ON TWO NEW ESTIMATORS FOR THE CMS THROUGH EXTENSIONS OF OLS

A Dissertation
Submitted to
the Temple University Graduate Board

---

in Partial Fulfillment
of the Requirements for the Degree of
DOCTOR OF PHILOSOPHY

---

by
Yongxu Zhang
MAY, 2017

Examining Committee Members:

Dr. Yuexiao Dong, Advisory Chair, Statistics
Dr. William Wei, Statistics
Dr. Pallavi Chitturi, Statistics
Dr. Yang Yang, School of Sport Tourism and Hospitality Management

ABSTRACT

ON TWO NEW ESTIMATORS FOR THE CMS THROUGH EXTENSIONS OF OLS

Yongxu Zhang

DOCTOR OF PHILOSOPHY

Temple University, MAY, 2017

Dr. Yuexiao Dong, Chair

As a useful tool for multivariate analysis, sufficient dimension reduction (SDR) aims to reduce the predictor dimensionality while simultaneously keeping the full regression information, or some specific aspects of the regression information, between the response and the predictor. When the goal is to retain the information about the regression mean, the target of the inference is known as the central mean space (CMS). Ordinary least squares (OLS) is a popular estimator of the CMS, but it has the limitation that it can recover at most one direction in the CMS. In this dissertation, we introduce two new estimators of the CMS: the sliced OLS and the hybrid OLS. Both estimators can estimate multiple directions in the CMS.

The dissertation is organized as follows. Chapter 1 provides a literature review about basic concepts and some traditional methods in SDR. Motivated from the popular SDR method called sliced inverse regression, sliced OLS is proposed as the first extension of OLS in Chapter 2. The asymptotic properties of sliced OLS, order determination, as well as testing predictor contribution through sliced OLS are studied in Chapter 2 as well. It is well-known that slicing methods such as sliced inverse regression may lead to different results with different number of slices. Chapter 3 proposes hybrid OLS as the second extension. Hybrid OLS shares the benefit of sliced OLS and recovers multiple directions in the CMS. At the same time, hybrid OLS improves over sliced OLS by avoiding slicing. Extensive numerical results are provided to demonstrate the desirable performances of the proposed estimators. We conclude the dissertation with some discussions about the future work in Chapter 4.

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude to my advisor Dr. Yuexiao Dong, who has been a tremendous mentor to me. His guidance and inspiration have benefited me enormously in all the time of research and writing of this thesis.

I would also like to thank the committees members Dr. William Wei and Dr. Pallavi Chitturi for their guidance not only on my research but also my life. A special thanks to Dr. Yang Yang for his expertise and insights.

Last but not least, I would like to thank my parents, my husband and my son for their love and support during this memorable journey.

TABLE OF CONTENTS

ABSTRACT **iii**

ACKNOWLEDGEMENT **v**

LIST OF TABLES **viii**

LIST OF FIGURES **ix**

1 LITERATURE REVIEW **1**
1.1 Central Space and Central Mean Space 1
1.2 Classical Methods for Dimension Reduction 3
1.2.1 Ordinary Least Squares 3
1.2.2 Principal Hessian Directions 4
1.2.3 Sliced Inverse Regression 5
1.3 Order Determination 7
1.4 Marginal Coordinate Test 8

2 SLICED OLS, THE PROPOSED FIRST METHOD **10**
2.1 Motivation 10
2.2 SOLS Estimator for the CMS 11
2.3 Asymptotic Properties of SOLS 13
2.4 Sequential Test for Order Determination 15
2.5 Marginal Coordinate Test for Variable Selection 17
2.6 Numerical Study 19
2.6.1 CMS Estimation 19
2.6.2 Order Determination 23
2.6.3 Marginal Coordinate Test 24

3 HYBRID OLS, THE PROPOSED SECOND METHOD **27**
3.1 Motivation 27
3.2 HOLS Estimator for the CMS 28

3.3 Multivariate Response 29
3.4 Numerical Study 31

4 CONCLUSION AND FUTURE WORK **35**
4.1 Summary 35

BIBLIOGRAPHY **37**

LIST OF TABLES

2.1 *Mean and Variance of the Distance r for the Sine Model (I), $d = 2$, $p = 10$ with 100 repetitions. $H_{SIR} = 5$ for SIR, $H = 2$ for SOLS.* 21

2.2 *Mean and Variance of the Distance r for the Rational Model (II), $d = 2$, $p = 10$, $H = 2$ with 100 repetitions.* 22

2.3 *Mean and Variance of the Distance r for the Hybrid Model (III), $d = 2$, $p = 10$, $H = 2$ with 100 repetitions.* 22

2.4 *Mean and Variance of the Distance r for Model I, Model II and Model III, $n = 800$, $p = 10$ with 100 repetitions. $H = 2$, 5, 8.* . . 23

2.5 *Report on Probabilities of Rejectiong $H_0^{(l)} : d = l$ based on Sequential Test of Order Determination with $H = 5,\ 10$ and $n = 400,\ 800$.* 24

2.6 *Report on Probabilities of Rejecting $H_0^{[k]} : k \in \mathcal{A}^c$ based on Asymptotic Test on SOLS with $H = 2$, $n = 200,\ 400,\ 800$.* . . . 26

3.1 *Mean and Variance of the Distance r for Model (I), Model (II) and Model (III), $d = 2$, $p = 10$ with 100 repetitions.* 32

3.2 *Mean and Variance of the Distance r for the multivariate response model, $p = 10$, $d = 3$ with 100 repetitions.* 33

LIST OF FIGURES

3.1 *(a) Scatter plot of* Y_1 *versus* $\hat{\boldsymbol{\beta}}_1'\mathbf{X}$*, (b) Scatter plot of* Y_1 *versus* $\hat{\boldsymbol{\beta}}_2'\mathbf{X}$*,(c) Scatter plot of* Y_1 *versus* $\hat{\boldsymbol{\beta}}_3'\mathbf{X}$*,(d) Scatter plot of* Y_2 *versus* $\hat{\boldsymbol{\beta}}_1'\mathbf{X}$*,(e) Scatter plot of* Y_2 *versus* $\hat{\boldsymbol{\beta}}_2'\mathbf{X}$*,(f) Scatter plot of* Y_2 *versus* $\hat{\boldsymbol{\beta}}_3'\mathbf{X}$*, for multivariate response model.* 34

CHAPTER 1

LITERATURE REVIEW

1.1 Central Space and Central Mean Space

Consider a univariate response Y and a p-dimensional predictor $\mathbf{X}$. The goal for dimension reduction (Li, 1991; Li, 1992; Cook and Weisberg, 1991; Cook, 1998) is to find a matrix $\boldsymbol{\beta} \in \mathbb{R}^{p\times d}$, with $d < p$, such that

$$Y \perp\!\!\!\perp \mathbf{X}|\boldsymbol{\beta}'\mathbf{X}, \tag{1.1}$$

where $\perp\!\!\!\perp$ denotes independence.

Remark 1.1.1. *For any non-singular matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$, given $Y \perp\!\!\!\perp \mathbf{X}|\boldsymbol{\beta}'\mathbf{X}$, $\boldsymbol{\beta}\mathbf{A}$ will satisfy $Y \perp\!\!\!\perp \mathbf{X}|(\boldsymbol{\beta}\mathbf{A})'\mathbf{X}$. Denote* $\mathrm{Span}(\boldsymbol{\beta})$ *as the column space of $\boldsymbol{\beta}$. Then it is easy to see $Span(\boldsymbol{\beta}) = Span(\boldsymbol{\beta}\mathbf{A})$.*

Definition 1.1.1. *For $\boldsymbol{\beta}$ satisfying (1.1), we define $Span(\boldsymbol{\beta})$ as a dimension* *reduction space. The central space (CS) for regressing between Y and $\mathbf{X}$ is the intersection of all dimension reduction spaces, denoted as $\mathcal{S}_{Y|\mathbf{X}}$.*

Let $\boldsymbol{\beta}$ be the basis of the $\mathcal{S}_{Y|\mathbf{X}}$. Then $Span(\boldsymbol{\beta})$ is also the smallest among all $\boldsymbol{\beta}$'s satisfying $F(Y|\mathbf{X}) = F(Y|\boldsymbol{\beta}'\mathbf{X})$, where F denotes the cumulative distribution function. In other words, the conditional distribution of Y given $\mathbf{X}$ is the same as the conditional distribution of Y given $\boldsymbol{\beta}'\mathbf{X}$.

Cook and Li (2002) introduced the concept of CMS, which aims to find $\boldsymbol{\beta}$ with the smallest column space such that $\mathrm{E}(Y|\mathbf{X}) = \mathrm{E}(Y|\boldsymbol{\beta}'\mathbf{X})$. More specifically, CMS is defined as follows.

Definition 1.1.2. *For $\boldsymbol{\beta}$ satisfying $Y \perp\!\!\!\perp \mathrm{E}(Y|\mathbf{X})|\boldsymbol{\beta}'\mathbf{X}$, we define $Span(\boldsymbol{\beta})$ as a mean dimension reduction space. The central mean space (CMS) for regressing between Y and $\mathbf{X}$ is the intersection of all mean dimension reduction subspace, denoted as $\mathcal{S}_{\mathrm{E}(Y|\mathbf{X})}$.*

Remark 1.1.2. *It is easy to see $Y \perp\!\!\!\perp \mathbf{X}|\boldsymbol{\beta}'\mathbf{X}$ implies $Y \perp\!\!\!\perp \mathrm{E}(Y|\mathbf{X})|\boldsymbol{\beta}'\mathbf{X}$. With Definition 1.1.1 and Definition 1.1.2 we can conclude $\mathcal{S}_{\mathrm{E}(Y|\mathbf{X})} \subseteq \mathcal{S}_{Y|\mathbf{X}}$.*

An important property for CMS is the invariance property below.

Proposition 1.1.1. *Let $\mathbf{A} \in \mathbb{R}^{p\times p}$ be non-singular, $\mathbf{b} \in \mathbb{R}^p$, and $\mathbf{X} = \mathbf{AZ}+\mathbf{b}$. Then, $\mathcal{S}_{\mathrm{E}(Y|\mathbf{X})} = (\mathbf{A}^{-1})'\mathcal{S}_{\mathrm{E}(Y|\mathbf{Z})}$.*

Note that the invariance property is also valid for $\mathcal{S}_{Y|\mathbf{X}}$.

Remark 1.1.3. *Let* $\mathrm{E}(\mathbf{X}) = \mu$, $\mathrm{Var}(\mathbf{X}) = \mathbf{\Sigma}$. *Then* $\mathbf{Z} = \mathbf{\Sigma}^{-1/2}(\mathbf{X} - \boldsymbol{\mu})$ *is the standardized predictor. From Proposition 1.1.1, we have* $\mathcal{S}_{\mathrm{E}(Y|\mathbf{X})} = \mathbf{\Sigma}^{-1/2}\mathcal{S}_{\mathrm{E}(Y|\mathbf{Z})}$.

1.2 Classical Methods for Dimension Reduction

In this section, we review some classical methods in dimension reduction.

1.2.1 Ordinary Least Squares

Li and Duan (1989) proposed OLS as an estimator of the CS. Cook and Li (2002) demonstrated the direction that OLS recovered is in the CMS.

Theorem 1.2.1. *Suppose* $\mathrm{E}(\mathbf{Z}) = \mathbf{0}$, $\mathrm{Var}(\mathbf{Z}) = \mathbf{I}_p$, *and Span*$(\boldsymbol{\eta}) = \mathcal{S}_{\mathrm{E}(Y|\mathbf{Z})}$. *In addition, assume*

$$\mathrm{E}(\mathbf{Z}|\boldsymbol{\eta}'\mathbf{Z}) \textit{ is a linear function of } \boldsymbol{\eta}'\mathbf{Z}. \tag{1.2}$$

Then $\mathrm{E}(\mathbf{Z}Y) \in \mathcal{S}_{\mathrm{E}(Y|\mathbf{Z})}$.

Remark 1.2.1. *From Proposition 1.1.1:* $\mathbf{\Sigma}^{-1/2}\mathcal{S}_{\mathrm{E}(Y|\mathbf{Z})} = \mathcal{S}_{\mathrm{E}(Y|\mathbf{X})}$, *together with Theorem 1.2.1, we have* $\mathbf{\Sigma}^{-1}\mathrm{E}\big((\mathbf{X}-\boldsymbol{\mu})Y\big) = \mathbf{\Sigma}^{-1/2}\mathrm{E}(\mathbf{Z}Y) \in \mathbf{\Sigma}^{-1/2}\mathcal{S}_{\mathrm{E}(Y|\mathbf{Z})} = \mathcal{S}_{\mathrm{E}(Y|\mathbf{X})}$. *Let* $\boldsymbol{\beta}_0 = \mathbf{\Sigma}^{-1}\mathrm{E}\big((\mathbf{X} - \boldsymbol{\mu})Y\big)$, *and we have shown* $\boldsymbol{\beta}_0 \in \mathcal{S}_{\mathrm{E}(Y|\mathbf{X})}$.

At the sample level, let $\bar{\mathbf{X}} = \mathrm{E}_n\mathbf{X} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i$, $\hat{\mathbf{\Sigma}}^{-1} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})'$, and $\mathrm{E}_n\big((\mathbf{X} - \boldsymbol{\mu})Y\big) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})Y_i$. Then the sample level OLS estimator can be formed as $\hat{\boldsymbol{\beta}}_0 = \hat{\mathbf{\Sigma}}^{-1}\mathrm{E}_n\big((\mathbf{X} - \boldsymbol{\mu})Y\big)$.

One disadvantage of the OLS method is that it can estimate at most one direction in $\mathcal{S}_{\mathrm{E}(Y|\mathbf{X})}$. Another disadvantage is that OLS does not perform well when there is symmetric link function between the response and the predictor.

1.2.2 Principal Hessian Directions

Li (1992) proposed the principal Hessian directions (PHD) and Cook and Li (2002) further demonstrate that PHD recovers the CMS.

Theorem 1.2.2. *Let* $\mathrm{E}(\mathbf{Z}) = \mathbf{0}$, *and* $\mathrm{Var}(\mathbf{Z}) = \mathbf{I}_p$, $\mathrm{E}(\mathbf{Y}) = \mathbf{0}$, *and* $Span(\boldsymbol{\eta}) = \mathcal{S}_{\mathrm{E}(Y|\mathbf{Z})}$. *In addition to assumption (1.2) in Theorem 1.2.1, we assume* $\mathrm{Var}(\mathbf{Z}|\boldsymbol{\eta}'\mathbf{Z})$ *is a constant matrix. Then* $Span\big(\mathrm{E}(Y\mathbf{Z}\mathbf{Z}')\big) \subseteq \mathcal{S}_{\mathrm{E}(Y|\mathbf{Z})}$.

We define the kernel matrix of PHD as $\mathbf{M}_{\mathrm{PHD}} = \mathrm{E}(Y\mathbf{Z}\mathbf{Z}')\mathrm{E}(Y\mathbf{Z}'\mathbf{Z})$, then $Span(\mathbf{M}_{\mathrm{PHD}}) \subseteq \mathcal{S}_{\mathrm{E}(Y|\mathbf{Z})}$.

Remark 1.2.2. *From Proposition 1.1.1:* $\boldsymbol{\Sigma}^{-1/2}\mathcal{S}_{\mathrm{E}(Y|\mathbf{Z})} = \mathcal{S}_{\mathrm{E}(Y|\mathbf{X})}$, *together with Theorem 1.2.2 and Remark 1.1.1, we have* $Span\Big(\boldsymbol{\Sigma}^{-1}\mathrm{E}\big(Y(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})'\big)\Big) = Span\Big(\boldsymbol{\Sigma}^{-1}\mathrm{E}\big(Y(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})\big)'\boldsymbol{\Sigma}^{-1/2}\Big) = Span\big(\boldsymbol{\Sigma}^{-1/2}\mathrm{E}(Y\mathbf{Z}\mathbf{Z}')\big) \subseteq \boldsymbol{\Sigma}^{-1/2}\mathcal{S}_{\mathrm{E}(Y|\mathbf{Z})} = \mathcal{S}_{\mathrm{E}(Y|\mathbf{X})}$.

Suppose $\mathcal{S}_{\mathrm{E}(Y|\mathbf{X})} = Span(\boldsymbol{\beta})$ with $\boldsymbol{\beta} \in \mathbb{R}^{p\times d}$. The sample level algorithm is formally described as follows.

Algorithm

i Let $\hat{\mathbf{\Sigma}}$ and $\bar{\mathbf{X}}$ be the sample covariance matrix and sample mean respectively of $\mathbf{X}$, then $\hat{\mathbf{Z}}_i = \hat{\mathbf{\Sigma}}^{-1/2}(\mathbf{X}_i - \bar{\mathbf{X}})$.

ii Let $\hat{Y}_i = Y_i - \bar{Y}$, where $\bar{Y}$ is the sample mean of Y.

iii Calculate the kernel matrix $\hat{\mathbf{M}}_{\text{PHD}} = \mathrm{E}_n(Y\mathbf{ZZ}')\mathrm{E}_n(Y\mathbf{Z}'\mathbf{Z})$, where $\mathrm{E}_n(Y\mathbf{ZZ}') = \frac{1}{n}\sum_{i=1}^{n}\hat{Y}_i\hat{\mathbf{Z}}_i\hat{\mathbf{Z}}_i'$.

iv Conduct an eigenvalue decomposition on the kernel matrix $\hat{\mathbf{M}}_{\text{PHD}}$ to find the eigenvectors $\hat{\boldsymbol{\eta}}_k, k = 1, ..., d$, corresponding to the d largest eigenvalues.

v By Proposition 1.1.1, transfer back to get $\hat{\boldsymbol{\beta}}_k = \hat{\mathbf{\Sigma}}^{-1/2}\hat{\boldsymbol{\eta}}_k$, and $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \ldots, \hat{\boldsymbol{\beta}}_d)$.

Remark 1.2.3. *Unlike OLS, which can estimate at most one direction, PHD can recover more than one direction in the CMS. Another difference between OLS and PHD is that OLS does not work well when quadratic trend exists, but works well when a linear trend exists. PHD is the opposite. See, for example, Li (1992), Yu et al. (2010) for more details.*

1.2.3 Sliced Inverse Regression

Li (1991) proposed the sliced inverse regression (SIR), which to recover the central space based on the following fact.

Theorem 1.2.3. *Let* $Span(\boldsymbol{\eta}) = \mathcal{S}_{Y|\mathbf{Z}}$. *In addition, assume*

$$\mathrm{E}(\mathbf{Z}|\boldsymbol{\eta}'\mathbf{Z}) \textit{ is a linear function of } \boldsymbol{\eta}'\mathbf{Z}. \tag{1.3}$$

Then $\mathrm{E}(\mathbf{Z}|Y) \in \mathcal{S}_{Y|\mathbf{Z}}$.

Note that when OLS is reviewed in Section 1.2.1, the assumption (1.2) assigns $Span(\boldsymbol{\eta})$ as the basis of the CMS, while in assumption (1.3), $Span(\boldsymbol{\eta})$ denotes the basis of the CS.

Corollary 1.2.1. *Under the same condition of Theorem 1.2.3, we define the kernel matrix of SIR as* $\mathbf{M}_{SIR} = \mathrm{Var}\big(\mathrm{E}(\mathbf{Z}|Y)\big)$, *then* $Span(\mathbf{M}_{SIR}) \subseteq \mathcal{S}_{Y|\mathbf{Z}}$.

Suppose $\mathcal{S}_{Y|\mathbf{X}} = Span(\boldsymbol{\beta})$ with $\boldsymbol{\beta} \in \mathbb{R}^{p\times d}$. The sample level algorithm is shown as follows.

Algorithm

i Let $\hat{\boldsymbol{\Sigma}}$ and $\bar{\mathbf{X}}$ be the sample covariance matrix and sample mean respectively of $\mathbf{X}$, then $\hat{\mathbf{Z}}_i = \hat{\boldsymbol{\Sigma}}^{-1/2}(\mathbf{X}_i - \bar{\mathbf{X}})$.

ii Divide the range of Y into H slices and let the probability of Y_i falling into the h^{th} slice be $\hat{p}_h = (1/n)\sum_{i=1}^{n} I(Y_i \in J_h)$, $I(Y_i \in J_h) = 1$ when Y_i is in the h^{th} slice, and 0 otherwise.

iii Within the h^{th} slice, calculate $\hat{\mathbf{m}}_h = (1/n\hat{p}_h)\sum_{\hat{Y}_i \in J_h} \hat{\mathbf{Z}}_i$. Let the kernel matrix $\hat{\mathbf{M}}_{SIR} = \sum_{h=1}^{H} \hat{p}_h \hat{\mathbf{m}}_h \hat{\mathbf{m}}_h'$.

iv Conduct an eigenvalue decomposition on the kernel matrix $\hat{\mathbf{M}}_{SIR}$ to find the eigenvectors $\hat{\boldsymbol{\eta}}_k$, k = 1, ..., d, corresponding to the d largest eigenvalue.

v By Proposition 1.1.1, transfer back to get $\hat{\boldsymbol{\beta}}_k = \hat{\boldsymbol{\Sigma}}^{-1/2}\hat{\boldsymbol{\eta}}_k$. Then $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \ldots, \hat{\boldsymbol{\beta}}_d)$.

Note that when Y is discrete, this denotes the possible values of Y that are in the set $(\xi_1, \ldots, \xi_H)$. Then the h^{th} intraslice mean will be $\mathbf{m}_h = \mathrm{E}(\mathbf{Z}|Y = \xi_h)$, $h = 1, 2, ..., H$. The corresponding sample level algorithm is similar and is thus omitted.

Remark 1.2.4. *Note that* $\sum_{h=1}^{H} \hat{p_h}\hat{\mathbf{m}}_h = \frac{1}{n}\sum_{h=1}^{H}\sum_{Y\in I_h}\hat{\mathbf{Z}}_{\mathbf{i}}$ $= \frac{1}{n}\boldsymbol{\Sigma}^{-1/2}\sum_{h=1}^{H}\sum_{Y\in I_h}(\mathbf{X}_{\mathbf{i}} - \bar{\mathbf{X}}) = \frac{1}{n}\boldsymbol{\Sigma}^{-1/2}\sum_{i=1}^{n}(\mathbf{X}_{\mathbf{i}} - \bar{\mathbf{X}}) = \mathbf{0}$. *Thus we have* $\hat{\mathbf{m}}_1, \hat{\mathbf{m}}_2, \ldots, \hat{\mathbf{m}}_H$ *are linearly dependent. From the definition of* $\hat{\mathbf{M}}_{SIR} = \Sigma_{h=1}^{H}\hat{p}_h\hat{\boldsymbol{m}}_h\hat{\boldsymbol{m}}_h'$, *we know* $d = rank(\hat{\mathbf{M}}_{SIR}) \leq H - 1$. *Thus we have proved SIR can estimate at most* $H - 1$ *directions in the CS.*

1.3 Order Determination

We review the sequential test approach as follows. Let $\mathbf{M}$ be the kernel matrix of a dimension reduction methods. For example $\mathbf{M}_{\mathrm{SIR}} = \mathrm{Var}\big(\mathrm{E}(\mathbf{Z}|Y\big)$. Let $\hat{\mathbf{M}}$ be the sample estimator of $\mathbf{M}$. Suppose $rank(\mathbf{M}) = d$. Order determination aims to estimate d based on $\hat{\mathbf{M}}$.

Consider $H_0^{(l)} : d = l$ v.s. $H_a^{(l)} : d > l$. We then estimate d as $\hat{d} = \text{argmin}\{l : H_0^{(l)}$ is not rejected$\}$. Specifically, let $\hat{T}_{n,l} = n\sum_{j=l+1}^{p}\hat{\lambda}_j$, where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p$ are eigenvalues of the kernel matrix $\hat{\mathbf{M}}$. Then we reject $H_0^{(l)}$ for large values of $\hat{T}_{n,l}$. The threshold for $\hat{T}_{n,l}$ is decided by the asymptotic distribution of $\hat{T}_{n,l}$ under $H_0^{(l)}$. For example, see Li (1991; 1992), Cook and Li (2004), Li and Wang (2007), and Dong and Yu (2012).

1.4 Marginal Coordinate Test

For $\mathbf{X} = (X_1, \ldots, X_p)'$, let $\mathbf{X}_{-k} = (X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_p)'$ for $k \in \{1, \ldots, p\}$. A marginal coordinate test (Cook, 2004; Shao et al., 2007; Yu and Dong, 2016) considers $H_0^{[k]} : Y \perp\!\!\!\perp \mathbf{X}|\mathbf{X}_{-k}$ v.s. $H_a^{[k]} : Y \not\perp\!\!\!\perp \mathbf{X}|\mathbf{X}_{-k}$. Here $\not\perp\!\!\!\perp$ denotes non-independence. The following result is due to Dong et al. (2016).

Proposition 1.4.1. *Let* $\mathbf{e}_k \in \mathbb{R}^p$, $k = 1, \ldots, p$, *where the* k^{th} *element of* $\mathbf{e}_k$ *is* 1 *and all the other elements are zero. Suppose* $\mathcal{S}_{Y|\mathbf{X}} = \text{Span}(\boldsymbol{\beta})$ *for* $\boldsymbol{\beta} \in \mathbb{R}^{p\times d}$. *Then* $\mathbf{e}_k'\boldsymbol{\beta} = 0$ *if and only if* $Y \perp\!\!\!\perp \mathbf{X}|\mathbf{X}_{-k}$.

Recall $\mathbf{M}_{\text{SIR}} = \text{Var}\big(\text{E}(\mathbf{Z}|Y)\big)$. Let $T_k^{\text{SIR}} = \mathbf{e}_k'\boldsymbol{\Sigma}^{-1/2}\mathbf{M}_{\text{SIR}}\boldsymbol{\Sigma}^{-1/2}\mathbf{e}_k$. From Proposition 1.4.1, we have the following result from the SIR-based marginal coordinate test.

Proposition 1.4.2. *Assume Span*$(\boldsymbol{\Sigma}^{-1/2}\mathbf{M}_{SIR}\boldsymbol{\Sigma}^{-1/2}) = \mathcal{S}_{Y|\mathbf{X}}$. *Then* $T_k^{SIR} = 0$ *if and only if* $Y \perp\!\!\!\perp \mathbf{X}|\mathbf{X}_{-k}$.

Recall $\mathbf{M}_{\text{SIR}} = \text{Var}\big(\text{E}(\mathbf{Z}|\mathbf{Y})\big) = \boldsymbol{\Sigma}^{-1/2}\text{E}\Big(\text{E}\big((\mathbf{X}-\boldsymbol{\mu})|\mathbf{Y}\big)\text{E}\big((\mathbf{X}-\boldsymbol{\mu})|\mathbf{Y}\big)'\Big)\boldsymbol{\Sigma}^{-1/2} \subseteq \mathcal{S}_{Y|\mathbf{Z}}$. It follows that $Span(\boldsymbol{\Sigma}^{-1/2}\mathbf{M}_{\text{SIR}}\boldsymbol{\Sigma}^{-1/2}) \subseteq \mathcal{S}_{Y|\mathbf{X}}$. We need a stronger assumption that $Span(\boldsymbol{\Sigma}^{-1/2}\mathbf{M}_{\text{SIR}}\boldsymbol{\Sigma}^{-1/2}) = \mathcal{S}_{Y|\mathbf{X}}$ in Proposition 1.4.2.

At the sample level, let $\hat{T}^{\text{SIR}}_{n,k}$ be the sample estimator of T^{SIR}_{k}, then we reject $H_0^{[k]} : Y \perp\!\!\!\perp \mathbf{X}|\mathbf{X}_{-k}$ for large values of $\hat{T}^{\text{SIR}}_{n,k}$. The threshold for $\hat{T}^{\text{SIR}}_{n,k}$ is decided by the asymptotic distribution of $\hat{T}^{\text{SIR}}_{n,k}$ under $H_0^{[k]}$. See Cook (2004) for details.

CHAPTER 2

SLICED OLS, THE PROPOSED FIRST METHOD

2.1 Motivation

Along with the advancement of sciences and technologies, high-dimensional data has the tendency of been collected across various fields. To facilitate the problem of reduce many predictors to a few predictors, researchers mainly developed two different approaches in the statistical literature. One is a variable selection method, like LASSO (Tibshirani, 1996), where researchers believe that among all the predictors, only a few of them truly are related to the response. Another approach is (sufficient) dimension reduction, which assumes the information of response variables can be compressed by a linear combination of the predictors.

The rest of this chapter is constructed as follows. In section 2.2, we derive a new SOLS estimator of the CMS. In the following we study the asymptotic properties of the SOLS estimator in section 2.3. In section 2.4, we introduce a sequential test of order determination. In section 2.5, we propose a new variable selection method which utilizes the SOLS estimator in a marginal coordinate test. In the last section, we study the aforementioned topics through extensive numerical study.

2.2 SOLS Estimator for the CMS

Inspired by the sliced method from Li (1991), we define $W = \boldsymbol{\beta}_0'\mathbf{X}$ and propose a new method that can recover the CMS by dividing the range of W into $J_1, \ldots, J_H$ slices. We call the proposed method Sliced OLS (SOLS). Recall that in Remark 1.2.1, we have $\boldsymbol{\beta}_0 = \boldsymbol{\Sigma}^{-1}\mathrm{E}\big((\mathbf{X} - \boldsymbol{\mu})Y\big) \in \mathcal{S}_{\mathrm{E}(Y|\mathbf{X})}$. Without loss of generality, we assume $\mathrm{E}(Y) = 0$, $\mathrm{E}(\mathbf{X}) = \mathbf{0}$ throughout the rest of the chapter, and the corresponding $\boldsymbol{\beta}_0 = \boldsymbol{\Sigma}^{-1}\mathrm{E}(\mathbf{X}Y)$ from now on. Then the theorem of SOLS is shown as follows.

Theorem 2.2.1. *Let Span*$(\boldsymbol{\beta}) = \mathcal{S}_{\mathrm{E}(Y|\mathbf{X})}$ *and* $\boldsymbol{\beta}_0 = \boldsymbol{\Sigma}^{-1}\mathrm{E}(\mathbf{X}Y)$*. In addition, assume* $\mathrm{E}(\mathbf{X}) = \mathbf{0}$ *and*

$$\mathrm{E}(\mathbf{X}|\boldsymbol{\beta}'\mathbf{X}) \text{ is a linear function of } \boldsymbol{\beta}'\mathbf{X}. \tag{2.1}$$

Then, $\mathbf{\Sigma}^{-1}\mathrm{E}(\mathbf{X}Y|\boldsymbol{\beta}_0'\mathbf{X}) \in \mathcal{S}_{\mathrm{E}(Y|\mathbf{X})}$.

Note that assumption (2.1) here is different from assumption (1.2). Assumption (2.1) is based on $\mathbf{X}$-scale while assumption (1.2) is based on $\mathbf{Z}$-scale. They imply each other but they are different.

Corollary 2.2.1. *Under the same assumption of Theorem 2.2.1, and* $W = \boldsymbol{\beta}_0'\mathbf{X}$ *, define* $\mathbf{M}_{SOLS} = \mathbf{\Sigma}^{-1}\mathrm{E}\big(\mathrm{E}(\mathbf{X}Y|W)\mathrm{E}(\mathbf{X}'Y|W)\big)\mathbf{\Sigma}^{-1}$*, then* $Span(\mathbf{M}_{SOLS}) \subseteq \mathcal{S}_{\mathrm{E}(Y|\mathbf{X})}$.

The proof of Theorem 2.2.1 can be found in the Appendix.

Remark 2.2.1. *Compared with SIR, which can estimate at most* $H-1$ *directions in the CS, SOLS can estimate at most* H *directions in the CMS. The proof is similar to Remark 1.2.4 and can be found in the Appendix.*

We define $\mathbf{U}_h = \mathrm{E}(\mathbf{X}YR_h),\ h = 1, \ldots, H$, where $R_h = I(W \in J_h)$ be the indicator function of W belonging to the h^{th} slice. Let $p_h = \mathrm{E}(R_h)$, the kernel matrix of SOLS can be re-written as

$$\mathbf{M}_{\mathrm{SOLS}} = \sum_{h=1}^{H} p_h^{-1}\mathbf{\Sigma}^{-1}\mathbf{U}_h\mathbf{U}_h'\mathbf{\Sigma}^{-1}.$$

Given $\{(\mathbf{X}_i, Y_i), i = 1, \ldots, n\}$ as a random sample of $(\mathbf{X}, Y)$. Parallel to the SIR algorithm, we conduct a step by step sample level algorithm.

Algorithm

i Let $\hat{W}_i = \hat{\boldsymbol{\beta}}_0'(\hat{\mathbf{X}}_i - \bar{\mathbf{X}})$, $\hat{\boldsymbol{\beta}}_0 = n^{-1}\hat{\boldsymbol{\Sigma}}^{-1}\sum_{i=1}^{n}(\hat{\mathbf{X}}_i - \bar{\mathbf{X}})Y_i$.

ii Divide the range of $\hat{W}$ into H slices. Let the probability of $\hat{W}_i$ falling into the h^{th} slice be $\hat{p}_h = n^{-1}\Sigma_{i=1}^{n} I(\hat{W}_i \in J_h)$.

iii Calculate $\hat{\mathbf{U}}_h = n^{-1}\sum_{i=1}^{n}(\hat{\mathbf{X}}_i - \bar{\mathbf{X}})Y_i I(\hat{W}_i \in J_h)$. Let the kernel matrix $\hat{\mathbf{M}}_{SOLS} = \sum_{h=1}^{H}\hat{p}_h^{-1}\hat{\boldsymbol{\Sigma}}^{-1}\hat{\mathbf{U}}_h\hat{\mathbf{U}}_h'\hat{\boldsymbol{\Sigma}}^{-1}$.

iv Conduct an eigenvalue decomposition on the kernel matrix $\hat{\mathbf{M}}_{SOLS}$ to find the eigenvectors $\hat{\boldsymbol{\beta}}_k,\ k = 1, ..., d$, corresponding to the d largest eigenvalue. Then $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_k)$.

Unlike the algorithms in Chapter 1, no transformation is needed at the last step because the kernel matrix is defined at the **Z**-scale before, and at the **X**-scale here.

2.3 Asymptotic Properties of SOLS

Let $\mathbf{V} = (\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_H)$, where $\mathbf{V}_h = p_h^{-1/2}\boldsymbol{\Sigma}^{-1}\mathbf{U}_h$, and then $\mathbf{M}_{\text{SOLS}} = \mathbf{V}\mathbf{V}'$. At the sample level, define $\hat{p}_h$ as the proportion of $\hat{W} = \hat{\boldsymbol{\beta}}_{\mathbf{0}}'\mathbf{X}$ that falls into the h^{th} slice, denoted as $\hat{p}_h = (1/n)\Sigma_{i=1}^{n} I(\hat{W}_i \in J_h)$. With $\hat{\mathbf{U}}_h = \text{E}_n(\mathbf{X}YR_h)$, $\hat{\mathbf{V}}_h = \hat{p}_h^{-1/2}\hat{\boldsymbol{\Sigma}}^{-1}\hat{\mathbf{U}}_h$, we have $\hat{\mathbf{V}} = (\hat{\mathbf{V}}_1, \hat{\mathbf{V}}_2, \ldots, \hat{\mathbf{V}}_H)$ and $\hat{\mathbf{M}}_{\text{SOLS}} = \hat{\mathbf{V}}\hat{\mathbf{V}}'$. We are interested in deriving the asymptotic distribution of $\sqrt{n}(\big(vec(\hat{\mathbf{M}}_h) - vec(\mathbf{M}_h)\big)$, but we will begin by illustrating the asymptotic distribution of $\sqrt{n}\big(vec(\hat{\mathbf{V}}_h) - vec(\mathbf{V}_h)\big)$ first. The asymptotic expansions will have the following basic form. Let F_n be the empirical measure based on the iid sample. Let S be a real- or matrix- valued functional on $\mathcal{F}$, where $\mathcal{F}$ is a convex set of distributions that includes empirical distributions and the true distribution F.

$$S(F_n) = S(F) + \mathrm{E}_n S^*(F) + O_p(n^{-1}), \tag{2.2}$$

where $\mathrm{E}S^*(F) = 0$, then $\mathrm{E}_n S^*(F) = O_p(n^{-1/2})$ (Hampel (1974); Fernholz (1983); Serfling (1980); Bickel et al. (1993)). Here S^* is known as the Frechet derivative.

Lemma 2.3.1. *Suppose the entries of* $\mathrm{E}(\mathbf{X}Y|W)$ *have finite second moments. Then the the expansions of* $\hat{p}_h$, $\hat{\mathbf{\Sigma}}^{-1}$, $\hat{\mathbf{U}}_h$ *have the form of equation (2.2), where* $(p_h^{-1/2})^*$, $(\mathbf{\Sigma}^{-1})^*$, $\mathbf{U}_h^*$ *take the place of* $S^*(F)$,

$$(p_h^{-1/2})^* = -\frac{1}{2}p_h^{-3/2}(R_h - p_h),$$

$$(\mathbf{\Sigma}^{-1})^* = -\mathbf{\Sigma}^{-1}\big(\mathbf{X}\mathbf{X}' - \mathrm{E}(\mathbf{X}\mathbf{X}')\big)\mathbf{\Sigma}^{-1},$$

$$\mathbf{U}_h^* = \mathbf{X}YR_h - \mathrm{E}(\mathbf{X}YR_h) - \mathbf{X}\mathrm{E}(YR_h).$$

We can now write the asymptotic distribution of $\mathrm{vec}(\hat{\mathbf{V}})$, where $\mathrm{vec}(\hat{\mathbf{V}})$ is the vectorization of matrix $\hat{\mathbf{V}}$, which is the concatenation of columns of $\hat{\mathbf{V}}$.

Theorem 2.3.1. *Under the assumption (1.2), and let* $\mathbf{\Gamma} = \mathrm{Cov}(vec(\mathbf{A}))$, *where* $\mathbf{A} = (\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_H) \in \mathbb{R}^{p\times H}$,

$$\sqrt{n}\big(vec(\hat{\mathbf{V}}) - vec(\mathbf{V})\big) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \mathbf{\Gamma}),$$

with $\mathbf{A}_h = (p_h^{-1/2})^*\mathbf{\Sigma}^{-1}\mathbf{U}_h + p_h^{-1/2}(\mathbf{\Sigma}^{-1})^*\mathbf{U}_h + p_h^{-1/2}\mathbf{\Sigma}^{-1}\mathbf{U}_h^*$, *where* $(p_h^{-1/2})^*$, $(\mathbf{\Sigma}^{-1})^*$, $\mathbf{U}_h^*$ *defined in Lemma 2.3.1.*

2.4 Sequential Test for Order Determination

In the previous development of the SOLS estimator, we assume that the order d is known. In practice, d is usually unknown and the estimation of the order d is needed. There are mainly two approaches to the order determination. One needs to derive the asymptotic distribution of the test statistic and the other one is an empirical method based on the bootstrap re-sampling process. For both approaches, deriving the asymptotic distribution of the test statistic is critical.

Recall the sequential test we introduced in Chapter 1. Suppose $d = \mathrm{rank}(\mathbf{M}_{\mathrm{SOLS}})$. We parallel the asymptotic test as: $H_0^{(l)} : d = l$ v.s. $H_a^{(l)} : d > l$. Define the test statistic

$$\hat{T}_{n,l}^{\mathrm{SOLS}} = n \sum_{j=l+1}^{p} \hat{\lambda}_j,$$

where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p$ are eigenvalues of the kernel matrix $\hat{\mathbf{M}}_{SOLS}$. We will need to find the asymptotic distribution of the test statistic under the null hypothesis. $H_0^{(l)} : d = l$. Through singular value decomposition of the matrix $\mathbf{V}$, we can get following form:

$$\mathbf{V} = \begin{pmatrix} \mathbf{S}_1 & \mathbf{S}_0 \end{pmatrix} \begin{pmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{T}_1' \\ \mathbf{T}_0' \end{pmatrix}, \tag{2.3}$$

where $\mathbf{S}_1 \in \mathbb{R}^{p\times l}$, $\mathbf{S}_0 \in \mathbb{R}^{p\times(p-l)}$,$\mathbf{T}_1 \in \mathbb{R}^{H\times l}$,$\mathbf{T}_0 \in \mathbb{R}^{H\times(H-l)}$, and $\mathbf{D} \in \mathbb{R}^{l\times l}$.

Here we introduced the asymptotic testing method, which is based on the following corollary.

Theorem 2.4.1. *Let* $\mathbf{A}$ *be defined as Theorem 2.3.1 and under the same condition of Theorem 2.3.1, define* $\mathbf{H} = \mathbf{S}_0'\mathbf{A}\mathbf{T}_0$*, and* $\boldsymbol{\Lambda} = \mathrm{Cov}(vec(\mathbf{H}))$*, where* $\mathbf{S}_0$ *and* $\mathbf{T}_0$ *are defined as in (2.3). Under* $H_0^{(l)}$*,*

$$\hat{T}_{n,l}^{SOLS} \xrightarrow{D} \sum_{j=1}^{(p-l)(H-l)} \nu_j \chi_j^2(1),$$

where $\nu_1, \nu_2, \ldots$ *are the eigenvalues of* $\boldsymbol{\Lambda}$*.*

Recall $\hat{d} = \mathrm{argmin}\{l : H_0^{(l)} \text{ is not rejected}\}$, for testing $H_0^{(l)} : d = l$ v.s. $H_a^{(l)} : d > l$, we describe the sample level algorithm here.

Algorithm

i Conduct a singular value decomposition on $\hat{\mathbf{V}} = (\hat{\mathbf{V}}_1, \hat{\mathbf{V}}_2, \ldots, \hat{\mathbf{V}}_H)$, where $\hat{\mathbf{V}}_h = \hat{p}_h^{-1/2}\hat{\boldsymbol{\Sigma}}^{-1}\hat{\mathbf{U}}_h$. Get corresponding $\hat{\mathbf{S}}_0$, $\hat{\mathbf{T}}_0$, where $\hat{\mathbf{S}}_0$ and $\hat{\mathbf{T}}_0$ are the sample estimator of $\mathbf{S}_0$, $\mathbf{T}_0$ in equation 2.3. Calculate $\hat{\mathbf{H}} = \hat{\mathbf{S}}_0'\hat{\mathbf{A}}\hat{\mathbf{T}}_0$.

ii Apply the eigenvalue decomposition on $\hat{\boldsymbol{\Lambda}} = Cov(vec(\hat{\mathbf{H}}))$ to get eigenvalues $\hat{\nu}_j$'s, $j = 1, \ldots, (p-l)(H-l)$.

iii Generate N realizations from a distribution $\Omega_l = \sum_{j=1}^{(p-l)(H-l)} \hat{\nu}_j \chi_j^2(1)$. Here Ω_l is a weighted $\chi^2(1)$ distribution, in which the weights are the eigenvalues $\hat{\nu}_j$'s, $j = 1, \ldots, (p-l)(H-l)$. Denote the upper 5th percentile as C_l.

iv Reject the null hypothesis of $H_0^{(l)} : d = 1$ when $\hat{T}_{n,l}^{\text{SOLS}} > C_l$ and accept otherwise.

2.5 Marginal Coordinate Test for Variable Selection

Now we apply the SOLS estimator to test the predictor contribution. Recall the marginal coordinate hypothesis test in Chapter 1. We conduct following hypothesis test to test the contribution of an individual predictor to the regression mean in the presence of all the other predictors. $H_0^{[k]} : Y \perp\!\!\!\perp \mathrm{E}(Y|\mathbf{X})|\mathbf{X}_{-k}$ v.s. $H_a^{[k]} : Y \not\perp\!\!\!\perp \mathrm{E}(Y|\mathbf{X})|\mathbf{X}_{-k}$. The null hypothesis implies that X_k has no additional contribution to Y given the other predictors. Parallel to Proposition 1.4.1, we have the below proposition.

Proposition 2.5.1. *Suppose* $\mathcal{S}_{\mathrm{E}(Y|\mathbf{X})} = \mathrm{Span}(\boldsymbol{\beta})$ *for* $\boldsymbol{\beta} \in \mathbb{R}^{p\times d}$*. Then* $\mathbf{e}_k'\boldsymbol{\beta} = 0$ *if and only if* $Y \perp\!\!\!\perp \mathrm{E}(Y|\mathbf{X})|\mathbf{X}_{-k}$.

Knowing that the kernel matrix $M_{SOLS} = \boldsymbol{\Sigma}^{-1}\mathrm{E}\big(\mathrm{E}(\mathbf{X}Y|W)\mathrm{E}(\mathbf{X}'Y|W)\big)\boldsymbol{\Sigma}^{-1}$, we have below proposition.

Proposition 2.5.2. *Assume* $Span(\mathbf{M}_{SOLS}) = \mathcal{S}_{\mathrm{E}(Y|\mathbf{X})}$*. Let* $T_k^{SOLS} = \mathbf{e}_k' M_{SOLS}\mathbf{e}_k$*, then* $T_k^{SOLS} = 0$ *if and only if* $Y \perp\!\!\!\perp \mathrm{E}(Y|\mathbf{X})|\mathbf{X}_{-k}$.

Recall from Theorem 2.2.1 that $\mathrm{E}(\mathbf{X}Y|W) \in \mathcal{S}_{\mathrm{E}(Y|\mathbf{X})}$, with $R_h = I(W \in$

J_h), $p_h = \mathrm{E}(R_h)$, and $U_h = \mathrm{E}(\mathbf{X}YR_h)$. The test statistics at the sample level can be written as

$$\hat{T}_{n,k}^{SOLS} = n\sum_{h=1}^{H} \mathbf{e}_k' \hat{\boldsymbol{\Sigma}}^{-1} \hat{p}_h^{-1} \hat{\mathbf{U}}_h \hat{\mathbf{U}}_h' \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{e}_k.$$

Applying the result from Theorem 2.3.1, we can derive the asymptotic distribution of the test statistics and summarize the conclusion as follows.

Theorem 2.5.1. *Let* $\mathbf{A}$ *be defined as Theorem 2.3.1 and under the same condition of Theorem 2.3.1,*

$$\hat{T}_{n,k}^{SOLS} \xrightarrow{D} \sum_{h=1}^{H} \gamma_{kh}\chi_h^2(1),$$

where $\gamma_1, \gamma_2, \ldots$ *are the eigenvalues of* $\mathrm{Cov}(\mathbf{G}_k)$, *where* $\mathbf{G}_k = (\mathbf{g}_{k,2}, \mathbf{g}_{k,1}, \ldots, \mathbf{g}_{k,H})'$, $\mathbf{g}_{k,h} = p_h^{-1/2}\mathbf{e}_k'(\boldsymbol{\Sigma}^{-1})^*\mathbf{U}_h + p_h^{-1/2}\mathbf{e}_k'\boldsymbol{\Sigma}^{-1}\mathbf{U}_h^*$.

The sample level algorithm is shown as follows.

Algorithm

i Apply the eigenvalue decomposition to matrix $\mathrm{Cov}(\hat{\mathbf{G}}_k)$ to get the eigenvalues $\hat{\gamma}_j$'s, $j = 1, \ldots, (p-l)(H-l)$, where $\hat{\mathbf{G}}_k = (\hat{\mathbf{g}}_{k,1}, \hat{\mathbf{g}}_{k,2}, \ldots, \hat{\mathbf{g}}_{k,H})$, $\hat{\mathbf{g}}_{k,h} = \hat{p}_h^{-1/2}\mathbf{e}_k'(\hat{\boldsymbol{\Sigma}}^{-1})^*\hat{\mathbf{U}}_h + \hat{p}_h^{-1/2}\mathbf{e}_k'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\mathbf{U}}_h^*$.

ii Generate N realizations from a distribution $\Psi_k = \sum_{h=1}^{H}\hat{\gamma}_{kh}\chi_h^2(1)$. Here Ψ is a weighted $\chi^2(1)$ distribution, in which the weights are the eigenvalues $\hat{\gamma}_{kh}$'s, $h = 1, \ldots, H$. Denote the upper 5th percentile as D_k.

iii Reject the null hypothesis $H_0^{[k]} : Y \perp\!\!\!\perp \mathrm{E}(Y|\mathbf{X})|\mathbf{X}_{-k}$ when $\hat{T}_{n,k}^{SOLS}(e_k) > D_k$ and accept $H_0^{[k]}$ otherwise.

2.6 Numerical Study

2.6.1 CMS Estimation

To demonstrate the performance of the proposed SOLS method, we simulate data from the following models. Furthermore, we compare the SOLS with classical methods in parallel to further evaluate the advantages and disadvantages of SOLS in various scenarios.

Model (I) — Sine Link Function: $Y = sin(X_1) + (X_2 + 1) \cdot \varepsilon$;

Model (II) — Rational Link Function: $Y = \frac{X_1}{0.5+(X_2+1.5)^2} + 0.1 \cdot \varepsilon$;

Model (III) — Hybrid Cubic and Quadratic Link Function:

$Y = (X_1 + X_2)^3 + (X_3 + X_4)^2 + 0.1 \cdot \varepsilon$.

Fist, let $\mathbf{X} \in \mathbb{R}^{10}$. $\mathbf{X}_i$'s ($i = 1, \ldots, 10$) and ε are generated from $N(0, 1)$. All the simulations will be run 100 times. The mean and the variance of the distances between the estimated spaces and the true directions will be reported. Here is defined the distance r between two spaces Span($\mathbf{A}$) and Span($\mathbf{B}$), where $\mathbf{A} \in \mathbb{R}^{p\times d}$ and $\mathbf{B} \in \mathbb{R}^{p\times D}$ are orthogonal matrices, with $d \geq D$, the same as the criterion m in Xia et al. (2002).

$$r(\mathbf{A}, \mathbf{B}) = ||(\mathbf{I}_p - \mathbf{A}\mathbf{A}')\mathbf{B}||,$$

where $||\cdot||$ denotes the Frobenius norm. If Span($\mathbf{B}$) belongs to Span($\mathbf{A}$), the corresponding criterion r will be zero. The closer to zero, the smaller the distance between Span($\mathbf{A}$) and Span($\mathbf{B}$).

Model (I) Comparison of SIR and SOLS. Here $\mathbf{X} \in \mathbb{R}^{10}$. X_i's ($i = 1, \ldots, 10$) and ε are generated from $N(0, 1)$.

Here $\boldsymbol{\beta}_1' = (1, 0, 0, \ldots)$ is in the CMS, but $\boldsymbol{\beta}_2' = (0, 1, 0, \ldots)$ is not in the CMS. Both $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ are in the CS. Table 2.1 shows the comparison between SIR and SOLS when $n = 200,\ 400$ and 800. The numbers in the first row of each n are the mean of criterion r with 100 repetitions.The numbers in parentheses in the other rows represent the variance of corresponding criterion r. Denote the estimator of $\boldsymbol{\beta}$ from SIR and SOLS as $\hat{\boldsymbol{\beta}}_{\mathrm{SIR}}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{SOLS}}$ separately. Assuming $d = 2$, the results in table 2.1 show that SIR finds two directions in the CS. On the other hand, SOLS can find direction $\boldsymbol{\beta}_1$ in CMS, but not $\boldsymbol{\beta}_2$. The accuracy of both approaches improve along with an increasing n.

Model (II) Comparison between OLS and SOLS. Here $d = 2$ and $\boldsymbol{\beta}_1' = (1, 0, 0, \ldots), \boldsymbol{\beta}_2' = (0, 1, 0, \ldots)$ are in the CMS. The results in table 2.2 show that OLS gets only one direction, because the distance between $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ and $\boldsymbol{\beta}_1$ is close to zero, and the distance between $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ and $\boldsymbol{\beta}_2$ is large. In contrast, SOLS is able to get both directions, and results improve as sample size n

Table 2.1: *Mean and Variance of the Distance r for the Sine Model (I), $d = 2$, $p = 10$ with 100 repetitions. $H_{SIR} = 5$ for SIR, $H = 2$ for SOLS.*

n	SIR		SOLS	
	$r(\hat{\boldsymbol{\beta}}_{\text{SIR}}, \boldsymbol{\beta}_1)$	$r(\hat{\boldsymbol{\beta}}_{\text{SIR}}, \boldsymbol{\beta}_2)$	$r(\hat{\boldsymbol{\beta}}_{\text{SOLS}}, \boldsymbol{\beta}_1)$	$r(\hat{\boldsymbol{\beta}}_{\text{SOLS}}, \boldsymbol{\beta}_2)$
200	0.382	0.562	0.425	0.884
	(0.009)	(0.026)	(0.011)	(0.014)
400	0.277	0.398	0.324	0.893
	(0.004)	(0.011)	(0.006)	(0.016)
800	0.202	0.282	0.239	0.891
	(0.002)	(0.005)	(0.006)	(0.015)

increases, which is as expected.

Model (III) Comparison between PHD and SOLS. Recall the hybrid model. We can see the true dimensions $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, with $\boldsymbol{\beta}_1' = (1/\sqrt{2}, 1/\sqrt{2}, 0, \dots)$ and $\boldsymbol{\beta}_2' = (0, 0, 1/\sqrt{2}, 1/\sqrt{2}, 0, \dots)$, which are both in the CMS. Results from table 2.3 are compatible with the inference drawn from the population level, which is that SOLS can get a direction with a linear link function, $\boldsymbol{\beta}_2$, but cannot recover a direction with the quadratic link function $\boldsymbol{\beta}_1$. PHD is on the opposite.

We summarize the simulation results of all models with just SOLS in table 2.4. And here we want to test SOLS with a different number of slices $H = 2,\ 5,\ 10$ and with correlated predictors, where $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_p)$ is generated from multivariate normal with a mean of zero and covariance matrix

Table 2.2: *Mean and Variance of the Distance r for the Rational Model (II), $d = 2,\ p = 10,\ H = 2$ with 100 repetitions.*

n	OLS		SOLS	
	$r(\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}, \boldsymbol{\beta}_1)$	$r(\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}, \boldsymbol{\beta}_2)$	$r(\hat{\boldsymbol{\beta}}_{\mathrm{SOLS}}, \boldsymbol{\beta}_1)$	$r(\hat{\boldsymbol{\beta}}_{\mathrm{SOLS}}, \boldsymbol{\beta}_2)$
200	0.204 (0.002)	0.996 (0.000)	0.185 (0.002)	0.521 (0.018)
400	0.142 (0.001)	0.999 (0.000)	0.131 (0.001)	0.347 (0.007)
800	0.099 (0.001)	0.999 (0.000)	0.092 (0.001)	0.251 (0.005)

Table 2.3: *Mean and Variance of the Distance r for the Hybrid Model (III), $d = 2,\ p = 10,\ H = 2$ with 100 repetitions.*

n	PHD		SOLS	
	$r(\hat{\boldsymbol{\beta}}_{\mathrm{PHD}}, \boldsymbol{\beta}_1)$	$r(\hat{\boldsymbol{\beta}}_{\mathrm{PHD}}, \boldsymbol{\beta}_2)$	$r(\hat{\boldsymbol{\beta}}_{\mathrm{SOLS}}, \boldsymbol{\beta}_1)$	$r(\hat{\boldsymbol{\beta}}_{\mathrm{SOLS}}, \boldsymbol{\beta}_2)$
200	0.712 (0.025)	0.399 (0.032)	0.213 (0.001)	0.830 (0.026)
400	0.663 (0.025)	0.287 (0.017)	0.144 (0.003)	0.866 (0.020)
800	0.649 (0.021)	0.207 (0.005)	0.103 (0.001)	0.831 (0.025)

Table 2.4: *Mean and Variance of the Distance r for Model I, Model II and Model III, $n = 800$, $p = 10$ with 100 repetitions. $H = 2, 5, 8$.*

	H=2		H=5		H=10	
	$r(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_1)$	$r(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_2)$	$r(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_1)$	$r(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_2)$	$r(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_1)$	$r(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_2)$
Model I	0.269	0.919	0.334	0.883	0.366	0.861
	(0.005)	(0.008)	(0.010)	(0.013)	(0.011)	(0.015)
Model II	0.106	0.265	0.166	0.287	0.232	0.301
	(0.001)	(0.004)	(0.003)	(0.006)	(0.005)	(0.011)
Model III	0.132	0.832	0.146	0.883	0.179	0.923
	(0.001)	(0.016)	(0.002)	(0.011)	(0.002)	(0.008)

$\boldsymbol{\Sigma} = (\sigma_{ij})_{p\times p}$, $\sigma_{ij} = 0.5^{|i-j|}$.

By testing different numbers of slices, we found that SOLS is sensitive to the slice number. The result is better when $H = 2$ for all three models. As we expected, the SOLS method is able to recover the CMS with predictors correlation present.

2.6.2 Order Determination

Now we run a simulation to determine the order d. As we mentioned before, a sequential hypothesis testing of $H_0^{(l)} : d = l$ v.s. $H_a^{(l)} : d > l$ is conducted for Models I, II, III and the results are reported in Table 2.5. The probability of rejecting the null hypothesis is reported based on 100 repetitions with nominal level 0.05. Recalling that SOLS recovers directions only in the CMS. Thus, for model I , the order is $d = 1$ with only one direction in the CMS, $d = 2$ for

Table 2.5: *Report on Probabilities of Rejectiong* $H_0^{(l)}: d = l$ *based on Sequential Test of Order Determination with* $H = 5,\ 10$ *and* $n = 400,\ 800.$

		H=5			H=10		
	n	$\hat{d}=0$	$\hat{d}=1$	$\hat{d}=2$	$\hat{d}=0$	$\hat{d}=1$	$\hat{d}=2$
Model I	400	**1.00**	0.00	0.00	**0.89**	0.04	0.00
	800	**1.00**	0.04	0.00	**1.00**	0.04	0.01
Model II	400	**1.00**	**0.99**	0.13	**1.00**	**0.71**	0.12
	800	**1.00**	**1.00**	0.16	**1.00**	**1.00**	0.15
Model III	400	**0.99**	0.04	0.10	**0.99**	0.07	0.10
	800	**1.00**	0.03	0.13	**1.00**	0.04	0.12

model II. For model III, since SOLS cannot estimate the quadratic direction, we are expecting test order $d = 1$. We boldface the entries that reject the null hypothesis for easy reference. We can see that the proposed SOLS method can estimate the order d while rejecting the corresponding null hypothesis at around 1 when $n = 200$, or equal to 1 when $n = 800$, and accept the null hypothesis at around nominal level 0.05.

2.6.3 Marginal Coordinate Test

We apply the marginal coordinate test of SOLS in favor of variable selection through Models I to III. The results are shown in table 2.6 with the probabilities of rejecting the null hypothesis. Here rejecting a null hypothesis is equivalent to the tested predictor contributing to the function. Ideally, we want those X_k's in the model to have probability of been rejected as high as 1 and those not contributed has probability at the nominal level. Here we compare the results of variable selections through SIR and SOLS at nominal level $\alpha = 0.05$ and the potion of been rejected serve the estimators of the probabilities, which based on 100 repetitions,

Overall, the results show that both SIR and SOLS asymptotic tests are able to test predictor contribution, but the difference is SOLS only tests those predictors that contributed to CMS, while SIR accounts for all of the predictors in CS. This phenomenon can be seen in Model I, where X_2 can be tested only by SIR, because it does not contribute to the CMS. In the case of Model II, both SIR and are SOLS reject the null hypothesis with probabilities around 1, and not able to reject the null hypothesis for X_2 to X_{10} at probabilities around $\alpha = 0.05$. The results of Model III show us both SIR and SOLS cannot test predictors X_3, X_4 which has a quadratic trend.

Table 2.6: *Report on Probabilities of Rejecting* $H_0^{[k]} : k \in \mathcal{A}^c$ *based on Asymptotic Test on SOLS with* $H = 2$, $n = 200,\ 400,\ 800$.

	n		X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
Model I	200	SIR	**1.00**	**0.92**	0.12	0.05	0.06	0.07	0.07	0.06	0.11	0.05
		SOLS	**0.99**	0.12	0.08	0.06	0.06	0.09	0.03	0.07	0.05	0.03
	400	SIR	**1.00**	**1.00**	0.05	0.07	0.10	0.07	0.05	0.03	0.03	0.03
		SOLS	**1.00**	0.09	0.03	0.06	0.03	0.05	0.07	0.05	0.04	0.05
	800	SIR	**1.00**	**1.00**	0.07	0.06	0.04	0.03	0.04	0.03	0.09	0.11
		SOLS	**1.00**	0.03	0.02	0.01	0.04	0.06	0.04	0.07	0.06	0.09
Model II	200	SIR	**1.00**	**1.00**	0.09	0.05	0.07	0.09	0.06	0.07	0.07	0.07
		SOLS	**1.00**	**0.99**	0.04	0.08	0.03	0.05	0.06	0.09	0.06	0.07
	400	SIR	**1.00**	**1.00**	0.07	0.02	0.07	0.03	0.07	0.06	0.03	0.05
		SOLS	**1.00**	**1.00**	0.05	0.07	0.08	0.03	0.02	0.05	0.02	0.05
	800	SIR	**1.00**	**1.00**	0.08	0.05	0.05	0.03	0.05	0.05	0.02	0.06
		SOLS	**1.00**	**1.00**	0.04	0.06	0.05	0.08	0.09	0.08	0.03	0.07
Model III	200	SIR	**1.00**	**1.00**	0.12	0.08	0.05	0.01	0.09	0.06	0.05	0.04
		SOLS	**1.00**	**0.99**	0.16	0.11	0.03	0.10	0.12	0.12	0.05	0.05
	400	SIR	**1.00**	**1.00**	0.15	0.15	0.08	0.05	0.08	0.05	0.02	0.04
		SOLS	**1.00**	**1.00**	0.09	0.15	0.04	0.08	0.09	0.05	0.02	0.04
	800	SIR	**1.00**	**1.00**	0.19	0.22	0.07	0.07	0.06	0.05	0.04	0.11
		SOLS	**1.00**	**1.00**	0.18	0.23	0.09	0.04	0.10	0.07	0.03	0.04

CHAPTER 3

HYBRID OLS, THE PROPOSED SECOND METHOD

3.1 Motivation

Recall that the disadvantage of all the sliced methods is the sensitivity to the number of slices H. In fact, all sliced methods involve a pre-determination of the number of slices H, which leads to a question of choosing the best H. The number of slices also relates to the maximum number of directions found, and SOLS can estimate at most H directions of CMS. Slice methods are limited by the number of possible response value when the response is a categorical variable. For example, when Y is categorical data with 4 possible values, then the possible number of slices will be no more than 4. In this chapter, we propose a new approach named Hybrid Ordinary Least Squares (HOLS) which does not need to take slices.

3.2 HOLS Estimator for the CMS

Theorem 3.2.1. *Let* $\mathbf{Z}$ *have mean* $\mathrm{E}(\mathbf{Z}) = \mathbf{0}$*, and* $\mathrm{Var}(\mathbf{Z}) = \mathbf{I}_p$*,* $(\tilde{\mathbf{Z}}, \tilde{Y})$ *be the independent copy of* $(\mathbf{Z}, Y)$*, where* $\mathbf{Z}, \tilde{\mathbf{Z}} \in \mathbb{R}^p$ *and* $Y, \tilde{Y} \in \mathbb{R}$*. Under the assumption (1.2),*

$$\mathrm{Span}\Big(\mathrm{E}(\mathbf{Z}\tilde{\mathbf{Z}}'Y\tilde{Y}|\boldsymbol{\beta}_0'\mathbf{Z} - \boldsymbol{\beta}_0'\tilde{\mathbf{Z}}|)\Big) \subseteq \mathcal{S}_{\mathrm{E}(Y|\mathbf{Z})},$$

where $\boldsymbol{\beta}_{0,\mathbf{Z}} = E(\mathbf{Z}\mathbf{Y}') \in \mathbb{R}^{p\times q}$ *denotes the OLS estimator* $\boldsymbol{\beta}_{OLS}$ *of* $\mathbf{Z}$*-scale.*

With the invariance property, for $\mathbf{X}$ with $\mathrm{Var}(\mathbf{X}) = \boldsymbol{\Sigma}$, we can get the conclusion $\boldsymbol{\Sigma}^{-1}\mathrm{E}(\mathbf{X}\tilde{\mathbf{X}}'Y\tilde{Y}\ |\boldsymbol{\beta}_0'\mathbf{X} - \boldsymbol{\beta}_0'\tilde{\mathbf{X}}|)\boldsymbol{\Sigma}^{-1} \subseteq \mathcal{S}_{\mathrm{E}(Y|\mathbf{X})}$.

We conclude the algorithm below, which breaks down the method at the sample level. Let $\mathbf{X}_i \in \mathbb{R}^p$, where $i = 1, \ldots, n$, be the independent sample of p-dimensional vectors. Let $Y_i \in \mathbb{R}$ be the random univariate response variable.

Algorithm

i Let $\hat{\mathbf{Z}}_i = \hat{\boldsymbol{\Sigma}}^{-1/2}(\mathbf{X}_i - \hat{\boldsymbol{\mu}})$.

ii Calculate $\hat{\boldsymbol{\beta}}_0 = \frac{1}{n}\sum_{i=1}^n \hat{\mathbf{Z}}_i Y_i$.

iii Let $\hat{\mathbf{m}}_i = \hat{\mathbf{Z}}_i Y_i$, $\hat{\mathbf{m}}_j = \hat{\mathbf{Z}}_j Y_j$, and $g(\hat{\mathbf{Z}}_i, \hat{\mathbf{Z}}_j) = |\hat{\boldsymbol{\beta}}_0'\hat{\mathbf{Z}}_i - \hat{\boldsymbol{\beta}}_0'\hat{\mathbf{Z}}_j|$, where $i, j = 1, \ldots, n$. Then define $\hat{\mathbf{V}} = \frac{1}{n(n-1)}\sum_{i,j=1}^n \hat{\mathbf{m}}_i\hat{\mathbf{m}}_j' g(\hat{\mathbf{Z}}_i, \hat{\mathbf{Z}}_j)$.

iv Conduct an eigenvalue decomposition on the kernel matrix $\widehat{\mathbf{M}}_{HOLS}$, where $\widehat{\mathbf{M}}_{HOLS} = \hat{\mathbf{V}}\hat{\mathbf{V}}'$, to find the largest d eigenvalue corresponding eigenvectors $\hat{\boldsymbol{\eta}}_k, k = 1, ..., d$.

v By Proposition 1.1.1, transfer back to get $\hat{\boldsymbol{\beta}}_k = \hat{\boldsymbol{\Sigma}}^{-1/2}\hat{\boldsymbol{\eta}}_k$. Then $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \ldots, \hat{\boldsymbol{\beta}}_d)$.

3.3 Multivariate Response

Real world analysis involves not only univariate responses but also multivariate responses. Let $\mathbf{Y} \in \mathbb{R}^q$ be a q-dimensional response variable. Let $\mathbf{X} \in \mathbb{R}^p$ be a p-dimensional predictor. The dimension reduction problem for multivariate response is to find a reduced column space of $\boldsymbol{\beta}'\mathbf{X}$ conditionally on which the q-dimensional response $\mathbf{Y}$ is independent of the p-dimensional predictor $\mathbf{X}$. Denote as $\mathbf{Y} \perp\!\!\!\perp \mathbf{X}|\boldsymbol{\beta}'\mathbf{X}$. Furthermore, finding $\boldsymbol{\beta}'\mathbf{X}$ reduces the regression mean, which makes the equation below hold: $\mathrm{E}(\mathbf{Y}|\mathbf{X}) = \mathrm{E}(\mathbf{Y}|\boldsymbol{\beta}'\mathbf{X})$. In this section, we extend the HOLS method into a multivariate response case and show the results in a numerical study.

Recall Theorem 3.2.1, $\mathrm{E}(\mathbf{Z}\tilde{\mathbf{Z}}'Y\tilde{Y}||\boldsymbol{\beta}_0'\mathbf{Z} - \boldsymbol{\beta}_0'\tilde{\mathbf{Z}}||) \in \mathcal{S}_{\mathrm{E}(Y|\mathbf{Z})}$. Similarly we can get Theorem 3.3.1 when$\mathbf{Y}$ is a multivariate response. The proof is similar and thus omitted.

Theorem 3.3.1. *Let* $(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}})$ *be the independent copy of* $(\mathbf{X}, \mathbf{Y})$*, where* $\mathbf{X}, \tilde{\mathbf{X}} \in$

$\mathbb{R}^p$ *and* $\mathbf{Y}, \tilde{\mathbf{Y}} \in \mathbb{R}^q$*. Under the assumption (1.2),* $\mathrm{Span}\big(\boldsymbol{\Sigma}^{-1}\mathrm{E}(\mathbf{X}\mathbf{Y}'\tilde{\mathbf{Y}}\tilde{\mathbf{X}}'||\boldsymbol{\beta}_0'\mathbf{X} - \boldsymbol{\beta}_0'\tilde{\mathbf{X}}||)\boldsymbol{\Sigma}^{-1}\big) \subseteq \mathcal{S}_{\mathrm{E}(\mathbf{Y}|\mathbf{X})}$*, where* $\boldsymbol{\beta}_{0,\mathbf{Z}} = E(\mathbf{Z}\mathbf{Y}') \in \mathbb{R}^{p\times q}$ *denotes the OLS estimator* $\boldsymbol{\beta}_{OLS}$ *of* $\mathbf{Z}$*-scale.*

The sample level algorithm is listed below.

Algorithm

i Let $\hat{\mathbf{Z}}_i = \hat{\boldsymbol{\Sigma}}^{-1/2}(\mathbf{X}_i - \hat{\boldsymbol{\mu}})$.

ii Let $\tilde{\mathbf{Y}}_i' = (\tilde{Y}_{i,1}, \tilde{Y}_{i,2}, \ldots, \tilde{Y}_{i,q})$, where $\tilde{Y}_{i,s} = \sigma_s^{-1}(Y_{i,s} - \bar{Y}_s)$ with $s = 1, \ldots, q$, $\bar{Y}_s$ and σ_s^2 are mean and variance of Y_s respectively.

iii Calculate $\hat{\boldsymbol{\beta}}_0 = \frac{1}{n}\sum_{i=1}^n \mathbf{Z}_i\tilde{\mathbf{Y}}_i$, where $\hat{\boldsymbol{\beta}}_0 \in \mathbb{R}^{p\times q}$.

iv Let $\hat{\mathbf{m}}_i = \hat{\mathbf{Z}}_i\tilde{\mathbf{Y}}_i'$, $\hat{\mathbf{m}}_j = \hat{\mathbf{Z}}_j\tilde{\mathbf{Y}}_j'$, and $\hat{g}(\hat{\mathbf{Z}}_i, \hat{\mathbf{Z}}_j) = \|\hat{\boldsymbol{\beta}}_0'\hat{\mathbf{Z}}_i - \hat{\boldsymbol{\beta}}_0'\hat{\mathbf{Z}}_j\|$, where $i, j = 1, \ldots, n$. Then get the kernel matrix $\hat{\mathbf{M}}_{\text{m-HOLS}} = \sum_{i,j=1}^n \hat{\mathbf{m}}_i\hat{\mathbf{m}}_j'\hat{g}(\hat{\mathbf{Z}}_i, \hat{\mathbf{Z}}_j)$.

v Conduct an eigenvalue decomposition on the kernel matrix $\hat{\mathbf{M}}_{\text{m-HOLS}}$ for the largest d eigenvalue corresponding eigenvectors $\hat{\boldsymbol{\eta}}_k (k = 1, ..., d)$.

vi With the invariance property, transfer back to get the $\mathbf{X}$-scale basis $\hat{\boldsymbol{\beta}}_k = \hat{\boldsymbol{\Sigma}}^{-1/2}\hat{\boldsymbol{\eta}}_k$, and $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_d)$.

3.4 Numerical Study

We have introduced two new dimension reduction methods based on the regression mean. To demonstrate how HOLS works, we walk through the same models in Chapter 2 and compare HOLS with SOLS side by side.

Recall the models in Chapter 2:

Model (I) — Sine Link Function: $Y = sin(X_1) + (X_2 + 1) \cdot \varepsilon$;

Model (II) — Rational Link Function: $Y = \frac{X_1}{0.5+(X_2+1.5)^2} + 0.1 \cdot \varepsilon$;

Model (III) — Hybrid Cubic and Quadratic Link Function:

$Y = 0.5 \cdot (X_1 + X_2)^3 + (X_3 + X_4)^2 + 0.1 \cdot \varepsilon$.

In the comparison of SOLS and HOLS, the number of observations is fixed to $n = 800$ and the simulation is run over 100 times. SOLS is applied with a different number of slices $H = 2,\ 5$ and 10. All $\mathbf{X}_i$'s, $i = 1, \ldots, 10$, and ε are generated from the standard multivariate normal $N(0, \Sigma)$. The means and the variances of the distance r, which are defined in Chapter 2 are reported in table 3.1. In SOLS, $\hat{\boldsymbol{\beta}}$ represents $\hat{\boldsymbol{\beta}}_{\text{SOLS}}$ and in HOLS, $\hat{\boldsymbol{\beta}}$ represents $\hat{\boldsymbol{\beta}}_{\text{HOLS}}$.

In the results from table 3.1, we can see that HOLS estimator behaves similar as the SOLS estimator. The results from model I shows that both methods recovery the CMS, not the CS. And by extend the OLS, both successfully find 2 directions in the CMS in model II. As we expect, HOLS enjoys the advantage of OLS, which estimate well when linear trend exist, but facing the challenge of quadratic trend. The results of model (II) also finds SOLS is

Table 3.1: *Mean and Variance of the Distance r for Model (I), Model (II) and Model (III), $d = 2$, $p = 10$ with 100 repetitions.*

Model	SOLS H=2 $r(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_1)$	SOLS H=2 $r(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_2)$	SOLS H=5 $r(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_1)$	SOLS H=5 $r(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_2)$	SOLS H=10 $r(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_1)$	SOLS H=10 $r(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_2)$	HOLS $r(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_1)$	HOLS $r(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_2)$
I	0.239	0.891	0.236	0.909	0.236	0.926	0.239	0.856
	(0.006)	(0.015)	(0.004)	(0.011)	(0.005)	(0.007)	(0.004)	(0.019)
II	0.092	0.251	0.092	0.576	0.093	0.838	0.102	0.239
	(0.001)	(0.005)	(0.001)	(0.023)	(0.001)	(0.020)	(0.001)	(0.004)
III	0.103	0.831	0.105	0.930	0.112	0.931	0.125	0.881
	(0.001)	(0.025)	(0.001)	(0.007)	(0.001)	(0.008)	(0.001)	(0.017)

sensitive to the number of slices. This phenomenon gives us an incentive to use HOLS over SOLS in certain circumstances.

We further demonstrate HOLS by looking at a multivariate response numerical study.

Let $\mathbf{X} \in \mathbb{R}^{10}$ with mean $\mathbf{0}$ and covariance matrix $\mathbf{I}_{10}$. Let $\mathbf{Y}' = (Y_1, Y_2)$ be a multivariate response variable with $\mathbf{Y} \in \mathbb{R}^2$. Conduct the following model.

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ \\ Y_2 \end{bmatrix} \in \mathbb{R}^2 \qquad \mathbf{X} = \begin{bmatrix} X_1 \\ \\ \vdots \\ \\ X_{10} \end{bmatrix} \in \mathbb{R}^{10}$$

$$\begin{cases} Y_1 = \frac{X_1}{0.5+(X_2+1.5)^2} + 0.1 \cdot \varepsilon \\ \\ Y_2 = sin(X_3) + 0.1 \cdot \varepsilon \end{cases}$$

$\varepsilon \sim N(0,1)$. True dimension reduction space is Span$(\boldsymbol{\beta}) \in \mathbb{R}^{10\times 3}$, where $\boldsymbol{\beta} =$

Table 3.2: *Mean and Variance of the Distance r for the multivariate response model, $p = 10$, $d = 3$ with 100 repetitions.*

n	$r(\hat{\boldsymbol{\beta}}_{\text{HOLS}}, \boldsymbol{\beta}_1)$	$r(\hat{\boldsymbol{\beta}}_{\text{HOLS}}, \boldsymbol{\beta}_2)$	$r(\hat{\boldsymbol{\beta}}_{\text{HOLS}}, \boldsymbol{\beta}_3)$
200	0.185	0.479	0.083
	(0.002)	(0.017)	(0.001)
400	0.139	0.338	0.059
	(0.002)	(0.010)	(0.000)
800	0.094	0.235	0.041
	(0.001)	(0.004)	(0.000)

$(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3)$ with $\boldsymbol{\beta}_1' = (1, 0, 0, \dots)$, $\boldsymbol{\beta}_2' = (0, 1, 0, \dots)$ and $\boldsymbol{\beta}_3' = (0, 0, 1, \dots)$.

Applying the HOLSthe HOLS to a multivariate study. We report the distance r in table 3.2. With the distance between estimate space $\hat{\boldsymbol{\beta}}_{\text{HOLS}}$ and true directions close to zero, HOLS is able to find all three directions. By increasing the sample size n, the results improve. HOLS proves to be good at estimating the CMS when response variable $\mathbf{Y}$ is a multivariate variable.

Data visualization can be done by a set of scatter plots of marginal relationship of Y_s versus $\hat{\boldsymbol{\beta}}_k \mathbf{X}$, $s = 1, \dots, q,\ k = 1, \dots, d$. With $q = 2$ and $d = 3$, six plots are shown here in figure (3.1). The first column of scatter plots (a)-(c) are plots for Y_1, and we find (b) and (c) show strong patterns. Similarly for column (d)-(f), plot (d) shows a relationship between Y_2 and $\hat{\boldsymbol{\beta}}_1 \mathbf{X}$.

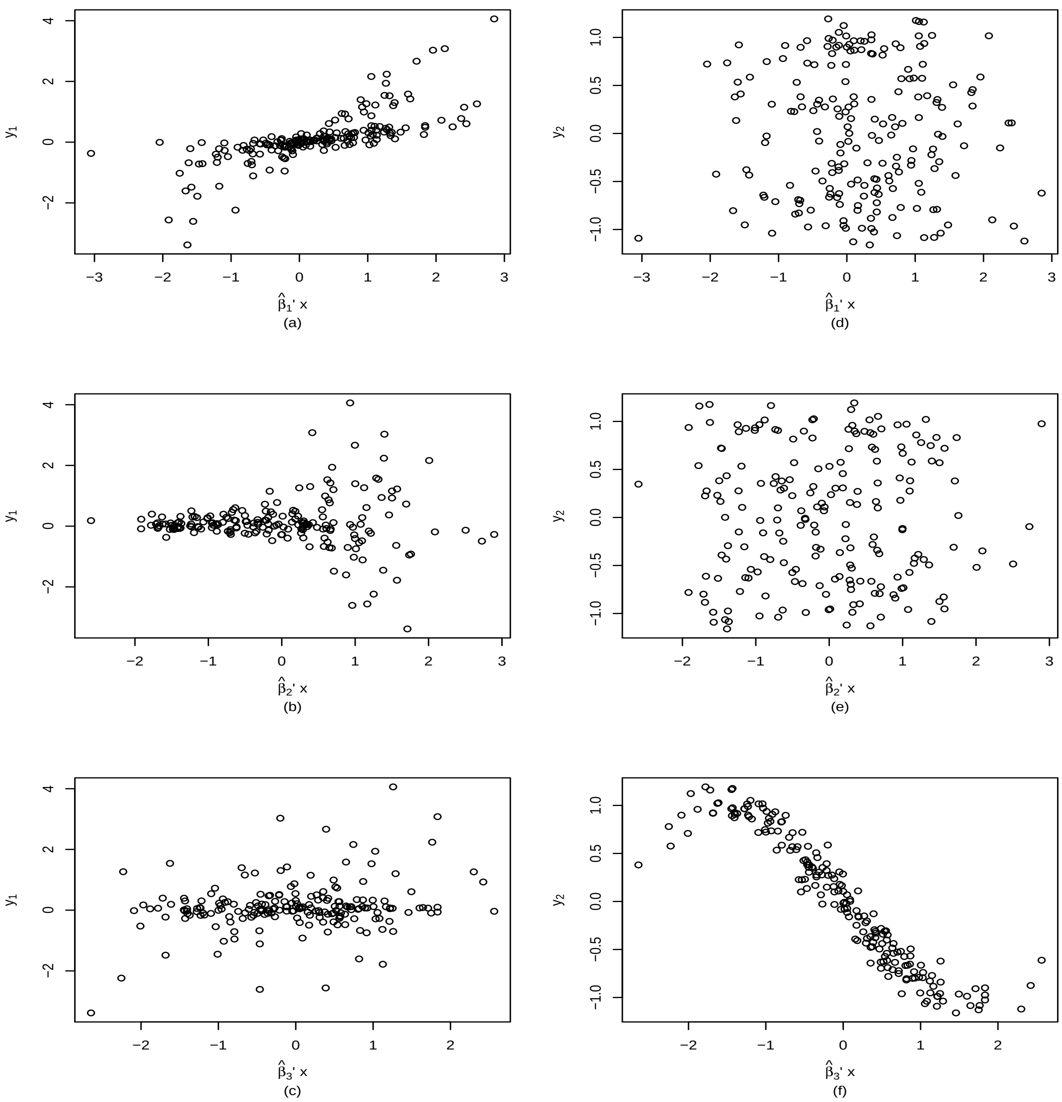

Figure 3.1: *(a) Scatter plot of* Y_1 *versus* $\hat{\boldsymbol{\beta}}_1'\mathbf{X}$*, (b) Scatter plot of* Y_1 *versus* $\hat{\boldsymbol{\beta}}_2'\mathbf{X}$*,(c) Scatter plot of* Y_1 *versus* $\hat{\boldsymbol{\beta}}_3'\mathbf{X}$*,(d) Scatter plot of* Y_2 *versus* $\hat{\boldsymbol{\beta}}_1'\mathbf{X}$*,(e) Scatter plot of* Y_2 *versus* $\hat{\boldsymbol{\beta}}_2'\mathbf{X}$*,(f) Scatter plot of* Y_2 *versus* $\hat{\boldsymbol{\beta}}_3'\mathbf{X}$*, for multivariate response model.*

CHAPTER 4

CONCLUSION AND FUTURE WORK

4.1 Summary

Dimension reduction of the regression mean has been an important topic in the field of dimension reduction. With the development of database management system (DBMS) and other technologies, more and more high dimensional data is stored and needs to be analyzed after reducing its dimension. Many of these analyses only yield the mean of the data.

This proposal introduced two regression methods: SOLS in Chapter 2 and HOLS in Chapter 3. Both methods are expanded from OLS. SOLS is regression with slicing. By taking slices, the SOLS algorithm is more computationally effective than a non-slice method. But the result can vary depending on the number of slices H. Furthermore, the maximum number of directions SOLS can get is H. HOLS can estimate the regression mean without taking slices. Also, HOLS can be more robust when the response variable is multivariate, because slicing a multivariate response causes the exponential increase of the number of slices. Neither method can estimate CMS when the link function is quadratic. Or in other words, when the first derivative of the conditional expectation of Y given $\mathbf{Z}$ is zero, both methods fail. But when compared with PHD, they are successful when the link function is linear.

We further utilize SOLS to conduct a sequential test for order determination and a marginal coordinate test for variable selection. Further expansion of HOLS for these tests is necessary.

BIBLIOGRAPHY

[1] Bickel, P. J., Klaassen, C. A., Ritov, Y., and Wellner, J. A. (1993), Efficient and Adaptive Estimation for Semiparametric Models, Baltimore: Johns Hopkins University Press.

[2] Cook, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions through Graphics*. New York: Wiley.

[3] Cook, R.D., Li, B. (2002). Dimension Reduction for the Conditional Mean in Regression. *The Annals of Statistics*, **30**, 455–474.

[4] Cook, R. D. and Li, B. (2004). Determining the Dimension of Iterative Hessian Transformation. *The Annals of Statistics*, **32**, 2501âĂŞ2531.

[5] Cook, R. D. and Weisberg, S. (1991). Discussion of "Sliced inverse regression for dimension reduction" by K. C. Li. *Journal of the American Statistical Association*, **86**, 328–332.

[6] Dong, Y. Yang, C. Yu, Z (2016). On permutation tests for predictor contribution in sufficient dimension reduction. J*ournal of Multivariate Analysis* **149** 81–91

[7] Dong, Y. and Yu, Z. (2012). Dimension Reduction for the Conditional kth Moment via Central Solution Space. *Journal of Multivariate Analysis*, **112**, 207-218.

[8] Fernholz, L. T. (1983), von Mises Calculus for Statistical Functionals, New York: Springer.

[9] Hampel, F. R. (1974), The Influence Curve and Its Role in Robust Estimation, *Journal of the American Statistical Association*, **69**, 383-397.

[10] Li, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association*, **86**, 316–342.

[11] Li, K. C. (1992). On principal Hessian directions for data visualization and dimension reduction: Another application of SteinâĂŹs lemma. *Journal of the American Statistical Association*, **87**, 1025-1039.

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

[13] Li, B. and Wang, S. (2007). On Directional Regression for Dimension

Reduction. *Journal of the American Statistical Association*, **102**, 997–1008.

[14] Shao, Y., Cook, R. D., and Weisberg, S. (2007). Marginal tests with sliced average variance estimation. *Biometrika*, **94**, 285–296.

[15] Serfling, R. (1980), Approximation Theorems of Mathematical Statistics, New York: Wiley.

[16] Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso *Journal of the Royal Statistical Society*, Series B (Methodological) **58**, No. 1, 267-288

[17] Xia, Q. and Dong, Y. (2017) On a New Hybrid Estimator for the Central Mean Space. *Journal of Systems Science and Complexity*, **30**, 111-121.

[18] Xia, Y., Tong, H., Li, W. K., and Zhu, L.-X. (2002). An Adaptive Estimation of Dimension Reduction Space. *Journal of the Royal Statistical Society*, Series. B, **64**,363-410.

one of a new class of dimension reduction methods. *Journal of the American Statistical Association*, **98**, 968–979.

[19] Yu, Z. and Dong, Y. (2016). Model-free coordinate test and variable selection via directional regression. *Statistica Sinica*, **26**,1159–1174.

APPENDIX

Proof of Theorem 1.2.3. Let Span($\boldsymbol{\eta}$) $= \mathcal{S}_{Y|\mathbf{Z}}$. Then, $\mathrm{E}(\mathbf{Z}|Y) = \mathrm{E}[\mathrm{E}(\mathbf{Z}|Y, \boldsymbol{\eta}'\mathbf{Z})|Y] = \mathrm{E}[\mathrm{E}(\mathbf{Z}|\boldsymbol{\eta}'\mathbf{Z})|Y]$. Let $P_{\boldsymbol{\eta}} = \boldsymbol{\eta}(\boldsymbol{\eta}'\boldsymbol{\eta})^{-1}\boldsymbol{\eta}'$ be the Euclidean projection. Under assumption (1.2), $\mathrm{E}(\mathbf{Z}|\boldsymbol{\eta}'\mathbf{Z}) = \mathcal{P}_{\boldsymbol{\eta}}\mathbf{Z}$, we can get $\mathrm{E}[\mathrm{E}(\mathbf{Z}|\boldsymbol{\eta}'\mathbf{Z})|Y] = \mathrm{E}(\mathcal{P}_{\boldsymbol{\eta}}\mathbf{Z}|Y) = \mathcal{P}_{\boldsymbol{\eta}}\mathrm{E}(\mathbf{Z}|Y) = \boldsymbol{\eta}(\boldsymbol{\eta}'\boldsymbol{\eta})^{-1}\boldsymbol{\eta}'\mathrm{E}(\mathbf{Z}|Y)$.

Proof of Theorem 1.2.1. By the law of total expectation $\mathrm{E}(\mathbf{Z}Y) = \mathrm{E}[\mathrm{E}(\mathbf{Z}Y|\mathbf{Z})] = \mathrm{E}[\mathbf{Z}\mathrm{E}(Y|\mathbf{Z})]$. Let Span($\boldsymbol{\eta}$) $= \mathcal{S}_{\mathrm{E}(Y|\mathbf{Z})}$, under the conditional independence of means implied by Definition 1.1.2, $\mathrm{E}(Y|\mathbf{Z}) = \mathrm{E}(Y|\boldsymbol{\eta}'\mathbf{Z})$. Above equation can be rewritten as, $\mathrm{E}[\mathbf{Z}\mathrm{E}(Y|\mathbf{Z})] = \mathrm{E}[\mathbf{Z}\mathrm{E}(Y|\boldsymbol{\eta}'\mathbf{Z})]$. Followed by the self-adjoint property of projection operator, we can get below equation. See the proof of self-adjoint property. $\mathrm{E}[\mathbf{Z}\mathrm{E}(Y|\boldsymbol{\eta}'\mathbf{Z})] = \mathrm{E}[Y\mathrm{E}(\mathbf{Z}|\boldsymbol{\eta}'\mathbf{Z})] = \mathrm{E}(Y\mathcal{P}_{\boldsymbol{\eta}}\mathbf{Z}) = \mathcal{P}_{\boldsymbol{\eta}}\mathrm{E}(\mathbf{Z}Y)$. The last step was derived from assumption (1.2), which is $\mathrm{E}(\mathbf{Z}|\boldsymbol{\eta}'\mathbf{Z}) = \mathcal{P}_{\boldsymbol{\eta}}\mathbf{Z}$.

Proof of Theorem 1.2.2. By the law of total expectation, we can get $\mathrm{E}(Y\mathbf{Z}\mathbf{Z}') = \mathrm{E}[\mathrm{E}(Y\mathbf{Z}\mathbf{Z}'|'\mathbf{Z})] = \mathrm{E}[\mathrm{E}(Y|\mathbf{Z})\mathbf{Z}\mathbf{Z}']$. Let Span($\boldsymbol{\eta}$) $= \mathcal{S}_{\mathrm{E}(Y|\mathbf{Z})}$, then under Definition 1.1.2 , $\mathrm{E}(Y|\mathbf{Z}) = \mathrm{E}(Y|\boldsymbol{\eta}'\mathbf{Z})$ and self-adjoint property,

$\mathrm{E}[\mathrm{E}(Y|\boldsymbol{\eta}'\mathbf{Z})\mathbf{Z}\mathbf{Z}'] = \mathrm{E}[Y\mathrm{E}(\mathbf{Z}\mathbf{Z}'|\boldsymbol{\eta}'\mathbf{Z})]$, we have $\mathrm{E}[\mathrm{E}(Y|\mathbf{Z})\mathbf{Z}\mathbf{Z}'] = \mathrm{E}[\mathrm{E}(Y|\boldsymbol{\eta}'\mathbf{Z})\mathbf{Z}\mathbf{Z}'] = \mathrm{E}[Y\mathrm{E}(\mathbf{Z}\mathbf{Z}'|\boldsymbol{\eta}'\mathbf{Z})] = \mathrm{E}\{Y[\mathrm{Var}(\mathbf{Z}|\boldsymbol{\eta}'\mathbf{Z}) - \mathrm{E}(\mathbf{Z}|\boldsymbol{\eta}'\mathbf{Z})\mathrm{E}(\mathbf{Z}'|\boldsymbol{\eta}'\mathbf{Z})]\}$.

Recall that $\mathrm{E}(Y) = 0$. Under assumption $\mathrm{Var}(\mathbf{Z}|\boldsymbol{\eta}'\mathbf{Z})$ is a constant matrix, let $\mathrm{Var}(\mathbf{Z}|\boldsymbol{\eta}'\mathbf{Z}) = \mathbf{Q}_{\boldsymbol{\eta}}$, where $\mathbf{Q}_{\boldsymbol{\eta}} = \mathbf{I} - \mathcal{P}_{\boldsymbol{\eta}}$, we further break down above equation as follows. $\mathrm{E}\{Y[\mathrm{Var}(\mathbf{Z}|\boldsymbol{\eta}'\mathbf{Z}) - \mathrm{E}(\mathbf{Z}|\boldsymbol{\eta}'\mathbf{Z})\mathrm{E}(\mathbf{Z}'|\boldsymbol{\eta}'\mathbf{Z})]\} = \mathrm{E}\{Y[\mathbf{Q}_{\boldsymbol{\eta}} - \mathcal{P}_{\boldsymbol{\eta}}\mathbf{Z}(\mathcal{P}_{\boldsymbol{\eta}}\mathbf{Z})']\} = \mathrm{E}(Y\mathbf{Q}_{\boldsymbol{\eta}}) - \mathrm{E}(\mathcal{P}_{\boldsymbol{\eta}}Y\mathbf{Z}\mathbf{Z}'\mathcal{P}_{\boldsymbol{\eta}}') = \mathcal{P}_{\boldsymbol{\eta}}\mathrm{E}(Y\mathbf{Z}\mathbf{Z}')\mathcal{P}_{\boldsymbol{\eta}}'$. Thus, we have proved the column space from Hessian matrix $\boldsymbol{M}_{\mathrm{PHD}} = \mathrm{E}(Y\mathbf{Z}\mathbf{Z}')$ is subspace of $S_{\mathrm{E}(Y|\mathbf{Z})}$. Similarly, when $\mathrm{Var}(\mathbf{X}) = \boldsymbol{\Sigma}$, above theorem can be updated as $Span\{\mathrm{E}(Y\mathbf{X}\mathbf{X}')\} = Span\{\boldsymbol{\Sigma}^{-1}\mathrm{E}(Y\mathbf{X}\mathbf{X}')\boldsymbol{\Sigma}^{-1}\} \in \mathcal{S}_{\mathrm{E}(Y|\mathbf{X})}$.

Proof of Proposition 1.4.1. This proof can be found in Dong et al. (2016) and thus omitted.

Lemma 4.1.1. *Let U, V and W be random variables. Then the self-adjoint property of the condition mean is* $\mathrm{E}[\mathrm{E}(U|W)V] = \mathrm{E}[U\mathrm{E}(V|W)]$.

Proof of lemma 4.1.1. Proof of the self-adjoint property, $U \in \mathbb{R}$, $V \in \mathbb{R}$ and $W \in \mathbb{R}$. We have $\mathrm{E}[\mathrm{E}(U|W)V] = \mathrm{E}[\mathrm{E}[\mathrm{E}(U|W)V|W] = \mathrm{E}[\mathrm{E}(U|W)|\mathrm{E}(V|W)] = \mathrm{E}[\mathrm{E}[U\mathrm{E}(V|W)|W]] = \mathrm{E}[U\mathrm{E}(V|W)]$.

Proof of Theorem 2.2.1. Define $\boldsymbol{\beta}_{0,\mathbf{Z}} = \mathrm{E}(\mathbf{Z}Y)$. By the law of total expectation, $\mathrm{E}(\mathbf{Z}Y|\boldsymbol{\beta}_{0,\mathbf{Z}}'\mathbf{Z}) = \mathrm{E}[\mathrm{E}(\mathbf{Z}Y|\boldsymbol{\beta}_{0,\mathbf{Z}}'\mathbf{Z}, \mathbf{Z})|\boldsymbol{\beta}_{0,\mathbf{Z}}'\mathbf{Z}] = \mathrm{E}[\mathrm{E}(\mathbf{Z}Y|\mathbf{Z})|\boldsymbol{\beta}_{0,\mathbf{Z}}'\mathbf{Z}]$. Assuming $\mathrm{Span}(\boldsymbol{\eta}) = \mathcal{S}_{\mathrm{E}(Y|\mathbf{Z})}$, by Definition 1.1.2, $\mathrm{E}(Y|\mathbf{Z}) = \mathrm{E}(Y|\boldsymbol{\eta}'\mathbf{Z})$. Above equation can be rewritten as, $\mathrm{E}[\mathrm{E}(\mathbf{Z}Y|\mathbf{Z})|\boldsymbol{\beta}_{0,\mathbf{Z}}'\mathbf{Z}] = \mathrm{E}[\mathbf{Z}\mathrm{E}(Y|\boldsymbol{\eta}'\mathbf{Z})|\boldsymbol{\beta}_{0,\mathbf{Z}}'\mathbf{Z}]$.

Also, recall that $\boldsymbol{\beta}_{0,\mathbf{Z}} \in \mathcal{S}_{\mathrm{E}(Y|\mathbf{Z})}$, and self-adjoint property, $\mathrm{E}[\mathbf{Z}\mathrm{E}(Y|\boldsymbol{\eta}'\mathbf{Z})|\boldsymbol{\beta}_{0,\mathbf{Z}}'\mathbf{Z}] =$

$\mathrm{E}[\mathbf{Z}\mathrm{E}(Y|\boldsymbol{\eta}'\mathbf{Z},\boldsymbol{\beta}_{0,\mathbf{Z}}'\mathbf{Z})|\boldsymbol{\beta}_{0,\mathbf{Z}}'\mathbf{Z}] = \mathrm{E}[\mathrm{E}(\mathbf{Z}|\boldsymbol{\eta}'\mathbf{Z},\boldsymbol{\beta}_{0,\mathbf{Z}}'\mathbf{Z})Y|\boldsymbol{\beta}_{0,\mathbf{Z}}'\mathbf{Z}] = \mathrm{E}[\mathrm{E}(\mathbf{Z}|\boldsymbol{\eta}'\mathbf{Z})Y|\boldsymbol{\beta}_{0,\mathbf{Z}}'\mathbf{Z}]$. Assumption (1.2) implies $\mathrm{E}(\mathbf{Z}|\boldsymbol{\eta}'\mathbf{Z}) = \mathcal{P}_{\boldsymbol{\eta}}\mathbf{Z}$. We can get $\mathrm{E}[\mathbf{Z}Y|\boldsymbol{\beta}_{0,\mathbf{Z}}'\mathbf{Z}] = \mathrm{E}[\mathrm{E}(\mathbf{Z}|\boldsymbol{\eta}'\mathbf{Z})Y|\boldsymbol{\beta}_{0,\mathbf{Z}}'\mathbf{Z}] = \mathcal{P}_{\boldsymbol{\eta}}\mathrm{E}(\mathbf{Z}Y|\boldsymbol{\beta}_{0,\mathbf{Z}}'\mathbf{Z})$. Thus, we have proved that $\mathrm{E}(\mathbf{Z}Y|\boldsymbol{\beta}_{0,\mathbf{Z}}'\mathbf{Z}) \in \mathcal{S}_{\mathrm{E}(Y|\mathbf{Z})}$.

When $\mathbf{X}$ has variance $\boldsymbol{\Sigma}$, by the invariance property, we have $\boldsymbol{\Sigma}^{-1}\mathrm{E}(\mathbf{X}Y|\boldsymbol{\beta}_0'\mathbf{X}) = \boldsymbol{\Sigma}^{-1}\mathrm{E}(\mathbf{X}Y|\boldsymbol{\beta}_{0,\mathbf{Z}}'\mathbf{X}) \subseteq \boldsymbol{\Sigma}^{-1/2}\mathcal{S}_{\mathrm{E}(Y|\mathbf{Z})} = \mathcal{S}_{\mathrm{E}(Y|\mathbf{X})}$.

Proof of Corollary 2.2.1. Denote $\boldsymbol{\mu_h} = \mathrm{E}(\mathbf{Z}Y|\boldsymbol{\beta}_{0,\mathbf{Z}}'\mathbf{Z} \in J_h)$, by Theorem 2.2.1 $\boldsymbol{\mu_h} \in \mathrm{Span}(\boldsymbol{\eta}) = \mathcal{S}_{\mathrm{E}(\mathbf{Z}|Y)}$. Let $\boldsymbol{\mu_h} = \boldsymbol{\eta}\mathbf{c_h}$, where $\mathbf{c_h} \in \mathbb{R}^d$. Then, $\mathrm{E}[\mathrm{E}(\mathbf{Z}Y|\boldsymbol{\beta}_{0,\mathbf{Z}}'\mathbf{Z})\mathrm{E}'(\mathbf{Z}Y|\boldsymbol{\beta}_{0,\mathbf{Z}}'\mathbf{Z})] = \sum_{h=1}^H P(\boldsymbol{\beta}_{0,\mathbf{Z}}'\mathbf{Z} \in J_h)\boldsymbol{\mu_h}\boldsymbol{\mu_h}' = \sum_{h=1}^H P(\boldsymbol{\beta}_{0,\mathbf{Z}}'\mathbf{Z} \in J_h)\boldsymbol{\eta}\mathbf{c_h}(\boldsymbol{\eta}\mathbf{c_h})' = \boldsymbol{\eta}[\sum_{h=1}^H P(\boldsymbol{\beta}_{0,\mathbf{Z}}'\mathbf{Z} \in J_h)\mathbf{c_h}\mathbf{c_h}']\boldsymbol{\eta}'$, which is in the form of eigendecomposition. Then $Span\{\mathrm{E}[\mathrm{E}(\mathbf{Z}Y|\boldsymbol{\beta}_{0,\mathbf{Z}}'\mathbf{Z})\mathrm{E}'(\mathbf{Z}Y|\boldsymbol{\beta}_{0,\mathbf{Z}}'\mathbf{Z})]\} \subseteq \mathcal{S}_{\mathrm{E}(Y|\mathbf{Z}}$. By the invariance property, we have $Span\{\boldsymbol{\Sigma}^{-1}\mathrm{E}[\mathrm{E}(\mathbf{X}Y|\boldsymbol{\beta}_0'\mathbf{X})\mathrm{E}'(\mathbf{X}Y|\boldsymbol{\beta}_0'\mathbf{X})]\boldsymbol{\Sigma}^{-1}\} \subseteq \mathcal{S}_{\mathrm{E}(Y|\mathbf{X}}$..

Proof Remark 2.2.1.

Recall Remark 1.2.4 that SIR can estimate at most $H-1$ directions, because the restriction on conditional expectation took 1 degree of freedom. The restriction is

$$\sum_{h=1}^{H}\mathrm{E}(\mathbf{Z}|Y \in J_h)P(Y \in J_h) = \mathrm{E}(\mathrm{E}(\mathbf{Z}|Y)) = \mathrm{E}(\mathbf{Z}) = \mathbf{0}.$$

However, $\mathrm{E}(\mathbf{Z}Y) \neq \mathbf{0}$, we will not have this restriction in SOLS, which makes the maximum number of directions SOLS can reach H.

Proof of Lamma 2.3.1. For the notion of Frechet derivative, see, for example, Fernholz (1983). We refer to $G^*(F)$ as the Frechet derivative of $G(F)$. Following Yu and Dong (2016), we have $p_h^* = R_h - p_h$, $\Sigma^* = \mathbf{XX}' - \mathrm{E}(\mathbf{XX}')$, $(p_h^{-1/2})^* = -\frac{1}{2}p_h^{-3/2}p_h^* = -\frac{1}{2}p_h^{-3/2}(R_h - p_h)$, $(\mathbf{\Sigma}^{-1})^* = -\mathbf{\Sigma}^{-1}\mathbf{\Sigma}^*\mathbf{\Sigma}^{-1} = -\mathbf{\Sigma}^{-1}[\mathbf{XX}' - \mathrm{E}(\mathbf{XX}')]\mathbf{\Sigma}^{-1}$, $U_h^* = \mathrm{E}^*[(\mathbf{X} - \mathrm{E}(\mathbf{X})YR_h] = \mathrm{E}^*(\mathbf{X}YR_h) - \mathrm{E}^*(\mathbf{X})\mathrm{E}(YR_h) - \mathrm{E}(\mathbf{X})\mathrm{E}^*(YR_h) = \mathbf{X}YR_h - \mathrm{E}(\mathbf{X}YR_h) - \mathbf{X}\mathrm{E}(YR_h)$.

Proof of Theorem 2.3.1. Followed by the result of Lamma 2.3.1. Let $\mathbf{A}_h = \mathbf{V}_h^*$, Theorem 2.3.1 can be derived directly by Yu and Dong (2016),

$$\mathbf{V}_h^* = (p_h^{-1/2})^*\mathbf{\Sigma}^{-1}\mathbf{U}_h + p_h^{-1/2}(\mathbf{\Sigma}^{-1})^*\mathbf{U}_h + p_h^{-1/2}\mathbf{\Sigma}^{-1}\mathbf{U}_h^*$$

Proof of Theorem 2.4.1. The proof is similar to Theorem 3 in Xia and Dong (2016), and thus omitted.

Proof of Proposition 2.5.1. The proof is similar to Proposition 3.1. in Dong et al. (2016), and thus omitted.

Proof of Proposition 2.5.2. The proof is similar to Proposition 3.2. in Dong et al. (2016), and thus omitted.

Proof of Theorem 2.5.1. Let $\mathbf{t}_k = (t_{k,1}, t_{k,2}, \ldots, t_{k,H})'$ and $\hat{\mathbf{t}}_k = (\hat{t}_{k,1}, \hat{t}_{k,2}, \ldots, \hat{t}_{k,H})'$, where $t_{k,h} = \mathbf{e}_k'\mathbf{V}_h = p_h^{-1/2}\mathbf{e}_i'\mathbf{\Sigma}^{-1}\mathbf{U}_h$ and $\hat{t}_{k,h} = \mathbf{e}_k'\hat{\mathbf{V}}_h = \hat{p}_h^{-1/2}\mathbf{e}_k'\hat{\mathbf{\Sigma}}^{-1}\hat{\mathbf{U}}_h$. Then we can write $(t_{k,h})^*$ as follows: $(t_{k,h})^* = (p_h^{-1/2})^*\mathbf{e}_k'\mathbf{\Sigma}^{-1}\mathbf{U}_h + p_h^{-1/2}\mathbf{e}_k'(\mathbf{\Sigma}^{-1})^*\mathbf{U}_h + p_h^{-1/2}\mathbf{e}_k'\mathbf{\Sigma}^{-1}(\mathbf{U}_h)^*$. Knowing when $k \in \mathcal{A}^c$, $\mathbf{e}_k'\mathbf{\Sigma}^{-1}\mathbf{U}_h = 0$, then under H_0, $(t_{k,h})^* = g_{kh} = p_h^{-1/2}\mathbf{e}_k'(\mathbf{\Sigma}^{-1})^*\mathbf{U}_h + p_h^{-1/2}\mathbf{e}_k'\mathbf{\Sigma}^{-1}(\mathbf{U}_h)^*$. Thus, $\sqrt{n}(\hat{\mathbf{t}}_k - \mathbf{t}_k) \xrightarrow{D} \mathcal{N}(0, \mathrm{Cov}(\mathbf{G}_k))$. Substitute $(\mathbf{t}_k)^* = \mathbf{G}_k = (g_{k1}, g_{k2}, \ldots, g_{kh})$ into the test statistic of SOLS, we can get $T_k^{SOLS}(e_k) = \mathbf{t}_k'\mathbf{t}_k$. Since $\mathbf{t}_k = 0$

under H_0: $k \in \mathcal{A}^c$, $T_k^{SOLS}(e_k) \xrightarrow{D} \sum_{h=1}^{H} \gamma_{kh}\chi_h^2(1)$, where $(\gamma_1, \gamma_2, \ldots, \gamma_H)$ are eigenvalues of $\mathrm{Cov}(\mathbf{G}_k)$.

Proof of Theorem 3.2.1. By the law of total expectation, $\mathrm{E}[\mathbf{Z}\tilde{\mathbf{Z}}'Y\tilde{Y}|\boldsymbol{\beta}_0'\mathbf{Z}-\boldsymbol{\beta}_0'\tilde{\mathbf{Z}}|] = \mathrm{E}[\mathrm{E}(\mathbf{Z}\tilde{\mathbf{Z}}'Y\tilde{Y}\ |\boldsymbol{\beta}_0'\mathbf{Z}-\boldsymbol{\beta}_0'\tilde{\mathbf{Z}}|\ |\boldsymbol{\beta}_0'\mathbf{Z}, \boldsymbol{\beta}_0'\tilde{\mathbf{Z}})] = \mathrm{E}[\mathrm{E}(\mathbf{Z}\tilde{\mathbf{Z}}'Y\tilde{Y}|\boldsymbol{\beta}_0'\mathbf{Z}, \boldsymbol{\beta}_0'\tilde{\mathbf{Z}})\ |\boldsymbol{\beta}_0'\mathbf{Z}-\boldsymbol{\beta}_0'\tilde{\mathbf{Z}}|] = \mathrm{E}[\mathrm{E}(\mathbf{Z}Y|\boldsymbol{\beta}_0'\mathbf{Z})\mathrm{E}(\tilde{\mathbf{Z}}'\tilde{Y}|\boldsymbol{\beta}_0'\tilde{\mathbf{Z}})\ |\boldsymbol{\beta}_0'\mathbf{Z}-\boldsymbol{\beta}_0'\tilde{\mathbf{Z}}|]$. Recall Theorem 2.2.1, $\mathrm{E}(\mathbf{Z}Y|\boldsymbol{\beta}_0'\mathbf{Z}) = \mathcal{P}_\eta \mathrm{E}(\mathbf{Z}Y|\boldsymbol{\beta}_0'\mathbf{Z})$, where $\mathrm{Span}(\eta) = \mathcal{S}_{\mathrm{E}(Y|\mathbf{Z})}$. Then above equation can be re-written as $\mathrm{E}[\mathcal{P}_{\boldsymbol{\eta}}\mathrm{E}(\mathbf{Z}Y|\boldsymbol{\beta}_0'\mathbf{Z})\mathrm{E}(\tilde{\mathbf{Z}}'\tilde{Y}|\boldsymbol{\beta}_0'\tilde{\mathbf{Z}})\mathcal{P}_{\boldsymbol{\eta}}'\ |\boldsymbol{\beta}'(\mathbf{Z}-\tilde{\mathbf{Z}})|] = \mathcal{P}_{\boldsymbol{\eta}}\mathrm{E}(\mathbf{Z}\tilde{\mathbf{Z}}'Y\tilde{Y}\ |\boldsymbol{\beta}_0'\mathbf{Z}-\boldsymbol{\beta}_0'\tilde{\mathbf{Z}}|)\mathcal{P}_{\boldsymbol{\eta}}'$. It follows that $\mathrm{Span}\{\mathrm{E}(\mathbf{Z}\tilde{\mathbf{Z}}'Y\tilde{Y}|\boldsymbol{\beta}_0'\mathbf{Z} - \boldsymbol{\beta}_0'\tilde{\mathbf{Z}}|)\} \subseteq \mathcal{S}_{\mathrm{E}(Y|\mathbf{Z})}$.