Statistics and Its Interface Volume 12 (2019) 489–500

A martingale-difference-divergence-based estimation of central mean subspace[*]

Yu Zhang, Jicai Liu, Yuesong Wu, and Xiangzhong Fang[†]

In this article, we propose a new method for estimating the central mean subspace via the martingale difference divergence. This method enjoys a model free property and does not need any nonparametric estimation. These advantages enable our method to work effectively when many discrete or categorical predictors exist. Under mild conditions, we show that our estimator is root-n consistent. To determine the structural dimension of the central mean subspace, a consistent Bayesian-type information criterion is developed. Simulation studies and a real data example are given to illustrate the proposed estimation methodology.

Keywords and phrases: Central mean subspace, Distance covariance, Martingale difference divergence, Multiple index models, Sufficient dimension reduction.

1. INTRODUCTION

Sufficient dimension reduction (SDR) (Li [8], Cook [2]) is a popular approach to tackle the challenges of high dimensional data analysis. It is aimed at seeking a matrix $\mathbf{B} \in \mathbb{R}^{p\times d}$ with rank d ($d < p$), such that $Y \perp\!\!\!\perp \mathbf{X}|\mathbf{B}^T\mathbf{X}$, or equivalently,

$$(1) \qquad P\{Y \le y|\mathbf{X}\} = P\{Y \le y|\mathbf{B}^T\mathbf{X}\}, \text{ for all } y \in \mathbb{R},$$

where Y is the response, $\mathbf{X} = (X_1, \cdots, X_p)^T$ is the predictor vector and $\perp\!\!\!\perp$ indicates conditional independence. The column space of $\mathbf{B}$ satisfying (1) is called the SDR subspace. The intersection of all such subspaces, if itself satisfies (1), is called the central subspace (CS), denoted by $\mathcal{S}_{Y|\mathbf{X}}$. The column dimension d of $\mathbf{B}$ is called the structural dimension of $\mathcal{S}_{Y|\mathbf{X}}$. There is a huge literature on estimating $\mathcal{S}_{Y|\mathbf{X}}$, for instance, Li [8], Cook and Weisberg [4], Li and Wang [6], Ma and Zhu [11], among others.

The SDR model (1) concerns about all aspects of the conditional distribution of Y given $\mathbf{X}$. However, in many applications, certain characteristics of the conditional distribution may often be of special interest. For example, we might be only interested in the conditional expectation $E\{Y|\mathbf{X}\}$ in regression analysis. For this purpose, Cook and Li [3] introduced the following model

$$(2) \qquad E\{Y|\mathbf{X}\} = E\{Y|\mathbf{B}^T\mathbf{X}\},$$

where $\mathbf{B} \in \mathbb{R}^{p\times d}$. The minimum column space of $\mathbf{B}$ satisfying (2) is called the central mean subspace (CMS), denoted by $\mathcal{S}_{E\{Y|\mathbf{X}\}}$. Various methods have been developed to estimate $\mathcal{S}_{E\{Y|\mathbf{X}\}}$, among with ordinary least squares method (OLS, Li and Duan [10]), principle Hessian direction (pHd, Li [9]), iterative Hessian transformation (ITH, Cook and Li [3]), minimum average variance estimation (MAVE, Xia et al. [23]), Fourier transformation method (FMN, Zhu and Zeng [25]) and semiparametric approach [13]. See, Ma and Zhu [12] for recent review.

These existing CMS methods often rely on nonparametric smoothing techniques or impose strong conditions on predictors, such as linearity condition or constant covariance condition [12]. Different from the traditional methods, this paper proposes a new method based on the martingale difference divergence (MDD, Shao and Zhang [17]). Our proposed method is a model-free procedure, which can recover $\mathcal{S}_{E\{Y|\mathbf{X}\}}$ without prespecifying any models, without smoothing techniques and work effectively under different kinds of predictors.

Recently, Sheng and Yin [18, 19] used the distance covariance (DCOV, Székely et al. [21, 22]) to estimate $\mathcal{S}_{Y|\mathbf{X}}$ for model (1). Our proposed MDD method is motivated from Sheng and Yin [18, 19]. Notice that DCOV is to measure (in)dependence between random variables, while MDD is for the conditional mean (in)dependence. This is the main reason why we should use MDD instead of DCOV in model (2). In the simulation studies, we can see that the DCOV procedure may fail to identify the central mean subspace $\mathcal{S}_{E\{Y|\mathbf{X}\}}$ in many settings, for example, in models with heteroscedastic errors.

It is noteworthy that Propositions 1 and 2 in Sheng and Yin [19] are fundamental to ensure their method work. However, since MDD does not inherit all the properties of DCOV, the proof strategies used by Sheng and Yin [19] are not adaptive to our method. Thus, how to obtain similar propositions for model (2) is an important but challenging problem. To deal with this difficulty, more complicated techniques are needed. More details can been found in the proofs of Propositions 1 and 2 in the Appendix.

[*]The research was supported in part by National Natural Science Foundation of China (11426156, 11501372).

[†]Corresponding author.

The rest of the paper is organized as follows. Section 2 describes our method and its asymptotic properties. In Section 3, we introduce a BIC criterion for the CMS dimension determination. In Sections 4–5, a Monte Carlo simulation study and a real data application are used to illustrate the proposed methodology. In Section 6, some conclusion remarks are given. All the regularity conditions and the technical proofs are deferred to the Appendix.

2. METHODOLOGY

2.1 A brief review of MDD

For two random vectors $\mathbf{X} \in \mathbb{R}^p$ and $Y \in \mathbb{R}$, Shao and Zhang [17] introduced the following conditional mean independence of Y on $\mathbf{X}$

$$E\{Y|\mathbf{X}\} = E\{Y\}, \quad \text{almost surely.} \tag{3}$$

The relationship (3) plays an important role in statistics. To measure this relationship, Shao and Zhang [17] proposed the martingale difference divergence (MDD), given by

$$\mathrm{MDD}(Y|\mathbf{X})^2 = \int_{\mathcal{R}^p} |E\{Ye^{i<\mathbf{s},\mathbf{X}>}\} - E\{Y\}E\{e^{i<\mathbf{s},\mathbf{X}>}\}|^2 \omega(\mathbf{s})ds,$$

where $i = \sqrt{-1}$ is the imaginary unit, $\omega(\mathbf{s}) = 1/\{\|\mathbf{s}\|^{1+p}c_p\}$ and $c_p = \pi^{(p+1)/2}/\Gamma((p+1)/2)$. Throughout the paper, $\|\cdot\|$ denotes the (possibly complex) Euclidean norm, defined by $\|\mathbf{u}\| = \sqrt{\mathbf{u}^H\mathbf{u}}$, where $\mathbf{u}^H$ denotes the conjugate transpose of $\mathbf{u} \in \mathbb{C}^p$. If $\mathbf{u} \in \mathbb{C}^1$, its modulus is simply denoted as $|u|$.

The MDD has an attractive property that $\mathrm{MDD}(Y|\mathbf{X}) = 0$ if and only if (3) holds. That is, the MDD can be used to characterize the conditional mean independence. Furthermore, Theorem 1 in Shao and Zhang [17] suggests that

$$\mathrm{MDD}(Y|\mathbf{X})^2 = -E\{(Y - E\{Y\})(Y' - E\{Y'\})\|\mathbf{X} - \mathbf{X}'\|\}, \tag{4}$$

where $(Y', \mathbf{X}')$ is an independent copy of $(Y, \mathbf{X})$, if $E\{|Y|^2 + \|\mathbf{X}\|_2^2\} < \infty$. This equation only involves the expectations of $(Y, \mathbf{X})$ and thus it is easy to be computed and estimated.

Given n independent observations $\{(Y_k, \mathbf{X}_k), k = 1, \cdots, n\}$ from the joint distribution of $(Y, \mathbf{X})$, we adopt the idea of $\mathcal{U}$-centring in Park et al. [15] to construct an unbiased estimator for $\mathrm{MDD}(Y|\mathbf{X})^2$. Define $\Phi = (\Phi_{kl})_{k,l=1}^n$ and $\Psi = (\Psi_{kl})_{k,l=1}^n$, where $\Phi_{kl} = \|\mathbf{X}_k - \mathbf{X}_l\|$ and $\Psi_{kl} = (Y_k - Y_l)^2/2$. The $\mathcal{U}$-centred versions of Φ_{ij} and Ψ_{ij} are defined respectively

$$\widetilde{\Phi}_{kl} = \Phi_{kl} - \frac{1}{n-2}\sum_{j=1}^n \Phi_{kj} - \frac{1}{n-2}\sum_{j=1}^n \Phi_{jl} + \frac{1}{(n-1)(n-2)}\sum_{i,j=1}^n \Phi_{ij},$$

$$\widetilde{\Psi}_{kl} = \Psi_{kl} - \frac{1}{n-2}\sum_{j=1}^n \Psi_{kj} - \frac{1}{n-2}\sum_{j=1}^n \Psi_{jl} + \frac{1}{(n-1)(n-2)}\sum_{i,j=1}^n \Psi_{ij}.$$

Then, an unbiased estimator of $\mathrm{MDD}(Y|\mathbf{X})^2$ is given by

$$\mathrm{MDD}_n(Y|\mathbf{X})^2 = \frac{1}{n(n-3)}\sum_{k\neq l} \widetilde{\Phi}_{kl}\widetilde{\Psi}_{kl}. \tag{5}$$

2.2 Estimating the central mean subspace

Let the columns of $\mathbf{B}_0 = (\beta_{01}, \cdots, \beta_{0d_0})$ be a basis of $\mathcal{S}_{E\{Y|\mathbf{X}\}}$ with $\mathbf{B}_0^T\Sigma_{\mathbf{x}}\mathbf{B}_0 = \mathbf{I}_{d_0}$, where d_0 ($< p$) is the true structural dimension, $\Sigma_{\mathbf{x}}$ is the covariance matrix of $\mathbf{X}$ and $\mathbf{I}_{d_0}$ is the identity matrix. Assume that $\mathbf{P}_{\mathbf{B}_0} = \mathbf{B}_0(\mathbf{B}_0^T\Sigma_{\mathbf{x}}\mathbf{B}_0)^{-1}\mathbf{B}_0^T\Sigma_{\mathbf{x}}$ and $\mathbf{Q}_{\mathbf{B}_0} = \mathbf{I} - \mathbf{P}_{\mathbf{B}_0}$. In this section, we focus on the estimation of $\mathbf{B}_0$ when d_0 is known.

The following proposition suggests the MDD measure can be used to identify the central mean subspace under some conditions.

Proposition 1. *Assume that $\mathbf{P}_{\mathbf{B}_0}^T\mathbf{X} \perp\!\!\!\perp \mathbf{Q}_{\mathbf{B}_0}^T\mathbf{X}$. For any $p \times d_0$ matrix $\mathbf{A}$ with $\mathbf{A}^T\Sigma_{\mathbf{x}}\mathbf{A} = \mathbf{I}_{d_0}$, we have*

$$\mathrm{MDD}(Y|\mathbf{A}^T\mathbf{X})^2 \leq \mathrm{MDD}(Y|\mathbf{B}_0^T\mathbf{X})^2. \tag{6}$$

Moreover, the equality holds if and only if there exists an orthogonal matrix $\mathbf{C}_1 \in \mathbb{R}^{d_0 \times d_0}$ such that $\mathbf{A} = \mathbf{B}_0\mathbf{C}_1$.

If $\mathbf{X}$ is normal, we have $\mathrm{Cov}(\mathbf{P}_{\mathbf{B}_0}^T\mathbf{X}, \mathbf{Q}_{\mathbf{B}_0}^T\mathbf{X}) = 0$, which indicates that $\mathbf{P}_{\mathbf{B}_0}^T\mathbf{X} \perp\!\!\!\perp \mathbf{Q}_{\mathbf{B}_0}^T\mathbf{X}$. In general, a distribution with "the linear condtional mean condition" or "constant covariance conditions" used in the SDR literature, does not necessarily satisfy such condition. When p is large, Sheng and Yin [18] showed that the independence condition is not as stringent as it seems to be. Our simulations indicate that the proposed method still works well when it is non-normal so the method is widely applicable.

Let $\mathcal{S}(\mathbf{B})$ be the linear subspace spanned by the columns vectors of any matrix $\mathbf{B}$. Proposition 1 suggests that, if $\mathcal{S}(\mathbf{B}_0) \neq \mathcal{S}(\mathbf{A})$ for $\mathbf{A}^T\Sigma_{\mathbf{x}}\mathbf{A} = \mathbf{I}_{d_0}$ holds, $\mathrm{MDD}(Y|\mathbf{A}^T\mathbf{X})^2$ is strictly less than $\mathrm{MDD}(Y|\mathbf{B}_0^T\mathbf{X})^2$. Furthermore, if $\mathbf{A}$ is another basis of the central mean subspace, i.e., $\mathcal{S}(\mathbf{A}) = \mathcal{S}(\mathbf{B}_0)$, we have $\mathrm{MDD}(Y|\mathbf{A}^T\mathbf{X})^2 = \mathrm{MDD}(Y|\mathbf{B}_0^T\mathbf{X})^2$. Thus, Proposition 1 implies that

$$\mathbf{B}_0 = \underset{\mathbf{B}^T\Sigma_{\mathbf{x}}\mathbf{B} = \mathbf{I}_{d_0}}{\arg\max} \ \mathrm{MDD}(Y|\mathbf{B}^T\mathbf{X})^2. \tag{7}$$

Therefore, the MDD measure can be used to identify the central mean subspace.

Suppose that $\{(\mathbf{X}_i, Y_i), i = 1, \cdots, n\}$ are independent identically distributed from model (2). Based on (5), $\mathrm{MDD}(Y|\mathbf{B}^T\mathbf{X})^2$ could be estimated by the following form

$$\mathrm{MDD}_n(Y|\mathbf{B}^T\mathbf{X})^2 = \frac{1}{n(n-3)}\sum_{k\neq l} \widetilde{\Phi}_{kl}(\mathbf{B})\widetilde{\Psi}_{kl}, \tag{8}$$

where $\Phi_{kl}(\mathbf{B}) = \|\mathbf{B}^T(\mathbf{X}_k - \mathbf{X}_l)\|$ and

$$\begin{aligned}\widetilde{\Phi}_{kl}(\mathbf{B}) = \Phi_{kl}(\mathbf{B}) &- \frac{1}{n-2}\sum_{j=1}^{n}\Phi_{kj}(\mathbf{B}) \\ &- \frac{1}{n-2}\sum_{j=1}^{n}\Phi_{jl}(\mathbf{B}) + \frac{1}{(n-1)(n-2)}\sum_{i,j=1}^{n}\Phi_{ij}(\mathbf{B}).\end{aligned}$$

Let $\hat{\Sigma}_{\mathbf{x}}$ be the sample covariance of $\mathbf{X}$. From (7), $\mathbf{B}_0$ can be estimated by

$$\widehat{\mathbf{B}}_n = \underset{\mathbf{B}^T\hat{\Sigma}_{\mathbf{x}}\mathbf{B}=\mathbf{I}_{d_0}}{\arg\max}\ \mathrm{MDD}_n(Y|\mathbf{B}^T\mathbf{X})^2. \tag{9}$$

We have the following asymptotic result for $\widehat{\mathbf{B}}_n$.

Theorem 1. *Assume that* $E\{|Y|^2 + \|\mathbf{X}\|_2^2\} < \infty$*. If* $\mathbf{P}_{\mathbf{B}_0}^T\mathbf{X} \perp\!\!\!\perp \mathbf{Q}_{\mathbf{B}_0}^T\mathbf{X}$*, then we have*

$$\left\|\widehat{\mathbf{B}}_n\widehat{\mathbf{B}}_n^T - \mathbf{B}_0\mathbf{B}_0^T\right\|_F = O_p(n^{-1/2}), \tag{10}$$

where $\|\cdot\|_F$ *is the Frobenius norm of a matrix.*

In Theorem 1, the moment condition $E\{|Y|^2 + \|\mathbf{X}\|_2^2\} < \infty$ is commonly used in Székely et al. [21] and Shao and Zhang [17]. Theorem 1 shows that $\widehat{\mathbf{B}}_n$ converges at a root-n consistency rate.

There are several optimization algorithms to obtain the estimator $\widehat{\mathbf{B}}_n$ in (9). Sheng and Yin [18] and Sheng and Yin [19] recommended the Sequential Quadratic Programming method (SQP, Nocedal and Wright [14]) to solve similar optimization problems. Xue et al. [24] advocated a projection pursuit type of sufficient searching algorithm, which searches and estimates one direction at a time. Cowley et al. [5] solved similar problems by projected gradient descent with backtracking line search.

In our numerical studies, we use the SQP method. It solves a sequence of optimization subproblems, each of which optimizes a quadratic programming subproblem. The SQP algorithm is similar to the algorithm in Sheng and Yin [18], hence we here omit it. Our numerical results indicate the algorithm is accurate and easy to implement. In this article, we use the MAVE method to estimate the initials.

3. ESTIMATING THE STRUCTURAL DIMENSION

In practice, one may have little prior knowledge about the true structural dimension d_0. In this section, we propose a Bayesian-type information criterion to estimate d_0. The following proposition ensures the MDD measure can be used in selecting d_0.

Proposition 2. *Assume that* $\mathbf{A}$ *is any* $p \times d$ *matrix with* $\mathbf{A}^T\Sigma_{\mathbf{x}}\mathbf{A} = \mathbf{I}_d$*, for any* $d \in \{1, \cdots, p\}$*. If* $\mathbf{P}_{\mathbf{B}_0}^T\mathbf{X} \perp\!\!\!\perp \mathbf{Q}_{\mathbf{B}_0}^T\mathbf{X}$ *holds, then*

(i) for any $d < d_0$*, we have*

$$\mathrm{MDD}(Y|\mathbf{A}^T\mathbf{X})^2 < \mathrm{MDD}(Y|\mathbf{B}_0^T\mathbf{X})^2; \tag{11}$$

(ii) for any $d > d_0$ *and* $\mathcal{S}(\mathbf{B}_0) \nsubseteq \mathcal{S}(\mathbf{A})$*, then* (11) *still holds;*

(iii) for any $d > d_0$ *and* $\mathcal{S}(\mathbf{B}_0) \subseteq \mathcal{S}(\mathbf{A})$*, then we obtain*

$$\mathrm{MDD}(Y|\mathbf{A}^T\mathbf{X})^2 \leqslant \mathrm{MDD}(Y|\mathbf{B}_0^T\mathbf{X})^2.$$

Proposition 2(iii) indicates that the MDD measure may fail to distinguish the true structural dimension from overfitted ones. For this reason, a modified Bayesian-type information criterion is developed to determine d_0. Specifically, for an arbitrary working dimension d, we define the following BIC criterion

$$G_n(d) = -\log(\mathrm{MDD}_n(Y|\hat{\mathbf{B}}_d^T\mathbf{X})^2) + C_n d, \tag{12}$$

where the second term is the penalty term, C_n is a penalty constant and

$$\mathrm{MDD}_n(Y|\hat{\mathbf{B}}_d^T\mathbf{X})^2 = \max_{\mathbf{B}^T\hat{\Sigma}_{\mathbf{x}}\mathbf{B}=\mathbf{I}_d}\{\mathrm{MDD}_n(Y|\mathbf{B}^T\mathbf{X})^2\}.$$

Then, we can estimate the structural dimension d_0 by

$$\hat{d} = \underset{1\leq d\leq p}{\arg\min}\, G_n(d).$$

The penalty constant C_n can be selected by the following fact. If $d < d_0$, we can show that $\mathrm{MDD}_n(Y|\hat{\mathbf{B}}_d^T\mathbf{X})^2 - \mathrm{MDD}_n(Y|\hat{\mathbf{B}}_{d_0}^T\mathbf{X})^2 = O_p(1)$. If $d > d_0$, then $\mathrm{MDD}_n(Y|\hat{\mathbf{B}}_d^T\mathbf{X})^2 - \mathrm{MDD}_n(Y|\hat{\mathbf{B}}_{d_0}^T\mathbf{X})^2 = O_p(\frac{1}{n})$. Thus, we have the following theoretical result for $\hat{d}$.

Theorem 2. *Assume that* $E\{|Y|^2 + \|\mathbf{X}\|_2^2\} < \infty$ *and* $\mathbf{P}_{\mathbf{B}_0}^T\mathbf{X} \perp\!\!\!\perp \mathbf{Q}_{\mathbf{B}_0}^T\mathbf{X}$*. If* C_n *satisfies* $C_n \to 0$ *and* $nC_n \to \infty$*, then we have*

$$\lim_{n\to\infty} P\{\hat{d} = d_0\} = 1.$$

In the framework of model selection, a similar BIC criterion have been studied in Shao [16] and Shi and Tsai [20]. They proved that it is able to identify the true model consistently. Selecting an optimal C_n is a challenging problem, but from our limited experience, the choice of $C_n = \frac{\log n}{n}$ seems to work well in our numerical studies. To estimate the structural dimension, we can also use the bootstrap method in Sheng and Yin [19]. Generally speaking, the bootstrap method requires more computation time than the BIC method.

4. SIMULATION STUDIES

In this section, we conduct some simulations to illustrate the performance of our proposed method and compare it with several existing CMS methods: MAVE, FMN, IHT, OLS and r-pHd, as well as DCOV. Let $\mathbf{B}_0$ is a $p \times d_0$ matrix spanning $\mathcal{S}_{E\{Y|\mathbf{X}\}}$ and $\widehat{\mathbf{B}}$ is a $p \times d_0$ matrix to estimate $\mathbf{B}_0$.

Table 1. The mean (standard deviation) of $\mathrm{dist}(\hat{\mathcal{S}}_{E\{Y|\mathbf{X}\}}, \mathcal{S}_{E\{Y|\mathbf{X}\}})$ *for Example 1.*

Settings	Method	$n = 100$	$n = 200$	$n = 400$
Design (A)	MDD	0.8011 (0.1434)	0.6743 (0.1547)	0.5219 (0.1336)
	DCOV	0.8290 (0.1407)	0.7113 (0.1657)	0.5577 (0.1483)
	MAVE	0.8450 (0.1440)	0.7020 (0.1897)	0.4624 (0.1704)
	OLS	0.9287 (0.0525)	0.9044 (0.0549)	0.8937 (0.0483)
	r-pHd	0.9323 (0.0801)	0.9235 (0.0885)	0.9124 (0.1007)
	FMN	0.8119 (0.1317)	0.6823 (0.1472)	0.5255 (0.1287)
	IHT	0.8198 (0.1236)	0.6965 (0.1376)	0.5663 (0.1369)
Design (B)	MDD	0.8986 (0.1054)	0.8475 (0.1308)	0.7695 (0.1514)
	DCOV	0.8998 (0.1063)	0.8508 (0.1401)	0.7704 (0.1559)
	MAVE	0.9283 (0.0858)	0.8986 (0.1171)	0.7954 (0.1818)
	r-pHd	0.9811 (0.0211)	0.9859 (0.0114)	0.9867 (0.0070)
	OLS	0.9558 (0.0594)	0.9642 (0.0546)	0.9808 (0.0287)
	FMN	0.9395 (0.0804)	0.9401 (0.0904)	0.9511 (0.0783)
	IHT	0.9125 (0.0931)	0.8812 (0.1220)	0.8458 (0.1325)
Design (C)	MDD	0.8765 (0.1260)	0.8264 (0.1476)	0.7366 (0.1627)
	DCOV	0.8922 (0.1189)	0.8354 (0.1460)	0.7568 (0.1679)
	MAVE	0.9227 (0.0952)	0.8806 (0.1252)	0.7664 (0.1941)
	OLS	0.9805 (0.0193)	0.9828 (0.0123)	0.9843 (0.0064)
	r-pHd	0.9414 (0.0739)	0.9456 (0.0710)	0.9490 (0.0689)
	FMN	0.9231 (0.0970)	0.8991 (0.1220)	0.8568 (0.1516)
	IHT	0.8991 (0.1117)	0.8449 (0.1399)	0.7463 (0.1538)

To evaluate the estimation accuracy of $\widehat{\mathcal{S}}_{E\{Y|\mathbf{X}\}} = \mathcal{S}(\widehat{\mathbf{B}})$, we use the following distance measure [7], defined by

$$\mathrm{dist}(\widehat{\mathcal{S}}_{E\{Y|\mathbf{X}\}}, \mathcal{S}_{E\{Y|\mathbf{X}\}}) = \|\mathrm{P}_{\widehat{\mathbf{B}}} - \mathrm{P}_{\mathbf{B}_0}\|, \tag{13}$$

where $\mathrm{P}_{\mathbf{B}}$ is the projection operator in the standard inner product of any matrix $\mathbf{B}$. The smaller value of $\mathrm{dist}(\widehat{\mathcal{S}}_{E\{Y|\mathbf{X}\}}, \mathcal{S}_{E\{Y|\mathbf{X}\}})$, the better performance of $\widehat{\mathcal{S}}_{E\{Y|\mathbf{X}\}}$.

In the first three examples, we generate the predictors $\mathbf{X} = (X_1, \cdots, X_{10})^T$ from the following three different designs for each model to cover a variety of model assumptions:

1. Design (A): the predictors follow independently the standard normal distribution;
2. Design (B): $X_2, X_{10} \sim \mathrm{Unif}(-\sqrt{3}, \sqrt{3})$, $X_4, X_9 \sim \mathrm{Exp}(1)$ and other predictors follow independently the standard normal distribution;
3. Design (C): $X_3, X_9 \sim$ Possion(1) and other predictors follow independently the standard normal distribution.

For each scenario, the error term ε has the standard normal distribution, independent of $\mathbf{X}$, $\beta_1 = (1, 1, 1, 1, 0, 0, 0, 0, 0, 0)^T/2$ and $\beta_2 = (0, 0, 0, 0, 0, 0, 1, 1, 1, 1)^T/2$. We report the mean and standard deviation of the distances $\mathrm{dist}(\widehat{\mathcal{S}}_{E\{Y|\mathbf{X}\}}, \mathcal{S}_{E\{Y|\mathbf{X}\}})$ based on 500 replicates with $n = 100, 200$ and 400 for each model.

Example 1. In this example, the data are generated from the following model

$$Y = \frac{(\beta_1^T\mathbf{X})^2}{3 + (2 + \beta_2^T\mathbf{X})^2} + 0.2\varepsilon. \tag{14}$$

This model was investigated by Zhu and Zeng [25].

It can be seen from Table 1 that MDD approach outperforms other methods across almost all scenarios. Although $\mathcal{S}_{E\{Y|\mathbf{X}\}} = \mathcal{S}_{Y|\mathbf{X}}$ in model (14), we can see that MDD is slightly superior DCOV, probably owing to the efficiency gain by focusing on the mean model. FMN has similar performance as MDD in Design (A), whereas MDD performs noticeably better than FMN in Designs (B)–(C). This might be due to that FMN depends on the normality assumption of predictors. Also, it can be seen that the performance of MAVE improves substantially when n increases from 100 to 400, which is presumably related to the nonparametric estimation involved.

Example 2. Consider the following heteroscedastic model

$$Y = \beta_1^T\mathbf{X} + 4(\beta_2^T\mathbf{X})^2\varepsilon, \tag{15}$$

where the central mean subspace $\mathcal{S}_{E\{Y|\mathbf{X}\}}$ is equal to $\mathcal{S}(\beta_1)$. This model is similar to the model in Example 2 of Zhu and Zeng [25].

As seen from Table 2, MDD and OLS always perform the best, followed by FMN, and then MAVE, whereas DCOV, r-pHd and IHT fail. It can be seen that MDD and OLS deliver comparable results in the linear model (15), which is favored by OLS. Thus, it indicates that MDD can detect the linear dependence. In most cases, especially in Designs (B)-(C), MAVE has worse performance, perhaps because MAVE may not work well in some heteroscedastic error settings.

Table 2. The mean (standard deviation) of $\mathrm{dist}(\hat{\mathcal{S}}_{E\{Y|\mathbf{X}\}}, \mathcal{S}_{E\{Y|\mathbf{X}\}})$ *for Example 2.*

Settings	Method	$n = 100$	$n = 200$	$n = 400$
Design (A)	MDD	0.8484 (0.1348)	0.8245 (0.1334)	0.7312 (0.1504)
	DCOV	0.8672 (0.1569)	0.7951 (0.2244)	0.8270 (0.2366)
	MAVE	0.9138 (0.1038)	0.9183 (0.1043)	0.8889 (0.1300)
	OLS	0.8650 (0.1220)	0.8262 (0.1281)	0.7338 (0.1436)
	r-pHd	0.9615 (0.0570)	0.9658 (0.0451)	0.9684 (0.0452)
	FMN	0.9118 (0.1013)	0.8665 (0.1195)	0.7764 (0.1506)
	IHT	0.9589 (0.0583)	0.9573 (0.0548)	0.9575 (0.0587)
Design (B)	MDD	0.9022 (0.1103)	0.8801 (0.1166)	0.8344 (0.1253)
	DCOV	0.9467 (0.0899)	0.9735 (0.0635)	0.9901 (0.0210)
	MAVE	0.9264 (0.0861)	0.9332 (0.0835)	0.9282 (0.0949)
	OLS	0.9094 (0.0981)	0.8901 (0.1051)	0.8397 (0.1159)
	r-pHd	0.9693 (0.0463)	0.9735 (0.0425)	0.9744 (0.0373)
	FMN	0.9376 (0.0841)	0.9216 (0.0930)	0.8927 (0.1199)
	IHT	0.9671 (0.0459)	0.9729 (0.0379)	0.9718 (0.0463)
Design (C)	MDD	0.9067 (0.1052)	0.8645 (0.1262)	0.8272 (0.1341)
	DCOV	0.9431 (0.0945)	0.9725 (0.0672)	0.9914 (0.0142)
	MAVE	0.9375 (0.0802)	0.9351 (0.0880)	0.9254 (0.0955)
	OLS	0.9069 (0.1034)	0.8677 (0.1217)	0.8159 (0.1292)
	r-pHd	0.9670 (0.0447)	0.9703 (0.0407)	0.9689 (0.0412)
	FMN	0.9263 (0.0869)	0.9011 (0.1094)	0.8494 (0.1349)
	IHT	0.9628 (0.0518)	0.9680 (0.0479)	0.9677 (0.0451)

Note that, although the DCOV has small means of $\mathrm{dist}(\hat{\mathcal{S}}_{E\{Y|\mathbf{X}\}}, \mathcal{S}_{E\{Y|\mathbf{X}\}})$ in Designs (A), its standard deviation is very large. Thus, it may have a bad performance. To illustrate more clearly the results, we further present the scatterplots of the square of correlation coefficients: $corr^2(\beta_1\mathbf{X}, \hat{\beta}\mathbf{X})$ versus $corr^2(\beta_2\mathbf{X}, \hat{\beta}\mathbf{X})$. See, Figure 1. Intuitively, a good and consistent estimating method would yield that its scatters of the correlation coefficients are as much as possible on the bottom right corner.

Figure 1 presents the scatterplots of the correlation coefficients for Design (A) and Design (B) in Example 2 with $n = 400$ over 500 replicates. The scatterplots for Design (C) is similar to those for Design (B), hence we omit it. It can been seen from Figure 1 that DCOV often fails to estimate the central mean subspace $\mathcal{S}(\beta_1)$. This is due to that DCOV focuses on the conditional distribution of Y given $\mathbf{X}$.

Example 3. In the example, we consider the following model with non-smooth link functions

$$Y = \frac{1}{2}\log(|\beta_1^T\mathbf{X}|) + \frac{1}{2}\mathrm{sign}(\beta_2^T\mathbf{X}) + 2\varepsilon, \tag{16}$$

where $\mathrm{sign}(v)$ is the sign function.

From Table 3, we can see that MDD and FMN demonstrate superior performance over other methods when the link functions are not smooth. Although FMN slightly outperforms MDD in Design (A), MDD is slightly superior to FMN in the non-normal cases. MAVE is inferior to MDD and FMN, which may be due to estimating inaccurately the non-smooth link functions. Moreover, we can also see that MDD and FMN perform better than DCOV, which indicates reasonably that MDD and FMN capture effectively the information about the conditional mean.

Example 4. In this example, we would like to evaluate the performance of the Bayesian-type information criterion (12) in estimating the structural dimension of $\mathcal{S}_{E(Y|\mathbf{X})}$. We consider the following two-index model:

$$Y = (\beta_1^T\mathbf{X})^2 + (\beta_2^T\mathbf{X}) + 0.1\varepsilon, \tag{17}$$

where $\beta_1 = (1, 0, 0, 0, 0)^T$ and $\beta_2 = (0, 0, 0, 0, 1)^T$. The predictors are generated from the following two cases:

Case (i): $X_j \sim N(0, 1), j = 1, 2, \cdots, 5$;
Case (ii): $X_3, X_4 \sim \mathrm{Poisson}(1)$ and $X_j \sim N(0, 1), j \neq 3, 4$.

The error term ε follows $N(0, 1)$, independent of the covariates.

Table 4 reports the frequencies of the estimated structural dimensions in 500 replications with $C_n = \log n/n$. The simulation results indicate that the BIC works reasonably and the performance becomes better as the sample size increases.

To illustrate how to effect the estimation of the structural dimension for different C_n, we consider to estimate d_0 for Example 2 with $C_n = \log n/(2n)$, $\log n/n$ and $(p - d)\log n/n$ in Table 5. From Table 5, we can see that $C_n = (p - d)\log n/n$ is more better in this case. The results also confirm that the order of C_n in Theorem 2 is reasonable.

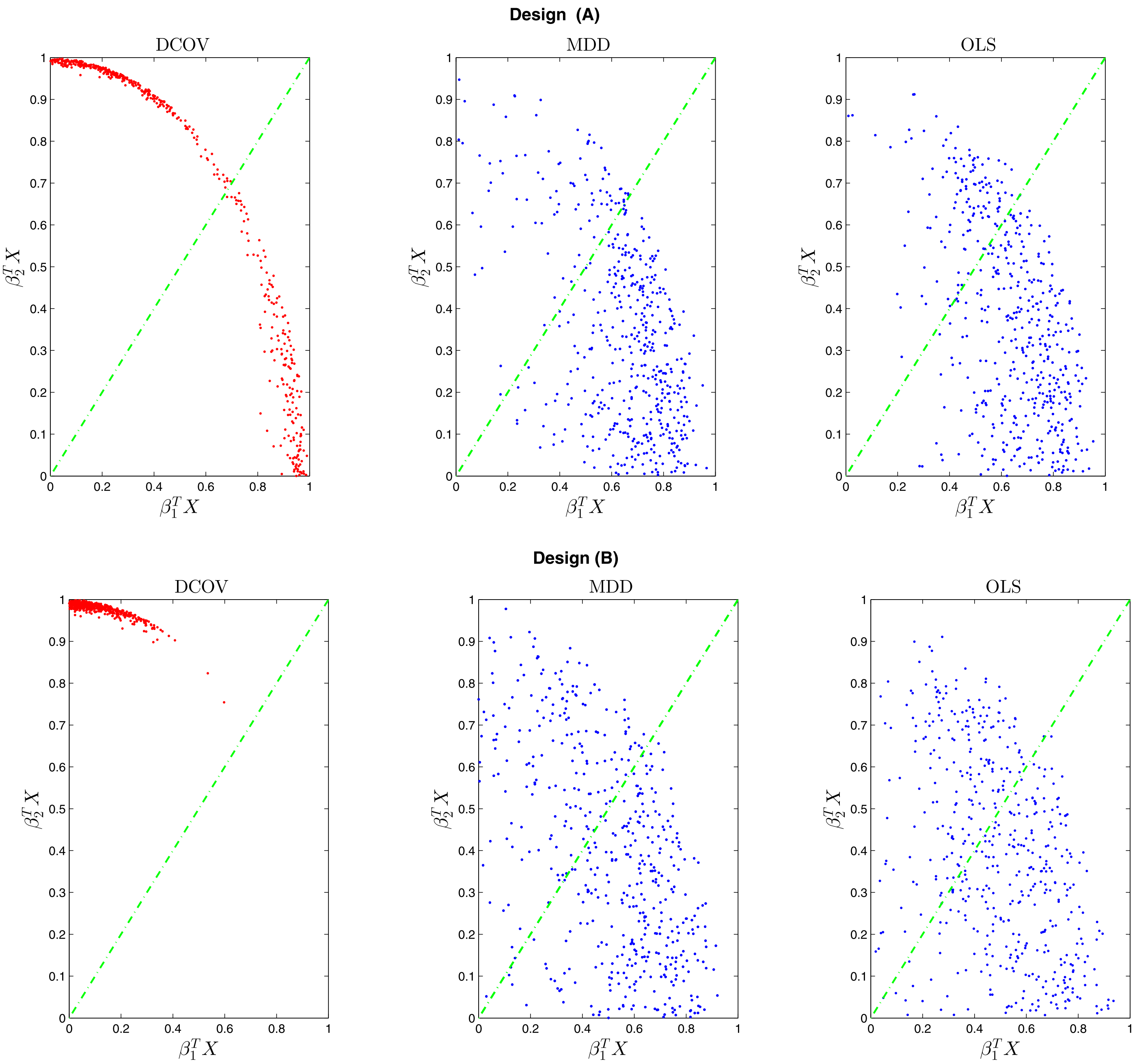

Figure 1. the square of correlation coefficients: $corr^2(\beta_1\mathbf{X}, \hat{\beta}\mathbf{X})$ versus $corr^2(\beta_2\mathbf{X}, \hat{\beta}\mathbf{X})$ for Design (A) and Design (B) in Example 2 with $n = 400$.

5. REAL DATA ANALYSIS

In this section, we apply our method to the Boston housing dataset. The data set consists of the median value of owner-occupied homes (MEDV) in 506 US census tracts in the Boston area in 1970, as well as thirteen other predictors. The dataset can be downloaded from `http://lib.stat.cmu.edu/datasets/boston`.

Among the thirteen predictors, there are two discrete variables: Charles River (CHAS) (= 1 if tract bounds river; 0 otherwise) and index of accessibility to radial highways (RAD). The other 11 predictors are respectively: crime rate (CRIM), proportion of area zoned with large lots (ZN), proportion of non-retail business acres per town (INDUS), nitric oxides concentration (NOX), average number of rooms per dwelling (RM), proportion of owner-occupied units built

Table 3. The mean (standard deviation) of $\mathrm{dist}(\hat{\mathcal{S}}_{E\{Y|\mathbf{X}\}}, \mathcal{S}_{E\{Y|\mathbf{X}\}})$ *for Example 3.*

Settings	Method	$n = 100$	$n = 200$	$n = 400$
Design (A)	MDD	0.9418 (0.0814)	0.8897 (0.1255)	0.7546 (0.1945)
	DCOV	0.9443 (0.0792)	0.9155 (0.1061)	0.8198 (0.1769)
	MAVE	0.9487 (0.0719)	0.9182 (0.1013)	0.8078 (0.1764)
	OLS	0.9165 (0.0553)	0.8939 (0.0515)	0.8763 (0.0443)
	r-pHd	0.9659 (0.0464)	0.9598 (0.0539)	0.9577 (0.0556)
	FMN	0.9464 (0.0706)	0.8747 (0.1209)	0.7370 (0.1539)
	IHT	0.9536 (0.0616)	0.9158 (0.0984)	0.8489 (0.1343)
Design (B)	MDD	0.9407 (0.0935)	0.9135 (0.0996)	0.8375 (0.1503)
	DCOV	0.9498 (0.0705)	0.9252 (0.0926)	0.8797 (0.1297)
	MAVE	0.9490 (0.0831)	0.9257 (0.0911)	0.8744 (0.1322)
	OLS	0.9381 (0.0744)	0.9365 (0.0499)	0.9350 (0.0453)
	r-pHd	0.9568 (0.0744)	0.9561 (0.0532)	0.9397 (0.0745)
	FMN	0.9488 (0.0641)	0.9245 (0.0843)	0.8481 (0.1268)
	IHT	0.9541 (0.0586)	0.9265 (0.0819)	0.8561 (0.1194)
Design (C)	MDD	0.9440 (0.0871)	0.9154 (0.0985)	0.8305 (0.1605)
	DCOV	0.9463 (0.0681)	0.9252 (0.0954)	0.8702 (0.1360)
	MAVE	0.9568 (0.0703)	0.9340 (0.0839)	0.8814 (0.1327)
	OLS	0.9413 (0.0656)	0.9322 (0.0523)	0.9326 (0.0450)
	r-pHd	0.9640 (0.0627)	0.9542 (0.0609)	0.9378 (0.0796)
	FMN	0.9466 (0.0684)	0.9180 (0.0890)	0.8439 (0.1273)
	IHT	0.9506 (0.0625)	0.9225 (0.0869)	0.8384 (0.1278)

Table 4. Frequency (%) of the estimated dimension d *for Example 4.*

Examples	Sample size	$\hat{d} = 1$	$\hat{d} = 2$	$\hat{d} \geq 3$
Case (i)	$n = 400$	0.1304	0.7600	0.1096
	$n = 600$	0.0825	0.8557	0.0618
	$n = 800$	0.0653	0.9067	0.0280
Case (ii)	$n = 400$	0.1605	0.7380	0.1015
	$n = 600$	0.1093	0.8400	0.0507
	$n = 800$	0.0560	0.9040	0.0400

prior to 1940 (AGE), weighted distances to five Boston employment centers (DIS), property tax rate (TAX), pupil-teacher ratio by town (PTRATIO), black population proportion town (B), and lower status population proportion (LSTAT).

By the BIC in (12), we use two directions to estimate the central mean subspace. To compare our method with other methods, we adopt the bootstrap method described in Sheng and Yin [19] and then calculate the distance in (13). Specifically, we first obtain the estimated subspace $\hat{\mathcal{S}}_{E\{Y|\mathbf{X}\}}$ based on all the data. Then, we randomly resample from the data to generate 500 bootstrap samples and derive the estimate of $\mathcal{S}_{E\{Y|\mathbf{X}\}}$ based on the bootstrap samples, denoted by $\hat{\mathcal{S}}^b_{E\{Y|\mathbf{X}\}}$, $b = 1, \cdots, 500$. Finally, we can obtain the distances $\mathrm{dist}(\hat{\mathcal{S}}_{E\{Y|\mathbf{X}\}}, \hat{\mathcal{S}}^b_{E\{Y|\mathbf{X}\}})$, $b = 1, \cdots, 500$. Table 6 reports the mean and standard deviation (SD) of $\mathrm{dist}(\hat{\mathcal{S}}_{E\{Y|\mathbf{X}\}}, \hat{\mathcal{S}}^b_{E\{Y|\mathbf{X}\}})$ over 500 simulations for each method. The results show that MDD has the smallest mean.

As seen from Table 6, although the mean for OLS is bigger, it is reasonable and the variability is very small. This

Table 5. Frequency (%) of the estimated dimension d *for Example 2.*

C_n	Sample size	$\hat{d} = 1$	$\hat{d} = 2$	$\hat{d} \geq 3$
$\log n/(2n)$	$n = 200$	0.8056 (0.3961)	0.1383 (0.3455)	0.0561 (0.2304)
	$n = 400$	0.8136 (0.3898)	0.1623 (0.3691)	0.0240 (0.1534)
$\log n/n$	$n = 200$	0.8397 (0.3673)	0.1263 (0.3325)	0.0341 (0.1816)
	$n = 400$	0.8437 (0.3635)	0.1363 (0.3434)	0.0200 (0.1403)
$(p-d)\log n/n$	$n = 200$	0.8938 (0.3084)	0.0882 (0.2838)	0.0180 (0.1332)
	$n = 400$	0.9038 (0.2952)	0.0782 (0.2687)	0.0180 (0.1235)

Table 6. Summary of $\mathrm{dist}(\hat{\mathcal{S}}_{E\{Y|\mathbf{X}\}}, \hat{\mathcal{S}}^b_{E\{Y|\mathbf{X}\}})$ *for the Boston housing data.*

Method	MDD	DCOV	MAVE	OLS	r-pHd	FMN	IHT
Mean	0.2972	0.3673	0.3675	0.4154	0.9859	0.3664	0.4937
SD	0.3191	0.2699	0.2883	0.0925	0.0269	0.2221	0.2269

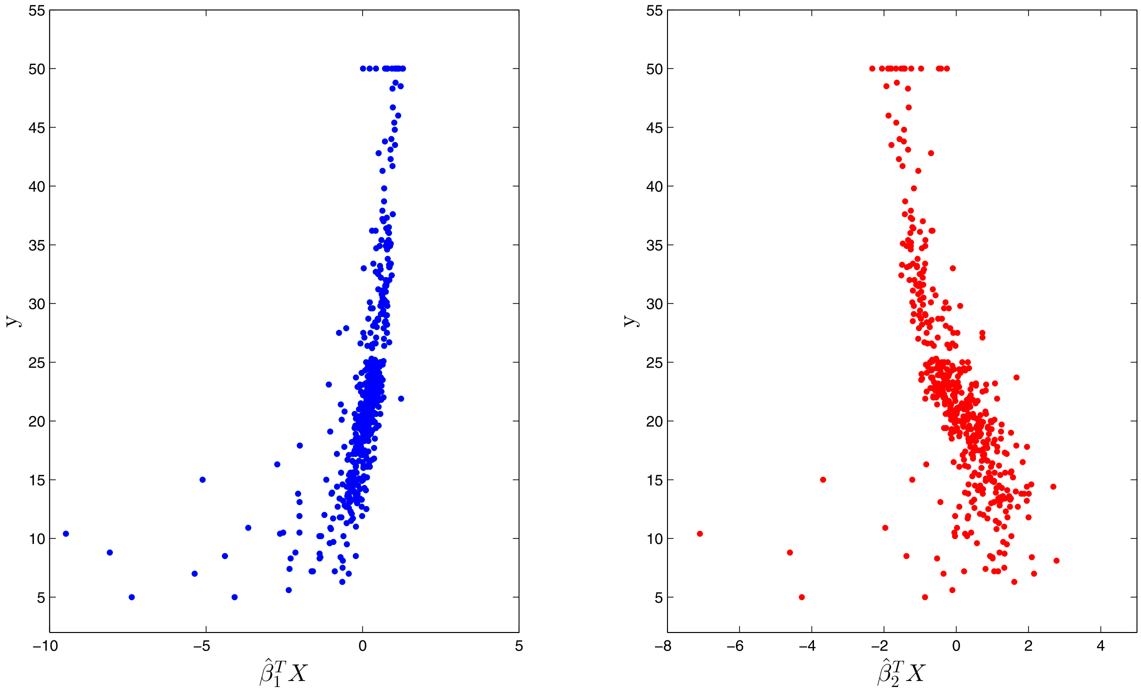

Figure 2. The response Y plotted against the first two MDD directions for the Boston housing data.

suggests that there may exist strong linear patterns between the response and the first two directions. To illustrate the relations, we present the scatterplots of the response versus the first two MDD directions in Figure 2. This results indicate the obvious linear trends between Y and $\hat{\beta}_1\mathbf{X}$ or $\hat{\beta}_2\mathbf{X}$.

6. CONCLUSIONS

This paper proposes a model-free SDR method based on the martingale difference divergence to recover the central mean subspace. This method enjoys many merits, including model free property and consistent in non-normal and discrete distributions. Under mild conditions, we establish asymptotic properties of the proposed estimators. A Bayesian-type information criterion is proposed to determine the structural dimension. However, how to choose an optimal C_n for the BIC is a challenging and open problem. The simulation studies and real data example show that MDD is able to extract efficiently information about the conditional mean dependence of the response on predictors

APPENDIX

Proof of Proposition 1. Let $\mathcal{S}^{\perp}(\mathbf{B}_0)$ be the orthogonal complement space of $\mathcal{S}(\mathbf{B}_0)$ and $\mathbf{B}_{0\perp} \in \mathbb{R}^{p\times(p-d_0)}$ be a basis matrix of $\mathcal{S}^{\perp}(\mathbf{B}_0)$ with $(\mathbf{B}_{0\perp})^T\Sigma_{\mathbf{x}}\mathbf{B}_{0\perp} = \mathbf{I}_{p-d_0}$ and $\mathbf{B}_0^T\Sigma_{\mathbf{x}}\mathbf{B}_{0\perp} = \mathbf{0}$.

For any $\mathbf{A} \in \mathbb{R}^{p\times d_0}$ with $\mathbf{A}^T\Sigma_{\mathbf{x}}\mathbf{A} = \mathbf{I}_{d_0}$, there exist two matrices $\mathbf{C}_1 \in \mathbb{R}^{d_0\times d_0}$ and $\mathbf{C}_2 \in \mathbb{R}^{(p-d_0)\times d_0}$, such that

$$\mathbf{A} = \mathbf{B}_0\mathbf{C}_1 + \mathbf{B}_{0\perp}\mathbf{C}_2.$$

Moreover, $\mathbf{C}_1^T\mathbf{C}_1 + \mathbf{C}_2^T\mathbf{C}_2 = \mathbf{I}_{d_0}$. Consider

(A.1)

$$\begin{aligned}
&\mathrm{MDD}(Y|\mathbf{A}^T\mathbf{X})^2\\
&\quad= \int_{\mathcal{R}^{d_0}} |E\{Ye^{i<\mathbf{s},\mathbf{A}^T\mathbf{X}>}\} - E\{Y\}E\{e^{i<\mathbf{s},\mathbf{A}^T\mathbf{X}>}\}|^2\omega(\mathbf{s})d\mathbf{s}\\
&\quad= \int_{\mathcal{R}^{d_0}} |E\{E\{Y|\mathbf{X}\}e^{i<\mathbf{s},\mathbf{A}^T\mathbf{X}>}\}\\
&\qquad - E\{Y\}E\{e^{i<\mathbf{s},\mathbf{A}^T\mathbf{X}>}\}|^2\omega(\mathbf{s})d\mathbf{s}\\
&\quad= \int_{\mathcal{R}^{d_0}} |E\{E\{Y|\mathbf{B}_0^T\mathbf{X}\}e^{i<\mathbf{s},\mathbf{X}^T(\mathbf{B}_0\mathbf{C}_1+\mathbf{B}_{0\perp}\mathbf{C}_2)>}\}
\end{aligned}$$

$$
\begin{aligned}
&- E\{Y\}E\{e^{i<\mathbf{s},\mathbf{X}^T(\mathbf{B}_0\mathbf{C}_1+\mathbf{B}_{0\perp}\mathbf{C}_2)>}\}|^2\omega(\mathbf{s})d\mathbf{s}\\
&= \int_{\mathcal{R}^{d_0}} |E\{e^{i<\mathbf{s},\mathbf{X}^T\mathbf{B}_{0\perp}\mathbf{C}_2>}\}|^2|E\{Ye^{i<\mathbf{s},\mathbf{X}^T\mathbf{B}_0\mathbf{C}_1>}\}\\
&\quad - E\{Y\}E\{e^{i<\mathbf{s},\mathbf{X}^T\mathbf{B}_0\mathbf{C}_1>}\}|^2\omega(\mathbf{s})d\mathbf{s}\\
&\le \int_{\mathcal{R}^{d_0}} |E\{Ye^{i<\mathbf{s},\mathbf{X}^T\mathbf{B}_0\mathbf{C}_1>}\}\\
&\quad - E\{Y\}E\{e^{i<\mathbf{s},\mathbf{X}^T\mathbf{B}_0\mathbf{C}_1>}\}|^2\omega(\mathbf{s})d\mathbf{s}\\
&= \mathrm{MDD}(Y|\mathbf{C}_1^T\mathbf{B}_0^T\mathbf{X})^2,
\end{aligned}
$$

where the third equality follows from model (2), which implies $E\{Y|\mathbf{X}\} = E\{Y|\mathbf{B}_0^T\mathbf{X}\}$, the fourth inequality follows the assumption $\mathbf{P}_{\mathbf{B}_0}^T\mathbf{X}\perp\!\!\!\perp\mathbf{Q}_{\mathbf{B}_0}^T\mathbf{X}$.

Consider the singular-value decomposition of $\mathbf{C}_1$, given by

$$\text{(A.2)} \qquad \mathbf{C}_1 = \mathbf{U}\Sigma\mathbf{V}^T,$$

where $\mathbf{U}$ and $\mathbf{V}$ are $d_0\times d_0$ orthogonal matrices and $\Sigma = \mathrm{diag}\{\lambda_1,\cdots,\lambda_{d_0}\}$. Note that $\mathbf{C}_1^T\mathbf{C}_1+\mathbf{C}_2^T\mathbf{C}_2 = \mathbf{I}_{d_0}$. Thus, $0\le\lambda_j\le 1$, for $j=1,2,\cdots,d_0$. Moreover, $\lambda_j = 1$ holds for all $j=1,2,\cdots,d_0$, if and only if $\mathbf{C}_2^T\mathbf{C}_2 = \mathbf{0}$. This, together with (4) and (A.2), yields that

$$
\begin{aligned}
&\mathrm{MDD}(Y|\mathbf{C}_1^T\mathbf{B}_0^T\mathbf{X})^2\\
&= -E\{[Y-E\{Y\}][Y'-E\{Y'\}]\|\mathbf{C}_1^T\mathbf{B}_0^T[\mathbf{X}-\mathbf{X}']\|\}\\
&= -E\Big\{[Y-E\{Y\}][Y'-E\{Y'\}]\\
&\quad\times\Big([\mathbf{X}-\mathbf{X}']^T\mathbf{B}_0\mathbf{U}\,\mathrm{diag}\{\lambda_1^2,\cdots,\lambda_{d_0}^2\}\mathbf{U}^T\mathbf{B}_0^T[\mathbf{X}-\mathbf{X}']\Big)^{1/2}\Big\}\\
&\le -E\Big\{[Y-E\{Y\}][Y'-E\{Y'\}]\\
&\quad\times\Big([\mathbf{X}-\mathbf{X}']^T\mathbf{B}_0\mathbf{B}_0^T[\mathbf{X}-\mathbf{X}']\Big)^{1/2}\Big\}\\
&= \mathrm{MDD}(Y|\mathbf{B}_0^T\mathbf{X})^2.
\end{aligned}
$$

Note that the equality holds if and only if $\lambda_j = 1$ for all $j=1,\cdots,d_0$, namely, $\mathbf{C}_1^T\mathbf{C}_1 = \mathbf{I}_{d_0}$ and thus $\mathbf{C}_2 = \mathbf{0}$. Combining this with (A.1), we obtain that the maximum of $\mathrm{MDD}(Y|\mathbf{A}^T\mathbf{X})^2$ is achieved if only and if $\mathbf{A} = \mathbf{B}_0\mathbf{C}_1$ with $\mathbf{C}_1^T\mathbf{C}_1 = \mathbf{I}_{d_0}$. □

In order to prove Theorem 1, a technique based on manifold theories will be used. This method is similar to that used by Chen et al. [1]. We first introduce some notation. Denote the Stiefel manifold $\mathrm{St}(p,d)$ as $\mathrm{St}(p,d) = \{\mathbf{\Gamma}\in\mathbb{R}^{p\times d}:\mathbf{\Gamma}^T\mathbf{\Gamma} = \mathbf{I}_d\}$. Define the projection operator $R:\mathbb{R}^{p\times d}\to\mathrm{St}(p,d)$ onto the manifold $\mathrm{St}(p,d)$ as

$$R(\mathbf{\Gamma}) = \underset{\mathbf{W}\in\mathrm{St}(p,d)}{\arg\min}\ \|\mathbf{\Gamma}-\mathbf{W}\|^2.$$

The tangent space $T_{\mathbf{\Gamma}}(p,d)$ at $\mathbf{\Gamma}\in\mathrm{St}(p,d)$ is

$$
\begin{aligned}
&T_{\mathbf{\Gamma}}(p,d) = \{\mathbf{Z}\in\mathbb{R}^{p\times d}:\\
&\quad\mathbf{Z} = \mathbf{\Gamma}\mathbf{A}+\mathbf{\Gamma}_\perp\mathbf{B}, \mathbf{A}\in\mathbb{R}^{d\times d}, \mathbf{A}+\mathbf{A}^T = 0, \mathbf{B}\in\mathbb{R}^{(p-d)\times d}\},
\end{aligned}
$$

where $\mathbf{\Gamma}_\perp\in\mathbb{R}^{p\times(p-d)}$ satisfies $[\mathbf{\Gamma},\mathbf{\Gamma}_\perp]^T[\mathbf{\Gamma},\mathbf{\Gamma}_\perp] = \mathbf{I}_p$.

Note that $\mathbf{\Gamma}\in\mathrm{St}(p,d)$ implies $\mathcal{S}(\mathbf{\Gamma})\in\mathrm{Gr}(p,d)$, where $\mathrm{Gr}(p,d)$ stands for the Grassmann manifold. Note that $L_n(\mathbf{\Gamma})$ satisfies $L_n(\mathbf{\Gamma}) = L_n(\mathbf{\Gamma}\mathbf{Q})$ for any orthogonal matrix $\mathbf{Q}\in\mathbb{R}^{d\times d}$. Thus, $L_n(\mathbf{\Gamma})$ should be minimized on $\mathrm{Gr}(p,d)$ rather than on $\mathrm{St}(p,d)$.

To prove Theorem 1, the following lemma is needed. See, Chen et al. [1].

Lemma 1. *Assume that* $\mathbf{Z}\in T_{\mathbf{\Gamma}}(p,d)$ *and* $\mathbf{\Gamma}\in\mathrm{St}(p,d)$, *then we have*

(i) $\mathrm{tr}(\mathbf{Z}^T\mathbf{\Gamma}\mathbf{C}) = 0$ *for any symmetric matrix* $\mathbf{C}\in\mathbb{R}^{d\times d}$*;*
(ii) $R(\mathbf{\Gamma}+t\mathbf{Z}) = \mathbf{\Gamma}+t\mathbf{Z}-\frac{1}{2}t^2\mathbf{\Gamma}\mathbf{Z}^T\mathbf{Z}+O(t^3)$.

Let $\mathbf{\Gamma} = \hat{\Sigma}_{\mathbf{x}}^{1/2}\mathbf{B}$ and $\mathbf{\Gamma}_0 = \Sigma_{\mathbf{x}}^{1/2}\mathbf{B}_0$. Thus, by Proposition 1, we have

$$\text{(A.3)}\quad \mathbf{\Gamma}_0 = \underset{\mathbf{\Gamma}^T\mathbf{\Gamma}=\mathbf{I}_{d_0}}{\arg\min}\{L(\mathbf{\Gamma})\} = \underset{\mathbf{\Gamma}^T\mathbf{\Gamma}=\mathbf{I}_{d_0}}{\arg\min}\{-\mathrm{MDD}(Y|\mathbf{\Gamma}^T\Sigma_{\mathbf{x}}^{-1/2}\mathbf{X})^2\},$$

where $L(\mathbf{\Gamma}) = -\mathrm{MDD}(Y|\mathbf{\Gamma}^T\Sigma_{\mathbf{x}}^{-1/2}\mathbf{X})^2$. Maximizing (9) is equivalent to search for a local minimizer $\hat{\mathbf{\Gamma}}$ of $L_n(\mathbf{\Gamma})$, given by

$$\text{(A.4)}\qquad L_n(\mathbf{\Gamma}) = -\mathrm{MDD}_n(Y|\mathbf{\Gamma}^T\hat{\Sigma}_{\mathbf{x}}^{-1/2}\mathbf{X})^2,$$

subject to $\mathbf{\Gamma}^T\mathbf{\Gamma} = \mathbf{I}_{d_0}$.

Proof of Theorem 1. To solve the Grassmann manifold optimization problem, it is necessary to define the concept of the neighborhood of $\mathcal{S}(\mathbf{\Gamma}_0)$. For an arbitrary matrix $\mathbf{W}\in\mathbb{R}^{p\times d_0}$ and scaler $\delta\in\mathbb{R}$, the perturbed point around $\mathcal{S}(\mathbf{\Gamma}_0)$ in Grassmann manifold can be expressed by $\mathcal{S}(R(\mathbf{\Gamma}_0+\delta\mathbf{W}))$. It follows from Chen et al. [1] that $\mathcal{S}(R(\mathbf{\Gamma}_0+\delta\mathbf{W})) = \mathcal{S}(R(\mathbf{\Gamma}_0+\delta\mathbf{\Gamma}_{0\perp}\mathbf{B}))$, where $\mathbf{B}\in\mathbb{R}^{(p-d_0)\times d_0}$ depends on $\mathbf{W}$. This implies that the movement from $\mathcal{S}(\mathbf{\Gamma}_0)$ in the near neighborhood only depends on the $\mathbf{\Gamma}_{0\perp}\mathbf{B}$. Furthermore, as suggested by Chen et al. [1], we can consider perturbed points like $R(\mathbf{\Gamma}_0+\delta\mathbf{Z})$ in the following proofs, where $\mathbf{Z}\in T_{\mathbf{\Gamma}_0}(p,d_0)$ satisfies $\|\mathbf{B}\|_F = C$ for some given C.

For any given $\varepsilon>0$, we will show that there exists a constant C such that

$$\text{(A.5)}\quad P\left\{\inf_{\mathbf{Z}\in T_{\mathbf{\Gamma}_0}(p,d_0):\|\mathbf{B}\|_F=C} L_n(R(\mathbf{\Gamma}_0+n^{-1/2}\mathbf{Z}))>L_n(\mathbf{\Gamma}_0)\right\}\ge 1-\varepsilon.$$

This implies with probability at least $1-\varepsilon$ that there exists a local minimizer $\hat{\mathbf{\Gamma}}$ of $L_n(\mathbf{\Gamma})$, such that $\|\hat{\mathbf{\Gamma}}-\mathbf{\Gamma}_0\|_F = O_p(n^{-1/2})$, which implies that $\|\widehat{\mathbf{B}}_n\widehat{\mathbf{B}}_n^T-\mathbf{B}_0\mathbf{B}_0^T\|_F = O_p(n^{-1/2})$. By Lemma 1 and Taylor expansion, we obtain

$$
\begin{aligned}
&\text{(A.6)}\\
&n\{L_n(R(\mathbf{\Gamma}_0+n^{-1/2}\mathbf{Z}))-L_n(\mathbf{\Gamma}_0)\}\\
&\quad = n\Big\{L_n(\mathbf{\Gamma}_0+n^{-1/2}\mathbf{Z}-\frac{1}{2}n^{-1}\mathbf{\Gamma}_0\mathbf{Z}^T\mathbf{Z})-L_n(\mathbf{\Gamma}_0)\Big\}\\
&\qquad\times\{1+o_p(1)\}
\end{aligned}
$$

$$= \Big\{ n^{1/2}\mathrm{tr}(\mathbf{Z}^T L_n'(\boldsymbol{\Gamma}_0)) - \frac{1}{2}\mathrm{tr}(\mathbf{Z}^T\mathbf{Z}\boldsymbol{\Gamma}_0^T L_n'(\boldsymbol{\Gamma}_0))\Big\} \\ \times \{1+o_p(1)\}.$$

Next, we consider the two terms in the right-hand side of (A.6). To study the order of $n^{1/2}\mathrm{tr}(\mathbf{Z}^T L_n'(\boldsymbol{\Gamma}_0))$, we first consider the following Lagrange function for the optimization problem (A.3)

$$\ell(\boldsymbol{\Gamma}, \Lambda) = L(\boldsymbol{\Gamma}) + \mathrm{tr}(\Lambda(\boldsymbol{\Gamma}^T\boldsymbol{\Gamma} - \mathbf{I}_{d_0})),$$

where $\Lambda \in \mathbb{R}^{d_0\times d_0}$ is the Lagrange multiplier. By Proposition 1, Lagrange multiplier technique suggests that there exists a symmetric matrix $\Lambda_0 \in \mathbb{R}^{d_0\times d_0}$, such that

$$L'(\boldsymbol{\Gamma}_0) + 2\boldsymbol{\Gamma}_0\Lambda_0 = 0.$$

By the similar argument of Proposition 3.4 in Park et al. [15], we can obtain that

$$\|E\{L_n'(\boldsymbol{\Gamma}_0)\} - L'(\boldsymbol{\Gamma}_0)\| = O(n^{-1}).$$

This, together with Lemma 1, yields that

(A.7)
$$\begin{aligned}
&n^{1/2}\mathrm{tr}(\mathbf{Z}^T L_n'(\boldsymbol{\Gamma}_0))\\
&\quad= n^{1/2}\mathrm{tr}(\mathbf{Z}^T[L_n'(\boldsymbol{\Gamma}_0) - E\{L_n'(\boldsymbol{\Gamma}_0)\}]) + n^{1/2}\mathrm{tr}(\mathbf{Z}^T L'(\boldsymbol{\Gamma}_0))\\
&\qquad+ n^{1/2}\mathrm{tr}(\mathbf{Z}^T[E\{L_n'(\boldsymbol{\Gamma}_0)\} - L'(\boldsymbol{\Gamma}_0)])\\
&\quad= n^{1/2}\mathrm{tr}(\mathbf{Z}^T[L_n'(\boldsymbol{\Gamma}_0) - E\{L_n'(\boldsymbol{\Gamma}_0)\}]) + n^{1/2}\mathrm{tr}(\mathbf{Z}^T L'(\boldsymbol{\Gamma}_0))\\
&\qquad+ O_p(n^{-1/2})\\
&\quad= n^{1/2}\mathrm{tr}(\mathbf{Z}^T[L_n'(\boldsymbol{\Gamma}_0) - E\{L_n'(\boldsymbol{\Gamma}_0)\}]) + 2n^{1/2}\mathrm{tr}(\mathbf{Z}^T\boldsymbol{\Gamma}_0\Lambda_0)\\
&\qquad+ O_p(n^{-1/2})\\
&\quad= n^{1/2}\mathrm{tr}(\mathbf{A}^T\boldsymbol{\Gamma}_0^T[L_n'(\boldsymbol{\Gamma}_0) - E\{L_n'(\boldsymbol{\Gamma}_0)\}])\\
&\qquad+ n^{1/2}\mathrm{tr}(\mathbf{B}^T\boldsymbol{\Gamma}_{0\perp}^T[L_n'(\boldsymbol{\Gamma}_0) - E\{L_n'(\boldsymbol{\Gamma}_0)\}])\\
&\qquad+ O_p(n^{-1/2})\\
&\quad= O_P(\|\mathbf{B}\|_F \times \|n^{1/2}(L_n'(\boldsymbol{\Gamma}_0) - E\{L_n'(\boldsymbol{\Gamma}_0)\})\|_F),
\end{aligned}$$

where $\mathbf{Z} = \boldsymbol{\Gamma}_0\mathbf{A} + \boldsymbol{\Gamma}_{0\perp}\mathbf{B}$. The last equality follows from $\mathrm{tr}(\mathbf{A}^T\boldsymbol{\Gamma}_0^T[L_n'(\boldsymbol{\Gamma}_0) - E\{L_n'(\boldsymbol{\Gamma}_0)\}]) = 0$, because $\boldsymbol{\Gamma}_0^T[L_n'(\boldsymbol{\Gamma}_0) - E\{L_n'(\boldsymbol{\Gamma}_0)\}]$ is symmetric and $\mathbf{A}$ is skew-symmetric.

By the Law of Large Numbers, we have

$$\begin{aligned}
\|L_n'(\boldsymbol{\Gamma}_0) - L'(\boldsymbol{\Gamma}_0)\|_F &\le \|L_n'(\boldsymbol{\Gamma}_0) - E\{L_n'(\boldsymbol{\Gamma}_0)\}\|_F\\
&\quad+ \|E\{L_n'(\boldsymbol{\Gamma}_0)\} - L'(\boldsymbol{\Gamma}_0)\|_F\\
&= o_p(1),
\end{aligned}$$

where

$$L'(\boldsymbol{\Gamma}_0) = \Sigma_{\mathbf{x}}^{-1/2}E\Big\{(Y - E[Y])(Y' - E[Y']) \\ \times \frac{(\mathbf{X}-\mathbf{X}')(\mathbf{X}-\mathbf{X}')^T}{|\boldsymbol{\Gamma}_0^T\Sigma_{\mathbf{x}}^{-1/2}(\mathbf{X}-\mathbf{X}')|}\Big\}\Sigma_{\mathbf{x}}^{-1/2}\boldsymbol{\Gamma}_0$$

and $(\mathbf{X}', Y')$ is independent of $(\mathbf{X}, Y)$. Following the argument in the proof of Theorem 1 in Shao and Zhang [17], it can be shown that

$$-\boldsymbol{\Gamma}_0^T L'(\boldsymbol{\Gamma}_0) = -\boldsymbol{\Gamma}_0^T\Sigma_{\mathbf{x}}^{-1/2}E\Big\{(Y - E[Y])(Y' - E[Y']) \\ \times \frac{(\mathbf{X}-\mathbf{X}')(\mathbf{X}-\mathbf{X}')^T}{|\boldsymbol{\Gamma}_0^T\Sigma_{\mathbf{x}}^{-1/2}(\mathbf{X}-\mathbf{X}')|}\Big\}\Sigma_{\mathbf{x}}^{-1/2}\boldsymbol{\Gamma}_0$$

is strictly positive definite matrix. Then, we have

(A.8)
$$\begin{aligned}
&-\frac{1}{2}\mathrm{tr}(\mathbf{Z}^T\mathbf{Z}\boldsymbol{\Gamma}_0^T L_n'(\boldsymbol{\Gamma}_0))\\
&\quad= -\frac{1}{2}\mathrm{tr}(\mathbf{Z}^T\mathbf{Z}\boldsymbol{\Gamma}_0^T L'(\boldsymbol{\Gamma}_0))\{1+o_p(1)\}\\
&\quad= \frac{1}{2}\Big\{-\mathrm{tr}(\mathbf{A}^T\mathbf{A}\boldsymbol{\Gamma}_0^T L'(\boldsymbol{\Gamma}_0)) - \mathrm{tr}(\mathbf{B}^T\mathbf{B}\boldsymbol{\Gamma}_0^T L'(\boldsymbol{\Gamma}_0))\Big\}\\
&\qquad\times\{1+o_p(1)\}\\
&\quad\ge \frac{1}{2}\lambda_{\min}\{-\boldsymbol{\Gamma}_0^T L'(\boldsymbol{\Gamma}_0)\}\|\mathbf{B}\|_F^2 + o_p(1),
\end{aligned}$$

where $\lambda_{\min}\{-\boldsymbol{\Gamma}_0^T L'(\boldsymbol{\Gamma}_0)\}(> 0)$ is the smallest eigenvalue of $-\boldsymbol{\Gamma}_0^T L'(\boldsymbol{\Gamma}_0)$.

By (A.7)–(A.8) and choosing a sufficiently large C, the first term in the right-hand side of (A.6) is dominated by the second term, which is positive. Hence, for the sufficiently large C, (A.5) holds. This completes the proof. □

Proof of Proposition 2. For $\mathbf{A} \in \mathbb{R}^{p\times d}$ with $\mathbf{A}^T\Sigma_{\mathbf{x}}\mathbf{A} = \mathbf{I}_d$, by a similar decomposition of $\mathbf{A}$ in the proof of Proposition 1, we have

(A.9)
$$\mathbf{A} = \mathbf{B}_0\mathbf{C}_1 + \mathbf{B}_{0\perp}\mathbf{C}_2,$$

with $\mathbf{C}_1 \in \mathbb{R}^{d_0\times d}$ and $\mathbf{C}_2 \in \mathbb{R}^{(p-d_0)\times d}$. Then, $\mathbf{C}_1^T\mathbf{C}_1 + \mathbf{C}_2^T\mathbf{C}_2 = \mathbf{I}_d$. Under the condition $\mathbf{P}_{\mathbf{B}_0}^T\mathbf{X} \perp\!\!\!\perp \mathbf{Q}_{\mathbf{B}_0}^T\mathbf{X}$, we have

(A.10)
$$\begin{aligned}
&\mathrm{MDD}(Y|\mathbf{A}^T\mathbf{X})^2\\
&\quad= \int_{\mathcal{R}^{d_0}} |E\{e^{i<\mathbf{s},\mathbf{X}^T\mathbf{B}_{0\perp}\mathbf{C}_2>}\}|^2|E\{Ye^{i<\mathbf{s},\mathbf{X}^T\mathbf{B}_0\mathbf{C}_1>}\}\\
&\qquad- E\{Y\}E\{e^{i<\mathbf{s},\mathbf{X}^T\mathbf{B}_0\mathbf{C}_1>}\}|^2\omega(\mathbf{s})d\mathbf{s}\\
&\quad\le \int_{\mathcal{R}^{d_0}} |E\{Ye^{i<\mathbf{s},\mathbf{X}^T\mathbf{B}_0\mathbf{C}_1>}\}\\
&\qquad- E\{Y\}E\{e^{i<\mathbf{s},\mathbf{X}^T\mathbf{B}_0\mathbf{C}_1>}\}|^2\omega(\mathbf{s})d\mathbf{s}\\
&\quad= \mathrm{MDD}(Y|\mathbf{C}_1^T\mathbf{B}_0^T\mathbf{X})^2.
\end{aligned}$$

(i) We first prove that (11) holds for any $d < d_0$. By the singular-value decomposition theorem, there exist matrices $\mathbf{U} \in \mathbb{R}^{d_0\times d_0}$ and $\mathbf{V} \in \mathbb{R}^{d\times d}$ such that

$$\mathbf{C}_1 = \mathbf{U}\Sigma\mathbf{V}^T, \quad \text{with } \Sigma = \begin{pmatrix} \mathrm{diag}\{\lambda_1, \cdots, \lambda_d\} \\ \mathbf{0}_{(d_0-d)\times d} \end{pmatrix} \in \mathbb{R}^{d_0\times d}.$$

Then, we have $\mathbf{C}_1\mathbf{C}_1^T = \mathbf{U}\Sigma\Sigma^T\mathbf{U}^T$. Since $\mathbf{C}_1^T\mathbf{C}_1 + \mathbf{C}_2^T\mathbf{C}_2 = \mathbf{I}_d$, we obtain $0 \le \lambda_j \le 1$, $j = 1, 2, \cdots, d$. This leads to

$$\begin{aligned}
&\mathrm{MDD}(Y|\mathbf{C}_1^T\mathbf{B}_0^T\mathbf{X})^2\\
&= -E\{[Y - E\{Y\}][Y' - E\{Y'\}]\|\mathbf{C}_1^T\mathbf{B}_0^T[\mathbf{X} - \mathbf{X}']\|\}\\
&= -E\Big\{[Y - E\{Y\}][Y' - E\{Y'\}]\Big([\mathbf{X} - \mathbf{X}']^T\mathbf{B}_0\mathbf{U}\\
&\quad \times \mathrm{diag}\{\lambda_1^2, \cdots, \lambda_d^2, \underbrace{0, \cdots, 0}_{d_0 - d}\}\mathbf{U}^T\mathbf{B}_0^T[\mathbf{X} - \mathbf{X}']\Big)^{\frac{1}{2}}\Big\}\\
&\le -E\Big\{[Y - E\{Y\}][Y' - E\{Y'\}]\\
&\quad \times \Big([\mathbf{X} - \mathbf{X}']^T\mathbf{B}_{01}\mathbf{B}_{01}^T[\mathbf{X} - \mathbf{X}']\Big)^{\frac{1}{2}}\Big\}\\
&< -E\Big\{[Y - E\{Y\}][Y' - E\{Y'\}]\\
&\quad \times \Big([\mathbf{X} - \mathbf{X}']^T\mathbf{B}_0\mathbf{B}_0^T[\mathbf{X} - \mathbf{X}']\Big)^{\frac{1}{2}}\Big\}\\
&= \mathrm{MDD}(Y|\mathbf{B}_0^T\mathbf{X})^2,
\end{aligned} \tag{A.11}$$

where $\mathbf{B}_0 = [\mathbf{B}_{01}, \mathbf{B}_{02}]$ with $\mathbf{B}_{01} \in \mathbb{R}^{p\times d}$ and $\mathbf{B}_{02} \in \mathbb{R}^{p\times(d_0-d)}$. Then, combining (A.10) and (A.11), we have $\mathrm{MDD}(Y|\mathbf{A}^T\mathbf{X})^2 < \mathrm{MDD}(Y|\mathbf{B}_0^T\mathbf{X})^2$ for any $d < d_0$.

(ii) We now prove that (11) holds in the setting of $d > d_0$ and $\mathcal{S}(\mathbf{B}_0) \nsubseteq \mathcal{S}(\mathbf{A})$. By the singular-value decomposition of $\mathbf{C}_1$ in (A.9), we have

$$\mathbf{C}_1 = \mathbf{U}\Sigma\mathbf{V}^T,$$

where $\mathbf{U} \in \mathbb{R}^{d_0\times d_0}$, $\mathbf{V} \in \mathbb{R}^{d\times d}$ and $\Sigma = (\mathrm{diag}\{\lambda_1, \cdots, \lambda_{d_0}\}, \mathbf{0}_{d_0\times(d-d_0)}) \in \mathbb{R}^{d_0\times d}$ with $\lambda_j \ge 0$, $j = 1, 2, \cdots, d_0$. Then, we obtain that

$$\mathbf{C}_1\mathbf{C}_1^T = \mathbf{U}\,\mathrm{diag}\{\lambda_1^2, \cdots, \lambda_{d_0}^2\}\mathbf{U}^T. \tag{A.12}$$

When $d > d_0$ and $\mathcal{S}(\mathbf{B}_0) \nsubseteq \mathcal{S}(\mathbf{A})$. In this setting, we have $\mathrm{rank}(\mathbf{C}_1) < d_0$. Thus, there exists at least one singular value, denoted by λ_{j^*}, such that $\lambda_{j^*} = 0$. In addition, it follows from $\mathbf{C}_1^T\mathbf{C}_1 + \mathbf{C}_2^T\mathbf{C}_2 = \mathbf{I}_d$ that $0 \le \lambda_j \le 1$, $j = 1, 2, \cdots, d_0$. This, together with (A.10)–(A.12), yields that

$$\begin{aligned}
&\mathrm{MDD}(Y|\mathbf{C}_1^T\mathbf{B}_0^T\mathbf{X})^2\\
&= -E\{[Y - E\{Y\}][Y' - E\{Y'\}]\|\mathbf{C}_1^T\mathbf{B}_0^T[\mathbf{X} - \mathbf{X}']\|\}\\
&= -E\Big\{[Y - E\{Y\}][Y' - E\{Y'\}]\Big([\mathbf{X} - \mathbf{X}']^T\mathbf{B}_0\mathbf{U}\\
&\quad \times \mathrm{diag}\{\lambda_1^2, \cdots, \lambda_{j^*}^2, \cdots, \lambda_{d_0}^2\}\mathbf{U}^T\mathbf{B}_0^T[\mathbf{X} - \mathbf{X}']\Big)^{\frac{1}{2}}\Big\}\\
&< -E\Big\{[Y - E\{Y\}][Y' - E\{Y'\}]\\
&\quad \times \Big([\mathbf{X} - \mathbf{X}']^T\mathbf{B}_0\mathbf{B}_0^T[\mathbf{X} - \mathbf{X}']\Big)^{\frac{1}{2}}\Big\}\\
&= \mathrm{MDD}(Y|\mathbf{B}_0^T\mathbf{X})^2.
\end{aligned} \tag{A.13}$$

Thus, combining (A.10) with (A.13), we can obtain that $\mathrm{MDD}(Y|\mathbf{A}^T\mathbf{X})^2 < \mathrm{MDD}(Y|\mathbf{B}_0^T\mathbf{X})^2$.

(iii) When $d > d_0$ and $\mathcal{S}(\mathbf{B}_0) \subseteq \mathcal{S}(\mathbf{A})$. In this setting, we have $\mathrm{rank}(\mathbf{C}_1) = d_0$. By a similar argument of above (A.13), we can obtain that

$$\mathrm{MDD}(Y|\mathbf{C}_1^T\mathbf{B}_0^T\mathbf{X})^2 \leqslant \mathrm{MDD}(Y|\mathbf{B}_0^T\mathbf{X})^2. \tag{A.14}$$

Thus, combining (A.10) with (A.14), we have $\mathrm{MDD}(Y|\mathbf{A}^T\mathbf{X})^2 \leqslant \mathrm{MDD}(Y|\mathbf{B}_0^T\mathbf{X})^2$. □

Proof of Theorem 2. When $d < d_0$,

$$\begin{aligned}
G_n(d) - G_n(d_0) &= \log\{\mathrm{MDD}_n(Y|\hat{\mathbf{B}}_{d_0}^T\mathbf{X})^2\}\\
&\quad - \log\{\mathrm{MDD}_n(Y|\hat{\mathbf{B}}_d^T\mathbf{X})^2\} + C_n\{d - d_0\}\\
&= \log\{\mathrm{MDD}(Y|\hat{\mathbf{B}}_{d_0}^T\mathbf{X})^2\}\\
&\quad - \log\{\mathrm{MDD}(Y|\hat{\mathbf{B}}_d^T\mathbf{X})^2\} + o_p(1)\\
&\quad + C_n\{d - d_0\}.
\end{aligned} \tag{A.15}$$

By Proposition 2(i), we have

$$\log\{\mathrm{MDD}(Y|\hat{\mathbf{B}}_d^T\mathbf{X})^2\} < \log\{\mathrm{MDD}(Y|\hat{\mathbf{B}}_{d_0}^T\mathbf{X})^2\}.$$

This, together with (A.15), implies that

$$G_n(d) > G_n(d_0), \text{ for } d < d_0, \tag{A.16}$$

in probability, if $C_n \to 0$.

When $d > d_0$, the following Taylor expansion holds

$$\begin{aligned}
&G_n(d) - G_n(d_0)\\
&= -\log\left\{1 + \frac{\mathrm{MDD}_n(Y|\hat{\mathbf{B}}_d^T\mathbf{X})^2 - \mathrm{MDD}_n(Y|\hat{\mathbf{B}}_{d_0}^T\mathbf{X})^2}{\mathrm{MDD}_n(Y|\hat{\mathbf{B}}_{d_0}^T\mathbf{X})^2}\right\}\\
&\quad + C_n\{d - d_0\}\\
&= -\frac{\mathrm{MDD}_n(Y|\hat{\mathbf{B}}_d^T\mathbf{X})^2 - \mathrm{MDD}_n(Y|\hat{\mathbf{B}}_{d_0}^T\mathbf{X})^2}{\mathrm{MDD}(Y|\hat{\mathbf{B}}_{d_0}^T\mathbf{X})^2}\{1 + o_p(1)\}\\
&\quad + o_p(1) + C_n\{d - d_0\}.
\end{aligned} \tag{A.17}$$

By the triangle inequality, for $\mathbf{B} \in \mathbb{R}^{p\times d}$ with $\mathbf{B}^T\Sigma_{\mathbf{x}}\mathbf{B} = \mathbf{I}_d$, we have

$$\|\mathbf{B}^T(\mathbf{X} - \mathbf{X}')\| \le \|\mathbf{B}_1^T(\mathbf{X} - \mathbf{X}')\| + \|\mathbf{B}_2^T(\mathbf{X} - \mathbf{X}')\|,$$

where $\mathbf{B} = (\mathbf{B}_1, \mathbf{B}_2)$, $\mathbf{B}_1 \in \mathbb{R}^{p\times d_0}$ and $\mathbf{B}_2 \in \mathbb{R}^{p\times(d-d_0)}$. This, together with (4), yields that

$$\begin{aligned}
\mathrm{MDD}_n(Y|\hat{\mathbf{B}}_d^T\mathbf{X})^2
&\le \max_{\mathbf{B}_1^T\hat{\Sigma}_{\mathbf{x}}\mathbf{B}_1 = \mathbf{I}_{d_0}} \{\mathrm{MDD}_n(Y|\mathbf{B}_1^T\mathbf{X})^2\}\\
&\quad + \max_{\substack{\mathbf{B}_2^T\hat{\Sigma}_{\mathbf{x}}\mathbf{B}_2 = \mathbf{I}_{d-d_0},\\ \hat{\mathbf{B}}_{d_0}^T\hat{\Sigma}_{\mathbf{x}}\mathbf{B}_2 = \mathbf{0}}} \{\mathrm{MDD}_n(Y|\mathbf{B}_2^T\mathbf{X})^2\}
\end{aligned} \tag{A.18}$$

$$= \mathrm{MDD}_n(Y|\hat{\mathbf{B}}_{d_0}^T\mathbf{X})^2 + \max_{\mathbf{C}^T\mathbf{C}=\mathbf{I}_{d-d_0}} \{\mathrm{MDD}_n(Y|(\hat{\mathbf{B}}_{d_0\perp}\mathbf{C})^T\mathbf{X})^2\},$$

where $\mathbf{C} \in \mathbb{R}^{(p-d)\times(d-d_0)}$. By a similar argument of Proposition (A.1), we have

$$\mathrm{MDD}(Y|(\mathbf{B}_{0\perp}\mathbf{C})^T\mathbf{X})^2 = 0, \text{ for any } \mathbf{C} \in \mathbb{R}^{(p-d)\times(d-d_0)}.$$

Thus, by Theorem 4 in Shao and Zhang [17], we can show that

$$\mathrm{MDD}_n(Y|(\mathbf{B}_{0\perp}\mathbf{C})^T\mathbf{X})^2 = O_p(\frac{1}{n}).$$

This, together with (A.18), yields

$$\text{(A.19)} \quad \mathrm{MDD}_n(Y|\hat{\mathbf{B}}_d^T\mathbf{X})^2 - \mathrm{MDD}_n(Y|\hat{\mathbf{B}}_{d_0}^T\mathbf{X})^2 = O_p(\frac{1}{n}).$$

By (A.17) and (A.19), we have

$$\text{(A.20)} \qquad G_n(d) > G_n(d_0), \text{ for } d > d_0,$$

in probability, if $nC_n \to \infty$. Consequently, Theorem 2 follows from (A.16) and (A.20). □

Received 6 October 2018

REFERENCES

[1] Chen, X., Zou, C., Cook, R. D., et al. (2010). Coordinate-independent sparse sufficient dimension reduction and variable selection. *The Annals of Statistics*, 38(6):3696–3723. MR2766865

[2] Cook, R. D. (1998). *Regression graphics: ideas for studying regressions through graphics*. John Wiley & Sons. MR1645673

[3] Cook, R. D. and Li, B. (2002). Dimension reduction for conditional mean in regression. *The Annals of Statistics*, 30:455–474. MR1902895

[4] Cook, R. D. and Weisberg, S. (1991). Discussion of "sliced inverse regression for dimension reduction". *Journal of the American Statistical Association*, 86:28–33. MR1137117

[5] Cowley, B., Semedo, J., Zandvakili, A., Smith, M., Kohn, A., and Yu, B. (2017). Distance covariance analysis. In *Artificial Intelligence and Statistics*, pages 242–251.

[6] Li, B. and Wang, S. (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association*, 102(479):997–1008. MR2354409

[7] Li, B., Zha, H., and Chiaromonte, F. (2005). Contour regression: A general approach to dimension reduction. *The Annals of Statistics*, 33(4):1580–1616. MR2166556

[8] Li, K. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86:316–327. MR1137117

[9] Li, K.-C. (1992). On principal hessian directions for data visualization and dimension reduction: another application of stein's lemma. *Journal of the American Statistical Association*, 87(420):1025–1039. MR1209564

[10] Li, K.-C. and Duan, N. (1989). Regression analysis under link violation. *The Annals of Statistics*, 17:1009–1052. MR1015136

[11] Ma, Y. and Zhu, L. (2012). A semiparametric approach to dimension reduction. *Journal of the American Statistical Association*, 107:168–179. MR2949349

[12] Ma, Y. and Zhu, L. (2013). A review on dimension reduction. *International Statistical Review*, 81:134–150. MR3047506

[13] Ma, Y. and Zhu, L. (2014). On estimation efficiency of the central mean subspace. *Journal of the Royal Statistical Society: Series B*, 76:885–901. MR3271171

[14] Nocedal, J. and Wright, S. J. (2006). *Numerical optimization, 2nd*. Springer. MR2244940

[15] Park, T., Shao, X., and Yao, S. (2015). Partial martingale difference correlation. *Electronic Journal of Statistics*, 9:1492–1517. MR3367668

[16] Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica*, pages 221–242. MR1466682

[17] Shao, X. and Zhang, J. (2014). Martingale difference correlation and its use in high-dimensional variable screening. *Journal of the American Statistical Association*, 109(507):1302–1318. MR3265698

[18] Sheng, W. and Yin, X. (2013). Direction estimation in single-index models via distance covariance. *Journal of Multivariate Analysis*, 122:148–161. MR3189314

[19] Sheng, W. and Yin, X. (2016). Sufficient dimension reduction via distance covariance. *Journal of Computational and Graphical Statistics*, 25(1):91–104. MR3474038

[20] Shi, P. and Tsai, C.-L. (2002). Regression model selection-a residual likelihood approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(2):237–252. MR1904703

[21] Székely, G. J., Rizzo, M. L., Bakirov, N. K., et al. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794. MR2382665

[22] Székely, G. J., Rizzo, M. L., et al. (2009). Brownian distance covariance. *The Annals of Applied Statistics*, 3(4):1236–1265. MR2752127

[23] Xia, Y., Tong, H., Li, W., and Zhu, L. (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B*, 64:363–410. MR1924297

[24] Xue, Y., Zhang, N., Yin, X., and Zheng, H. (2017). Sufficient dimension reduction using Hilbert–Schmidt independence criterion. *Computational Statistics & Data Analysis*, 115:67–78. MR3683129

[25] Zhu, Y. and Zeng, P. (2006). Fourier methods for estimating the central subspace and the central mean subspace in regression. *Journal of the American Statistical Association*, 101(476):1638–1651. MR2279485

Yu Zhang
School of Mathematical Sciences,
Peking University, Beijing
China

Jicai Liu
College of Mathematics and Sciences,
Shanghai Normal University, Shanghai
China

Yuesong Wu
College of Mathematics and Sciences,
Shanghai Normal University, Shanghai
China

Xiangzhong Fang
School of Mathematical Sciences,
Peking University, Beijing
China
E-mail address: xzfang@math.pku.edu.cn