Computational Statistics and Data Analysis 140 (2019) 144–154

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Robust sufficient dimension reduction via ball covariance

Jia Zhang [a], Xin Chen [b,*]

[a] School of Statistics, Southwestern University of Finance and Economics, Chengdu 611130, China
[b] Department of Statistics & Data Science, Southern University of Science and Technology, Shenzhen 518055, China

ARTICLE INFO

Article history:
Received 3 September 2018
Received in revised form 15 June 2019
Accepted 15 June 2019
Available online 27 June 2019

Keywords:
Ball covariance
Central subspace
Robustness
Sufficient dimension reduction

ABSTRACT

Sufficient dimension reduction is an important branch of dimension reduction, which includes variable selection and projection methods. Most of the sufficient dimension reduction methods are sensitive to outliers and heavy-tailed predictors, and require strict restrictions on the predictors and the response. In order to widen the applicability of sufficient dimension reduction, we propose BCov-SDR, a novel sufficient dimension reduction approach that is based on a recently developed dependence measure: ball covariance. Compared with other popular sufficient dimension reduction methods, our approach requires rather mild conditions on the predictors and the response, and is robust to outliers or heavy-tailed distributions. BCov-SDR does not require the specification of a forward regression model and allows for discrete or categorical predictors and multivariate response. The consistency of the BCov-SDR estimator of the central subspace is obtained without imposing any moment conditions on the predictors. Simulations and real data studies illustrate the applicability and versatility of our proposed method.

1. Introduction

Sufficient dimension reduction is an important branch of dimension reduction research, which draws much attention in the era of big data. Early sufficient dimension reduction methods, including the seminal sliced inverse regression (Li, 1991), sliced average variance estimation (Cook and Weisberg, 1991), inverse regression (Cook and Ni, 2005) and directional regression (Li and Wang, 2007), require the linearity condition or the constant variance condition or both to hold, which are not easy to verify in practice. Although methods like minimum average variance estimation (Xia et al., 2002), sliced regression (Wang and Xia, 2008) and ensemble method (Yin and Li, 2011) can avoid the restrictions aforementioned, they need continuous predictors and smooth link functions instead. Besides, the Fourier transform method proposed by Zhu and Zeng (2006) and Zeng and Zhu (2010) and the Kullback–Leibler based approach proposed by Yin et al. (2008) also require the predictors to be continuous. Other methods of sufficient dimension reduction, such as the likelihood based dimension reduction method proposed by Cook and Forzani (2009), the method proposed by Cook and Li (2009) for exponential family predictors, and the methods suggested by Bura and Forzani (2015) and Bura et al. (2016) for elliptically contoured inverse predictors and exponential family inverse predictors respectively, impose different assumptions on the distribution of $\mathbf{X}|Y$, where $\mathbf{X}$ and Y denote the predictors and the response respectively. Sheng and Yin (2016) achieved sufficient dimension reduction via distance covariance, which does not need any conditions mentioned above and can work well when some predictors are discrete or categorical. However, their method requires the first order moments of $\mathbf{X}$ and Y to be bounded. Hence, this method seems to be sensitive to outliers and heavy-tailed predictors.

* Corresponding author.
E-mail addresses: jeanzhang9@2015.swufe.edu.cn (J. Zhang), xchen8587@gmail.com (X. Chen).

https://doi.org/10.1016/j.csda.2019.06.004
0167-9473/

Problems related to heavy tails arise frequently from a variety of applications, especially those in finance, economics, genomics and bio-imaging. In the field of variable selection, which is another line of dimension reduction, researchers have began to realize the importance of robust estimation. For example, Li et al. (2012a) suggested a robust screening method based on the Kendall-τ correlation coefficient, and Pan et al. (2018) proposed a generic sure independence screening procedure via ball correlation, which is robust to heavy-tailed predictors. Moreover, in principal component analysis, which is a vigorous unsupervised dimension reduction method, heavy-tailed phenomenon is drawing more and more attention. To name a few, Croux et al. (2013) proposed a robust space method for principal component analysis based on a robust measure of variance. Han and Liu (2018) proposed the elliptical component analysis, which is an alternative to principal component analysis, to tackle high dimensional elliptically distributed data. However, in sufficient dimension reduction, robustness has not received due attention, and few works have been done to address this problem. Dong et al. (2015) proposed two robust inverse regression methods, which are insensitive to outliers and heavy-tailed predictors. Zhou et al. (2015) proposed a robust version of the estimating equation based sufficient dimension reduction method by constructing a robust nonparametric regression estimator of the central subspace (Cook, 1994, 1996). Rekabdarkolaee et al. (2017) suggested a robust extension of the minimum average variance estimation by means of modal regression to alleviate the influence of outliers. Nevertheless, these methods are only extensions of existed sufficient dimension reduction approaches, thus they also rely heavily on the conditions mentioned above, which are usually difficult to test and validate. Hence, the corresponding conclusions may be misleading if some of the conditions are violated.

In this paper, we propose BCov-SDR, a novel sufficient dimension reduction method that is based on a recently developed dependence measure: ball covariance. Thanks to the robust property of the empirical ball covariance, we are able to obtain an estimator of the central subspace, which is robust to outliers and heavy-tailed predictors. In addition, we do not need to assume a number of conditions to hold, such as the linearity condition, the constant variance condition, the continuity condition of $\mathbf{X}$ and the special distribution condition of $\mathbf{X}|Y$ and so on, nor do we need to involve any complex nonparametric estimation. Specially, compared with the distance covariance based sufficient dimension reduction method (Sheng and Yin, 2016), which seems similar to ours, the proposed method avoids the moment restrictions on the predictors and the response, thus it enjoys robustness to heavy tailed variables. All these advantages mentioned above ensure the practicability and versatility of our proposed method. Under mild conditions, we prove the validity of BCov-SDR as a sufficient dimension reduction method and construct the consistency of the BCov-SDR estimator of the central subspace. Simulation and real data analysis are conducted to exhibit the priority of our proposed method.

The rest of the paper is organized as follows. In Section 2, we introduce the BCov-SDR method and investigate its asymptotic properties. Simulations are carried out to compare our method with other popular sufficient dimension reduction methods in Section 3, and real data analysis is conducted in Section 4. Section 5 concludes the paper, and all the proofs are deferred to the Appendix.

2. Method

2.1. Ball covariance

Ball covariance is a novel measure of the dependence between two random vectors proposed by Pan et al. (2018). Let U and V be two random vectors defined respectively in two separable Banach spaces $(\mathscr{U}, \zeta_U)$ and $(\mathscr{V}, \zeta_v)$, where ζ_U and ζ_V are norms defined on $\mathscr{U}$ and $\mathscr{V}$ respectively. Let θ, μ and ν be Borel probability measures on $\mathscr{U}\times\mathscr{V}$, $\mathscr{U}$ and $\mathscr{V}$, respectively. (U, V) is a Banach-valued random vector defined on a probability space $(\Omega, \mathcal{F}, P)$ such that $(U, V) \sim \theta$, $U \sim \mu$ and $V \sim \nu$. Denote a closed ball with center u_1 and radius $\zeta_U(u_1, u_2)$ in $\mathscr{U}$ by $\bar{B}_{\zeta_U}(u_1, u_2)$. $\bar{B}_{\zeta_V}(v_1, v_2)$ is then similarly defined in $\mathscr{V}$. Ball covariance is defined by

$$\mathrm{BCov}(U, V) = \left\{ \iint_{\mathscr{U}\times\mathscr{V}} [\theta - \mu\otimes\nu]^2(\bar{B}_{\zeta_U}(u_1, u_2)\times\bar{B}_{\zeta_V}(v_1, v_2))\theta(\mathrm{d}u_1, \mathrm{d}u_2)\theta(\mathrm{d}v_1, \mathrm{d}v_2)\right\}^{1/2},$$

where $\mu\otimes\nu$ is the product measure on $\mathscr{U}\times\mathscr{V}$.

Under some mild conditions, it can be shown that $\mathrm{BCov}(U, V) = 0$ if and only if U and V are independent. Hence, ball covariance can identify both linear and nonlinear correlations between two random vectors. Thanks to this property, we are able to achieve sufficient dimension reduction via ball covariance. Furthermore, ball covariance can detect non-Euclidean dependence. Compared with distance covariance (Székely et al., 2007) and Hilbert–Schmidt independence criterion (HSIC, Gretton et al. (2008)), which can also detect multivariate or nonlinear dependence, ball covariance requires less conditions and can be applied to more settings. Specifically, distance covariance and HSIC require strong negative type condition (Lyons, 2013) or positive type condition (Sejdinovic et al., 2013) to hold. Therefore, these methods can only be applied to spaces that can be embedded into some Hilbert spaces or some reproduced kernel Hilbert spaces. Ball covariance does not need such conditions and can be applied to various spaces.

Let $\{U_k, V_k\}_{k=1}^n$ be an i.i.d sample of (U, V). Define $\delta^U_{ij,k} = I\{U_k \in \bar{B}_{\zeta_U}(U_i, U_j)\}$, where $I(\cdot)$ denotes the indicator function, $\delta^U_{ij,kl} = \delta^U_{ij,k}\delta^U_{ij,l}$ and $\xi^U_{ij,klst} = (\delta^U_{ij,kl} + \delta^U_{ij,st} - \delta^U_{ij,ks} - \delta^U_{ij,lt})/2$. $\xi^V_{ij,klst}$ is defined in the same way as that of $\xi^U_{ij,klst}$. The empirical ball covariance is then defined by

$$\mathrm{BCov}_n(U, V) = \left(\frac{1}{n^6}\sum_{i,j,k,l,s,t=1}^{n}\xi^U_{ij,klst}\xi^V_{ij,klst}\right)^{1/2}.$$

Define the ball correlation

$$\mathrm{BCor}(\mathbf{X}, Y) = \frac{\mathrm{BCov}(\mathbf{X}, Y)}{\mathrm{BCov}^{1/2}(\mathbf{X}, \mathbf{X}) \times \mathrm{BCov}^{1/2}(Y, Y)},$$

and the sample ball correlation

$$\mathrm{BCor}_n(\mathbf{X}, Y) = \frac{\mathrm{BCov}_n(\mathbf{X}, Y)}{\mathrm{BCov}_n^{1/2}(\mathbf{X}, \mathbf{X}) \times \mathrm{BCov}_n^{1/2}(Y, Y)}.$$

Pan et al. (2018) shows that

- if $\mathbf{X}$ and Y are standard normal random variables, $\mathrm{BCor}(\mathbf{X}, Y)$ is a nondecreasing function of the absolute value of the Pearson correlation of $\mathbf{X}$ and Y.
- $\mathrm{BCor}(\mathbf{X}, Y) \in [0, 1]$ and $\mathrm{BCor}_n(\mathbf{X}, Y) \in [0, 1]$.
- $\mathrm{BCor}_n(\mathbf{X}, Y) = \mathrm{BCor}(\mathbf{X}, Y) = 1$, provided that there exists a vector $\mathbf{a}$, a nonzero real number b and an orthonormal matrix C such that $Y = \mathbf{a} + bC\mathbf{X}$ or $Y = \mathbf{a} + b\mathbf{X}$.

Notice that the empirical ball covariance is in fact a rank statistic, which is a complex function of indicator functions. Hence, compared with the sample distance covariance (Székely et al., 2007), the empirical ball covariance enjoys robustness to outliers and heavy-tailed random variables. Consequently, robust sufficient dimension reduction via ball covariance is promising.

2.2. Sufficient dimension reduction

Let Y be a random scaler and $\mathbf{X} = (X_1, \ldots, X_p)^{\mathrm{T}}$ be a p-dimensional random vector. If

$$Y \perp\!\!\!\perp \mathbf{X} | \mathbf{B}^{\mathrm{T}}\mathbf{X}, \tag{2.1}$$

where $\mathbf{B}$ is a $p \times q$ matrix with $q \leq p$, and $\perp\!\!\!\perp$ denotes independence, we say that $\mathbf{B}^{\mathrm{T}}\mathbf{X}$ is a sufficient dimension reduction of $\mathbf{X}$ regarding Y. The column subspace of $\mathbf{B}$, denoted by $S(\mathbf{B})$, is called a dimension-reduction subspace (Li, 1991). The central subspace is then defined by the intersection of all the dimension-reduction subspaces, see Cook (1994, 1996) for details. We assume that the central subspace exists, and its dimension is known to be $d(d \leq p)$. Our goal is to estimate a basis matrix of the $p \times d$-dimensional central subspace. Of note is that Y can also be a multi-dimensional random vector.

2.3. Ball covariance based sufficient dimension reduction

We try to estimate a basis matrix of the central subspace by solving the following optimization problem:

$$\boldsymbol{\eta} = \arg\max \mathrm{BCov}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, Y), \quad \text{subject to } \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\beta} = \mathbf{I}_d, \tag{2.2}$$

where $\boldsymbol{\beta}$ is an arbitrary matrix with dimension $p \times d$, and $\boldsymbol{\Sigma}_{\mathbf{x}}$ is the covariance matrix of $\mathbf{X}$. We call this method BCov-SDR. Let $\mathbf{B}$ denote a basis of the central subspace which satisfies $\mathbf{B}\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{B} = \mathbf{I}_d$, and (B_1, B_2) be any partition of $\mathbf{B}$. We need the following conditions to construct the theory.

(1) $\mathrm{BCov}(B_i^{\mathrm{T}}\mathbf{X}, Y) < \mathrm{BCov}(\mathbf{B}^{\mathrm{T}}\mathbf{X}, Y)$ for $i = 1, 2$.

(2) $P^{\mathrm{T}}_{\mathbf{B}(\boldsymbol{\Sigma}_{\mathbf{X}})}\mathbf{X} \perp\!\!\!\perp Q^{\mathrm{T}}_{\mathbf{B}(\boldsymbol{\Sigma}_{\mathbf{X}})}\mathbf{X}$, where $P_{\mathbf{B}(\boldsymbol{\Sigma}_{\mathbf{X}})} = \mathbf{B}(\mathbf{B}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{B})^{-1}\mathbf{B}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{X}}$ and $Q_{\mathbf{B}(\boldsymbol{\Sigma}_{\mathbf{X}})} = \mathbf{I} - P_{\mathbf{B}(\boldsymbol{\Sigma}_{\mathbf{X}})}$.

Condition (1) is assumed to guarantee an increasing order of the objective function when the effective dimension grows. This condition is not necessary but assumed here to facilitate the proof of Theorem 1. Nevertheless, we find that this condition holds for most of our simulation scenarios, except Model (D) where the link function is of the form of a fraction. Condition (2) is mild, and it asymptotically holds when the dimension of $\mathbf{X}$ gets reasonably high, see Sheng and Yin (2013) for details.

Theorem 1. *For any $\boldsymbol{\beta}$ satisfying $\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\beta} = \mathbf{I}_d$ and $S(\boldsymbol{\beta}) \neq S(\mathbf{B})$, where $\mathbf{B}$ is a basis of the central subspace which satisfies $\mathbf{B}\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{B} = \mathbf{I}_d$. If Conditions (1) and (2) hold, we have $\mathrm{BCov}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, Y) < \mathrm{BCov}(\mathbf{B}^{\mathrm{T}}\mathbf{X}, Y)$.*

Theorem 1 guarantees the validity of BCov-SDR as a sufficient dimension reduction method. When the dimension of the central subspace d is known, we can obtain a basis of the central subspace by solving the optimization problem (2.2).

In practice, we use the empirical ball covariance BCov_n to implement the optimization, *i.e.* ,

$$\boldsymbol{\eta}_n = \arg\max \mathrm{BCov}_n(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, Y), \quad \text{subject to } \boldsymbol{\beta}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}\boldsymbol{\beta} = \mathbf{I}_d, \tag{2.3}$$

where $\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}$ is the sample covariance matrix of $\mathbf{X}$.

Theorem 2. *For $\boldsymbol{\eta}$ and $\boldsymbol{\eta}_n$ defined in (2.2) and (2.3), we have $\boldsymbol{\eta}_n \to_p \boldsymbol{\eta}$, provided that Conditions (1) and (2) hold.*

Theorem 2 establishes the consistency of the BCov-SDR estimator of the central subspace. The convergence rate and the asymptotic normality of the estimator depend heavily on the differentiability of the objective function with respect to $\boldsymbol{\beta}$. As mentioned above, the empirical ball covariance 2.1 is a function of indicator functions, thus it is not differentiable. Hence, it is difficult to achieve some convergence rate and the asymptotic normality with a non-differentiable objective function. However, if we can smooth the empirical ball covariance by substituting the indicator function by some kernel alternatives, the convergence rate and the asymptotic normality can be achieved by some routine techniques. We leave this point for further research.

We solve the optimization problem (2.3) by R package "Rdonlp2", which is designed to solve constrained nonlinear programming problems, on "R x64 3.3.2". The basic algorithm behind the package is the sequential quadratic programming, which is one of the most successful methods for the numerical solution of constrained nonlinear optimization problems. The initial value is estimated by the distance covariance based sufficient dimension reduction method suggested by Sheng and Yin (2016). Despite the non-differentiability of the objective function, "Rdonlp2" is capable of finding a stable solution quickly.

Remark 1 (*On the Decision of the Dimension d Of the Central Subspace*)**.** The dimension, d, of the central subspace is assumed to be known above. In practice, we have to estimate d by some criterion or tests. There have been several methods in the literature which may also be applied to our BCov-SDR method. Li (1991, 1992) proposed a chi-squared sequential test. Following their work, Bura and Cook (2001) proposed a general weighted sequential test based on the chi-squared statistic. Zhu et al. (2006) suggested a procedure based on the Bayes information criterion to estimate the dimension of the central subspace. Permutation and bootstrap methods were employed to determine the dimension of the central subspace by Cook and Yin (2001), Ye and Weiss (2003), Zhu and Zeng (2006) and Sheng and Yin (2016). Chen et al. (2015) utilized conditional distance covariance to check the goodness-of-fit of a given dimension reduction subspace. Bootstrap method suggested by Sheng and Yin (2016) can be readily extended to our BCov-SDR, and it is employed to estimate the dimension d of the central subspace in the real data analysis below.

3. Simulation

In this section, we compare our method (BC, for simplicity) with 8 popular dimension-reduction methods in the literature: SIR (Li, 1991), SAVE (Cook and Weisberg, 1991), PHD (Li, 1992; Cook, 1996), IRE (Cook and Ni, 2005), PFC (Cook and Forzani, 2008), LAD (Cook and Forzani, 2009) and DC (Sheng and Yin, 2016). Among them, the PHD method includes PHDy and PHDres, using residuals and response respectively. Of note is that not all these methods can be applied to cases with a multivariate response. Hence, in Model (E) below, we only present the results of a part of the above methods. Here, we assume that the dimension d of the central subspace is known, and we choose its real value as the working dimension. Five models are investigated to illustrate the efficacy of our proposed method:

$$
\begin{aligned}
&\text{(A)} \quad Y = (\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X})^2 + \boldsymbol{\beta}_2^{\mathrm{T}}\mathbf{X} + 0.1\epsilon,\\
&\text{(B)} \quad Y = \operatorname{sign}(2\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X} + \epsilon_1) \times \log|2\boldsymbol{\beta}_2^{\mathrm{T}}\mathbf{X} + 4 + \epsilon_2|,\\
&\text{(C)} \quad Y = \exp(\boldsymbol{\beta}_3^{\mathrm{T}}\mathbf{X})\epsilon,\\
&\text{(D)} \quad Y = \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}/(0.5 + (\boldsymbol{\beta}_2^{\mathrm{T}}\mathbf{X} + 1.5)^2) + 0.1\epsilon, \quad \text{and}\\
&\text{(E)} \quad Y_1 = (\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X})^2 + \boldsymbol{\beta}_2^{\mathrm{T}}\mathbf{X} + 0.1\epsilon_1\\
&\qquad Y_2 = \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}/(0.5 + (\boldsymbol{\beta}_2^{\mathrm{T}}\mathbf{X} + 1.5)^2) + 0.1\epsilon_2,
\end{aligned}
$$

where in Model (E), $\mathbf{Y} = (Y_1, Y_2)^{\mathrm{T}}$ is a 2-dimensional responsible vector. These models frequently appear in the literature of sufficient dimension reduction (see *e.g.* , Li (1991), Sheng and Yin (2016)). Throughout all the five models, we set $(n, p) = (100, 6)$, $\boldsymbol{\beta}_1 = (1, 0, 0, 0, 0, 0)^{\mathrm{T}}$, $\boldsymbol{\beta}_2 = (0, 1, 0, 0, 0, 0)^{\mathrm{T}}$, and $\boldsymbol{\beta}_3 = (1, 0.5, 1, 0, 0, 0)^{\mathrm{T}}$. For each model, $M = 100$ data sets are simulated to calculate the mean and standard errors of Δ_m, which is a commonly-used measure of the distance between two spaces defined by Li et al. (2005). For each simulated data set, we calculate Δ_m by

$$\Delta_m(S_1, S_2) = \|P_{S_1} - P_{S_2}\|,$$

where S_i indicates a d-dimensional subspace of $\mathbb{R}^p$, P_{S_i} is the orthogonal projection on to S_i for $i = 1, 2$, and $\|\cdot\|$ denotes the biggest singular value of a matrix. The distance between the true central subspace and its estimated counterpart of a specific dimension reduction method is calculated to evaluate the efficacy of the method. Clearly, the smaller the distance is, the better the estimator is.

For each model, six different kinds of $\mathbf{X} = (X_1, \ldots, X_p)^{\mathrm{T}}$ are generated: Part (1), $\mathbf{X} \sim N(\mathbf{0}, \mathbf{I}_p)$; Part (2), non-normal covariates; Part (3), discrete covariates; Part (4), $\mathbf{X} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} = (\sigma_{ij})$ and $\sigma_{ij} = 0.5^{|i-j|}$ for $i, j = 1, \ldots, p$; Part (5) and (6), heavy tailed covariates, *i.e.* , $\mathbf{X} \sim t(\mathbf{0}, \boldsymbol{\Sigma}, 3)$ and $\mathbf{X} \sim t(\mathbf{0}, \boldsymbol{\Sigma}, 2)$ where $\boldsymbol{\Sigma} = (\sigma_{ij})$, $\sigma_{ij} = 0.5^{|i-j|}$ for $i, j = 1, \ldots, p$, and the last parameter (2 or 3) of the multivariate t distribution indicates the degree of the freedom. We specify Parts (2) and (3) for different models.

Model (A), (D) and (E): Part (2), $(X_i + 2)/5 \sim \mathit{Beta}(0.75, 1)$, $i = 1, \ldots, p$; Part (3), $X_i \sim \mathit{Poisson}(1)$, $i = 1, \ldots, p$.

Model (B): Part (2), $X_i \sim \mathit{Uniform}(-2, 2)$, $i = 1, \ldots, p$; Part (3), $X_i \sim \mathit{Binomial}(10, 0.1)$, $i = 1, \ldots, p$.

Table 1
Mean and standard errors of Δ_m in Model (A).

Method	Part(1)		Part(2)		Part(3)	
	mean	sd	mean	sd	mean	sd
SIR	0.87	0.18	0.57	0.22	0.43	0.18
SAVE	0.67	0.23	0.73	0.20	0.62	0.28
PHDy	0.80	0.18	0.82	0.17	0.79	0.24
PHDres	0.88	0.15	0.90	0.09	0.86	0.13
IRE	0.89	0.15	0.60	0.22	0.50	0.20
PFC	0.66	0.21	0.54	0.14	0.94	0.08
LAD	**0.35**	0.20	**0.33**	0.19	**0.28**	0.18
DC	**0.19**	0.10	**0.20**	0.06	***0.00***	0.01
BC	***0.15***	0.10	***0.18***	0.06	**0.09**	0.09
Method	Part(4)		Part(5)		Part(6)	
	mean	sd	mean	sd	mean	sd
SIR	0.80	0.22	0.90	0.13	0.94	0.07
SAVE	0.78	0.22	0.89	0.14	0.92	0.09
PHDy	0.83	0.18	0.90	0.12	0.92	0.09
PHDres	0.88	0.13	0.90	0.12	0.93	0.08
IRE	0.84	0.21	0.92	0.12	0.93	0.10
PFC	0.68	0.18	0.86	0.16	0.96	0.07
LAD	**0.38**	0.18	**0.58**	0.28	**0.82**	0.20
DC	**0.25**	0.09	**0.47**	0.24	**0.91**	0.12
BC	***0.18***	0.07	***0.24***	0.18	***0.54***	0.29

Mode (C): Part (2), $(X_i + 1)/2 \sim Beta(1.5, 1)$, $i = 1, \ldots, p$; Part (3), $X_i \sim Poisson(1)$ for $i = 1, \ldots, 5$, $X_6 \sim Binomial(10, 0.3)$ and $X_i \sim Poisson(1)$ for $i = 7, \ldots, p$.

The error term in each model is stand-normally distributed and independent of the covariates. In Model (B), ϵ_1 and ϵ_2 are mutually independent, but in Model (E) the two error terms are weakly correlated, and their correlation coefficient is set to be 0.3.

Simulation results are shown in Tables 1–5. Throughout these tables, the bold numbers indicate the top three methods with the italic one representing the best. In most cases, our proposed BC method outperforms the other 8 methods. Compared with DC, we find that BC performs much better in most heavy-tailed scenarios, which confirms our conjecture. Specifically, in Model (A), BC, DC and LAD rank the top three methods, and BC performs the best except in Part (3), where the covariates are discrete. In fact, DC estimates the central subspace exactly with the mean error of Δ_ms decreasing to 0 in this model. In Model (B), BC and DC hold the first two positions and BC performs better than DC in all parts except Part (3), where DC performs similarly to but a little better than BC. In Model (C), BC, DC and SIR are the best three methods in Parts (1)–(3), and they perform similarly with BC working slightly better. In Parts (4)–(6), SIR and LAD show their advantage, and BC maintains to be competitive. In Model (D), BC and DC rank the top two methods again, and BC outperforms DC in most scenarios. Moreover, SIR seems to work well in this model. In Model (E), BC dominates all the other methods except in Part (3). The efficiency gain of BC is significant in this model with a multivariate response. Throughout these models, BC exhibits significant robustness to heavy-tailed predictors (Parts 5–6) in comparison with DC. In summary, BC is a good robust alternative to DC, and shows practicality and superiority compared with other sufficient dimension reduction methods.

The proposed BCov-SDR method is quite general, and we cannot expect it to outperform the other methods in all cases. The fact is that some methods are in favor of some specific models. For example, SIR and LAD seems to work quite well in the single index model (Model (C)) while BC and DC lose their efficacy, and DC seems to work quite well in presence of discrete covariates. This may come from the comparatively high computational complexity of BC and DC.

Next, we do some changes to $\mathbf{X}$ in Model (A) to illustrate the robustness of our BCov-SDR method to outliers. First, we generate n samples of $\mathbf{X} \sim N(\mathbf{0}, \Sigma)$, where $\Sigma = (\sigma_{ij})$ and $\sigma_{ij} = 0.5^{|i-j|}$ for $i, j = 1, \ldots, p$. Then, we replace r% samples of X_1, X_3 and X_5 with $2 \cdot t(1)$ randomly, where $t(1)$ indicates the t distribution with the degree of freedom being 1. r% is set to be 10%, 30% and 50%, and the means and standard errors of Δ_m of SIR, DC and BC are given in Table 6.

From Table 6, we see that compared with SIR and DC, our BC method shows robustness to outliers. When more samples are contaminated by outliers, all the three methods tend to lose efficiency, but our BC method performs much better than DC and SIR. However, the standard error of Δ_m of BC is bigger than that of SIR and DC in cases where $r = 30$ and $r = 50$, which indicates that the performance of BC seems to be not stable in presence of outliers.

The stating value of the optimization procedure may influence the performance of the estimator, since the optimization procedure may suffer from the non-differentiability of the objective function. Different starting values are tried, including the DC based solution, and we find that our ball covariance (BC) based estimator is not sensitive to the choice of the starting value, but using the DC solution as the starting value can always give a satisfactory BC based estimator of the central subspace. A simulation is conducted to show the influence of starting value to the BC based estimator. We compare the means of Δ_m of SIR, DCsir, BCsir and BCdc in Models (A)–(E), where DCsir (BCsir) denotes the DC (BC) based estimator

Table 2
Mean and standard errors of Δ_m in Model (B).

Method	Part(1)		Part(2)		Part(3)	
	mean	sd	mean	sd	mean	sd
SIR	**0.30**	0.10	**0.29**	0.12	0.46	0.20
SAVE	0.87	0.15	0.67	0.23	0.86	0.17
PHDy	0.81	0.17	0.54	0.20	0.93	0.09
PHDres	0.72	0.19	0.46	0.14	0.90	0.12
IRE	0.37	0.13	0.33	0.13	0.49	0.19
PFC	0.38	0.17	0.38	0.19	0.41	0.15
LAD	0.37	0.15	0.33	0.12	**0.40**	0.26
DC	***0.27***	0.09	**0.22**	0.06	***0.25***	0.18
BC	***0.27***	0.09	***0.21***	0.06	**0.27**	0.17
Method	Part(4)		Part(5)		Part(6)	
	mean	sd	mean	sd	mean	sd
SIR	**0.37**	0.13	**0.53**	0.19	0.88	0.13
SAVE	0.90	0.13	0.97	0.04	0.96	0.06
PHDy	0.85	0.15	0.92	0.10	0.96	0.05
PHDres	0.80	0.16	0.88	0.14	0.92	0.09
IRE	0.42	0.13	0.57	0.20	**0.88**	0.12
PFC	0.54	0.22	0.65	0.22	0.90	0.13
LAD	0.45	0.17	0.68	0.23	0.88	0.15
DC	***0.35***	0.12	**0.43**	0.18	**0.82**	0.19
BC	***0.35***	0.12	***0.38***	0.15	***0.71***	0.22

Table 3
Mean and standard errors of Δ_m in Model (C).

Method	Part(1)		Part(2)		Part(3)	
	mean	sd	mean	sd	mean	sd
SIR	**0.19**	0.06	**0.31**	0.12	***0.18***	0.06
SAVE	0.83	0.21	0.92	0.11	0.93	0.12
PHDy	0.68	0.19	0.80	0.16	0.65	0.17
PHDres	0.68	0.20	0.79	0.16	0.68	0.19
IRE	0.22	0.09	0.36	0.14	0.23	0.09
PFC	0.50	0.17	0.43	0.14	0.54	0.15
LAD	0.22	0.09	0.38	0.14	0.20	0.07
DC	**0.19**	0.06	***0.30***	0.13	**0.19**	0.07
BC	***0.18***	0.06	***0.30***	0.13	***0.18***	0.06
Method	Part(4)		Part(5)		Part(6)	
	mean	sd	mean	sd	mean	sd
SIR	***0.25***	0.08	**0.35**	0.16	**0.69**	0.18
SAVE	0.80	0.23	0.97	0.04	0.96	0.05
PHDy	0.80	0.13	0.86	0.12	0.91	0.10
PHDres	0.79	0.15	0.85	0.13	0.91	0.10
IRE	0.31	0.10	**0.41**	0.16	0.74	0.18
PFC	0.60	0.13	0.77	0.17	0.82	0.16
LAD	**0.27**	0.10	***0.28***	0.11	***0.62***	0.31
DC	0.32	0.11	0.69	0.22	0.88	0.12
BC	**0.28**	0.10	0.49	0.24	**0.73**	0.28

with the starting value given by SIR, and BCdc denotes the BC based estimator with the starting value given by DC. Let $\mathbf{X} \sim t(0, \Sigma, 2)$ where $\Sigma = (\sigma_{ij})$ with $\sigma_{ij} = 0.5^{|i-j|}$ for $i, j = 1, \ldots, p$, where the parameter 2 denotes the degree of freedom of multivariate t distribution. Other simulation parameters stay the same as those in Section 3. The results are shown in Table 7.

4. Real data analysis

Auto MPG data.

We first employ the auto MPG data (MPG) to illustrate the advantage of our BCov-SDR method. The data concerns city-cycle fuel consumption in miles per gallon (mpg), to be predicted in terms of 3 multi-valued discrete and 4 continuous attributes: mpg, cylinders, displacement, horsepower, weight, acceleration, model year and origin, where cylinders, model year and origin are multi-valued discrete. As done in Sheng and Yin (2016), we avoid using "origin", because it correlates with "cylinders" closely. Missing values are deleted, and 392 observations are left for study.

Table 4
Mean and standard errors of Δ_m in Model (D).

Method	Part(1)		Part(2)		Part(3)	
	mean	sd	mean	sd	mean	sd
SIR	**0.28**	0.10	**0.19**	0.07	**0.80**	0.18
SAVE	0.94	0.09	0.88	0.12	0.94	0.07
PHDy	0.75	0.19	0.47	0.12	0.91	0.11
PHDres	0.70	0.20	0.40	0.96	0.93	0.09
IRE	0.80	0.18	0.52	0.20	0.92	0.10
PFC	0.70	0.17	0.50	0.15	0.91	0.11
LAD	0.86	0.14	0.52	0.25	0.93	0.11
DC	**0.20**	0.07	***0.14***	0.05	***0.27***	0.36
BC	***0.19***	0.06	***0.14***	0.05	**0.29**	0.35
Method	Part(4)		Part(5)		Part(6)	
	mean	sd	mean	sd	mean	sd
SIR	**0.43**	0.17	**0.50**	0.15	**0.63**	0.24
SAVE	0.95	0.06	0.96	0.06	0.95	0.07
PHDy	0.82	0.16	0.92	0.10	0.96	0.06
PHDres	0.80	0.17	0.89	0.14	0.94	0.08
IRE	0.88	0.14	0.86	0.13	0.94	0.10
PFC	0.88	0.13	0.87	0.14	0.94	0.08
LAD	0.87	0.14	0.92	0.09	0.94	0.07
DC	**0.36**	0.15	**0.34**	0.11	**0.43**	0.17
BC	***0.34***	0.12	***0.28***	0.10	***0.37***	0.12

Table 5
Mean and standard errors of Δ_m in Model (E).

Method	Part(1)		Part(2)		Part(3)	
	mean	sd	mean	sd	mean	sd
SIR	0.19	0.06	0.18	0.06	0.19	0.06
SAVE	0.65	0.28	0.33	0.19	0.74	0.26
DC	0.14	0.05	0.13	0.05	***0.00***	0.01
BC	***0.09***	0.03	***0.12***	0.04	0.10	0.11
Method	Part(4)		Part(5)		Part(6)	
	mean	sd	mean	sd	mean	sd
SIR	0.27	0.10	0.44	0.18	0.89	0.12
SAVE	0.71	0.22	0.97	0.04	0.94	0.09
DC	0.16	0.07	0.39	0.22	0.92	0.10
BC	***0.12***	0.04	***0.19***	0.15	***0.49***	0.31

Table 6
Mean and standard errors of Δ_m in Model (A) with outliers.

Method	10%		30%		50%	
	mean	sd	mean	sd	mean	sd
SIR	0.93	0.15	0.98	0.08	0.98	0.08
DC	0.42	0.34	0.81	0.28	0.90	0.17
BC	0.34	0.30	0.66	0.33	0.76	0.27

Table 7
Mean of Δ_m of BC based estimator with different starting values.

	Model (A)	Model (B)	Model (C)	Model (D)	Model (E)
SIR	0.92	0.66	0.46	0.69	0.59
DCsir	0.70	0.57	0.83	0.50	0.67
BCsir	0.28	0.51	0.53	0.55	0.22
BCdc	0.30	0.47	0.58	0.44	0.26

Fig. 1 presents the box plots of the scaled predictors. It can be seen clearly that outliers exist in the variables "horsepower" and "acceleration". "cylinders" and "displacement" are heavily left-skewed, thus they are far from the normal distributed variables. In addition, the discrete property of some covariates adds difficulty to the estimation. In order to investigate the city-cycle fuel consumption in miles per gallon, we assume that a sufficient dimension reduction structure indicated by (2.1) exists. The bootstrap method suggested by Sheng and Yin (2016) are used to determine the dimension d of the central subspace, which shows $d = 2$. Then, DC and BC are utilized to estimate the central subspace

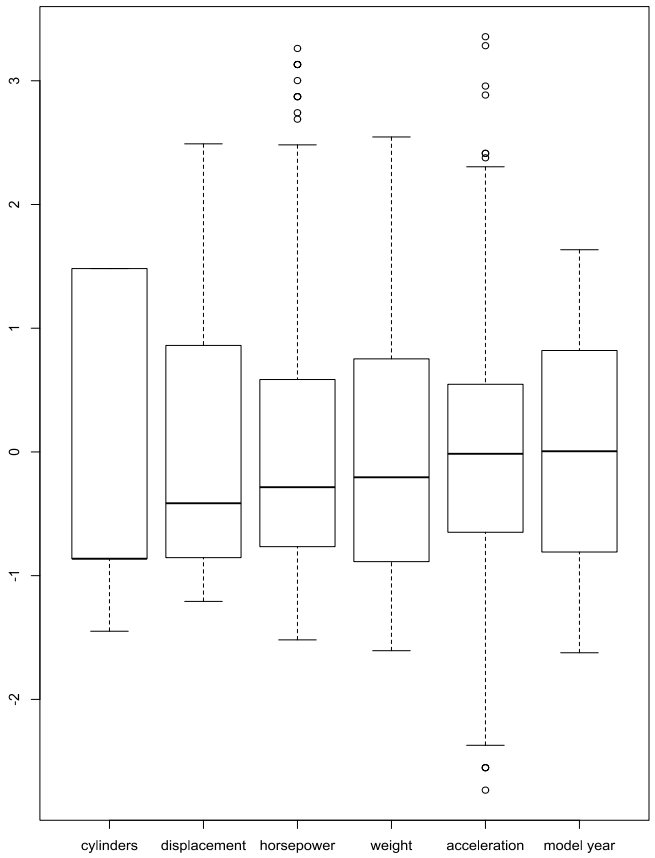

Fig. 1. Boxplots of the 6 covariates.

Table 8
Regression results of "mpg" against the derived indexes.

	DC		BC	
	linear model	nonlinear model	linear model	nonlinear model
Adjusted R-squared	0.19	0.51	0.64	0.85
F	46.48	81.78	343.80	446.20

and to formulate indexes for the following regression. The reason why only DC and BC are considered is that they do not require the conditions mentioned in Section 1, which are not easy to verify.

Let $\hat{\boldsymbol{\eta}} = (\hat{\eta}_1, \hat{\eta}_2)$ and $\tilde{\boldsymbol{\eta}} = (\tilde{\eta}_1, \tilde{\eta}_2)$ be the estimated basis matrices of the central subspace via DC and BC respectively. Denote the derived indexes by $\hat{Z}_1 = \hat{\eta}_1^{\mathrm{T}} X$, $\hat{Z}_2 = \hat{\eta}_2^{\mathrm{T}} X$, $\tilde{Z}_1 = \tilde{\eta}_1^{\mathrm{T}} X$ and $\tilde{Z}_2 = \tilde{\eta}_2^{\mathrm{T}} X$, where X represents the 6 predictors aforementioned. Fig. 2 shows the scatter plots of "mpg" against the four derived indexes. The patterns of "mpg" against the derived indexes are not very clear, so both linear and nonlinear models are tried to evaluate the efficiency of the two sufficient dimension reduction methods. In the linear model, "mpg" is regressed against the indexes $\hat{Z}_1$ ($\tilde{Z}_1$) and $\hat{Z}_2$ ($\tilde{Z}_2$). In the nonlinear model, we add the squared terms $\hat{Z}_1^2$ ($\tilde{Z}_1^2$), $\hat{Z}_2^2$ ($\tilde{Z}_2^2$) and $\hat{Z}_1 \cdot \hat{Z}_2$ ($\tilde{Z}_1 \cdot \tilde{Z}_2$) to the linear model.

Table 8 summarizes the regression results. Clearly, the indexes generated by BC fit much better to "mpg" than those generated by DC in both linear and nonlinear models. The significant increase of the values of the adjusted R-squared and F statistic from DC based regressions to BC based regressions demonstrates the efficacy of our BCov-SDR method for sufficient dimension reduction. The efficiency gain comes in part from the robustness of BC to the predictors with outliers. For some deeper reason, we think that BC may be a more linear measure than DC. Notice that when $\mathbf{X}$ and Y are standard normal random variables, BC and DC are both nondecreasing functions of the absolute value of their Pearson correlation. Recall that the empirical ball covariance is a rank statistic. The robustness of rank statistic may force heavy tailed distribution closer to normal distribution in a sense. Thus, we believe it is reasonable to think BC as a more linear measure than DC. Furthermore, it seems that BCov-SDR works pretty well in presence of discrete predictors.

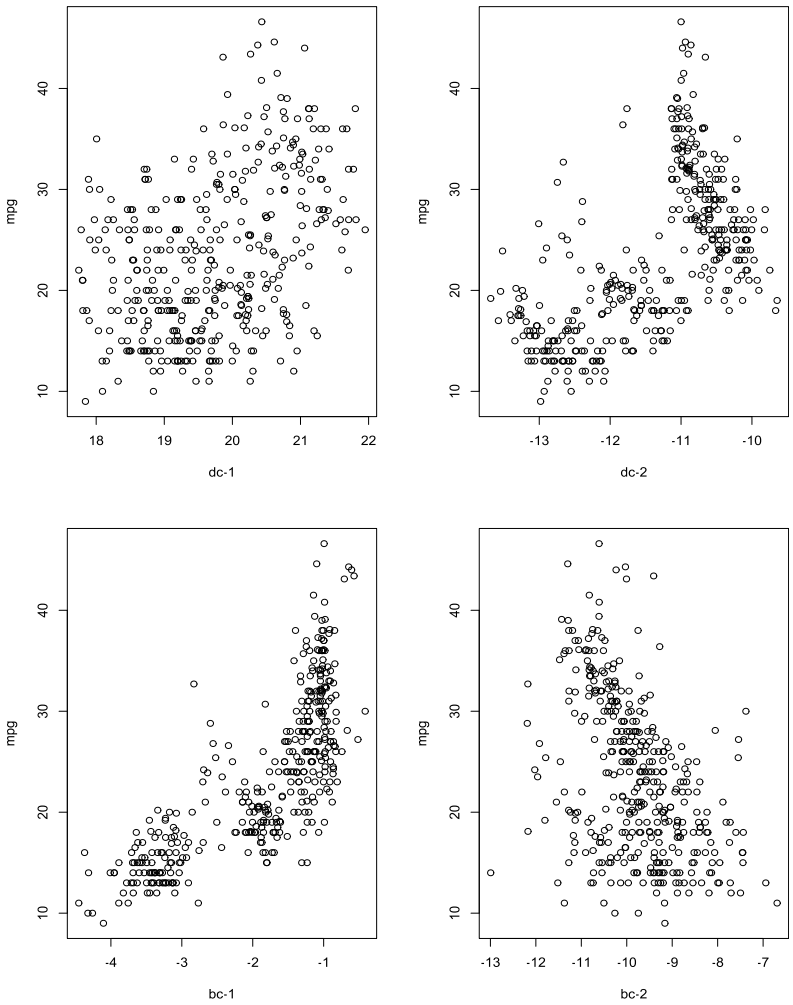

Fig. 2. Scatterplots of *mpg* versus the derived indexes produced by DC and BC: DC-1, DC-2, BC-1, and BC-2.

Cardiomyopathy Microarray Data.

Cardiomyopathy Microarray Data is used to evaluate the robustness of BCov-SDR (or BC, for simplicity) to heavy-tailed predictors. This data has been frequently studied (Segal et al., 2003; Zou and Yuan, 2008; Li et al., 2012b; Chen et al., 2018) to investigate the relationship between over expression of a G protein-coupled receptor (Ro1) in mice and 6319 related genes. Only 30 observations are collected, so we conduct feature screening as a first step. Three screening methods are employed: SIS (Fan and Lv, 2008), DCSIS (Li et al., 2012b), and BCSIS (Pan et al., 2018). SIS is quite popular in the literature, and DCSIS and BCSIS are both model-free feature screening methods. The basic idea of marginal feature screening is to rank the predictors by some utility measure between the response and predictors, and then to retain the predictors with top utilities for further study. Specifically, Let $U(Y, X_j)$ denote the utility measure between the response Y and the jth predictor X_j, the marginal feature screening chooses a set $\mathcal{A}$ of predictors:

$$\mathcal{A} = \{j : U(Y, X_j) > \tau, j = 1, \ldots, n\}$$

where τ is some threshold. Fan and Lv (2008) suggests retaining $\lfloor n/\log(n) \rfloor$ predictors for a sample with size n.

SIS uses the absolute value of Pearson correlation as the utility measure, DCSIS uses distance correlation, and BCSIS uses ball correlation. Different utility measures may select different predictors. We put all the predictors selected by SIS, DCSIS and BCSIS together to get a conservative set of predictors for the following sufficient dimension reduction. After each screening procedure, $\lfloor n/\log(n) \rfloor = 8$ variables or genes are screened out: Msa.2877.0, Msa.964.0, Msa.2134.0, Msa.741.0, Msa.3041.0, Msa.1590.0, Msa.15405.0 and Msa.1166.0 from SIS; Msa.2134.0, Msa.2877.0, Msa.26025.0, Msa.5583.0, Msa.1590.0, Msa.1166.0, Msa.2400.0 and Msa.15442.0 from DCSIS; and Msa.2134.0, Msa.5794.0, Msa.2877.0, Msa.1166.0, Msa.5727.0, Msa.15442.0, Msa.26025.0 and Msa.42131.0 from BCSIS. We then put the selected variables together and finally get 15 variables after removing the duplicates. At the significant level 0.1, 7 variables cannot pass the Shapiro–Wilk normality test, and most of them are characterized by heavy tails.

Table 9
Regression results of "Ro1" against the derived indexes.

	DC		BC	
	linear model	nonlinear model	linear model	nonlinear model
Adjusted R-squared	0.75	0.86	0.79	0.94
F	45.06	37.97	56.15	87.29

The following procedures are similar to those done in the auto MPG data. Two indexes are derived by DC and BC respectively, after which linear and nonlinear regressions are constructed to measure the goodness of fit of the two pair of indexes to "Ro1". The regression results are presented in Table 9. There is no surprise that the BC based indexes perform far better than the DC based indexes considering the heavy-tailed character of the predictors.

5. Conclusion

In this paper, we propose BCov-SDR, a robust sufficient dimension reduction method that is based on the ball covariance. This method is general and widely applicable, because it is not restricted to various conditions required by other popular sufficient dimension reduction approaches. Moreover, it is robust to heavy-tailed predictors and outliers, which is demonstrated by both the simulation studies and real data analysis. The asymptotic property of the BCov-SDR estimator of the central subspace is also investigated. Although our method is now under a traditional fixed dimension setting, it can be extended to high dimensional settings easily by conducting a screening procedure first, see Fan and Lv (2017) for reference.

Acknowledgments

We gratefully thank the Editor, the Associate Editor and two referees for all the questions, constructive comments and suggestions. This work was funded by the Fundamental Research Funds for the Central Universities, China (Grant No. JBK1707113) and the Joint Lab of Data Science and Business Intelligence at SWUFE. This research was also supported by the "China Scholarship Council ([2017]3109)". Dr. Jia Zhang (No. 201706980021) thanks the China Scholarship Council for financial support to visit National University of Singapore, Singapore.

Appendix

Lemma 1. *If* $(W_1, V_1) \perp\!\!\!\perp (W_2, V_2)$, *then* $\mathrm{BCov}^2(W_1 + W_2, V_1 + V_2) < \mathrm{BCov}^2(W_1, V_1) + \mathrm{BCov}^2(W_2, V_2)$.

Proof. Let (W_1, V_1) and (W_2, V_2) be two B-valued random vectors defined on a probability space $(\Omega, \mathcal{F}, P)$ such that $(W_1, V_1) \sim \theta_1$, $W_1 \sim \mu_1$, $V_1 \sim \nu_1$, and $(W_2, V_2) \sim \theta_2$, $W_2 \sim \mu_2$, $V_2 \sim \nu_2$, and $(W, V) = (W_1 + W_2, V_1 + V_2) \sim \theta$, $W \sim \mu$, $V \sim \nu$, where θ, θ_1, θ_2, μ, μ_1, μ_2, ν, ν_1 and ν_2 are Borel probability measures.

$$
\begin{aligned}
&\mathrm{BCov}^2(W, V)\\
&= \mathrm{E}\{[\theta - \mu \otimes \nu]^2(\bar{B}_{\zeta_W}(w^1, w^2) \times \bar{B}_{\zeta_V}(v^1, v^2)|(w^1, v^1), (w^2, v^2))\}\\
&\le \mathrm{E}\{[\theta_1(\bar{B}_{\zeta_{W_1}}(w_1^1, w_1^2) \times \bar{B}_{\zeta_{V_1}}(v_1^1, v_1^2))\theta_2(\bar{B}_{\zeta_{W_2}}(w_2^1, w_2^2) \times \bar{B}_{\zeta_{V_2}}(v_2^1, v_2^2))-\\
&\quad \mu_1(\bar{B}_{\zeta_{W_1}}(w_1^1, w_1^2))\mu_2(\bar{B}_{\zeta_{W_2}}(w_2^1, w_2^2))\nu_1(\bar{B}_{\zeta_{V_1}}(v_1^1, v_1^2))\nu_2(\bar{B}_{\zeta_{V_2}}(v_2^1, v_2^2))]^2\\
&\quad |(w_1^1, v_1^1), (w_1^2, v_1^2), (w_2^1, v_2^1), (w_2^2, v_2^2)\}\\
&\le \mathrm{E}\{[\theta_1 - \mu_1 \otimes \nu_1]^2(\bar{B}_{\zeta_{W_1}}(w_1^1, w_1^2) \times \bar{B}_{\zeta_{V_1}}(v_1^1, v_1^2))|(w_1^1, v_1^1), (w_1^2, v_1^2)\}+\\
&\quad \mathrm{E}\{[\theta_2 - \mu_2 \otimes \nu_2]^2(\bar{B}_{\zeta_{W_2}}(w_2^1, w_2^2) \times \bar{B}_{\zeta_{V_2}}(v_2^1, v_2^2))|(w_2^1, v_2^1), (w_2^2, v_2^2)\}\\
&\le \mathrm{BCov}^2(W_1, V_1) + \mathrm{BCov}^2(W_2, V_2),
\end{aligned}
$$

where the first inequality is obtained by the fact that $(W_1, V_1) \perp\!\!\!\perp (W_2, V_2)$. Thus, we complete the proof. □

Proof of Theorem 1

Proof. For $\boldsymbol{\beta}$ and $\mathbf{B}$ defined in Theorem 1, we can find a rotation matrix Q such that $\boldsymbol{\beta}Q = (B_a, B_b)$, $S(B_a) \subset S(\mathbf{B})$ and $S(B_b) \subset S(\mathbf{B})^{\perp}$, where $S(\mathbf{B})^{\perp}$ denotes the orthogonal space of $S(\mathbf{B})$.

By Condition (2) and the definition of $\mathbf{B}$, we get $(Y, \mathbf{B}^{\mathrm{T}}\mathbf{X}) \perp\!\!\!\perp B_b^{\mathrm{T}}\mathbf{X}$. Then by Proposition 4.3 of Cook (1998), $(Y, B_a^{\mathrm{T}}\mathbf{X}) \perp\!\!\!\perp B_b^{\mathrm{T}}\mathbf{X}$. Denote $W_1 = (B_a^{\mathrm{T}}\mathbf{X}, 0)$, $V_1 = Y$, $W_2 = (0, B_b^{\mathrm{T}}\mathbf{X})$ and $V_2 = 0$. Clearly, $(W_1, V_1) \perp\!\!\!\perp (W_2, V_2)$. Hence, by Lemma 1 we obtain $\mathrm{BCov}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, Y) < \mathrm{BCov}(B_a^{\mathrm{T}}\mathbf{X}, Y)$. According to Condition (1), we have $\mathrm{BCov}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, Y) < \mathrm{BCov}(\mathbf{B}^{\mathrm{T}}\mathbf{X}, Y)$ and complete the proof. □

Lemma 2. $\mathrm{BCov}_n(\mathbf{X}, Y) \to_{a.s.} \mathrm{BCov}(\mathbf{X}, Y)$.

Proof. This lemma comes from Lemma 3 of Pan et al. (2018). □

Proof of Theorem 2.

Proof. If $\boldsymbol{\eta}_n$ is not a consistent estimator of $\boldsymbol{\eta}$, there exists a subsequence of $\boldsymbol{\eta}_{n^*}$ of $\boldsymbol{\eta}_n$ such that $\boldsymbol{\eta}_{n^*} \to_p \boldsymbol{\eta}^*$, where $\boldsymbol{\eta}^{*\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}\boldsymbol{\eta}^* = \mathbf{I}_d$ and $\boldsymbol{\eta}^* \neq \boldsymbol{\eta}$. Let $\boldsymbol{\eta}_{n^*} = \boldsymbol{\eta}^* + \boldsymbol{\epsilon}_{n^*}$, then when $\boldsymbol{\epsilon}_{n^*}$ gets small enough, by the definition of BCov_n, we obtain $\mathrm{BCov}_n(\boldsymbol{\eta}_{n^*}^{\mathrm{T}}\mathbf{X}, Y) = \mathrm{BCov}_n(\boldsymbol{\eta}^{*\mathrm{T}}\mathbf{X}, Y)$. According to Lemma 2, $\mathrm{BCov}_n(\boldsymbol{\eta}^{*\mathrm{T}}\mathbf{X}, Y) \to_{a.s.} \mathrm{BCov}(\boldsymbol{\eta}^{*\mathrm{T}}\mathbf{X}, Y)$. Hence, $\mathrm{BCov}_n(\boldsymbol{\eta}_{n^*}^{\mathrm{T}}\mathbf{X}, Y) \to_{a.s.} \mathrm{BCov}(\boldsymbol{\eta}^{*\mathrm{T}}\mathbf{X}, Y)$.

On the other hand, by (2.2), we have $\mathrm{BCov}_n(\boldsymbol{\eta}_{n^*}^{\mathrm{T}}\mathbf{X}, Y) \geq \mathrm{BCov}_n(\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}, Y)$. Let $n \to \infty$, and we get $\mathrm{BCov}(\boldsymbol{\eta}^{*\mathrm{T}}\mathbf{X}, Y) \geq \mathrm{BCov}(\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}, Y)$, which contradicts to the definition of $\boldsymbol{\eta}$ (see (2.3)). Thus, $\boldsymbol{\eta}_n$ must be a consistent estimator of $\boldsymbol{\eta}$. □

References

Bura, E., Cook, R.D., 2001. Extending sliced inverse regression: The weighted chi-squared test. J. Amer. Statist. Assoc. 96 (455), 996–1003.
Bura, E., Duarte, S., Forzani, L., 2016. Sufficient reductions in regressions with exponential family inverse predictors. J. Amer. Statist. Assoc. 111 (515), 1313–1329.
Bura, E., Forzani, L., 2015. Sufficient reductions in regressions with elliptically contoured inverse predictors. J. Amer. Statist. Assoc. 110 (509), 420–434.
Chen, X., Cook, R.D., Zou, C., 2015. Diagnostic studies in sufficient dimension reduction. Biometrika 102 (3), 545–558.
Chen, X., Sheng, W., Yin, X., 2018. Efficient sparse estimate of sufficient dimension reduction in high dimension. Technometrics 60 (2), 161–168.
Cook, R.D., 1994. On the interpretation of regression plots. J. Amer. Statist. Assoc. 89 (425), 177–189.
Cook, R.D., 1996. Graphics for regressions with a binary response. J. Amer. Statist. Assoc. 91 (435), 983–992.
Cook, R.D., 1998. Regression Graphics: Ideas for Studying Regressions Through Graphics. Wiley, New York.
Cook, R.D., Forzani, L., 2008. Principal fitted components for dimension reduction in regression. Statist. Sci. 23 (4), 485–501.
Cook, R.D., Forzani, L., 2009. Likelihood-based sufficient dimension reduction. J. Amer. Statist. Assoc. 104 (485), 197–208.
Cook, R.D., Li, L., 2009. Dimension reduction in regressions with exponential family predictors. J. Comput. Graph. Statist. 18 (3), 774–791.
Cook, R.D., Ni, L., 2005. Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. J. Amer. Statist. Assoc. 100 (470), 410–428.
Cook, R.D., Weisberg, S., 1991. Sliced inverse regression for dimension reduction: Comment. J. Amer. Statist. Assoc. 86 (414), 328–332.
Cook, R.D., Yin, X., 2001. Theory & methods: Special invited paper: Dimension reduction and visualization in discriminant analysis (with discussion). Aust. N. Z. J. Stat. 43 (2), 147–199.
Croux, C., Filzmoser, P., Fritz, H., 2013. Robust sparse principal component analysis. Technometrics 55 (2), 202–214.
Dong, Y., Yu, Z., Zhu, L., 2015. Robust inverse regression for dimension reduction. J. Multivariate Anal. 134, 71–81.
Fan, J., Lv, J., 2008. Sure independence screening for ultrahigh dimensional feature space. J. R. Stat. Soc. Ser. B Stat. Methodol. 70 (5), 849–911.
Fan, J., Lv, J., 2017. Sure Independence Screening. Wiley StatsRef.
Gretton, A., Fukumizu, K., Teo, C.H., Song, L., Schölkopf, B., Smola, A.J., 2008. A kernel statistical test of independence. In: Advances in Neural Information Processing Systems. pp. 585–592.
Han, F., Liu, H., 2018. ECA: High-dimensional elliptical component analysis in non-Gaussian distributions. J. Amer. Statist. Assoc. 113 (521), 252–268.
Li, K.C., 1991. Sliced inverse regression for dimension reduction. J. Amer. Statist. Assoc. 86 (414), 316–327.
Li, K.C., 1992. On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. J. Amer. Statist. Assoc. 87 (420), 1025–1039.
Li, G., Peng, H., Zhang, J., Zhu, L., 2012a. Robust rank correlation based screening. Ann. Statist. 40 (3), 1846–1877.
Li, B., Wang, S., 2007. On directional regression for dimension reduction. J. Amer. Statist. Assoc. 102 (479), 997–1008.
Li, B., Zha, H., Chiaromonte, F., 2005. Contour regression: a general approach to dimension reduction. Ann. Statist. 33 (4), 1580–1616.
Li, R., Zhong, W., Zhu, L., 2012b. Feature screening via distance correlation learning. J. Amer. Statist. Assoc. 107 (499), 1129–1139.
Lyons, R., 2013. Distance covariance in metric spaces. Ann. Probab. 41 (5), 3284–3305.
Pan, W., Wang, X., Xiao, W., Zhu, H., 2018. A generic sure independence screening procedure. J. Amer. Statist. Assoc. 1–29, (just-accepted).
Rekabdarkolaee, H.M., Boone, E., Wang, Q., 2017. Robust estimation and variable selection in sufficient dimension reduction. Comput. Statist. Data Anal. 108, 146–157.
Segal, M.R., Dahlquist, K.D., Conklin, B.R., 2003. Regression approaches for microarray data analysis. J. Comput. Biol. 10 (6), 961–980.
Sejdinovic, D., Sriperumbudur, B., Gretton, A., Fukumizu, K., 2013. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. Ann. Statist. 41 (5), 2263–2291.
Sheng, W., Yin, X., 2013. Direction estimation in single-index models via distance covariance. J. Multivariate Anal. 122, 148–161.
Sheng, W., Yin, X., 2016. Sufficient dimension reduction via distance covariance. J. Comput. Graph. Statist. 25 (1), 91–104.
Székely, G.J., Rizzo, M.L., Bakirov, N.K., 2007. Measuring and testing dependence by correlation of distances. Ann. Statist. 35 (6), 2769–2794.
Wang, H., Xia, Y., 2008. Sliced regression for dimension reduction. J. Amer. Statist. Assoc. 103 (482), 811–821.
Xia, Y., Tong, H., Li, W.K., Zhu, L.X., 2002. An adaptive estimation of dimension reduction space. J. R. Stat. Soc. Ser. B Stat. Methodol. 64 (3), 363–410.
Ye, Z., Weiss, R.E., 2003. Using the bootstrap to select one of a new class of dimension reduction methods. J. Amer. Statist. Assoc. 98 (464), 968–979.
Yin, X., Li, B., 2011. Sufficient dimension reduction based on an ensemble of minimum average variance estimators. Ann. Statist. 39 (6), 3392–3416.
Yin, X., Li, B., Cook, R.D., 2008. Successive direction extraction for estimating the central subspace in a multiple-index regression. J. Multivariate Anal. 99 (8), 1733–1757.
Zeng, P., Zhu, Y., 2010. An integral transform method for estimating the central mean and central subspaces. J. Multivariate Anal. 101 (1), 271–290.
Zhou, J., Xu, W., Zhu, L., 2015. Robust estimating equation-based sufficient dimension reduction. J. Multivariate Anal. 134, 99–118.
Zhu, L., Miao, B., Peng, H., 2006. On sliced inverse regression with high-dimensional covariates. J. Amer. Statist. Assoc. 101 (474), 630–643.
Zhu, Y., Zeng, P., 2006. Fourier methods for estimating the central subspace and the central mean subspace in regression. J. Amer. Statist. Assoc. 101 (476), 1638–1651.
Zou, H., Yuan, M., 2008. Regularized simultaneous model selection in multiple quantiles regression. Comput. Statist. Data Anal. 52 (12), 5296–5304.