the American Statistical Association* 103.483 (2008), pp. 1177–1186.

[15] B. Li, H. Zha, and F. Chiaromonte. “Contour Regression: A general approach to dimension reduction”. In: *Annals of statistics* (2005), pp. 1580–1616.

[16] K.C. Li. “On principal hessian directions for data visualization and dimension reduction: another application of stein’s lemma”. In: *Journal of the American Statistical Association* 87.420 (1992), pp. 1025–1039.

[17] K.C. Li. “Sliced inverse regression for dimension reduction”. In: *Journal of the American Statistical Association* 86.414 (1991), pp. 316–327.

[18] K.C. Li et al. “Dimension reduction for multivariate response data”. In: *Journal of the American Statistical Association* 98.461 (2003), pp. 99–109.

[19] A. Mandal and A. Cichocki. “Non-linear canonical correlation analysis using alpha-beta divergence”. In: *Entropy* 15.7 (2013), pp. 2788–2804.

[20] C. M. Setodji and R. D. Cook. “K-means inverse regression”. In: *Technometrics* 46.4 (2004), pp. 421–429.

[21] W. Sheng and X. Yin. “Direction estimation in single-index models via distance covariance”. In: *Journal of Multivariate Analysis* 122 (2013), pp. 148–161.

[22] W. Sheng and X. Yin. “Sufficient dimension reduction via distance covariance”. In: *Journal of Computational and Graphical Statistics* just-accepted (2015).

[23] R.H. Shumway, A.S. Azari, and Y. Pawitan. “Modeling mortality fluctuations in Los Angeles as functions of pollution and weather effects”. In: *Environmental Research* 45.2 (1988), pp. 224–241.

[24] Z. Su and R. D. Cook. “Estimation of multivariate means with heteroscedastic errors using envelope models”. In: *Statistica Sinica* 23.1 (2013), pp. 213–230.

[25] Z. Su and R. D. Cook. “Inner envelopes: efficient estimation in multivariate linear regression”. In: *Biometrika* 99.3 (2012), pp. 687–702.

[26] Z. Su and R. D. Cook. “Partial envelopes for efficient estimation in multivariate linear regression”. In: *Biometrika* (2011), asq063.

[27] G. J. Székely, M. L. Rizzo, N. K. Bakirov, et al. “Measuring and testing dependence by correlation of distances”. In: *The Annals of Statistics* 35.6 (2007), pp. 2769–2794.

[28] G. J. Székely, M. L. Rizzo, et al. “Brownian distance covariance”. In: *The annals of applied statistics* 3.4 (2009), pp. 1236–1265.

[29] Q. Wang, X. Yin, and F. Critchley. "Dimension reduction based on the hellinger integral". In: *Biometrika* 102.1 (2015), pp. 95–106.

[30] Y. Xia et al. "An adaptive estimation of dimension reduction space". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64.3 (2002), pp. 363–410.

[31] Z. Ye and R. E. Weiss. "Using the bootstrap to select one of a new class of dimension reduction methods". In: *Journal of the American Statistical Association* (2011).

[32] X. Yin. "Canonical correlation analysis based on information theory". In: *Journal of Multivariate Analysis* 91.2 (2004), pp. 161–176.

[33] X. Yin and E. Bura. "Moment-based dimension reduction for multivariate response regression". In: *Journal of Statistical Planning and Inference* 136.10 (2006), pp. 3675–3688.

[34] X. Yin and T.N. Sriram. "Common canonical variates for independent groups using information theory". In: *Statistica Sinica* (2008), pp. 335–353.

[35] L. Zhu, L. Zhu, and S. Wen. "On dimension reduction in regressions with multivariate responses". In: *Statistica Sinica* (2010), pp. 1291–1307.

[36] Y. Zhu and P. Zeng. "Fourier methods for estimating the central subspace and the central mean subspace in regression". In: *Journal of the American Statistical Association* 101.476 (2006), pp. 1638–1651.

Chapter 5

Dual Variable Selection via Distance Covariance & Future Work[1]

[1] Chen, X. and Yin, X. *To be Submitted to [Technometrics].*

Abstract

Existing variable selection methods in regression setting have focused on choosing a subset of the predictor variables. However, for problems involving a multivariate response, selecting subset of the response vector is also important. In this article, we develop a new concept, termed the Dual Variable Selection (DVS), to propose a method for simultaneously selecting subsets for each of the two random vectors, by employing Distance Covariance (DCOV) method combined with LASSO (Tibshirani, 1996) penalty. This method is a model-free approach and does not need nonparametric smoothing. The performance of this method is investigated through simulation studies and a real data analysis.

Key Words: Distance Covariance; Dual Variable Selection; LASSO; Sufficient Dimension Reduction.

5.1 Introduction

The main difference between variable selection and sufficient dimension reduction is that variable selection selects important individual variables, while sufficient dimension reduction (SDR) creates new variables via using all the original ones. When people are interested in investigating the effect of original explanatory variables on the response, variable selection is desirable. Many methods have been proposed for variable selection. These include Model-based methods, for example, LASSO (Tibshirani, 1996), Smoothly Clipped Absolute Deviation (Fan and Li, 2001a), Dantzig selector (Candes and Tao, 2007), and Adaptive LASSO (Zou, 2006); Model-free methods (Li, Cook, and Nachtsheim, 2005), for example, Shrinkage inverse regression method (Bondell and Li, 2009); Methods for ultrahigh dimensional data, for example, SIS

(Fan and Lv, 2008), and Sequential SDR (Yin and Hilafu, 2015).

Chen, Sheng, and Yin (2016) developed a model free variable selection method using distance covariance. In their method, they used a LASSO penalty together with squared distance covariance in the objective function, and solved the optimization problem for the best direction. LASSO penalty forces the contribution of unimportant variable of the direction to zero.

In this article, we propose the idea of the Dual Variable Selection (DVS), where the response is multivariate as that of the explanatory variables and we are interested in selecting a subset of both sides. Inspired by DCOV variable selection method of Chen, Sheng, and Yin (2016), we propose a dual variable selection procedure using DCOV with a penalized approach. Section 5.2 describes the general idea and its algorithm. Section 5.3 sets up model simulations and application of our method to a real data, followed by a future work in Section 5.4.

5.2 Methodology

5.2.1 SDR and model-free variable selection

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 column vectors 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 on to $\mathcal{S}(\boldsymbol{B})$ with respect to the inner product $< \boldsymbol{a}, \boldsymbol{b} >= \boldsymbol{a}^T \boldsymbol{\Sigma} \boldsymbol{b}$. That is, $\mathbf{P}_{\boldsymbol{B}(\boldsymbol{\Sigma}_X)} = \boldsymbol{B}(\boldsymbol{B}^T \boldsymbol{\Sigma}_X \boldsymbol{B})^{-1} \boldsymbol{B}^T \boldsymbol{\Sigma}_X$. Let $\mathbf{Q}_{\boldsymbol{B}(\boldsymbol{\Sigma}_X)} = \mathbf{I} - \mathbf{P}_{\boldsymbol{B}(\boldsymbol{\Sigma}_X)}$, where $\mathbf{I}$ is the identity matrix.

The ultimate goal of sufficient dimension reduction is to search a number of linear combinations of $\mathbf{X}$, say $\boldsymbol{\beta}^T \mathbf{X}$, where $\boldsymbol{\beta}$ is a $p \times d$ matrix, $d < p$, such that Y depends on $\mathbf{X}$ only through $\boldsymbol{\beta}^T\mathbf{X}$. That is:

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

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 (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 central subspace is well-defined and unique. We assume central subspace exists throughout the paper.

Since SDR does not require any traditional model, Li, Cook, and Nachtsheim (2005) proposed a model-free variable selection method using SDR. One common idea is to add a penalty term to SDR method. Some methods are developed inspired by this idea, following Ni, Cook, and Tsai (2005). For instance, Li and Nachtsheim (2012) developed sparse SIR; Wang and Yin (2008) introduced sparse MAVE; Bondell and Li (2009) proposed a shrinkage inverse regression method; Sheng, Chen and Yin (2016) developed sparse DCOV method.

5.2.2 Dual variable selection

In this section, we define the Dual Sufficient Variable Selection (DSVS). Yin and Hilafu (2015) gave a formal Sufficient Variable Selection (SVS) definition. Suppose there is a $p \times p_0$ matrix $\boldsymbol{\beta}$, $(p_0 \leq p)$, where the columns of $\boldsymbol{\beta}$ consist of unit vectors of e_js with jth element 1 and 0 otherwise, such that

$$\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{X}|\boldsymbol{\beta}^\top\boldsymbol{X},$$

then the column space of $\boldsymbol{\beta}$ is called the Variable Selection Space. The Central Variable Selection Space is defined as the intersection of all such variable selection spaces, if itself satisfies the conditional independence condition above, denoted as $\mathcal{S}^V_{\mathbf{Y}|\mathbf{X}}$ with dimension s. $\mathcal{S}_{\mathbf{Y}|\mathbf{X}} \subset \mathcal{S}^V_{\mathbf{Y}|\mathbf{X}}$, and $d \leq s$. Also, if $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ exists, then $\mathcal{S}^V_{\mathbf{Y}|\mathbf{X}}$ exists and is unique.

In order to select a subset of $\boldsymbol{Y}$, we define the Central Variable Selection Space of $\boldsymbol{Y}$, denoted $\mathcal{S}^V_{\mathbf{X}|\mathbf{Y}}$, by simply interchanging the roles of $\boldsymbol{X}$ and $\boldsymbol{Y}$ in the above definition. That is suppose there is a $q \times q_0$ matrix $\boldsymbol{\alpha}$, $(q_0 \leq q)$, where the columns of $\boldsymbol{\alpha}$ consist of unit vectors of e_js with jth element 1 and 0 otherwise, such that $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}|\boldsymbol{\alpha}^\top\boldsymbol{Y}$. In the sense of selecting subsets of both sides, the two sets of variables can be treated equally and thus we term Dual Central Variable Selection Subspaces as the combination of $\mathcal{S}^V_{\mathbf{Y}|\mathbf{X}}$ and $\mathcal{S}^V_{\mathbf{X}|\mathbf{Y}}$. We assume the dual central variable selection subspaces exist throughout the paper.

5.2.3 Sparse DCOV as variable selection method

Chen, Sheng, and Yin (2016) developed sparse DCOV method for variable selection. Let $\boldsymbol{X}$ denote the predictor vector and Y denote a univariate response. DCOV method for SDR is to find a matrix under Grassman manifold where the maximum of squared distance covariance of the direction and response happens. By solving a nonlinear optimization problem, we can find an estimated basis matrix $\boldsymbol{\eta}_n$ of the central subspace. By adding a penalty term in the objective function, we can shrink some coefficients in $\boldsymbol{\eta}_n$ to zero. The problem is to solve:

$$\min_{\boldsymbol{\beta}^T\boldsymbol{\Sigma}_X\boldsymbol{\beta}=\mathbf{I}_d} -\mathcal{V}^2(\boldsymbol{\beta}^T\mathbf{X}, Y) + \rho_\lambda(\boldsymbol{\beta})$$

where $\rho_\lambda(\boldsymbol{\beta})$ is a penalty term on $\boldsymbol{\beta}$, which can be as mentioned above as LASSO, SCAD, or Dantzig selector; while λ is a tuning parameter. The differences between this method and usual penalized approach are that: (1), $\mathcal{V}^2$ does not have an explicit solution; (2), nonlinear constraint on $\boldsymbol{\beta}$. Although the penalty term is the same, this will certainly add difficulties in the algorithm.

Note that here $\rho_\lambda(\cdot)$ is not differentiable. In order to overcome this problem, Chen, Sheng, and Yin (2016) adopt the local quadratic approximation (Fan and Li 2001b; Chen, Zou, and Cook 2010), that is, the penalty function is approximated locally with quadratic function at every step.

5.2.4 DCOV as dual variable selection method

Now consider the response $\boldsymbol{Y}$ is a random vector. Some variables in the response vector may be related, so it is also important to remove redundant variables in the response vector. Adding another penalty term on the direction of $\boldsymbol{Y}$, we can force coefficients of redundant variables to zero. The optimization problem becomes:

$$\min_{\substack{\boldsymbol{\beta}^T\boldsymbol{\Sigma}_X\boldsymbol{\beta}=\mathbf{I}_d \\ \boldsymbol{\alpha}^T\boldsymbol{\Sigma}_Y\boldsymbol{\alpha}=\mathbf{I}_l}} -\mathcal{V}^2(\boldsymbol{\beta}^T\mathbf{X}, \boldsymbol{\alpha}^\top\mathbf{Y}) + \rho_{\lambda_1}(\boldsymbol{\beta}) + \rho_{\lambda_2}(\boldsymbol{\alpha}),$$

where $\rho_{\lambda_1}(\boldsymbol{\beta})$ is a penalty term on $\boldsymbol{\beta}$, and $\rho_{\lambda_2}(\boldsymbol{\alpha})$ is a penalty term on $\boldsymbol{\alpha}$. The penalty can be LASSO, SCAD, or Dantzig selector, and we choose LASSO in this paper; while λ_1 and λ_2 are tuning parameters. We also employ the local quadratic approximation (Fan and Li 2001b; Chen, Zou, and Cook 2010) to solve the non-differentiable problem of $\rho_{\lambda_1}(\cdot)$ and $\rho_{\lambda_2}(\cdot)$.

5.3 Numerical studies

In this section, we assess the efficiency of the proposed sparse DCOV methods for dual variable selection through simulations and application to a real data. Sample sizes n=100, 200, and 400 are run for each model. For each sample size n, the results of the mean $\bar{\Delta}_m$ and standard error (SE) by 100 replicates are reported. The following models come from the models to assess DCOV method for CDCA and DSC from Chen and Yin (2016):

Model 8 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:

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

where $\boldsymbol{\beta}_1 = (1, 0, 0, 0, 0, 0)$, $\boldsymbol{\beta}_2 = (0, 2, 1, 0, 0, 0)$, $\boldsymbol{\alpha}_1 = (1, 0, 0, 0)$, $\boldsymbol{\alpha}_2 = (0, 1, 0, 0)$, and $\boldsymbol{\epsilon} \sim N_4(\mathbf{0}, \boldsymbol{\Delta})$ with

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

Model 9 Let $p = 6, q = 5$, $\boldsymbol{X} \sim N(\mathbf{0}, \mathbf{I}_6)$. We increase the dimension of $\boldsymbol{Y}$ to five, which is generated as:

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

where $\boldsymbol{\alpha}_1 = (1,0,0,0,0)$, $\boldsymbol{\alpha}_2 = (0,1,0,0,0)$, $\boldsymbol{\alpha}_3 = (0,0,1,0,0)$, and $\boldsymbol{\epsilon} \sim N_4(\mathbf{0}, \boldsymbol{\Delta})$ with $\boldsymbol{\Delta} = diag(\boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2)$

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

Model 10 Let $p = 5, q = 4$, $\boldsymbol{X} \sim N(\mathbf{0}, \mathbf{I}_5)$. The four dimensional response random vector $\boldsymbol{Y}$ is generated as:

$$
\begin{aligned}
Y_1 &= 4\cos(\boldsymbol{B}_1^T \boldsymbol{X}) + 0.3\epsilon_1, \\
Y_2 &= \boldsymbol{B}_1^T \boldsymbol{X} + 0.5\epsilon_2, \\
Y_3 &= \epsilon_3, \\
Y_4 &= \epsilon_4.
\end{aligned}
$$

where $\boldsymbol{\alpha}_1 = (1,0,0,0)$, $\boldsymbol{\alpha}_2 = (0,1,0,0)$, $\boldsymbol{\alpha}_3 = (0,0,1,0)$ and $\boldsymbol{\epsilon} \sim N_4(\mathbf{0}, I_4)$.

5.4 Future work

In this paper, we introduce a concept of DVS, and propose a method of sparse DCOV to estimate subsets on the predictor and the response as well. We will continue working on the following:

1. Using Lasso penalty to develop an algorithm to solve the problem.
2. Establish asymptotic property for the algorithm as in Chen, Sheng, and Yin (2016).
3. Finish simulations and real data application.
4. Try SCAD, or Dantsig selector. Furthermore, we may also use the sequential SDR of (Yin and Hilafu, 2015) to deal with Large p Small n problems.

Bibliography

[1] H. D. Bondell and L. Li. “Shrinkage inverse regression estimation for model-free variable selection”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71.1 (2009), pp. 287–299.

[2] E. Candes and T. Tao. “The Dantzig selector: statistical estimation when p is much larger than n”. In: *The Annals of Statistics* (2007), pp. 2313–2351.

[3] X. Chen, W. Sheng, and X. Yin. “Efficient sparse estimate of sufficient dimension reduction in high dimension.” In: *submitted to Technometrics* (2016).

[4] X. Chen and X. Yin. “Sufficient dimension reduction via distance covariance with multivariate responses”. In: (2016).

[5] X. Chen, C. Zou, and R. D. Cook. “Coordinate-independent sparse sufficient dimension reduction and variable selection”. In: *The Annals of Statistics* (2010), pp. 3696–3723.

[6] R. D. Cook. “Graphics for regressions with a binary response”. In: *Journal of the American Statistical Association* 91.435 (1996), pp. 983–992.

[7] J. Fan and R. Li. “Variable selection via nonconcave penalized likelihood and its oracle properties”. In: *Journal of the American statistical Association* 96.456 (2001), pp. 1348–1360.

[8] J. Fan and R. Li. “Variable selection via nonconcave penalized likelihood and its oracle properties”. In: *Journal of the American statistical Association* 96.456 (2001), pp. 1348–1360.

[9] J. Fan and J. Lv. “Sure independence screening for ultrahigh dimensional feature space”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70.5 (2008), pp. 849–911.

[10] K.C. Li. “Sliced inverse regression for dimension reduction”. In: *Journal of the American Statistical Association* 86.414 (1991), pp. 316–327.

[11] L. Li, R. D. Cook, and C. J. Nachtsheim. “Model-free variable selection”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2 (2005), pp. 285–299.

[12] L. Li and C. J. Nachtsheim. “Sparse sliced inverse regression”. In: *Technometrics* (2012).

[13] L. Ni, R. D. Cook, and C.L. Tsai. “A note on shrinkage sliced inverse regression”. In: *Biometrika* 92.1 (2005), pp. 242–247.

[14] R. Tibshirani. “Regression shrinkage and selection via the lasso”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1996), pp. 267–288.

[15] Q. Wang and X. Yin. “A nonlinear multi-dimensional variable selection method for high dimensional data: sparse MAVE”. In: *Computational Statistics & Data Analysis* 52.9 (2008), pp. 4512–4520.

[16] X. Yin and H. Hilafu. “Sequential sufficient dimension reduction for large p, small n problems”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 77.4 (2015), pp. 879–892.

[17] X. Yin, B. Li, and R. D. Cook. “Successive direction extraction for estimating the central subspace in a multiple-index regression”. In: *Journal of Multivariate Analysis* 99.8 (2008), pp. 1733–1757.

[18] H. Zou. “The adaptive lasso and its oracle properties”. In: *Journal of the American statistical association* 101.476 (2006), pp. 1418–1429.