2020

MOMENT KERNELS FOR T-CENTRAL SUBSPACE

Weihang Ren
University of Kentucky, weihang.ren@gmail.com
Digital Object Identifier: https://doi.org/10.13023/etd.2020.252

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Weihang Ren, Student

Dr. Xiangrong Yin, Major Professor

Dr. Katherine Thompson, Director of Graduate Studies

MOMENT KERNELS FOR T-CENTRAL SUBSPACE

DISSERTATION

A dissertation submitted in partial fulfillment of
the requirements for the degree of Doctor of
Philosophy in the College of Arts and Sciences
at the University of Kentucky

By
Weihang Ren

Lexington, Kentucky

Directors: Dr. Xiangrong Yin, Professor of Statistics

Lexington, Kentucky

2020

ABSTRACT OF DISSERTATION

MOMENT KERNELS FOR T-CENTRAL SUBSPACE

The T-central subspace allows one to perform sufficient dimension reduction for any statistical functional of interest. We propose a general estimator using a third moment kernel to estimate the T-central subspace. In particular, in this dissertation we develop sufficient dimension reduction methods for the central mean subspace via the regression mean function and central subspace via Fourier transform, central quantile subspace via quantile estimator and central expectile subsapce via expectile estimator. Theoretical results are established and simulation studies show the advantages of our proposed methods.

KEYWORDS: Central expectile subspace, Central mean subspace, Central quantile subspace, Central subspace, Sufficient dimension reduction.

Weihang Ren

April 7, 2020

MOMENT KERNELS FOR T-CENTRAL SUBSPACE

By
Weihang Ren

Dr. Xiangrong Yin
Director of Dissertation
Dr. Katherine Thompson
Director of Graduate Studies
April 7, 2020
Date

I dedicate this dissertation to
my family and
my beloved wife, Xu
for their constant support and unconditional love.
I love you all dearly.

ACKNOWLEDGMENTS

I would like to express my gratitude to my Ph.D. advisor, Dr. Xiangrong Yin, for his support and guidance on my research during this past five years. His technical advise, valuable feedback and encouragement on statistic and life was essential to the completion my graduate study and has taught me insights on the academic research and life. My Ph.D. would not have been achievable without his assistance and expertise. He's the nicest advisor and one of the smartest person I know. I simply cannot imagine a better advisor.

Furthermore, I want to thank the members of my Ph.D. committee, Dr. David Fardo, Dr. Solomon Harrar, Dr. Arnold Stromberg, and Dr. Derek Young, for their helpful career advice and suggestion in general.

Moreover, thanks to the University of Kentucky Statistics Department for providing me with a great learning/research platform and highly enjoyable five years, and in particular, Dr. Arnold Stromberg for his generous support.

Last, but not least, I would like to thank my family. My parents, Chunping and Yu, receive my deepest gratitude and love for their dedication and many years of unconditional love and support. My wife, Xu, deserve special thanks for her understanding and love during the past few years. Her support and encouragement was in the end what made this dissertation possible.

This dissertation is supported in part by National Science Foundation grant (CIF-1813330) awarded to Dr. Xiangrong Yin at University of Kentucky.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1 Introduction

In this chapter, we will review some basic concepts and main methodologies in Sufficient Dimension Reduction.

1.1 Sufficient Dimension Reduction

With the increasing high demand for handling high dimensional data, Sufficient Dimension Reduction (SDR) has become a fast developing research field due to its wide range of applications: data visualization; predictive modeling; machine learning, etc. Similar to other dimension reduction methodologies, SDR aims to organize the high dimensional variables so as to preserve the core information. With the purpose of extracting all the information of understanding the response variable, SDR tries to search for a few linear combination of the predictors such that there is no information loss in conditional distribution while achieving a lower rank of the data. Sufficiency is derived from conditional independence which is a core concept of describing all the information needed to understand relationship between response and predictors.

SDR was introduced by Li (1991), Cook (1996). We use the following introduction.

Definition 1.1. *Let* $\mathbf{X} : \Omega_{\mathbf{X}} \to \mathbb{R}^p$ *be a Borel measurable random vector,* $Y : \Omega_Y \to \mathbb{R}$ *be a Borel measurable random variable and* $\mathbf{P}_{\mathcal{S}} : \mathbb{R}^p \to \mathbb{R}^p$ *be a linear operator with* $\mathbf{P}_{\mathcal{S}}^2 = \mathbf{P}_{\mathcal{S}}$ *and* $\langle \mathbf{P}_{\mathcal{S}}\mathbf{x}_1, \mathbf{x}_2 \rangle = \langle \mathbf{x}_1, \mathbf{P}_{\mathcal{S}}\mathbf{x}_2 \rangle$ *for* $\forall \mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^p$ *such that* $\mathbf{P}_{\mathcal{S}}$ *project a* p *dimensional vector to a* d *dimensional linear subspace* $(d \leq p)$*,* $\mathcal{S} \subseteq \mathbb{R}^p$*, with respect to standard inner product. Sufficient dimension reduction (SDR) is concerned with the situation where the distribution of* Y *depend on* $\mathbf{X}$ *only through a linear projection of* $\mathbf{X}$*. That is*

$$Y \perp\!\!\!\perp \mathbf{X} \mid \mathbf{P}_{\mathcal{S}}\mathbf{X}.$$

Following the definition of SDR, $\mathcal{S}$ is called *sufficient dimension reduction subspace (SDR subspace)*. However, it is worthwhile to point out that the SDR subspace, $\mathcal{S}$, is not uniquely defined, for example $\mathbb{R}^p$ itself will always be a trivial solution. Therefore, it's natural to consider the smallest $\mathcal{S}$ such that the conditional independence condition holds. Yin, Li and Cook (2008) proved that as long as $\mathbf{X}$ was supported by an M-set, then the intersection of two SDR subspaces would be an SDR subspace.

Definition 1.2. *Let $\mathfrak{S}$ be the collection of all SDR subspace*

$$\mathfrak{S} = \{\mathcal{S} \subseteq \mathbb{R}^p : Y \perp\!\!\!\perp \mathbf{X} \mid \mathbf{P}_{\mathcal{S}}\mathbf{X}\}.$$

We assume $\cap\{\mathcal{S} : \mathcal{S} \in \mathfrak{S}\} \in \mathfrak{S}$, then $\cap\{\mathcal{S} : \mathcal{S} \in \mathfrak{S}\}$ is called central subspace (CS), and denote as $\mathcal{S}_{Y|\mathbf{X}}$. The dimension of $\mathcal{S}_{Y|\mathbf{X}}$ is called the structural dimension.

Cook (2007) showed that there were in fact two more equivalent ways to define the SDR subspace other than Definition 1.1, which are given by the following proposition:

Proposition 1.1. *Let $\mathbf{X} : \Omega_{\mathbf{X}} \to \mathbb{R}^p$ be a Borel measurable random vector, $Y : \Omega_Y \to \mathbb{R}$ be a Borel measurable random variable and $\mathcal{S}$ be a d dimensional subspace of $\mathbb{R}^p$. Then $Y \perp\!\!\!\perp \mathbf{X} \mid \mathbf{P}_{\mathcal{S}}\mathbf{X}$ if and only if one of the following two points holds:*

1. *$Y \mid \mathbf{X}$ is distributed as the same as $Y \mid \mathbf{P}_{\mathcal{S}}\mathbf{X}$,*
2. *$\mathbf{X} \perp\!\!\!\perp (Y, \mathbf{P}_{\mathcal{S}}\mathbf{X})$ is distributed as the same as $\mathbf{X} \perp\!\!\!\perp \mathbf{P}_{\mathcal{S}}\mathbf{X}$.*

The proof of Proposition 1.1 could be found in Cook (2007). The implication of Proposition 1.1 is interesting: it classifies and establishes three main approaches for SDR. Methodologies derived from Definition 1.1 depends on the joint distribution of $(Y, \mathbf{X})$ so they are called the *joint approach*; methodologies derived from statement 1 in Proposition 1.1 correspond to forward regression which are named after *forward approach* and the last statements will lead to a class of methods that named as the *inverse approach*.

1.2 Inference Targets in Sufficient Dimension Reduction

CS is a target for analysis as it's the smallest subspace of $\mathbb{R}^p$ which contains all the information between Y and $\mathbf{X}$. However, with the goal of analysis was set to infer about the regression function, which is known as the conditional expectation $\mathrm{E}(Y \mid \mathbf{X})$, a different subject may be of more interests. Instead of studying CS, which capture all the information in the conditional distribution of $P_{Y|\mathbf{X}}$, *central mean subspace (CMS)*, introduced by Cook and Li (2002) that only captures information in the conditional expectation itself is more of one's interest. The definition of CMS is also given similarly based on the statement of independence.

Definition 1.3. *Let $\mathbf{X} : \Omega_{\mathbf{X}} \to \mathbb{R}^p$ be a Borel measurable random vector, $Y : \Omega_Y \to \mathbb{R}$ be a Borel measurable random variable and $\mathbf{P}_{\mathcal{S}}$ be the projection operator to $\mathcal{S}$, where $\mathcal{S}$ is a d dimensional linear subspace with $d \leq p$. Central mean subspace (CMS), written as $\mathcal{S}_{\mathrm{E}(Y|\mathbf{X})}$, is defined as*

$$\begin{aligned}\mathcal{S}_{\mathrm{E}(Y|\mathbf{X})} &= \cap\{\mathcal{S} \subseteq \mathbb{R}^p : Y \perp\!\!\!\perp \mathrm{E}(Y \mid \mathbf{X}) \mid \mathbf{P}_{\mathcal{S}}\mathbf{X}\} \\ &= \cap\{\mathcal{S} \subseteq \mathbb{R}^p : \mathrm{E}(Y \mid \mathbf{X}) = \mathrm{E}(Y \mid \mathbf{P}_{\mathcal{S}}\mathbf{X})\}.\end{aligned}$$

The $\mathcal{S}_{\mathrm{E}(Y|\mathbf{X})}$ is more preferable than $\mathcal{S}_{Y|\mathbf{X}}$ in some of the application such as nonparametric regression, single index model or multi-index model, because $\mathcal{S}_{Y|\mathbf{X}}$ in these cases contains possible redundant information and $\mathcal{S}_{\mathrm{E}(Y|\mathbf{X})}$ could achieve greater data reduction, which generates the most parsimonious results.

In other applications, where the people interest in other aspects of the conditional distribution rather than the conditional expectation. Similar targets could be constructed to achieve the greatest reduction in terms of the dimension. Examples are: *central k-th moment subspace* from Yin and Cook (2002); *central variance subspace* from Zhu and Zhu (2009) and *central quantile subspace* from Kong and Xia (2012). Luo, Li and Yin (2014) proposed a more general framework which unified the above

approaches by noticing that $\mathrm{E}(Y \mid \mathbf{X})$, $\mathrm{Var}(Y \mid \mathbf{X})$ and $Q_\tau(Y \mid \mathbf{X})$ are all conditional statistical functionals of the conditional distribution $P_{Y|\mathbf{X}}$, where $Q_\tau(\cdot)$ represents the τ-th quantile of a given distribution. They introduced the idea of T-central subspace, with $\mathcal{T}$ stands for any conditional statistical functional of interest. The following sections will provide definition and discussion in greater detail.

1.3 T-Central Subspace

Let $(\Omega_{\mathbf{X}}, \mathfrak{F}_{\mathbf{X}}, P_{\mathbf{X}})$ and $(\Omega_Y, \mathfrak{F}_Y, P_Y)$ be two probability spaces, with $\Omega_{\mathbf{X}} \in \mathbb{R}^p$ and $\Omega_Y \in \mathbb{R}^1$, $\mathfrak{F}_{\mathbf{X}}$ and $\mathfrak{F}_Y$ are the corresponding σ-algebra for each probability space, $P_{\mathbf{X}}$ and P_Y are the corresponding probability measure which are absolutely continuous with respect to some σ-finite measure $\mu_{\mathbf{X}}$ and μ_Y. Moreover, let $(\mathbf{X}, Y)$ be the random vector taken value in $(\Omega_{\mathbf{X}} \times \Omega_Y, \mathfrak{F}_{\mathbf{X}} \times \mathfrak{F}_Y, P_{\mathbf{X}} \times P_Y)$, with density $f_{\mathbf{X}Y}$ with respect to $\mu_{\mathbf{X}} \times \mu_Y$. Let $\mathcal{G}$ be the family of such density and furthermore, we assume that this is a semiparametric family i.e. there exist $\Theta \subset \mathbb{R}^r$ and a family $\mathcal{F}$ of functions $f : \Omega_{\mathbf{X}} \times \Omega_Y \times \Theta \to \mathbb{R}^1$ such that $\mathcal{G} = \cup\{\mathcal{G}_\theta, \theta \in \Theta\}$ where $\mathcal{G}_\theta = \{f(\cdot, \cdot, \theta), f \in \mathcal{F}\}$. Moreover, if we let $f_{\mathbf{X}}$ be the derivative of $P_{\mathbf{X}}$ with respect to $\mu_{\mathbf{X}}$, then $f(\mathbf{x}, y, \theta) = f_{\mathbf{X}}(\mathbf{x})\eta(\mathbf{x}, y, \theta)$. That is, $\eta(\mathbf{x}, y, \theta)$ is the conditional density of Y given $\mathbf{X}$ containing all the information of θ.

Now let

$$\mathcal{G}_{Y|\mathbf{X}} = \{\eta(\cdot, \cdot, \cdot) : f(\cdot, \cdot, \cdot) \in \mathcal{F}\}, \mathcal{G}_{\mathbf{X}} = \{f_{\mathbf{X}}(\cdot) : f(\cdot, \cdot, \cdot) \in \mathcal{F}\}$$

We assume that $\mathcal{G}$ contains the true density of $(\mathbf{X}, Y)$. That is, there exist $\theta_0 \in \Theta$, $f_{\mathbf{X}_0} \in \mathcal{G}_{\mathbf{X}}$, and $\eta_0 \in \mathcal{G}_{Y|\mathbf{X}}$ such that, $f_0 = f_{\mathbf{X}_0}\eta_0$ is the true density of $(\mathbf{X}, Y)$.

Let $\mathcal{G}_{Y|\mathbf{X}} = \{\eta(\cdot, \cdot, \cdot) : f(\cdot, \cdot, \cdot) \in \mathcal{F}\}$, and let $\mathcal{H}$ be a class of densities with respect to μ_Y that contains all $\eta(\mathbf{x}, \cdot, \theta)$ for $\eta \in \mathcal{G}_{Y|\mathbf{X}}, \theta \in \Theta$ and $x \in \Omega_{\mathbf{X}}$.

Definition 1.4. *Let $\mathcal{T} : \mathcal{H} \to \mathbb{R}^1$ be the statistical functional then it induces a random variable on $\Omega_{\mathbf{X}}$ as*

$$\mathbf{x} \mapsto \mathcal{T}(\eta(\mathbf{x}, \cdot, \theta)).$$

If there is a matrix $\gamma \in \mathbb{R}^{p\times d}$, with $d \leq p$, such that $\mathcal{T}(\eta_0(\mathbf{X}, \cdot, \theta_0))$ is measurable with respect to $\sigma(\gamma^T\mathbf{X})$, then we call $\mathcal{S}(\gamma)$ a sufficient dimension reduction subspace for $\mathcal{T}$. The intersection of all such spaces is called the central subspace for conditional functional $\mathcal{T}$, or the T-central subspace. We denote the T-central subspace by $\mathcal{S}_{T(Y|\mathbf{X})}$.

Luo, Li and Yin (2014) developed their semiparametric estimator by further categorized the conditional statistical functional into three categories: the linear functionals (L-functional), such as conditional means as conditional moments; the composite functionals (C-functional) such as conditional variance; the implicit funcitonals (I-functional) such as conditional quantile and conditional expectile. We will closely follow their discussion in further development.

The T-central subspace is invariant under affine transformation (Luo, Li and Yin 2014). That is, let $\mathbf{Z} = \mathbf{\Sigma_X}^{-1/2}(\mathbf{X} - \mathrm{E}(\mathbf{X}))$, where $\Sigma_{\mathbf{X}}$ stands for the covariance matrix of $\mathbf{X}$. Then $\mathcal{S}_{\mathcal{T}(Y|\mathbf{Z})} = \mathbf{\Sigma_X}^{1/2}\mathcal{S}_{\mathcal{T}(Y|\mathbf{X})}$. Hence, the T-central subspaces under $\mathbf{Z}$-scale and $\mathbf{X}$-scale are equivalent. We will work mainly with $\mathbf{Z}$ as the standardized version of $\mathbf{X}$ without loss of generality.

1.4 Methodologies in Sufficient Dimension Reduction

Section 1.1 introduced a proposition that classified the methodologies of SDR into three categories. And Section 1.2 further introduced some of the targets of interests in SDR. In this section, we are going to review some of the popular methodologies in SDR based on the former approach. However, these categories are not mutually exclusive, i.e. some of the methods will not strictly fall into one category. So the first approach here is just used as a guideline for the literature review.

There has been many existing methods over the past decades in sufficient dimension reduction, which could be typically classified into three categories: inverse, forward and joint approach. Famous representatives of inverse approach include Sliced Inverse Regression (SIR; Li 1991) which estimated the inverse moment by slicing; Sliced Average Variance Estimation (Cook and Weisberg 1991) locally estimated the conditional variance instead; Inverse Third Moment Estimator (Yin and Cook 2003) which targeted on even higher moments and Inverse Regression Estimator (IRE; Cook and Ni 2005) based on minimum discrepancy approach to improve the inverse approaches that depend on slicing. Fused method (Cook and Zhang 2014) provided solution to the slice issue. A more general inverse approach include Contour Regression (CR; Li et al. 2005) which used a small variation of the response; Directional Regression (DR; Li and Wang 2007) which depended on inverse second moment and likelihood based approach (LAD; Cook and Forzani 2009). These methods depend on singular value decomposition and hence run fast, but they typically require linear condition and constant variance condition on the predictors so as to avoid conditioning on high dimensional variable.

In joint approach, famous examples include principal Hessian direction (pHd; Li 1992; Cook 1998); Iterative Hessian Direction (IHT; Cook and Li 2002); Joint higher moments methods (Yin and Cook 2004); Kullback-Leibler Distance based Method (Yin et al. 2008); Fourier Transformation approach (Zhu and Zeng 2006; Zeng 2008) and correlation approaching (Fukumizu et al. 2004, 2009).

Forward approach includes several outstanding semiparametric and nonparametric methods, and hence mitigate the condition put on predictors. A series of work were focusing on semiparametric methods, such as Ma and Zhu (2012; 2013a; 2013b) used semiparametric theory to establish an efficient estimator for central subspace; (Luo and Li 2014) developed one step estimator for T-central subspace. For nonparametric methods, Outer Product Gradient (OPG; Xia et al. 2002) and Minimum

Average Variance Estimator (MAVE; Xia et al. 2002) that utilized local polynomial approximation to estimate CMS perhaps are the most famous among them. Other derived methods include: Density MAVE (Xia 2007) targeting conditional density; Slice Regression (Wang and Xia 2008) using indicator function and Family of MAVE (Yin and Li 2011) estimating CS rather than CMS exhaustively. Different from inverse approach, these methods typically do not have strong assumption on predictors, but have to deal with the issue of conditioning on high dimensional predictors which slows down the estimation dramatically.

Our proposed method belongs to the last category. By imposing several conditions on $\mathbf{X}$, we are able to get away with the issue of conditioning on high dimensional variable. And we provide a rather robust estimator by employing the regression fit as well as maintaining a fast computation speed by applying singular value decomposition.

1.5 Overview of the Dissertation

Throughout this dissertation, we will be mainly focusing on the development of a novel estimator for the T-central subspace, named as *cubic kernel estimator (CK)*. In Chapter 2, we will develop the CK estimator mainly for L-funcitonals. In particular, method for CMS will be developed through conditional expectation and CS will be developed through conditional Fourier transformation. In Chapter 3, development will be focused on I-funcitonals with emphasis on conditional quantile and conditional expectile. Moreover, we will propose some variable selection procedure that could be applied to our method. Simulation results for our CK estimator will be given in the relevant sections to study the advantage and disadvantage of our method. In Chapter 4, we will look at our estimators from a new perspective and provide an approach that could improve the performance. Other future developments will be discussed as well.

Chapter 2 Moment Kernels for Estimating Central Mean Subspace and Central Subspace

2.1 Introduction

Since Li (1991), there has been considerable interests in concentrating regression information about a response $Y \in \mathbb{R}$ in a low-dimensional projection of the random predictor $\mathbf{X} \in \mathbb{R}^p$ without loss of the regression information. Reducing the dimensionality of the predictor could be quite useful in building a better model, especially when the dimension of $\mathbf{X}$ is high. What constitutes a high-dimensional predictor depends on the goal of the analysis. When a comprehensive low-dimensional graphical display is desired, high dimension might mean $p > 3$. Or it could mean p in the tens or hundreds, as in contemporary vernacular. In this chapter, we will propose a new estimator based on the moment kernels to estimate T-central subspace, targeting L-functionals.

The rest of this chapter is organized as follow: Section 2.2 provides a general theory for our proposed method. Section 2.3 further develops our method for two specific functionals: CMS and CS. Section 2.4 explains a method of order determination. Section 2.5 proposes a variable selection procedure for our method. Section 2.6 demonstrates the advantages of our estimator via simulation example and real data analysis.

2.2 The Moment Kernels

The foundation of our approach rests on a loss function of the form

$$L(K(\mathbf{Z}), \mathcal{T}(Y)) = -\mathcal{T}(Y)K(\mathbf{Z}) + \phi(K(\mathbf{Z})),$$

where ϕ is a convex function and T is a transformation on the response that leads to the L-functional $\mathcal{T}$. Thus, the loss function is convex in $K(\mathbf{Z})$. Objective functions of this form correspond to natural exponential families and cover the early work by Li and Duan (1989) on ordinary least squares (OLS) and Cook and Li (2002) on mean functions that are quadratic in $\mathbf{Z}$.

Suppose that $a \in \mathbb{R}^1$ and $\mathbf{b} \in \mathbb{R}^p$. Let $\mathbf{C}$ be a $p \times p$ symmetric matrix, and let $\mathcal{D}$ be a $p \times p \times p$ array with k-th face $\mathbf{D}_k = (d_{ijk})$, $i, j, k = 1 \ldots, p$. We constrain the $\mathbf{D}_k$ to be symmetric and the d_{ijk} to be invariant under permutations of its index ijk. We may also treat $\mathcal{D}$ as a $p^2 \times p$ matrix.

Consider now a fit of the cubic kernel $K(\mathbf{Z}) = a + \mathbf{b}'\mathbf{Z} + \mathbf{Z}'\mathbf{C}\mathbf{Z} + (\mathbf{Z}' \otimes \mathbf{Z}')\mathcal{D}\mathbf{Z}$, and define the objective function as the following risk:

$$R(a, \mathbf{b}, \mathbf{C}, \mathcal{D}) = \mathrm{E}(L(a + \mathbf{b}'\mathbf{Z} + \mathbf{Z}'\mathbf{C}\mathbf{Z} + (\mathbf{Z}' \otimes \mathbf{Z}')\mathcal{D}\mathbf{Z}, \mathcal{T}(Y))).$$

Many regression functions are smooth, and so can be approximated by polynomials of different order. For example, OLS could be seen as a first-order polynomial approximation. However, lower-order approximation typically has trouble capturing useful directions: OLS fails to capture symmetric pattern while second-order polynomial models are reluctant to pick up the linear trend. In regression, traditionally a suggestion on using third-polynomial model is a good approximation, in addition to its interpretability. However, fitting a third-polynomial may be difficult even if p is moderate. For instance, the number of parameter for fitting a third-order polynomial is $(p+1)(1 + \frac{p}{2} + \frac{p(p+2)}{6})$. That is, for instance, when $n = 100, p = 10$, the model has 286 parameters, causing much trouble to fit a model.

Our goal is to construct a simple cubic kernel that could easily capture all impor-

tant directions. Let

$$\begin{aligned}
\boldsymbol{\beta}_{T\mathbf{Z}} &= \mathrm{E}(\mathcal{T}(Y)\mathbf{Z}), \quad \boldsymbol{\Sigma}_{T\mathbf{ZZ}} = \mathrm{E}[\{\mathcal{T}(Y) - \mathrm{E}(\mathcal{T}(Y))\}\mathbf{Z}\mathbf{Z}'], \\
\mathcal{M}_{T\mathbf{ZZZ}} &= \mathrm{E}\{\mathcal{T}(Y) - \mathrm{E}(\mathcal{T}(Y))\}\mathbf{Z} \otimes \mathbf{Z}\mathbf{Z}' - \mathrm{E}(\mathcal{T}(Y)\mathbf{Z}) \otimes I \\
&\quad - I \otimes \mathrm{E}(\mathcal{T}(Y)\mathbf{Z}) - \mathrm{vec}(I)\mathrm{E}(\mathcal{T}(Y)\mathbf{Z})'
\end{aligned}$$

where $\boldsymbol{\beta}_{T\mathbf{Z}}$ represents population OLS slope estimate for regression of $\mathcal{T}(Y)$ on $\mathbf{Z}$, $\boldsymbol{\Sigma}_{T\mathbf{ZZ}}$ represents the population kernel for principal Hessian directions, $\mathcal{M}_{T\mathbf{ZZZ}}$ is the kernel for third moment (Yin and Cook 2004).

The toy example next gives a clear message: Consider $Y = X_1 + X_2^3 + X_3^5 + \epsilon$, where $X_1, X_2, X_3, \epsilon \sim N(0,1)$ and jointly independent. In this case $\mathcal{S}_{E(Y|\mathbf{Z})} = \mathbb{R}^3$. With $\mathcal{T}(Y) = Y$, it is straight forward to verify that $\boldsymbol{\beta}_{T\mathbf{X}} = (1,3,15)'$ and $\boldsymbol{\Sigma}_{T\mathbf{ZZ}} = \mathbf{0}_{3\times 3}$. And

$$\mathcal{M}'_{T\mathbf{ZZZ}} = \begin{pmatrix} 3 & 3 & 15 & 3 & 3 & 0 & 15 & 0 & 3 \\ 3 & 3 & 0 & 3 & 15 & 15 & 0 & 15 & 3 \\ 15 & 0 & 3 & 0 & 15 & 15 & 3 & 15 & 105 \end{pmatrix}$$

The true underlying model in this example is a polynomial of fifth-order, therefore, we do expect some failure in the lower order kernels: linear kernel is only picking one direction while the quadratic kernel fails to recover any direction. The cubic kernel in this case works as expected, successfully recovering the higher order term. In general we believe, if a smooth function can only be approximated well by an order higher than 3, then those directions still can be highly correlated with the first three lower orders. Thus, the model fitting may not be correct, but the full set of directions will be captured. The use of the objective function does not imply the cubic is a true model or is a good approximation of the data. However, the minimization problem does provide a connection between the solution and $\mathcal{S}_{\mathcal{T}(Y|\mathbf{Z})}$ as shown in the following result.

Proposition 2.1. *Let* $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{\Delta} = \arg\min_{a,\mathbf{b},\mathbf{C},\mathcal{D}} R(a, \mathbf{b}, \mathbf{C}, \mathcal{D})$. *Let* $\mathcal{S}_{\mathcal{T}(Y|\mathbf{Z})}$ *have*

basis matrix $\boldsymbol{\gamma}$*, i.e.* $\mathcal{S}(\boldsymbol{\gamma}) = \mathcal{S}_{\mathcal{T}(Y|\mathbf{Z})}$*. Assume that* $\mathrm{E}(\mathbf{Z} \mid \boldsymbol{\gamma}'\mathbf{Z})$ *is linear,* $\mathrm{Var}(\mathbf{Z} \mid \boldsymbol{\gamma}'\mathbf{Z})$ *is constant and* $\mathcal{M}^{(3)}(\mathbf{Z} \mid \boldsymbol{\gamma}'\mathbf{Z}) = 0$*, where* $\mathcal{M}^{(3)}(\mathbf{Z} \mid \boldsymbol{\gamma}'\mathbf{Z}) = \mathrm{E}(\mathbf{Z} \otimes \mathbf{Z}\mathbf{Z}' \mid \boldsymbol{\gamma}'\mathbf{Z})$*. Then* $\mathcal{S}(\boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{\Delta}') \subseteq \mathcal{S}_{\mathcal{T}(Y|\mathbf{Z})}$.

Proposition 2.1 reveals that the column space spanned by population minimizer, $\mathcal{S}(\boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{\Delta}')$, lies in $\mathcal{S}_{\mathcal{T}(Y|\mathbf{Z})}$, the desired T-central subspace. However, as mentioned earlier, in practice, it would be rather difficult to directly solve the optimization problem, which makes this proposition seems less practical.

The result next establishing a link between the minimizer and our kernels provides a further connection that may lead to relatively easy computational algorithm. Let $\boldsymbol{\alpha}$ be a basis for $\mathcal{S}(\boldsymbol{\beta}_{T\mathbf{Z}}, \boldsymbol{\Sigma}_{T\mathbf{Z}\mathbf{Z}}, \mathcal{M}'_{T\mathbf{Z}\mathbf{Z}\mathbf{Z}})$.

Proposition 2.2. *If* $\mathrm{E}(\mathbf{Z} \mid \boldsymbol{\alpha}'\mathbf{Z})$ *is linear,* $\mathrm{Var}(\mathbf{Z} \mid \boldsymbol{\alpha}'\mathbf{Z})$ *is constant and* $\mathcal{M}^{(3)}(\mathbf{Z} \mid \boldsymbol{\alpha}'\mathbf{Z}) = 0$*, then*

$$\mathcal{S}(\boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{\Delta}') \subseteq \mathcal{S}(\boldsymbol{\beta}_{T\mathbf{Z}}, \boldsymbol{\Sigma}_{T\mathbf{Z}\mathbf{Z}}, \mathcal{M}'_{T\mathbf{Z}\mathbf{Z}\mathbf{Z}}).$$

If, in addition, $\mathrm{E}(\mathbf{Z} \mid \boldsymbol{\gamma}'\mathbf{Z})$ *is linear,* $\mathrm{Var}(\mathbf{Z} \mid \boldsymbol{\gamma}'\mathbf{Z})$ *is constant and* $\mathcal{M}^{(3)}(\mathbf{Z} \mid \boldsymbol{\gamma}'\mathbf{Z}) = 0$*, then*

$$\mathcal{S}(\boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{\Delta}') \subseteq \mathcal{S}(\boldsymbol{\beta}_{T\mathbf{Z}}, \boldsymbol{\Sigma}_{T\mathbf{Z}\mathbf{Z}}, \mathcal{M}'_{T\mathbf{Z}\mathbf{Z}\mathbf{Z}}) \subseteq \mathcal{S}_{\mathcal{T}(Y|\mathbf{Z})}.$$

Proposition 2.2 requires three mild conditions on $\boldsymbol{\alpha}'\mathbf{Z}$. However, it conveys an important message that the column space spanned by the population minimizer, $\mathcal{S}(\boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{\Delta}')$, provides information no more than the space spanned by columns of the kernels, $\mathcal{S}(\boldsymbol{\beta}_{T\mathbf{Z}}, \boldsymbol{\Sigma}_{T\mathbf{Z}\mathbf{Z}}, \mathcal{M}'_{T\mathbf{Z}\mathbf{Z}\mathbf{Z}})$. This information is valuable: we can rely on these kernels to recover the directions rather than focusing on solving the optimization problem.

2.3 T-Central Subspace for Particular Functionals

In this section, we develop two estimators for CMS and CS by using the mean functional and the functional induced by Fourier transformation.

Central Mean Subspace

Studying the conditional mean $\mathrm{E}(Y \mid \mathbf{Z})$ via $\mathcal{S}_{\mathrm{E}(Y|\mathbf{Z})}$, is the first and perhaps the most important step in regression. In such a case, $\mathcal{T}(Y) = Y$. Many methods have been proposed for study of the conditional mean, such as OLS (Li and Duan 1989), principal Hessian direction (Li 1992, Cook 1998), Iterative Hessian Transformation (Cook and Li 2002), Higher Moment Methods (Yin and Cook 2004), and rMAVE (Xia et al. 2002). It is well-known that OLS fails to pick up the symmetric patterns, while pHd is weak at detecting linear trends. However, Proposition 2.2 suggests a high-order kernel that can use linear kernel, quadratic kernel and third moment kernel in a concert. Such a combination mitigates the drawbacks of using individual kernel separately, and thus results in an overall advantage as we illustrated in the toy example. Nevertheless, each individual kernel may have different importance in the model. To overcome this, we will fit a cubic model to find the weight for each individual kernel. Thus, the combined space $\mathcal{S}(\beta_{Y\mathbf{Z}}, \mathbf{\Sigma}_{Y\mathbf{ZZ}}, \mathcal{M}'_{Y\mathbf{ZZZ}})$ is indeed very useful. Therefore, we call it generally the *cubic kernel (CK)* method and, in particular, we call it the *cubic kernel for mean (CKM)* for the central mean subspace.

The estimation of $\boldsymbol{\beta}_{Y\mathbf{Z}}, \mathbf{\Sigma}_{Y\mathbf{ZZ}}, \mathcal{M}_{Y\mathbf{ZZZ}}$ can be done consistently by substituting sample moments. However, there are still several caveats for implementing these moment methods.

1. Yin and Cook (2003) pointed out that the three dimensional array can be written as a $p^2 \times p$ matrix, where this matrix can be considered as a column of p blocks with each block being a $p \times p$ matrix. However, this $p^2 \times p$

matrix contains only $\frac{p(p+1)}{2}$ unique ones. We then construct a new unique kernel matrix, $(\tilde{\mathcal{M}}'_{Y\mathbf{ZZZ}})_{p\times\frac{p(p+1)}{2}}$ by removing the repeated columns, so that $\mathcal{S}(\mathcal{M}'_{Y\mathbf{ZZZ}}) = \mathcal{S}(\tilde{\mathcal{M}}'_{Y\mathbf{ZZZ}})$.

2. Due to correlations among the linear, quadratic and third moment kernels, when constructing the CK estimator, we may use the residuals instead of the original response. To be more specific, take $r_1 = Y - a_1 - b_1\mathbf{Z}'\boldsymbol{\beta}_{Y\mathbf{Z}}$, with $a_1, b_1 \in \mathbb{R}$ being the OLS coefficients of Y on $(1, \mathbf{Z}'\boldsymbol{\beta}_{Y\mathbf{Z}})$, then construct

$$\boldsymbol{\Sigma}_{r_1\mathbf{ZZ}} = \mathrm{E}(r_1\mathbf{ZZ}).$$

Take $r_2 = Y - a_2 - b_2\mathbf{Z}'\boldsymbol{\beta}_{Y\mathbf{Z}} - c_2\mathbf{Z}'\boldsymbol{\Sigma}_{r_1\mathbf{ZZ}}\mathbf{Z}$, with $a_2, b_2, c_2 \in \mathbb{R}$ being the OLS coefficients of Y on $(1, \mathbf{Z}'\boldsymbol{\beta}_{Y\mathbf{Z}}, \mathbf{Z}'\boldsymbol{\Sigma}_{r_1\mathbf{ZZ}}\mathbf{Z})$ and construct

$$\mathcal{M}_{r_2\mathbf{ZZZ}} = \mathrm{E}(r_2\mathbf{Z}\otimes\mathbf{ZZ}') - \mathrm{E}(r_2\mathbf{Z})\otimes I - I\otimes\mathrm{E}(r_2\mathbf{Z}) - \mathrm{vec}(I)\mathrm{E}(r_2\mathbf{Z})'$$

to eliminate possible bias. This idea is very much similar to the residual based pHd (Li 1992), where the residuals from OLS is removed to improve the results, A trivial generalization of propositions 3 and 4 in Yin and Cook (2004) shows $\mathcal{S}(\mathcal{M}'_{r_2\mathbf{ZZZ}}) \subseteq \mathcal{S}_{E(Y|\mathbf{Z})}$.

We are now ready to provide the details of our algorithm for CKM.

Algorithm for CMS: Let $\{\mathbf{X}_i, Y_i\}, i = 1, 2, \cdots, n$ be an i.i.d sample, and assume that the structural dimension d for the CMS is known. Then

1. Standardize the predictor $\hat{\mathbf{Z}}_i = \hat{\boldsymbol{\Sigma}}_{\mathbf{X}}^{-1/2}(\mathbf{X}_i - \bar{\mathbf{X}})$, for $i = 1\cdots, n$.

2. Construct kernel matrices:

$$\hat{\boldsymbol{\beta}}_{Y\mathbf{Z}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{Z}}_i Y_i, \hat{\boldsymbol{\Sigma}}_{r_1\mathbf{ZZ}} = \frac{1}{n}\sum_{i=1}^{n}\hat{r}_{1i}\hat{\mathbf{Z}}_i\hat{\mathbf{Z}}_i',$$

where $\hat{r}_{1i}$ is the estimated version of r_1 for ith observation.

3. Construct $\hat{\mathcal{M}}_{r_2\mathbf{ZZZ}}$ and $\hat{\hat{\mathcal{M}}}_{r_2\mathbf{ZZZ}}$. First we construct p^2 vectors for $j, k = 1, \cdots, p$:

$$\hat{\mathbf{m}}_{jk} = \frac{1}{n}\sum_{i=1}^{n} \hat{r}_{2i}\hat{\mathbf{Z}}_i(\mathbf{e}_\mathbf{j}'\hat{\mathbf{Z}}_i)(\mathbf{e}_\mathbf{k}'\hat{\mathbf{Z}}_i) - \frac{1}{n}\sum_{i=1}^{n} \hat{r}_{2i}(\mathbf{e}_\mathbf{k}'\hat{\mathbf{Z}}_i)\mathbf{e_j} - \frac{1}{n}\sum_{i=1}^{n} \hat{r}_{2i}(\mathbf{e}_\mathbf{j}'\hat{\mathbf{Z}}_i)\mathbf{e_k} - \frac{1}{n}\sum_{i=1}^{n} \hat{r}_{2i}(\mathbf{e}_\mathbf{j}'\mathbf{e_k})\hat{\mathbf{Z}}_\mathbf{i},$$

where $\mathbf{e}_i$ is a unit vector with ith element 1 and $\hat{r}_{2i}$ is ith element in the residual vector $\hat{r}_2$. Let $\hat{\mathcal{M}}'_{r_2\mathbf{ZZZ}} = (\hat{\mathbf{m}}_{11}, \cdots, \hat{\mathbf{m}}_{pp})$, which is a $p \times p^2$ matrix. Then take $\hat{\mathbf{m}}_{jk}$ with $j \geq k$ and let $\hat{\hat{\mathcal{M}}}'_{r_2\mathbf{ZZZ}} = (\hat{\mathbf{m}}_{11}, \hat{\mathbf{m}}_{21}, \cdots, \hat{\mathbf{m}}_{pp})$, which is a $p \times \frac{p(p+1)}{2}$ matrix consisting of all the unique columns for $\hat{\mathcal{M}}'_{r_2\mathbf{ZZZ}}$.

4. Construct $\mathbf{S} = (1, S_1, S_2, S_3)'$, with $S_{1i} = \hat{\mathbf{Z}}_i'\hat{\boldsymbol{\beta}}_{Y\mathbf{Z}}$, $S_{2i} = \hat{\mathbf{Z}}_i'\hat{\boldsymbol{\Sigma}}_{r_1\mathbf{ZZ}}\hat{\mathbf{Z}}_i$ and $S_{3i} = \hat{\mathbf{Z}}_i' \otimes \hat{\mathbf{Z}}_i'\hat{\mathcal{M}}_{r_2\mathbf{ZZZ}}\hat{\mathbf{Z}}_i$. Find the OLS fit of Y on $\mathbf{S}$ and denote the fitted coefficient as $\hat{\boldsymbol{w}} = (\hat{w}_0, \hat{w}_1, \hat{w}_2, \hat{w}_3)'$.

5. Construct $\hat{\mathbf{M}} = \hat{w}_1^2\hat{\boldsymbol{\beta}}_{Y\mathbf{Z}}\hat{\boldsymbol{\beta}}_{T\mathbf{Z}}' + \hat{w}_2^2\hat{\boldsymbol{\Sigma}}_{r_1\mathbf{ZZ}}\hat{\boldsymbol{\Sigma}}'_{r_1\mathbf{ZZ}} + \hat{w}_3^2\hat{\hat{\mathcal{M}}}'_{r_2\mathbf{ZZZ}}\hat{\hat{\mathcal{M}}}_{r_2\mathbf{ZZZ}}$, where $\hat{w}_i's$ are obtained in step 4. Perform the eigen decomposition on $\hat{\mathbf{M}}$, our estimator for the central mean subspace is given by

$$\mathcal{S}(\hat{\boldsymbol{\Sigma}}_\mathbf{X}^{-1/2}\hat{\mathbf{u}}_1, \hat{\boldsymbol{\Sigma}}_\mathbf{X}^{-1/2}\hat{\mathbf{u}}_2, \cdots, \hat{\boldsymbol{\Sigma}}_\mathbf{X}^{-1/2}\hat{\mathbf{u}}_d),$$

where $\hat{\mathbf{u}}_i, i = 1 \cdots d$ are the eigenvectors corresponding to the first d eigenvalues of $\hat{\mathbf{M}}$.

Note that the weights $\hat{w}_i's$ in step 4 are a key component of the CK estimator: They adapt the importance of the individual kernels, $(\hat{\boldsymbol{\beta}}_{T\mathbf{Z}}, \hat{\boldsymbol{\Sigma}}_{T\mathbf{ZZ}}, \hat{\mathcal{M}}_{T\mathbf{ZZZ}})$ in $\hat{\mathbf{M}}$. Therefore, bigger values of $\hat{w}_i's$ will indicate that the corresponding kernel is more important.

Central Subspace

To estimate CS, we use CK approach via Fourier transformation. Fourier transformation was used by Zhu and Zeng (2006), Zhu, Zhu and Wen (2010), Weng and Yin (2018) in sufficient dimension reduction. We study the conditional distribution of $Y \mid \mathbf{Z}$ through the conditional characteristic function, $\mathrm{E}(\exp(i\omega Y) \mid \mathbf{Z})$, for $\omega \in \mathbb{R}$, which leads to investigate $(\mathbf{Z}, \exp(i\omega' Y))$ for each ω in the regression mean. For ease of discussion, we only discuss the univariate response Y here.

Yin and Li (2011) had a thorough discussion on the characteristic function. That is, $\mathcal{F} = \{\exp(i\omega Y), \omega \in \mathbb{R}\}$ is a family of functions that is dense in $L_2(F)$. In addition, their Theorem 2.1 indicates that the union of all directions in each of $\mathrm{E}(\exp(i\omega Y) \mid \mathbf{Z})$ recovers the central subspace and their Theorem 2.2 suggests a finite number of ω's would be suffice to recover the central subspace. Combining these with CKM algorithm leads to our new algorithm that we call the *cubic kernel for Fourier transformation* (CKF).

Algorithm for CS: Let $\{\mathbf{X}_i, Y_i\}, i = 1, 2, \cdots, n$ be i.i.d sample. Assuming structural dimension of CS, d, is known.

1. Standardize the predictor $\hat{\mathbf{Z}}_i = \hat{\mathbf{\Sigma}}_{\mathbf{X}}^{-1/2}(\mathbf{X}_i - \bar{\mathbf{X}})$, for $i = 1 \cdots, n$.

2. Fix an integer H and a small number s, say 0.1, and generate H random variables ω_i from $N(0, \frac{s\pi^2}{\mathrm{median}(Y^2)})$. Then construct $2H$ $n \times 1$ vectors $\mathbf{S}_{ij} = (S_{ij1}, S_{ij2}, \ldots, S_{ijn})', i = 1, 2; j = 1, \ldots, H$, where $S_{1jk} = \cos(\omega_j Y_k)$ and $S_{2jk} = \sin(\omega_j Y_k)$ for $k = 1, 2, \ldots, n$.

3. For each ij, treat $\mathbf{S}_{ij}$ as a response vector and apply the CKM algorithm to determine the kernel matrices i.e. $\hat{\mathbf{M}}_{ij}$.

4. Construct $\hat{\mathbf{M}} = \sum_i^2 \sum_j^H \hat{\mathbf{M}}_{ij}$. Perform the eigen decomposition on $\hat{\mathbf{M}}$, our

estimator for CS is given by

$$\mathcal{S}(\hat{\mathbf{\Sigma}}_{\mathbf{X}}^{-1/2}\hat{\mathbf{u}}_1, \hat{\mathbf{\Sigma}}_{\mathbf{X}}^{-1/2}\hat{\mathbf{u}}_2, \cdots, \hat{\mathbf{\Sigma}}_{\mathbf{X}}^{-1/2}\hat{\mathbf{u}}_d),$$

where $\hat{\mathbf{u}}_i, i = 1 \cdots d$ are the eigenvectors corresponding to the first d eigenvalues of $\hat{\mathbf{M}}$.

In the following we describe our rational for generating the ω's as described in step 2. Zhu, Zhu and Wen (2010) suggested a criterion based on the periodicity of Fourier transformation, i.e. $\exp(i\omega Y) = \exp(i(\omega Y + 2\pi))$, form $\omega \sim N(0, \sigma^2)$, and they suggested $|\omega Y| < \pi$ with high probability i.e. $P(|\omega Y| > \pi) < s$, where s is a small number. Then applying Chebyshev's inequality leads to the upper bound $\frac{s\pi^2}{\mathrm{E}(Y^2)}$ on the variance of ω. However, Durrett (2010) pointed out that Chebyshev's inequality could be too loose, especially when it is applied to the tail probability of a distribution. If this is the case, problems may arise as the upper bound for the variance of ω would be relatively small, limiting the variation of ω and, in consequence, resulting in bad estimate. We suggest using $\frac{s\pi^2}{\mathrm{median}(Y^2)}$ in place of $\frac{s\pi^2}{\mathrm{E}(Y^2)}$, a modification that in our experience performs well in practice. On the other hand, the number of ω's from 20 to 100 result in quite robust estimates based on our limited simulations. Thus, we use 50 for the number of ω's as the rule of thumb.

It is conceptually straightforward to establish asymptotic properties of our proposals, though details could be tedious at times. Here, we provide a brief sketch of an argument to show that our methods provided root n consistent estimator for the central mean subspace and central subspace when their dimensions are known. For CKM, because $\hat{\boldsymbol{\beta}}_{T\mathbf{Z}}$, $\hat{\mathbf{\Sigma}}_{T\mathbf{ZZ}}$ and $\hat{\mathcal{M}}_{T\mathbf{ZZZ}}$ are method of moments estimators, and hence are $\sqrt{n}$-consistent. Let $\hat{\mathbf{B}}$ be a sample version of $\mathbf{B} = (\boldsymbol{\beta}_{T\mathbf{Z}}, \mathbf{\Sigma}_{T\mathbf{ZZ}}, \mathcal{M}'_{T\mathbf{ZZZ}})$, then its projection operator $\hat{\mathbf{B}}(\hat{\mathbf{B}}'\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}'$ is a smooth function of $\hat{\boldsymbol{\beta}}_{T\mathbf{Z}}$, $\hat{\mathbf{\Sigma}}_{T\mathbf{ZZ}}$ and $\hat{\mathcal{M}}_{T\mathbf{ZZZ}}$, involving only $\sqrt{n}$-consistent estimator of $\mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'$. As a consequence,

the estimated projection operator is indeed $\sqrt{n}$ consistent for the respective population projection operator, hence for $\mathcal{S}_{T(Y|\mathbf{X})}$. For CKF estimator on CS, asymptotic properties can by established based on Theorem 3.2 of Li, Wen and Zhu (2008) under some mild conditions, which are easily satisfied since $\hat{\mathbf{M}}_{ij}$ is $\sqrt{n}$-consistent.

2.4 Order Determination

Practically, the structural dimension, d, needs to be estimated. Many methods are available for order determination, including BIC-type criteria (Zhu and Zeng 2006) and a bootstrap approach (Ye and Weiss 2003). For our cubic kernel method, we adapt the ladle estimator (Luo and Li 2016) to estimate d.

The basic idea of the ladle plot is the following: for each possible working dimension $1 \leq k \leq p-1$, we construct an estimated basis $\hat{\mathbf{M}}_k$ for the subspace of dimension k using our CK method. Then we take bootstrap sample of size B, and for each bootstrap sample, we obtain the corresponding estimated basis matrix $\hat{\mathbf{M}}_k^b, b = 1, \cdots, B$. Define functions $u(k), v(k), w(k) : \{0, 1, \cdots, p-1\} \mapsto \mathbb{R}$, $w(k) = u(k) + v(k)$, with

$$u(k) = \frac{\hat{\lambda}_{k+1}}{1 + \sum_{i=0}^{p-1} \hat{\lambda}_{i+1}}, v(k) = \begin{cases} 0 & \text{if } k = 0 \\ \frac{B^{-1} \sum_{b=1}^{B} (1 - |\det(\hat{\mathbf{M}}_k' \hat{\mathbf{M}}_k^b)|)}{1 + B^{-1} \sum_{i=1}^{p-1} \sum_{b=1}^{B} (1 - |\det(\hat{\mathbf{M}}_i' \hat{\mathbf{M}}_i^b)|} & \text{if } k = 1, \cdots, p-1 \end{cases}$$

where $\hat{\lambda}_i$ denotes the ith largest eigenvalue of the basis matrix $\hat{\mathbf{M}}_p$, and $u(\cdot)$ represents a normalization of these eigenvalues. In $v(\cdot)$, $1 - |\det(\hat{\mathbf{M}}_k' \hat{\mathbf{M}}_k^b)|$ describes the discrepancy between the sample estimate $\hat{M}_k$ and the bootstrap estimate $\hat{M}_k^b$. Therefore the numerator of $v(\cdot)$, which is the bootstrap sample average of above discrepancy measure, could be seen as the variability of bootstrap estimates around the sample estimate. The denominator of $v(\cdot)$ is for normalization. The estimated d is the value of k that minimizes the target function $w(\cdot)$.

2.5 Variable Selection

Reducing dimension by linear combination is very useful, especially for the prediction, as once it is reduced to a small number of dimensions, many traditional methods can be applied to the reduced variables. However, it is often difficult to interpret the results, in particular, the reduced dimensions involve all the original variables. In some cases, one may want to make decision based on a few relevant predictors rather than all of them. In this section, we discuss how to perform variable selection base on CK method.

Large n small p

Li (2007) formulated a generalized eigen-decomposition problem in a regression framework with penalization. Chen, Zou and Cook (2010) proposed coordinate-independent sparse estimation (CISE) by introducing a coordinate-independent penalty function. We adapt the CISE procedure for our CK method.

Let $\mathbf{M}^{\frac{1}{2}} = (\mathbf{m}_1, \mathbf{m}_2, \cdots, \mathbf{m}_p)$, where $\mathbf{M}$ could be the CK kernel matrix given in step 5 of the CMS algorithm or step 4 of the CS algorithm. Let $\Gamma = (\gamma_1, \gamma_2, \cdots, \gamma_p)'$ be any $p \times d$ matrix with $\Gamma'\mathbf{\Sigma_X}\Gamma = I_d$, note $\gamma_i', i = 1, 2, \cdots, p$ represent the rows of Γ. Then the CISE for CK could be obtained by solving the following optimization problem:

$$\text{Minimize: } \sum_{i=1}^{p} \|\mathbf{\Sigma_X^{-1}}\mathbf{m}_i - \Gamma\Gamma'\mathbf{m}_i\|^2_{\mathbf{\Sigma_X}} + \theta_i\|\gamma_i\| \text{ subject to } \Gamma'\mathbf{\Sigma_X}\Gamma = I_d.$$

Here $\|\cdot\|_{\mathbf{\Sigma_X}}$ denotes the norm with respect to $\mathbf{\Sigma_X}$, $\|\cdot\|$ denotes the norm in Euclidean space and $\theta_i, i = 1, 2, \cdots, p$ denote the tunning parameters. Based on proposition 2 of Chen, Zou and Cook (2010), this can be reparametrized as a Grassmann manifold optimization problem and the nondifferentiability of the penalty term is addressed

by local quadratic approximation, so as to minimize

$$\mathrm{tr}(\Gamma'(-\mathbf{G}_n + \frac{1}{2}\hat{\mathbf{\Sigma}}_{\mathbf{X}}^{-1/2}\mathbf{D}^{(0)}\hat{\mathbf{\Sigma}}_{\mathbf{X}}^{-1/2})\Gamma)$$

with $\mathbf{D}^{(0)} = \mathrm{diag}(\frac{\theta_1}{\|(\hat{\mathbf{\Sigma}}_{\mathbf{X}}^{-1/2})_1'\Gamma^{(0)}\|}, \frac{\theta_2}{\|(\hat{\mathbf{\Sigma}}_{\mathbf{X}}^{-1/2})_2'\Gamma^{(0)}\|}, \cdots, \frac{\theta_p}{\|(\hat{\mathbf{\Sigma}}_{\mathbf{X}}^{-1/2})_p'\Gamma^{(0)}\|})$, where $(\cdot)_i'$ denotes the ith row vector of a matrix, and $\Gamma^{(0)}$ being the initial value. This optimization problem can be easily solved by eigenvalue decomposition of $\mathbf{G}_n - \frac{1}{2}\hat{\mathbf{\Sigma}}_{\mathbf{X}}^{-1/2}\mathbf{D}^{(0)}\hat{\mathbf{\Sigma}}_{\mathbf{X}}^{-1/2}$ and picking the first d principal components as the columns of $\Gamma^{(1)}$. Then repeat the procedure until it converges. Following the discussion in their paper, selection of tuning parameter is done adaptively using AIC-type or BIC-type criteria. We name such a procedure as CISE-CK.

Large p small n

When predictor has an ultra-high dimensions, i.e., $p >> n$, we add a variable screening procedure. There are many methods since Fan and Lv (2008) proposed their sure independence screening procedure (SIS). For instance, the procedure by Li, Zhong and Zhu (2012) who introduced DC-SIS via the distance correlation (DC; Székely, Rizzo and Bakiro 2007). Recently, the SIS approach was improved by a new sufficient variable selection procedure that was proposed by Yang, Yin and Zhang (2019). They described three algorithms, marginal screening (MS), one stage variable selection (SVS_1) and two stage variable selection (SVS_2), where MS approach is the SIS procedure. We incorporated these three algorithms of Yang, Yin and Zhang (2019) into our CK method for ultra-dimensional data analysis. For clarity, we detailed one of such an algorithms below, see Yang, Yin and Zhang (2019) for more details.

Algorithm for One Stage Variable Selection: Let $\{\mathbf{X}_i, Y_i\}, i = 1, 2, \cdots, n$ be i.i.d sample.

1. Let $\hat{r}_j = \hat{\mathrm{DC}}^2(X_j, Y)$ for $j = 1, \cdots, p$.

2. Let $\hat{r}_{-j} = \hat{\mathrm{DC}}^2(X_j, (X'_{-j}, Y)')$ for $j = 1, \cdots, p$, where X_{-j} denote the vector removing j's element.

3. Define $\hat{p} = \lfloor \frac{n}{\log n} \rfloor$. Select X_js correspond to the $\hat{p}_1 = \lfloor 0.95\hat{p} \rfloor$ largest $\hat{r}_j$ values, as well as $\hat{p}_2 = \lfloor 0.05\hat{p} \rfloor$ largest $\hat{r}_{-j}$ values that has not been selected. Denote the reduced the sample as $\{\mathbf{X}_i^*, Y_i\}, i = 1, 2, \cdots, n$.

4. Perform CISE-CK to $\{\mathbf{X}_i^*, Y_i\}, i = 1, 2, \cdots, n$.

Note that the two sufficient variable screening procedures (SVS_1, SVS_2) of Yang, Yin and Zhang (2019) was based on two sequential sufficient dimension reduction procedures proposed by Yin and Hilafu (2015). We can further use their two sufficient dimension reduction procedures combining the projective resampling idea developed by Li, Wen and Zhu (2008) for CK approach as well.

2.6 Simulations and Applications

In this section, we compare the performance of the cubic kernel methods CKM and CKF with the well-known sufficient dimension reduction methods SIR, SAVE, pHd, IHT, PFC, DR, and rMAVE. We conducted simulations to show the usefulness of ladle estimator on cubic kernels for the order determination. For accuracy comparison, we used the distance of Li, Zha and Chairomonte (2005) between two subspaces of $\mathbb{R}^p$: Let $\mathcal{S}_1$ and $\mathcal{S}_2$ be two subspaces of $\mathbb{R}^p$, and $\mathbf{P}_{\mathcal{S}_1}$ and $\mathbf{P}_{\mathcal{S}_2}$ be the corresponding orthogonal projection, respectively. Let

$$\Delta_F = \|\mathbf{P}_{\mathcal{S}_1} - \mathbf{P}_{\mathcal{S}_2}\|_F,$$

where $\|\cdot\|_F$ is the Frobenius norm. The smaller the Δ_F is, the closer the two subspaces are to each other.

Comparison of the Model accuracy

Define the $p \times 1$ vectors $\beta_1 = (1, 1, 1, 0, \cdots, 0)'$, $\beta_2 = (1, 0, 0, 0, 1, 3, 0, \cdots, 0)'$, $\beta_3 = (1, 0, \cdots, 0)'$, $\beta_4 = (0, 1, 0, \cdots, 0)'$, and $\beta_5 = (1, 0.5, 1, 0, \cdots, 0)'$. Let $\epsilon \sim N(0, I_n)$ that is independent of $\mathbf{X}$. The simulation was conducted to make comparison for combinations of the following configurations: $n = 100, 200, 400$, $p = 10, 15, 20, 60$, using $N = 200$ replicated datasets. On each generated dataset, we applied different methods and computed the distance Δ_F between these estimates and the true central mean subspace or the true central subspace. Then we calculate the average and standard error from the resulting N distances. The simulated three models are:

(A) $Y = 0.4(\beta_1'\mathbf{X})^2 + 3\sin(\beta_2'\mathbf{X}/4) + 0.2\epsilon$. This was mentioned in Li (2007). The CMS is $\mathcal{S}(\beta_1, \beta_2)$. The first term in the model is symmetric about 0, so SIR may fail to estimate this direction but not for SAVE, pHd and IHT. The second term is roughly monotone, so SIR may recover it.

(B) $Y = \cos(2\beta_3'\mathbf{X}) - \cos(\beta_4'\mathbf{X}) + 0.5\epsilon$. The CMS is $\mathcal{S}(\beta_3, \beta_4)$. The model, which was used by Li (1992), tends to have a symmetric pattern. SIR may fail to estimate both directions, but not for SAVE, pHd and IHT. We will use Model (A) and (B) to demonstrate the efficacy of CKM

(C) $Y = 0.1(\beta_5'\mathbf{X}) - \exp(\beta_5'\mathbf{X})\epsilon$. The CS for this model is $\mathcal{S}(\beta_5)$. Methods targeting the CMS were removed from comparison while methods targeting the CS such as csOPG and SR (Wang and Xia 2008) were added for comparison. Both terms are monotone, so it is to the advantage of SIR. And we will illustrate the efficacy of CKF.

For each model, we consider two different simulation setups for $\mathbf{X}$. The first setup is $\mathbf{X} \sim N(0, I_p)$, so that this is an elliptic contour distribution while the other setup is $\frac{\mathbf{X}+2}{5} \sim \text{Beta}(0.75, 1)$ so that the linearity condition is violated. In these simulations,

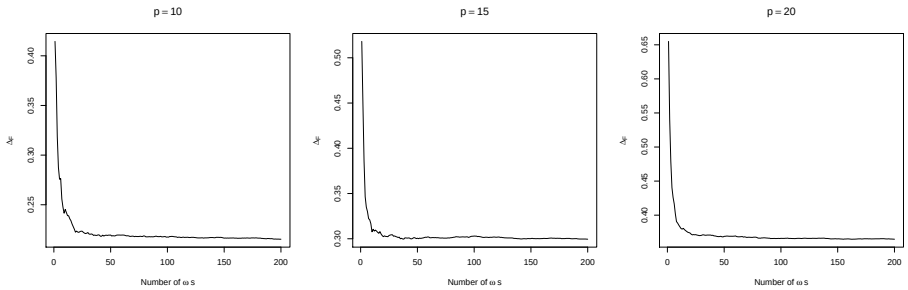


Figure 2.1: Simulation results for Δ_F vs Number of ωs based on Model (C).

Table 2.1: Accuracy for Model (A)

	Normal			Non-normal		
	p=10	p=20	p=60	p=10	p=20	p=60
n=200						
RND	1.770 (0.132)	1.909 (0.059)	1.967 (0.019)	1.770 (0.132)	1.909 (0.059)	1.967 (0.019)
SIR	1.298 (0.169)	1.410 (0.062)	1.551 (0.048)	1.078 (0.272)	1.287 (0.156)	1.599 (0.071)
PFC	1.0613 (0.271)	1.203 (0.191)	1.509 (0.087)	1.8372 (0.218)	1.000 (0.200)	1.553 (0.106)
SAVE	0.759 (0.300)	1.523 (0.124)	1.867 (0.085)	0.937 (0.313)	1.512 (0.088)	1.876 (0.097)
pHd	1.361 (0.094)	1.455 (0.064)	1.762 (0.053)	1.369 (0.089)	1.457 (0.045)	1.768 (0.064)
IHT	1.238 (0.122)	1.235 (0.117)	1.462 (0.117)	1.345 (0.107)	1.380 (0.113)	1.575 (0.076)
DR	0.553 (0.109)	0.903 (0.181)	1.582 (0.088)	0.590 (0.125)	0.997 (0.186)	1.634 (0.128)
rMAVE	0.136 (0.029)	0.220 (0.033)	0.958 (0.193)	0.200 (0.038)	0.337 (0.059)	1.291 (0.198)
CKM	0.448 (0.096)	0.709 (0.113)	1.361 (0.101)	0.549 (0.141)	0.851 (0.130)	1.552 (0.114)
n=400						
RND	1.770 (0.132)	1.909 (0.059)	1.967 (0.019)	1.770 (0.132)	1.909 (0.059)	1.967 (0.019)
SIR	1.224 (0.169)	1.341 (0.156)	1.459 (0.039)	0.856 (0.288)	1.012 (0.186)	1.381 (0.124)
PFC	0.933 (0.302)	1.089 (0.238)	1.398 (0.097)	0.600 (0.156)	0.744 (0.161)	1.275 (0.115)
SAVE	0.359 (0.073)	0.669 (0.117)	1.696 (0.062)	0.456 (0.084)	1.045 (0.271)	1.657 (0.093)
pHd	1.326 (0.098)	1.399 (0.074)	1.587 (0.045)	1.395 (0.036)	1.432 (0.027)	1.582 (0.031)
IHT	1.167 (0.143)	1.190 (0.147)	1.374 (0.107)	1.335 (0.082)	1.371 (0.066)	1.450 (0.122)
DR	0.371 (0.069)	0.566 (0.079)	1.148 (0.098)	0.388 (0.086)	0.625 (0.089)	1.177 (0.093)
rMAVE	0.077 (0.013)	0.124 (0.016)	0.498 (0.050)	0.125 (0.028)	0.186 (0.029)	0.707 (0.072)
CKM	0.312 (0.066)	0.481 (0.067)	0.972 (0.098)	0.410 (0.091)	0.591 (0.078)	1.113 (0.093)

we also assume that the true dimension d is known. For comparison purpose, we also include a benchmark method by randomly choosing d directions in $\mathbb{R}^p$, denote it as RND, where d denotes the true dimensions corresponding to each model.

For CKF, Figure 2.1 shows that the estimates for the simulations based on Model (C) is quite stable when the number of ωs goes beyond 50 or so. Thus, in our simulations, we choose the number of ωs to be 50.

From Table 2.1 and Table 2.2 CKM is consistently better than the rest of methods

Table 2.2: Accuracy for Model (B)

	Normal			Non-normal		
	p=10	p=20	p=60	p=10	p=20	p=60
n=200						
RND	1.790 (0.105)	1.897 (0.065)	1.961 (0.025)	1.790 (0.105)	1.897 (0.065)	1.961 (0.025)
SIR	1.820 (0.105)	1.908 (0.065)	1.970 (0.021)	1.524 (0.130)	1.711 (0.096)	1.909 (0.042)
PFC	1.648 (0.175)	1.749 (0.131)	1.946 (0.033)	1.382 (0.172)	1.511 (0.164)	1.852 (0.056)
SAVE	0.923 (0.250)	1.397 (0.206)	1.914 (0.067)	1.465 (0.148)	1.754 (0.103)	1.962 (0.024)
pHd	1.474 (0.047)	1.589 (0.086)	1.922 (0.044)	1.724 (0.106)	1.864 (0.070)	1.966 (0.023)
IHT	1.375 (0.182)	1.501 (0.239)	1.764 (0.226)	0.970 (0.210)	1.183 (0.133)	1.501 (0.204)
DR	1.109 (0.266)	1.582 (0.162)	1.935 (0.044)	1.437 (0.169)	1.748 (0.105)	1.944 (0.038)
rMAVE	0.822 (0.463)	1.431 (0.144)	1.934 (0.054)	0.841 (0.303)	1.317 (0.219)	1.902 (0.053)
CKM	0.760 (0.207)	1.332 (0.252)	1.917 (0.054)	1.189 (0.203)	1.550 (0.140)	1.888 (0.043)
n=400						
RND	1.790 (0.105)	1.897 (0.065)	1.961 (0.025)	1.790 (0.105)	1.897 (0.065)	1.961 (0.025)
SIR	1.777 (0.129)	1.903 (0.068)	1.973 (0.022)	1.419 (0.148)	1.600 (0.105)	1.815 (0.062)
PFC	1.651 (0.166)	1.774 (0.132)	1.944 (0.035)	1.210 (0.190)	1.369 (0.147)	1.749 (0.064)
SAVE	0.609 (0.148)	0.928 (0.174)	1.532 (0.149)	1.331 (0.143)	1.580 (0.085)	1.922 (0.049)
pHd	1.446 (0.025)	1.501 (0.050)	1.718 (0.082)	1.617 (0.097)	1.811 (0.100)	1.968 (0.025)
IHT	1.349 (0.169)	1.472 (0.212)	1.762 (0.245)	0.807 (0.171)	1.041 (0.188)	1.345 (0.216)
DR	0.678 (0.153)	1.098 (0.234)	1.697 (0.097)	1.360 (0.140)	1.587 (0.109)	1.909 (0.061)
rMAVE	0.303 (0.236)	1.109 (0.397)	1.630 (0.138)	0.301 (0.114)	0.890 (0.277)	1.737 (0.119)
CKM	0.501 (0.120)	0.805 (0.164)	1.689 (0.082)	0.853 (0.186)	1.214 (0.166)	1.771 (0.059)

except for rMAVE in Model (A) and (B) with normal distribution. However, when the dimension of predictor is relatively high, CKM perform better than or closer to rMAVE. From Table 2.3, CKF performs similarly in Model (C). Note that local methods, such as rMAVE behave better as the sample size increases, but the computation cost for rMAVE is also increasing. In Table 2.4 we compare the computational time for rMAVE and our method, and we suggest that when the sample size is getting larger, our method might be a better choice, as the accuracy of two methods are approximately same while CKM saves a lot of computing time.

For the non-normal cases, which indicate there is violation of assumptions, Model (A) suggests CKM is still the best among global methods, Model (B) suggests IHT in this case perform a little better than CKM. Model (C) shows that CKF is quite robust against departure from normality.

Overall, our method could provide a stable and well-rounded performance, while

Table 2.3: Accuracy for Model (C)

	Normal			Non-normal		
	p=10	p=20	p=60	p=10	p=20	p=60
n=200						
RND	1.344 (0.090)	1.364 (0.075)	1.381 (0.045)	1.344 (0.090)	1.364 (0.075)	1.381 (0.045)
SIR	0.217 (0.059)	0.370 (0.082)	0.655 (0.061)	0.231 (0.053)	0.373 (0.065)	0.655 (0.061)
PFC	0.272 (0.067)	0.745 (0.080)	0.682 (0.062)	0.263 (0.060)	0.687 (0.059)	0.682 (0.062)
SAVE	0.399 (0.170)	1.351 (0.095)	1.407 (0.010)	0.471 (0.247)	1.384 (0.040)	1.407 (0.010)
pHd	0.908 (0.204)	1.120 (0.162)	1.352 (0.036)	1.021 (0.171)	1.171 (0.105)	1.352 (0.036)
DR	0.201 (0.053)	0.422 (0.080)	0.686 (0.083)	0.227 (0.066)	0.351 (0.052)	0.686 (0.083)
SR	0.180 (0.047)	0.407 (0.170)	1.096 (0.301)	0.174 (0.053)	0.355 (0.237)	1.096 (0.301)
csOPG	0.213 (0.077)	0.564 (0.256)	1.228 (0.146)	0.222 (0.075)	0.599 (0.300)	1.228 (0.146)
rMAVE	0.779 (0.291)	1.023 (0.221)	1.325 (0.093)	0.896 (0.250)	1.059 (0.188)	1.325 (0.093)
CKF	0.210 (0.056)	0.365 (0.071)	0.689 (0.067)	0.222 (0.065)	0.352 (0.065)	0.689 (0.067)
n=400						
RND	1.344 (0.090)	1.364 (0.075)	1.381 (0.045)	1.344 (0.090)	1.364 (0.075)	1.381 (0.045)
SIR	0.165 (0.052)	0.274 (0.053)	0.517 (0.052)	0.184 (0.043)	0.247 (0.048)	0.386 (0.041)
PFC	0.220 (0.029)	0.323 (0.049)	0.504 (0.069)	0.217 (0.032)	0.300 (0.042)	0.464 (0.038)
SAVE	0.212 (0.068)	0.617 (0.245)	1.405 (0.012)	0.213 (0.045)	0.493 (0.178)	1.403 (0.014)
pHd	0.893 (0.224)	1.115 (0.128)	1.300 (0.046)	1.008 (0.168)	1.175 (0.091)	1.313 (0.049)
DR	0.148 (0.049)	0.262 (0.046)	0.511 (0.042)	0.146 (0.037)	0.226 (0.041)	0.400 (0.052)
SR	0.118 (0.039)	0.213 (0.055)	1.088 (0.316)	0.110 (0.034)	0.159 (0.030)	1.334 (0.371)
csOPG	0.128 (0.049)	0.256 (0.063)	1.268 (0.134)	0.125 (0.036)	0.212 (0.063)	1.124 (0.201)
rMAVE	0.760 (0.301)	0.969 (0.220)	1.230 (0.141)	0.798 (0.218)	0.952 (0.211)	1.298 (0.114)
CKF	0.160 (0.054)	0.274 (0.049)	0.510 (0.064)	0.169 (0.043)	0.240 (0.043)	0.420 (0.038)

Table 2.4: Comparison of Computation Time of rMave and CKM for Model (A)

	n=200				n=400				n=600			
	p=10	p=15	p=20	p=60	p=10	p=15	p=20	p=60	p=10	p=15	p=20	p=60
rMAVE	1.507 (0.209)	1.950 (0.103)	2.583 (0.147)	11.184 (0.490)	3.919 (0.059)	5.657 (0.049)	7.930 (0.067)	38.977 (0.314)	7.812 (0.141)	14.688 (0.929)	20.260 (1.185)	84.538 (1.715)
CKM	0.061 (0.016)	0.134 (0.056)	0.251 (0.073)	7.053 (0.491)	0.107 (0.011)	0.260 (0.055)	0.485 (0.107)	12.972 (0.966)	0.154 (0.024)	0.458 (0.088)	0.768 (0.134)	21.026 (2.661)

maintaining a lower computational cost. Other methods may be superior for specific combinations of model and predictor distribution, but none is consistently so and our methods provide a balanced approach that is relatively robust to these concerns.

Comparison of Order Determination

We now provide results on order determination for the previous three models. Simulations were conducted based on the following configurations: $n = 400$, $p = 10$ using $N = 200$ random samples. We applied different methods with their original methodology to estimate d. To be more specific, pHd, rMAVE, DR, SIR, SAVE, csOPG and

Table 2.5: Order Determination for Model (A)

	d=0	d=1	d=2	d=3	d=4	d≥5
n=400,p=10						
SIR	0.000	0.675	0.310	0.015	0.000	0.000
SAVE	0.040	0.880	0.075	0.005	0.000	0.000
pHd	0.000	0.940	0.045	0.010	0.000	0.005
IHT	0.000	1.000	0.000	0.000	0.000	0.000
DR	0.035	0.950	0.015	0.000	0.000	0.000
rMAVE	0.000	0.000	1.000	0.000	0.000	0.000
CKM	0.000	0.000	1.000	0.000	0.000	0.000

SR were all tested using sequential test in the respective papers. For IHT, although the author suggested sequential test, the paper did not provide the actual test statistic in detail. So we applied the ladle estimator for IHT, CKM and CKF. Simulation results can been found in Table 2.5, Table 2.6 and Table 2.7.

We can see that the ladle estimator in three cases worked fairly well. For Model (A), CHM and rMAVE picked up the correct dimension perfectly, while most other methods tended to underestimate it. In Model (B), our method had a bit underestimate when the sample size is not large enough, but it still provided a well-rounded performance, comparing with rMAVE which is the best. And in Model (C), CKF provided the best performance among other competitors, while no other method reached to a satisfactorily level. In our simulation studies, we noticed that as p gets larger, the ladle methods will not provide desirable results for some specific models, such as Model (B). This may be partly due to the fact that the consistency of ladle estimator is developed under $n \to \infty$, requiring that n relative large to p, and Model (B) is relatively harder to estimate than the other twos. In summary, the ladle estimator for CKM and CKF performs very well and is quite stable.

Comparison of Variable Selection

We now provide results for variable selection. Simulations were conducted based on the following configurations: $n = 200, 400, p = 20$ using $N = 200$ random samples. On each sample, we used CISE (Chen, Zou and Cook 2010) for kernel matrix-based

Table 2.6: Order Determination for Model (B)

	d=0	d=1	d=2	d=3	d=4	d≥5
n=400,p=10						
SIR	0.875	0.110	0.015	0.000	0.000	0.000
SAVE	0.030	0.805	0.155	0.010	0.000	0.000
pHd	0.000	0.000	0.925	0.065	0.005	0.005
IHT	0.005	0.995	0.000	0.000	0.000	0.000
DR	0.250	0.650	0.085	0.015	0.000	0.000
rMAVE	0.000	0.000	1.000	0.000	0.000	0.000
CKM	0.010	0.005	0.985	0.000	0.000	0.000

Table 2.7: Order Determination for Model (C)

	d=0	d=1	d=2	d=3	d=4	d≥5
n=400,p=10						
SIR	0.000	0.880	0.100	0.020	0.000	0.000
SAVE	0.920	0.040	0.040	0.000	0.000	0.000
pHd	0.940	0.040	0.000	0.020	0.000	0.000
DR	1.000	0.000	0.000	0.000	0.000	0.000
SR	0.000	0.100	0.600	0.300	0.000	0.000
csOPG	0.000	0.080	0.620	0.300	0.000	0.000
rMAVE	0.000	0.100	0.520	0.220	0.140	0.020
CKF	0.000	1.000	0.000	0.000	0.000	0.000

methods such as SIR, PFC, SAVE, pHd and CK methods. We use Wang and Yin (2008) for rMAVE, as CISE is not applicable to non-kernel method. We report three statistics: Specificity, fraction of irrelevant variables that are not being selected; Sensitivity, fraction of relevant variables that are being selected, and Run, fraction of runs in which the methods select both relevant and irrelevant covariates exactly right (Chen, Zou and Cook 2010), in Tables 2.8, 2.9 and 2.10.

In Model (A), when the predictor is normally distributed, we can see CKM is the best among all CISE based methods. And its performance is comparable to the local method, rMAVE. When the required assumption is violated, CKM is still quite good in terms of the sensitivity, which is what the researchers more care about, as one does not want to ignore the important signals.

In Model (B), where only symmetric terms are involved, pHd based on CISE is the most accurate when the predictor is nicely distributed. In that case, CKM provides comparable performance to pHd but outperforms other methods. However,

Table 2.8: Variable Selection for Model (A)

	Normal			Non-normal		
	sensitivity	specificity	runs	sensitivity	specificity	runs
n=200, p=20						
SIR	0.336	0.958	0.000	0.379	0.978	0.005
PFC	0.418	0.974	0.020	0.532	0.997	0.050
SAVE	0.586	0.866	0.000	0.513	0.942	0.000
pHd	0.765	0.653	0.005	0.760	0.630	0.000
rMAVE	1.000	0.762	0.055	0.990	0.747	0.065
CKM	0.950	0.743	0.040	0.982	0.250	0.000
n=400, p=20						
SIR	0.546	0.962	0.000	0.604	0.990	0.200
PFC	0.619	0.977	0.030	0.849	0.999	0.495
SAVE	0.790	0.803	0.025	0.708	0.967	0.005
pHd	0.802	0.548	0.000	0.697	0.863	0.005
rMAVE	1.000	0.792	0.080	1.000	0.772	0.025
CKM	0.999	0.838	0.240	0.998	0.248	0.000

Table 2.9: Variable Selection for Model (B)

	Normal			Non-normal		
	sensitivity	specificity	runs	sensitivity	specificity	runs
n=200, p=20						
SIR	0.100	0.896	0.000	0.530	0.941	0.180
PFC	0.180	0.908	0.000	0.630	0.956	0.320
SAVE	0.850	0.924	0.580	0.360	0.871	0.020
pHd	0.980	0.997	0.940	0.580	0.939	0.140
rMAVE	0.963	0.979	0.715	0.747	0.962	0.335
CKM	0.910	0.990	0.820	0.790	0.967	0.500
n=400, p=20						
SIR	0.100	0.900	0.020	0.770	0.968	0.420
PFC	0.180	0.908	0.000	0.910	0.987	0.780
SAVE	0.900	0.918	0.660	0.510	0.874	0.020
pHd	1.000	1.000	1.000	0.730	0.946	0.340
rMAVE	1.000	0.979	0.740	0.973	0.976	0.615
CKM	1.000	1.000	1.000	0.990	0.996	0.920

Table 2.10: Variable Selection for Model (C)

	Normal			Non-normal		
	sensitivity	specificity	runs	sensitivity	specificity	runs
n=200, p=20						
SIR	0.900	0.996	0.640	0.973	1.000	0.920
PFC	0.893	0.999	0.660	0.920	0.999	0.740
SAVE	0.280	0.724	0.000	0.180	0.834	0.000
pHd	0.900	0.158	0.000	0.980	0.082	0.000
rMAVE	0.932	0.429	0.000	0.972	0.227	0.000
CKF	0.893	1.000	0.680	0.927	1.000	0.005
n=400, p=20						
SIR	0.980	0.999	0.920	1.000	1.000	1.000
PFC	0.987	1.000	0.960	1.000	0.999	0.980
SAVE	0.367	0.673	0.000	0.393	0.836	0.100
pHd	0.920	0.202	0.000	1.000	0.067	0.000
rMAVE	0.977	0.220	0.000	0.985	0.121	0.000
CKF	0.973	0.996	0.860	0.987	1.000	0.960

Table 2.11: Sparse Variable Selection for Model (A)

	Normal			Non-normal		
	sensitivity	specificity	runs	sensitivity	specificity	runs
n=200, p=2000						
SIR	0.500	0.999	0.000	0.692	1.000	0.285
PFC	0.590	1.000	0.160	0.928	1.000	0.735
SAVE	0.800	0.999	0.220	0.730	0.999	0.270
pHd	0.650	0.999	0.200	0.662	1.000	0.280
CKM	1.000	1.000	1.000	0.975	1.000	0.665

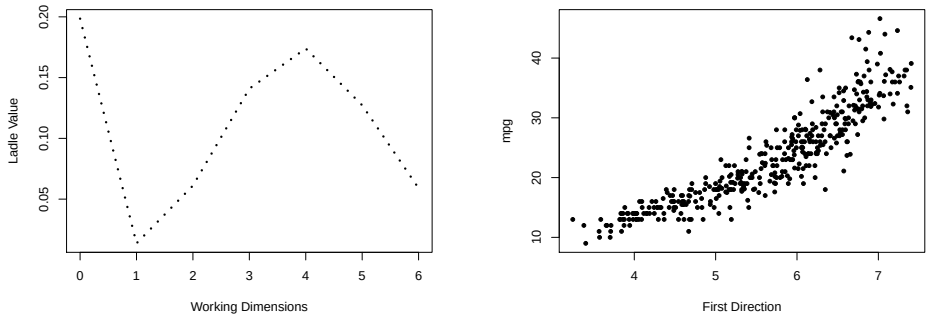
when the assumption is violated, CKM remains a solid performance while pHd shows vulnerability against such violation.

In Model (C), where the signal is monotone, CKF based on CISE provides good performance comparable to SIR and PFC no matter whether the assumption is met or not.

In addition, we also conducted variable selection under the large $p = 2000$ small $n = 200$ setup, and result is shown in Table 2.11. From the table, we can see that CKM approach is again very good and stable. Overall, our methods have stable and consistent performances across these models.

Auto MPG Data

We now consider a real dataset on 392 cases with complete record from UCI machine learning repository (`https://archive.ics.uci.edu/ml/datasets/auto+mpg`). The dataset was commonly used in predicting miles per gallon "mpg", and it was studied using sufficient dimension reduction methods (DCOV) by Sheng and Yin (2016). The continuous response Y is miles per gallon (mpg) attribute. And $\mathbf{X}$ is a 7 dimensional predictor vector consisting of: cylinders, displacement, horsepower, weight, acceleration, model year and origin. Unlike DCOV, we did not drop the origin but treating it as continuous variable, since cars made in United States (origin=1) tends to have the most cylinders, and cars made in Japan(origin=3) tend to have the fewest cylinders, while cars made in Europe(origin=2) tend to lie somewhere in the middle which indicate the origin does have an ordered relationship. Also, we did not perform transformation of predictors for normality to meet the assumption, since our simulation suggested CKF was robust to departure from linear condition.



(a) Ladle Plot (b) Response vs First Direction

Figure 2.2: Summary Plot for Auto MPG Data

The ladle plot results (Figure 2.2(a)) suggests the estimated d is 1. Figure 2.2(b) shows the scatter plot of the response variable against the estimated CKF direction. This direction indicates a strong linear trend, with some curvature. The result is slightly different from DCOV, because we include one more variable (origin) in the

Table 2.12: Average Distance Using Bootstrap Sample

Methods	pHd	SR	DR	SIR	SAVE	rMAVE	CKF
Δ_F	0.594 (0.354)	0.574 (0.259)	0.601 (0.257)	0.536 (0.256)	0.719 (0.299)	0.529 (0.247)	0.480 (0.257)

Table 2.13: Variable Selection on Bootstrap Sample

	SIR	PFC	SAVE	pHd	rMAVE	CKF
sensitivity	1.000	1.000	0.000	0.198	1.000	1.000
specificity	0.999	1.000	0.797	0.690	0.498	1.000
runs	0.995	1.000	0.000	0.000	0.095	1.000

analysis, which may reduce the dimension due to the association between origin and rest of the variables.

For accuracy comparison, we used bootstrap method to generate 200 datasets and calculated the distance between estimates from the data and from the bootstrap sample. Results (Table 2.12) report the mean and error from these distance for each method, and show that cubic kernel has a very good/stable performance. This may be due to the fact that our method is robust against mild violation of the commonly used conditions.

We use the following scheme to assess the accuracy of variable selection. We generated the 200 datasets as follows: First, we set the sample of size 392 with weight and model year, as weight and model year are two variables selected by most methods. And then we independently selected 392 data points from the remaining predictors and combined them to make one generated dataset. As a result, we forced the relevant variables as weight and model year in the generated datset. Simulation results for selecting the two variables: weight and model year, is shown in Table 2.13, and CKF provided the best performance.

2.7 Discussion

In this chapter, we proposed a cubic kernel for estimating L-functional of the T-central subspace, in particularly, we developed details for estimating central mean subspace and central subspace and the respective variable selection. Simulation results show the strength and usefulness of cubic kernel methods: CKM and CKF, especially, that they are quite robust against distributions of the predictors and their performances are not only very good but also stable.

There are also some interesting extension of our methods. For order determination, ladle estimator is quite effective for large n relatively to small p, as its consistent result is developed under $n \to \infty$. Otherwise, results will be broken down. Thus developing new approach for Ladle estimator under $p >> n$ may be interesting, such as those developed by Zhu, Miao and Peng (2006) adapting information theory. For CMS, we can simply extend univariate Y to multivariate $\mathbf{Y} \in \mathbb{R}^q$ as $\mathcal{S}_{\mathrm{E}(\mathbf{Y}|\mathbf{X})} = \oplus_{i=1}^{q} \mathcal{S}_{\mathrm{E}(Y_i|\mathbf{X})}$, where Y_i represent the ith element of $\mathbf{Y}$, see Cook and Setodji (2003). And for CS, multivariate responses $\mathbf{Y} \in \mathbb{R}^q$, could also easily extended by considering multivariate Fourier transformation. Note also, in stead of conditional characteristic function, other transformation may be used to recover the CS. For instance, $\{I(Y < y), \mathbf{X}\}$ for each $y \in \mathbb{R}$, which leads to slicing methods (Wang and Xia 2008). However, its development will be very similar to our CKF.

Chapter 3 Cubic Kernel Method for Implicit T-Central Subspace

3.1 Introduction

How to properly summarize high dimensional data is always an important topic in statistics. Simplifying the high dimension data to a low dimension structure could provide benefit for multiple statistical topics such as: data exploration, non-parametric estimation, data visualization, etc. Therefore, dimension reduction is a useful topic in statistics and many methods have been proposed to achieve such a goal. The most important goal for reducing dimension is to preserve the information of interests: co-variance structure of the predictors; relative distance between points; correlation structures etc. Concentrating regression information about a response might be the most important goal of regression, and SDR (Cook, 1996) offers such a tool to achieve this goal. In this chapter, we will follow the work from Chapter 2 to extend our estimator for T-central subspace targeting I-functional.

The rest of this chapter is organized as follow: Section 3.2 provides a general theory for our proposed method. Section 3.3 further develops our method for two specific functional: conditional quantile and conditional expectile. Section 3.4 explains methods of order determination. Section 3.5 establishes the asymptotic properties for our proposed methodology. Sections 3.6 and 3.7 demonstrate the advantages of our estimator via simulation example and real data analysis.

3.2 Cubic Kernel for I-functional

Similar to Chapter 2, we would like to make improvement to the CK estimator, which only targets L-functional, so that I-functional could also be handled. We will still restrict our attention to the loss function $L(\cdot, \cdot)$ that depends on $\mathbf{Z}$ only through the

kernel function $K(\mathbf{Z}) : \mathbb{R}^p \to \mathbb{R}^1$:

$$L(K(\mathbf{Z}), \mathcal{T}(Y \mid \mathbf{Z})) = -\mathcal{T}(Y \mid \mathbf{Z})K(\mathbf{Z}) + \phi(K(\mathbf{Z})),$$

where $\phi(\cdot)$ is a convex function and $\mathcal{T}(Y \mid \mathbf{Z})$ is the conditional statistical functional induced by an I-functional, $\mathcal{T}$. So the loss function is convex in $K(\mathbf{Z})$. $K(\mathbf{Z})$ is a polynomial defined by

$$K(\mathbf{Z}) = a + \mathbf{b}'\mathbf{Z} + \mathbf{Z}'\mathbf{C}\mathbf{Z} + (\mathbf{Z}' \otimes \mathbf{Z}')\mathcal{D}\mathbf{Z},$$

with $a \in \mathbb{R}^1$, $\mathbf{b} \in \mathbb{R}^p$ and $\mathbf{C}$ being a $p \times p$ symmetric matrix and $\mathcal{D}$ being a $p \times p \times p$ array such that if $\mathbf{D}_k$ be the k-th face of $\mathcal{D}$ and let d_{ijk} be the element of i-th row and j-th column in the $\mathbf{D}_k$. Then all d_{ijk} are the same for any permutation of the index ijk, and $\mathbf{D}_k' = \mathbf{D}_k$, we may treat $\mathcal{D}$ as a $p^2 \times p$ matrix. Objective functions of this form typically correspond to natural exponential families. We will develop our theory with this more general forms but for our algorithms, we will mainly use square loss.

$K(\mathbf{Z})$ is the cubic kernel designed for summarizing the data. $K(\mathbf{Z})$ itself does not represent the belief that we hold for the true underlying regression function, but rather an sufficient approximation to the truth. Polynomial with different orders could be a good approximation for many regression function, as lots of them are smooth. As the example in Chapter 2 showed, information picked by higher order polynomials are often correlated with information picked by the cubic kernels. So we believe a cubic kernel will be a perfect balance between the complexity of the summarizing function and deepness of information that could be captured.

To handle the I-functional, the major change on the loss function that we made lies in the target: $T(Y)$, the transformation induced on Y by L-functional $\mathcal{T}$, is replaced by $\mathcal{T}(Y \mid \mathbf{Z})$, the random variable induced by the conditional I-functional,

which makes the loss function measurable to $\sigma(\mathbf{Z})$.

On the population level, the estimate is assumed to be based on the risk:

$$R(a, \mathbf{b}, \mathbf{C}, \mathcal{D}) = \mathrm{E}\{L[a + \mathbf{b}'\mathbf{Z} + \mathbf{Z}'\mathbf{C}'\mathbf{Z} + (\mathbf{Z}' \otimes \mathbf{Z}')\mathcal{D}\mathbf{Z}, \mathcal{T}(Y \mid \mathbf{Z})]\}.$$

Based on above risk function, we reproduce Proposition 2.1 as follows:

Proposition 3.1. *Let* $\alpha, \boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{\Delta} = \arg\min_{a,\mathbf{b},\mathbf{C},\mathcal{D}} R(a, \mathbf{b}, \mathbf{C}, \mathcal{D})$. *Let* $\mathcal{S}_{T(Y|\mathbf{Z})}$ *have basis matrix* $\boldsymbol{\gamma}$, *i.e.* $\mathcal{S}(\boldsymbol{\gamma}) = \mathcal{S}_{T(Y|\mathbf{Z})}$. *Assume that* $\mathrm{E}(\mathbf{Z} \mid \boldsymbol{\gamma}'\mathbf{Z})$ *is linear,* $\mathrm{Var}(\mathbf{Z} \mid \boldsymbol{\gamma}'\mathbf{Z})$ *is constant and* $\mathcal{M}^{(3)}(\mathbf{Z} \mid \boldsymbol{\gamma}'\mathbf{Z}) = 0$, *where* $\mathcal{M}^{(3)}(\mathbf{Z} \mid \boldsymbol{\gamma}'\mathbf{Z}) = \mathrm{E}(\mathbf{Z} \otimes \mathbf{Z}\mathbf{Z}' \mid \boldsymbol{\gamma}'\mathbf{Z})$. *Then* $\mathcal{S}(\boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{\Delta}') \subseteq \mathcal{S}_{T(Y|\mathbf{Z})}$.

To better assist the development of our methodology, we need to point out that estimation of $\mathcal{T}(Y \mid \mathbf{Z})$ can not be avoided, which indicates that we need to evaluate $\hat{\mathcal{T}}(Y \mid \mathbf{Z})$ using some existing method. However, p, the dimension of $\mathbf{Z}$, is typically very high and lead to the curse of dimensionality. The high dimension of $\mathbf{Z}$ would not only lead to a problematic results due to sparsity in the high volume space, but also prevent the estimator from being efficacious. To improve the accuracy, reducing the dimension before implementing our algorithm is necessary. In stead of including $\mathcal{T}(Y \mid \mathbf{Z})$ in the loss function, we would consider $\mathcal{T}(Y \mid \mathbf{B}'_{CS}\mathbf{Z})$ instead, where $\mathbf{B}_{CS} \in \mathbb{R}^{p \times d}$ stands for a basis for the central subspace $\mathcal{S}_{Y|\mathbf{Z}}$, i.e. $\mathcal{S}(\mathbf{B}_{CS}) = \mathcal{S}_{Y|\mathbf{Z}}$. In such a case, the prediction problem is much easier to solve, because of the lower dimension. It's not hard to figure that we have no loss of information, as it's clear that $\mathcal{T}(Y \mid \mathbf{B}'_{CS}\mathbf{Z}) = \mathcal{T}(Y \mid \mathbf{Z})$ a.s.. In light of this, we also need to provide another version of Proposition 3.1 so that it works under the framework with reduced dimension of $\mathbf{Z}$.

Denote $\mathbf{Z}^* = \mathbf{B}'_{CS}\mathbf{Z}$. On the population level, the estimate is assumed to be based

on the risk:

$$R^*(a, \mathbf{b}, \mathbf{C}, \mathcal{D}) = \mathrm{E}\{L[a + \mathbf{b}'\mathbf{Z}^* + \mathbf{Z}^{*\prime}\mathbf{C}'\mathbf{Z}^* + (\mathbf{Z}^{*\prime} \otimes \mathbf{Z}^{*\prime})\mathcal{D}\mathbf{Z}^*, T(Y|\mathbf{Z}^*)]\}.$$

with $a \in \mathbb{R}^1$, $\mathbf{b} \in \mathbb{R}^d$, $\mathbf{C}$ being a $d \times d$ symmetric matrix and $\mathcal{D}$ being a $d \times d \times d$ array such that all d_{ijk} are the same for any permutation of the index ijk, and $\mathbf{D}'_k$ is a symmetric matrix.

The following proposition could be easily carried out.

Proposition 3.2. *Let* $\alpha^*, \boldsymbol{\beta}^*, \boldsymbol{\Gamma}^*, \boldsymbol{\Delta}^* = \arg\min_{a,\mathbf{b},\mathbf{C},\mathcal{D}} R^*(a, \mathbf{b}, \mathbf{C}, \mathcal{D})$. *Under the assumption of Proposition 3.1, we have*

$$\mathcal{S}(\mathbf{B}_{CS}\boldsymbol{\beta}^*, \mathbf{B}_{CS}\boldsymbol{\Gamma}^*\mathbf{B}'_{CS}, \mathbf{B}_{CS}\boldsymbol{\Delta}^{*\prime}(\mathbf{B}'_{CS} \otimes \mathbf{B}'_{CS})) \subseteq \mathcal{S}_{T(Y|\mathbf{Z})}.$$

The proof could be completed by using Proposition 3.1 along with the fact that $\mathcal{T}(Y \mid \mathbf{B}'_{CS}\mathbf{Z}) = \mathcal{T}(Y \mid \mathbf{Z})$ a.s.. Proposition 3.2 reveals that our two-stage estimator given by $\mathcal{S}(\mathbf{B}_{CS}\boldsymbol{\beta}^*, \mathbf{B}_{CS}\boldsymbol{\Gamma}^*\mathbf{B}'_{CS}, \mathbf{B}'_{CS}\boldsymbol{\Delta}^{*\prime}(\mathbf{B}_{CS} \otimes \mathbf{B}_{CS}))$ does lie in $\mathcal{S}_{T(Y|\mathbf{Z})}$, the desired T-central subspace. However, in practice, given the estimator for $\mathcal{T}(Y \mid \mathbf{Z}^*)$, it's still quite difficult to solve the estimation problem. Difficulty arises in the following perspective: This optimization problem has parameters of the order of $\mathcal{O}(d^3)$, which requires searching for solution on a tensor space. On the other hand, although by assumption, the objective function is convex in the polynomial $K(\mathbf{Z}^*)$, it's not necessarily convex in the parameters, which making this proposition less useful. However, we could get away with such obstacles by introducing the following kernels. Let

$$\begin{aligned}
\boldsymbol{\beta}_{T\mathbf{Z}} &= \mathrm{E}[\mathcal{T}(Y \mid \mathbf{Z}^*)\mathbf{Z}^*] \quad \boldsymbol{\Sigma}_{T\mathbf{Z}\mathbf{Z}} = \mathrm{E}\left\{\{\{\mathcal{T}(Y \mid \mathbf{Z}^*) - \mathrm{E}[\mathcal{T}(Y \mid \mathbf{Z}^*)]\}\,\mathbf{Z}^*\mathbf{Z}^{*\prime}\}\right\} \\
\mathcal{M}_{T\mathbf{Z}\mathbf{Z}\mathbf{Z}} &= \mathrm{E}\left\{\mathcal{T}(Y \mid \mathbf{Z}^*) - \mathrm{E}[\mathcal{T}(Y \mid \mathbf{Z}^*)]\right\}\mathbf{Z}^* \otimes \mathbf{Z}^*\mathbf{Z}^{*\prime} - \mathrm{E}[\mathcal{T}(Y \mid \mathbf{Z}^*)\mathbf{Z}^*] \otimes I \\
&\quad - I \otimes \mathrm{E}[\mathcal{T}(Y \mid \mathbf{Z}^*)\mathbf{Z}^*] - \mathrm{vec}(I)\mathrm{E}[\mathcal{T}(Y \mid \mathbf{Z}^*)\mathbf{Z}^*]'
\end{aligned}$$

where $\boldsymbol{\beta}_{T\mathbf{Z}}$ represents OLS slope estimate for regression of $\mathcal{T}(Y \mid \mathbf{Z}^*)$ on $\mathbf{Z}^*$, $\boldsymbol{\Sigma}_{T\mathbf{ZZ}}$ represents the kernel for principal Hessian directions, $\mathcal{M}_{T\mathbf{ZZZ}}$ is the kernel for third moment (Yin and Cook 2004). The next result provides a further connection that may lead to an easy computational algorithm.

Proposition 3.3. *Suppose* $\boldsymbol{\alpha}$ *serves as a basis for* $\mathcal{S}(\boldsymbol{\beta}_{T\mathbf{Z}}, \boldsymbol{\Sigma}_{T\mathbf{ZZ}}, \mathcal{M}'_{T\mathbf{ZZZ}})$. *Assume that* $\mathrm{E}(\mathbf{Z}^* \mid \boldsymbol{\alpha}'\mathbf{Z}^*)$ *is linear,* $\mathrm{Var}(\mathbf{Z}^* \mid \boldsymbol{\alpha}'\mathbf{Z}^*)$ *is constant and* $\mathcal{M}^{(3)}(\mathbf{Z}^* \mid \boldsymbol{\alpha}'\mathbf{Z}^*) = 0$. *Then*

$$\mathcal{S}(\boldsymbol{\beta}^*, \boldsymbol{\Gamma}^*, \boldsymbol{\Delta}^*) \subseteq \mathcal{S}(\boldsymbol{\beta}_{T\mathbf{Z}}, \boldsymbol{\Sigma}_{T\mathbf{ZZ}}, \mathcal{M}_{T\mathbf{ZZZ}})$$

If in addition, $\mathrm{E}(\mathbf{Z} \mid \mathbf{B}'_{CS}\mathbf{Z})$ *is linear,* $\mathrm{Var}(\mathbf{Z} \mid \mathbf{B}'_{CS}\mathbf{Z})$ *is constant and* $\mathcal{M}^{(3)}(\mathbf{Z} \mid \mathbf{B}'_{CS}\mathbf{Z}) = \mathbf{0}$, *then*

$$\begin{aligned}
&\mathcal{S}(\mathbf{B}_{CS}\boldsymbol{\beta}^*, \mathbf{B}_{CS}\boldsymbol{\Gamma}^*\mathbf{B}'_{CS}, \mathbf{B}'_{CS}\boldsymbol{\Delta}^{*\prime}(\mathbf{B}_{CS} \otimes \mathbf{B}_{CS})) \\
\subseteq &\mathcal{S}(\mathbf{B}_{CS}\boldsymbol{\beta}_{T\mathbf{Z}}, \mathbf{B}_{CS}\boldsymbol{\Sigma}_{T\mathbf{ZZ}}\mathbf{B}'_{CS}, \mathbf{B}_{CS}\mathcal{M}'_{T\mathbf{ZZZ}}(\mathbf{B}'_{CS} \otimes \mathbf{B}'_{CS})) \\
\subseteq &\mathcal{S}_{T(Y|\mathbf{Z})}
\end{aligned}$$

Proof for Proposition 3.3 is similar to Proposition 2.2 in Chapter 2. It conveys an important message that the population minimizer provides information no more than the existing kernels $(\boldsymbol{\beta}_{T\mathbf{Z}}, \boldsymbol{\Sigma}_{T\mathbf{ZZ}}, \mathcal{M}_{T\mathbf{ZZZ}})$. This information combining Propositions 3.2 and 3.3 is valuable: we could rely on these kernels to recover the desired T-central subspace and then utilize proposition 3.3 to perform a quadratic fit to improve the result without providing extra information outside of the space spanned by the kernel matrices.

The estimation of $\boldsymbol{\beta}_{T\mathbf{Z}}, \boldsymbol{\Sigma}_{T\mathbf{ZZ}}, \mathcal{M}_{T\mathbf{ZZZ}}$ can be done consistently by substituting with the sample moments. However, there are still several caveats for implementing the moment methods.

1. In theory, we do need a consistent estimator $\hat{\mathbf{B}}_{CS}$ for the basis of CS. In practice, we suggest choosing a consistent methodology but with a higher structural dimension. As we want the estimate $\hat{\mathcal{T}}(Y \mid \hat{\mathbf{B}}_{CS}\mathbf{Z})$ as close to $\mathcal{T}(Y \mid \mathbf{Z})$ as possible. So overfitting is the favor of this application. A more practical suggestion is to treat the dimension of CS as a constant, as we do not need precise knowledge of the structural dimension for CS, but rather prefer overfitting. Moreover, we could save computational time, as the order determination for CS typically much slower than the method itself. For our simulation study in Section 3.6, we fixed the "guess" of structural dimension of CS as 3, as all the models we used did not go beyond 2 dimensions.

2. Similar to Chapter 2, constructing kernel matrices using the residual could improve the efficiency of the estimator because of the correlations among the linear, quadratic and third moment kernels. When constructing the CK estimator, we may use the residuals instead of the original response. To be more specific, take $r_1 = \mathcal{T}(Y \mid \mathbf{Z}^*) - a_1 - b_1\mathbf{Z}^{*\prime}\boldsymbol{\beta}_{Y\mathbf{Z}}$, with $a_1, b_1 \in \mathbb{R}$ being the OLS coefficients of $\mathcal{T}(Y \mid \mathbf{Z}^*)$ on $(1, \mathbf{Z}^{*\prime}\boldsymbol{\beta}_{Y\mathbf{Z}})$, then construct

$$\boldsymbol{\Sigma}_{r_1\mathbf{ZZ}} = \mathrm{E}(r_1\mathbf{Z}^*\mathbf{Z}^{*\prime}).$$

   Take $r_2 = \mathcal{T}(Y \mid \mathbf{Z}^*) - a_2 - b_2\mathbf{Z}^{*\prime}\boldsymbol{\beta}_{Y\mathbf{Z}} - c_2\mathbf{Z}^{*\prime}\boldsymbol{\Sigma}_{r_1\mathbf{ZZ}}\mathbf{Z}^*$, with $a_2, b_2, c_2 \in \mathbb{R}$ being the OLS coefficients of $\mathcal{T}(Y \mid \mathbf{Z}^*)$ on $(1, \mathbf{Z}^{*\prime}\boldsymbol{\beta}_{Y\mathbf{Z}}, \mathbf{Z}^{*\prime}\boldsymbol{\Sigma}_{r_1\mathbf{ZZ}}\mathbf{Z}^*)$ and construct

$$\mathcal{M}_{r_2\mathbf{ZZZ}} = \mathrm{E}(r_2\mathbf{Z}^* \otimes \mathbf{Z}^*\mathbf{Z}^{*\prime}) - \mathrm{E}(r_2\mathbf{Z}^*) \otimes \mathbf{I} - \mathbf{I} \otimes \mathrm{E}(r_2\mathbf{Z}^*) - \mathrm{vec}(\mathbf{I})\mathrm{E}(r_2\mathbf{Z}^*)'.$$

3. The three dimensional array can be written as a $d_{CS}^2 \times d_{CS}$ matrix, where this matrix can be considered as a column of d_{CS} blocks with each block being a $d_{CS} \times d_{CS}$ matrix. However, this $d_{CS}^2 \times d_{CS}$ matrix contains only $\frac{d_{CS}(d_{CS}+1)}{2}$

unique ones. We then construct a new unique kernel matrix, $(\tilde{\mathcal{M}}'_{T\mathbf{ZZZ}})_{d_{CS}\times\frac{d_{CS}(d_{CS}+1)}{2}}$ by removing the repeated columns, so that $\mathcal{S}(\mathcal{M}'_{T\mathbf{ZZZ}}) = \mathcal{S}(\tilde{\mathcal{M}}'_{T\mathbf{ZZZ}})$.

We are now ready to provide the details of our algorithm for estimating the T-central subspace, which we named as *Cubic Kernel for T-Central Subspace (CKT).*

Algorithm for CKT: Let $\{\mathbf{X}_i, Y_i\}, i = 1, 2, \cdots, n$ be an i.i.d sample, and assume that the structural dimension d for the T-central subspace is known. Then

1. Standardize the predictor $\hat{\mathbf{Z}}_i = \hat{\mathbf{\Sigma}}_{\mathbf{X}}^{-1/2}(\mathbf{X}_i - \bar{\mathbf{X}})$, for $i = 1\cdots, n$.

2. Choose $d_{CS} \geq d$ as a constant. Apply existing methodology for finding CS to obtain the esitmated basis $\hat{\mathbf{B}}_{CS;p\times d_{CS}}$ for CS. Denote $\hat{\mathbf{Z}}_i^* = \hat{\mathbf{B}}'_{CS}\hat{\mathbf{Z}}_i$.

3. Use existing nonparametric estimator to obtain an estimate of $\hat{\mathcal{T}}(Y_i \mid \hat{\mathbf{B}}_{CS}\hat{\mathbf{Z}}_i) = \hat{\mathcal{T}}(Y_i \mid \hat{\mathbf{Z}}_i^*)$.

4. Construct kernel matrices:

$$\hat{\boldsymbol{\beta}}_{T\mathbf{Z}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{Z}}_i^*\hat{\mathcal{T}}(Y_i \mid \hat{\mathbf{Z}}_i^*), \hat{\mathbf{\Sigma}}_{r_1\mathbf{ZZ}} = \frac{1}{n}\sum_{i=1}^{n}\hat{r}_{1i}\hat{\mathbf{Z}}_i^*\hat{\mathbf{Z}}_i^{*\prime},$$

   where $\hat{r}_{1i}$ is the estimated version of r_1 for ith observation.

5. Construct $\hat{\mathcal{M}}_{r_2\mathbf{ZZZ}}$ and $\hat{\tilde{\mathcal{M}}}_{r_2\mathbf{ZZZ}}$. First we construct p^2 vectors for $j, k = 1, \cdots, p$:

$$\begin{aligned}\hat{\mathbf{m}}_{jk} = &\frac{1}{n}\sum_{i=1}^{n}\hat{r}_{2i}\hat{\mathbf{Z}}_i^*(\mathbf{e}_j'\hat{\mathbf{Z}}_i^*)(\mathbf{e}_k'\hat{\mathbf{Z}}_i^*)\\ &-\frac{1}{n}\sum_{i=1}^{n}\hat{r}_{2i}(\mathbf{e}_k'\hat{\mathbf{Z}}_i^*)\mathbf{e}_j - \frac{1}{n}\sum_{i=1}^{n}\hat{r}_{2i}(\mathbf{e}_j'\hat{\mathbf{Z}}_i^*)\mathbf{e}_k - \frac{1}{n}\sum_{i=1}^{n}\hat{r}_{2i}(\mathbf{e}_j'\mathbf{e}_k)\hat{\mathbf{Z}}_i^*,\end{aligned}$$

   where $\mathbf{e}_i$ is a unit vector with ith element 1 and $\hat{r}_{2i}$ is ith element in the residual vector $\hat{r}_2$. Let $\hat{\mathcal{M}}'_{r_2\mathbf{ZZZ}} = (\hat{\mathbf{m}}_{11}, \cdots, \hat{\mathbf{m}}_{pp})$, which is a $p \times p^2$ matrix. Then take

$\hat{\mathbf{m}}_{jk}$ with $j \geq k$ and let $\hat{\tilde{\mathcal{M}}}'_{r_2\mathbf{ZZZ}} = (\hat{\mathbf{m}}_{11}, \hat{\mathbf{m}}_{21}, \cdots, \hat{\mathbf{m}}_{pp})$, which is a $p \times \frac{p(p+1)}{2}$ matrix consisting of all the unique columns for $\hat{\mathcal{M}}'_{r_2\mathbf{ZZZ}}$.

6. Construct $\mathbf{S}_i = (1, S_{1i}, S_{2i}, S_{3i})'$, with $S_{1i} = \hat{\mathbf{Z}}_i^{*\prime}\hat{\boldsymbol{\beta}}_{Y\mathbf{Z}}$, $S_{2i} = \hat{\mathbf{Z}}_i^{*\prime}\hat{\boldsymbol{\Sigma}}_{r_1\mathbf{ZZ}}\hat{\mathbf{Z}}_i^{*}$ and $S_{3i} = \hat{\mathbf{Z}}_i^{*\prime} \otimes \hat{\mathbf{Z}}_i^{*\prime}\hat{\mathcal{M}}_{r_2\mathbf{ZZZ}}\hat{\mathbf{Z}}_i^{*}$. Find the OLS fit of $\hat{\mathcal{T}}(Y_i \mid \hat{\mathbf{Z}}_i^{*})$ on $\mathbf{S}_i$ and denote the fitted coefficient as $\hat{\boldsymbol{w}} = (\hat{w}_0, \hat{w}_1, \hat{w}_2, \hat{w}_3)'$.

7. Construct $\hat{\mathbf{M}} = \hat{w}_1^2\hat{\boldsymbol{\beta}}_{Y\mathbf{Z}}\hat{\boldsymbol{\beta}}'_{T\mathbf{Z}} + \hat{w}_2^2\hat{\boldsymbol{\Sigma}}_{r_1\mathbf{ZZ}}\hat{\boldsymbol{\Sigma}}'_{r_1\mathbf{ZZ}} + \hat{w}_3^2\hat{\tilde{\mathcal{M}}}'_{r_2\mathbf{ZZZ}}\hat{\tilde{\mathcal{M}}}_{r_2\mathbf{ZZZ}}$, where $\hat{w}'_i s$ are obtained in step 4. Perform the eigen decomposition on $\hat{\mathbf{M}}$, our estimator for the T-central subspace is given by

$$\mathcal{S}(\hat{\boldsymbol{\Sigma}}_{\mathbf{X}}^{-1/2}\hat{\mathbf{B}}_{CS}\hat{\mathbf{u}}_1, \hat{\boldsymbol{\Sigma}}_{\mathbf{X}}^{-1/2}\hat{\mathbf{B}}_{CS}\hat{\mathbf{u}}_2, \cdots, \hat{\boldsymbol{\Sigma}}_{\mathbf{X}}^{-1/2}\hat{\mathbf{B}}_{CS}\hat{\mathbf{u}}_d),$$

where $\hat{\mathbf{u}}_i, i = 1 \cdots d$ are the eigenvectors corresponding to the first d eigenvalues of $\hat{\mathbf{M}}$.

Note that the weights $\hat{w}'_i s$ in step 4 are a key component of the CK estimator: They adapt the importance of the individual kernels, $(\hat{\boldsymbol{\beta}}_{T\mathbf{Z}}, \hat{\boldsymbol{\Sigma}}_{T\mathbf{ZZ}}, \hat{\mathcal{M}}_{T\mathbf{ZZZ}})$ in $\hat{\mathbf{M}}$. Therefore, bigger values of $\hat{w}'_i s$ will indicate that the corresponding kernel is more important.

3.3 T-Central Subspace for Particular Functionals

In this section, we are going to apply the theory developed in previous section on several implicit conditional functional such as quantile and expectile. Luo, Li and Yin (2014) provided an efficient estimator, which is not necessarily the best for finite sample performance. Moreover, their estimator requires the derivation of the efficient score and the efficient information for the semiparametric problem. In their paper, formula for conditional mean, conditional variance and conditional quantile were derived, but conditional expectile was left alone and not developed. In this section,

we are going to show that, under the framework of our methodology, its rather easy to implement our algorithm as one does not need to derived different formulas for different type of conditional statistical functional. As it was pointed out in previous section, these approaches depend on an estimate of the conditional quantile or expectile. For those parts, we are going to rely on some of the existing nonparametric estimators.

Conditional Quantile

Quantile regression (Koenker and Bassett, 1978) has been playing a prominent role in a wide range of statistical applications and being a popular tool for many statistical studies in the past few decades. Comparing to ordinary least square, quantile regression is very useful when the behavior at different levels is of the main interests rather than the mean behavior. This allows one to explore the changes in effect of predictors at different levels. It also enjoys the benefit of robustness against extreme values which makes it stand out from the traditional regression in mean. Being able to handle heteroscedasticity in regression makes quantile regression earn its role in modern statistical applications.

For any fixed $\tau \in (0,1)$, then the τ-th conditional quantile of Y given $\mathbf{X} = \mathbf{x} \in \mathbb{R}^p$ is defined as

$$Q_\tau(Y \mid \mathbf{X} = \mathbf{x}) = \inf\{y : P(Y \leq y \mid \mathrm{X} = \mathbf{x}) \geq \tau\} = \arg\min_m \mathrm{E}[\rho_\tau(Y - m) \mid \mathbf{X} = \mathbf{x}],$$

with $\rho_\tau(t) = (2\tau - 1)t + |t|$.

Many existing literatures considered the estimation of the conditional quantile funciton, see for example: Koenker and Bassett (1978) for a direct approach; Chaudhuri (1991), Hong (2003) for local polynomial estimators with Bahadur representation and Takeuchi et al. (2006) for nonparametric estimator for conditional quantile.

In addition to its wide range of statistical application, conditional quantile is also being studied in SDR. In fact, the idea of using conditional quantile to recover CS was first introduced by Kong and Xia (2014) to address the inefficiency of the some embedded estimation procedure using the conditional density or conditional distribution function. Using conditional quantile estimate could alleviate the negative effect brought up by the non-linearity of the conditional density, as the conditional quantile function is at least piecewise linear in some of those cases, which results in the larger bandwidth (more data to use). By applying conditional quantile in their paper, they were able to gain advantage of: minimal assumption; exhaustively estimation of CS; robust against outliers, etc. Aside from recovering CS, quantile regression was also studied in Luo, Li and Yin (2014) where a one step estimator under semi-parametric framework was proposed to extract the direction in a specific quantile.

Let $(\Omega_Y, P_Y, \mathfrak{F}_Y), (\Omega_{\mathbf{Z}}, P_{\mathbf{Z}}, \mathfrak{F}_{\mathbf{Z}})$ be the probability space associate with Y and $\mathbf{Z}$, respectively. Let $\mathcal{F}_{Y|\mathbf{Z}} = \{f_{\mathbf{z}}, \mathbf{z} \in \Omega_{\mathbf{Z}} : P_{Y|\mathbf{Z}=\mathbf{z}}(A) = \int_A f_{\mathbf{z}}\, dP_Y, \forall A \in \mathfrak{F}_Y\}$ be a class of densities with respect to P_Y of the conditional distribution $P_{Y|\mathbf{Z}}$. If we consider the functional $\mathcal{T}_\tau^{(1)} : \mathcal{F} \to \mathbb{R}$ that is defined by the value of m that equates

$$\int_{\Omega_Y} \rho_\tau(t-m) f_{\mathbf{z}} P_Y(dt)$$

to zero. Then each $f_{\mathbf{z}} \in \mathcal{F}$ uniquely defines the mapping

$$\mathbf{z} \mapsto \mathcal{T}_\tau^{(1)}(f_{\mathbf{z}}),$$

which is the conditional τ-th quantile of $Y \mid \mathbf{Z}$. Then, $\mathcal{T}_\tau^{(1)}(f_{\mathbf{z}})$ is a version of conditional τ-th quantile $Q_\tau(Y \mid \mathbf{Z})$. Then the T-central subspace that induced by above functional is *τ-th Central Quantile Subspace (τ-th CQS)*, denote as $\mathcal{S}_{Q_\tau(Y|\mathbf{Z})}$. That is, all the information of the conditional quantile of $Y \mid \mathbf{Z}$ is summarized in the given

subspace.

Since the conditional statistical functional has been defined, we could apply our algorithm in Section 3.2 to obtain $\mathcal{S}_{Q_\tau(Y|\mathbf{Z})}$. As it was mentioned in the algorithm, an estimator for $\hat{Q}_\tau(Y \mid \hat{\mathbf{B}}'_{CS}\mathbf{Z})$ is necessary. In this chapter, we adopt the nonparametic estimator proposed by Takeuchi et al. (2006), which is given as below:

$$\begin{aligned}\hat{Q}_\tau(Y \mid \hat{\mathbf{B}}'_{CS}\mathbf{Z}) &= \arg\inf_f R_{\text{reg}}[f] \\ &= \frac{1}{n}\sum_{i=1}^{n}\rho_\tau(Y_i - f(\hat{\mathbf{B}}'_{CS}\mathbf{Z}_i)) + \frac{\lambda}{2}\|g\|^2_{\mathcal{H}}, \text{ with } f = g + b \text{ and } b \in \mathbb{R},\end{aligned}$$

where $\|\cdot\|^2_{\mathcal{H}}$ is RKHS norm and $g \in \mathcal{H}$. Then the algorithm could be carried out by simply taking $\hat{\mathcal{T}}((Y \mid \hat{\mathbf{B}}'_{CS}\mathbf{Z}))$ as $\hat{Q}_\tau(Y \mid \hat{\mathbf{B}}'_{CS}\mathbf{Z})$. We name such a procedure as *Cubic Kernel for Quantile (CKQ)*.

Remark: The algorithm for τ-th CQS provides an sample estimate of the basis. It naturally generalizes the idea of quantile and to think of stacking the information across different τ-th CQS. As τ traversing $(0, 1)$, quantile will contain all the information of a certain distribution. One would expect that by choosing sufficiently dense $\tau \in (0, 1)$, the union of all such τ-th CQS could exhaustively recover the CS.

This provides us a new method for estimating the CS:

Algorithm for CS: Let $\{\mathbf{X}_i, Y_i\}, i = 1, 2, \cdots, n$ be i.i.d sample.

1. Choose H points equally spaced points on $(0, 1)$ as $\tau_1, \tau_2, \cdots, \tau_H$.

2. For each $\tau_i, i = 1, 2, \cdots, H$, construct kernel matrices using CKT for conditional quantile, denote as $\hat{\mathbf{M}}_{\tau_i}$.

3. Let $\hat{\mathbf{V}} = \sum_{i=1}^{H}\hat{\mathbf{M}}_{\tau_i}$, and choose d_{CS} eigenvectors of $\hat{\mathbf{V}}$ that corresponding to the first d_{CS} largest eigenvalues, $\hat{\mathbf{v}}_1, \hat{\mathbf{v}}_2, \cdots, \hat{\mathbf{v}}_{d_{CS}}$, where d_{CS} represents the

dimensionality for CS. Then

$$\mathcal{S}(\hat{\mathbf{\Sigma}}_{\mathbf{X}}^{-1/2}\hat{\mathbf{B}}_{CS}\hat{\mathbf{v}}_1, \hat{\mathbf{\Sigma}}_{\mathbf{X}}^{-1/2}\hat{\mathbf{B}}_{CS}\hat{\mathbf{v}}_2, \cdots, \hat{\mathbf{\Sigma}}_{\mathbf{X}}^{-1/2}\hat{\mathbf{B}}_{CS}\hat{\mathbf{v}}_{d_{CS}}),$$

provides an estimate for CS.

Conditional Expectile

In light of the previous algorithm for recovering the CS, statistical functional such as conditional expectile could also be applied to recover the CS, because of the equivalency between conditional quantile and conditional distribution. In this subsection, we will take a look at another implicit conditional statistical functional: conditional expectile functional.

Expectile, as an extension of mean, is another type of statistical functional of interest which contains information of the full distribution for a random variable. It incorporates the information of the expectation of an random variable Y conditional on Y being in the tail of its distribution (Newey and Powell; 1987). Before we go further, we introduce expectile and expectile regression. For $\tau \in (0, 1)$, the *τ-th expectile*, denoted by $\mathrm{E}_\tau(Y)$, is defined by

$$\mathrm{E}_\tau(Y) = \arg\min_m \mathrm{E}(|\tau - I(Y < m)|(Y - m)^2)$$

for a random variable Y. Note that the 0.5-expectile is the mean.

Moreover, expectile is another form of quantile. It's the quantile of a distribution that is related to the original distribution (Jones; 1994). To be more specific, τth-expectile of Y is the τth-quantile of $\tilde{Y}$ where $\tilde{Y}$ has a cumulative distribution function given by

$$F_{\tilde{Y}}(y) = \frac{\nu(y) - yF_Y(y)}{\{(2\nu(y) - yF_Y(y)) + (y - \mathrm{E}(Y))\}^2},$$

where $\nu(y) = \int_{-\infty}^{y} tP_Y(dt)$ and $F_Y(y) = \int_{-\infty}^{y} P_Y(dt)$.

Consequently, the τ-th expectile of Y given a random vector $\mathbf{Z}$ is defined as

$$\mathrm{E}_\tau(Y \mid \mathbf{Z}) = \arg\min_m \mathrm{E}(|\tau - I(Y < m)|(Y - m)^2 \mid \mathbf{Z}).$$

Expectile regression, typically viewed as an extension of standard regression analysis, provides an effective diagnostic tool such as testing heteroscedasticity. Geometrically, $\{\mathbf{Z}_i, Y_i\}, i = 1, 2, \cdots, n$ is a point cloud in Euclidean space. If we viewed regression as describing the middle of point cloud as a function of $\mathbf{Z}$, then expectile regression can be viewed as an investigation of the higher or lower region of the the point cloud by introducing unequal weights to different regions (Newey and Powell 1987; Efron 1991).

To consider the problem of expectile regression in our framework. First consider the functional $\mathcal{T}_\tau^{(2)} : \mathcal{F} \to \mathbb{R}$ that is defined by the value of m that equates

$$\int_{\Omega_Y} \psi_\tau(t - m) f_{\mathbf{z}} P_Y(dt)$$

to zero with $\psi_\tau(t) = (\tau - I(t \leq 0))|t|$. Then each $f_{\mathbf{z}} \in \mathcal{F}$ uniquely defines the mapping

$$\mathbf{z} \mapsto \mathcal{T}_\tau^{(2)}(f_{\mathbf{z}})$$

which is the conditional τ-th expectile of $Y \mid \mathbf{Z}$, i.e. $\mathrm{E}_\tau(Y \mid \mathbf{Z})$.

Let's denote the T-central subspace induced by this functional as *τ-th Central Expectile Subspace, (τ-th CES)*. Remark that if $\tau = 0.5$, this should reduce to the CMS. Using our previously stated propositions, an algorithm could be easily carried out if we can provide an estimator for $\hat{\mathbf{B}}_{CS}$, the basis matrix for CS and $\hat{\mathrm{E}}_\tau(Y \mid \hat{\mathbf{B}}'_{CS}\mathbf{Z})$. And for the conditional expectile, Yang and Zou (2015) and Yang, Zhang and Zou (2018) established such an estimator through tree method and RKHS, which could

provide the sample estimate of $\hat{\mathrm{E}}_\tau(Y \mid \hat{\mathbf{B}}'_{CS}\mathbf{Z})$. The detail is given as below

$$\begin{aligned}\hat{\mathrm{E}}_\tau(Y \mid \hat{\mathbf{B}}'_{CS}\mathbf{Z}) &= \arg\inf_f R_{\mathrm{reg}}[f] \\ &= \frac{1}{n}\sum_{i=1}^{n}\psi_\tau(Y_i - f(\hat{\mathbf{B}}'_{CS}\mathbf{Z}_i)) + \frac{\lambda}{2}\|g\|^2_{\mathcal{H}}, \text{ with } f = g + b \text{ and } b \in \mathbb{R},\end{aligned}$$

where $\|\cdot\|^2_{\mathcal{H}}$ is RKHS norm and $g \in \mathcal{H}$. Then the algorithm could be carried out by simply taking $\hat{\mathcal{T}}(Y \mid \hat{\mathbf{B}}'_{CS}\mathbf{Z})$ as $\hat{\mathrm{E}}_\tau(Y \mid \hat{\mathbf{B}}'_{CS}\mathbf{Z})$. We name such a procedure as *Cubic Kernel for Expectile (CKE).*

3.4 Order Determination

In the previous sections, we have assumed the true structural dimension of the desired T-central subspace is known. But in practice, this is a quantity that need to be estimated. In SDR, many methods were proposed to provide an estimate for the true dimensions, those methods could be roughly categorized in the following categories: Sequential test methods, examples including: Bura and Cook (2001), Li (1991); Information criteria methods, such as: Zhu, Miao and Peng (2006), Zhu, Zhu and Wen (2010); cross validation type methods, which including, Wang and Xia (2008), Xia et al. (2002); permutation and bootstrap type of methods: Yin, Li and Cook (2008), Ye and Weiss (2003) and Luo and Li (2016). Sequantial testing methods requires the knowledge of the asymptotic distribution which in our cases is hard to obtain. Although it is hard for us to require the asymptotic distribution information of our estimator, consistency is fairly easy to prove, as we will show in this section. Given the consistency, information type of criteria and cross validation type methods are guaranteed to provide consistent estimator of the structural dimension.

In this chapter, we are going to use the BIC information criteria and ladle estimate (Luo and Li 2016) to provide estimate for the structural dimension. The advantage that BIC type of criteria enjoys lies in the computational speed, as BIC criteria is just

a function of the eigenvalues, which requires much less computational power comparing to cross validation. Moreover, our estimator requires estimating the structural dimension of the CS and T-central subspace, so computational speed is crucial for our methods. However, ladle estimator relies on bootstrap so it might require a bit effort to calculate, but it typically provides better estimate based on our experiments.

BIC type of criteria

Denote the estimated kernel matrix from the CKT algorithm as $\hat{\mathbf{M}}$. Let $\hat{\lambda}_i, i = 1 \cdots p$ be the corresponding eigenvalues, with $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p$ and define

$$\mathrm{BIC}_n(k) = n\frac{\sum_{i=1}^{k}\hat{\lambda}_i^2}{\sum_{i=1}^{p}\hat{\lambda}_i^2} - C_n\{\frac{k(k+1)}{2}\},$$

where $C_n/n \to 0$ as $n \to \infty$ and $C_n \to \infty$. C_n is typically chose as $2n^{3/4}/p$. The true dimension is estimated by $\arg\max_k \mathrm{BIC}_n(k)$(Li, Artemiou and Li 2011).

Ladle Estimator

Section 2.4 applied the ladle estimator to CK estimators, where the simulation results showed that ladle estimator works quite well when estimating the CS and CMS. We will skip the details here as the construction is very similar. Based on our past experiences, however, the ladle estimator suffers from the expensive computational cost, because it relies on bootstrap to assess the variability of eigenvectors. For our CKT algorithm in this chapter, we choose to fix the CS dimension as a small constant, d_{CS}, in the first step. By implementing this step, we could significantly reduce the size of kernel matrix and computational time because our estimator has computational complexity of $O(p^4 + np^3)$. Instead of considering a p dimensional problem, we are considering a d_{CS} dimensional problem instead. So it would be beneficial for us to consider using ladle estimator due to its accuracy and reduced computational time.

3.5 Asymptotic

Assumption 3.1. *Let $\mathbf{B}_{CS}$ be the basis for desired T-central subspace and $\hat{\mathbf{B}}_{CS}$ be the corresponding $\sqrt{n}$ estimator, $\mathcal{X}$ be the compact support for $\mathbf{X}$. Then for conditional statistical functional estimator, $\hat{\mathcal{T}}(Y \mid \hat{\mathbf{B}}'_{CS}\mathbf{X})$, we have*

$$\sup_{\mathbf{x}\in\mathcal{X}} \|\hat{T}(Y \mid \hat{\mathbf{B}}'_{CS}\mathbf{x}) - \mathcal{T}(Y \mid \mathbf{B}'_{CS}\mathbf{x})\| = O_p(1).$$

Note: Our estimator requires an initial estimate for the desired statistical functional of interests, the asymptotic behavior depends on the choice of methodologies applied. For commonly seen conditional statistical functionals such as conditional quantile, this assumption would be easily met if we choose to use the right version of the estimator. In the next paragraph, we will provide a simple argument to verify that conditional quantile estimator that we used meets this assumption.

Let $Q_\tau(Y \mid \mathbf{B}'_{CS}\mathbf{X})$ represent the τ-th quantile and $\hat{Q}_\tau(Y \mid \hat{\mathbf{B}}'_{CS}\mathbf{X})$ be the corresponding conditional quantile estimator. Observe that

$$\begin{aligned}
&\|Q_\tau(Y \mid \mathbf{B}'_{CS}\mathbf{X}) - \hat{Q}_\tau(Y \mid \hat{\mathbf{B}}'_{CS}\mathbf{X})\| \\
\leq &\|\hat{Q}_\tau(Y \mid \hat{\mathbf{B}}'_{CS}\mathbf{X}) - \hat{Q}_\tau(Y \mid \mathbf{B}'_{CS}\mathbf{X})\| \\
&+ \|\hat{Q}_\tau(Y \mid \mathbf{B}'_{CS}\mathbf{X}) - Q_\tau(Y \mid \mathbf{B}'_{CS}\mathbf{X})\|.
\end{aligned}$$

$\sup_{\mathbf{x}\in\mathcal{X}}\|\hat{Q}_\tau(Y \mid \hat{\mathbf{B}}'_{CS}\mathbf{X}) - \hat{Q}_\tau(Y \mid \mathbf{B}'_{CS}\mathbf{X})\|$ is $O_p(1)$, because of the Bahadur representation (Chaudhuri 1991) and $\sqrt{n}$-consistency of $\hat{\mathbf{B}}_{CS}$. And the second term $\sup_{\mathbf{x}\in\mathcal{X}}\|\hat{Q}_\tau(Y \mid \mathbf{B}'_{CS}\mathbf{X}) - Q_\tau(Y \mid \mathbf{B}'_{CS}\mathbf{X})\|$ is also $O_p(1)$, based on Guerre and Sabbah (2012).

Assumption 3.2. *Let $\mathfrak{X} = \left(1, \mathbf{X}', \mathbf{X}' \otimes \mathbf{X}', \mathbf{X}' \otimes \mathbf{X}' \otimes \mathbf{X}'\right)'$, then we have the fol-*

lowing moment conditions:

$$\mathrm{E}\|\mathfrak{X}\mathfrak{X}'\|_\infty < \infty, \quad \mathrm{E}[T(Y \mid \mathbf{B}_{CS}'\mathbf{X})^2\|\mathfrak{X}\mathfrak{X}'\|_\infty] < \infty,$$

where $\|A\|_\infty = \max_{ij} |a_{ij}|$ *denotes the max norm of the matrix.*

Given the assumptions above, the following Proposition 3.4 will demonstrate the consistency of our proposed estimator.

Proposition 3.4. *Let* $\boldsymbol{\alpha}$ *be the basis for* $\mathcal{S}_{T(Y|\mathbf{X})}$ *and* $\hat{\mathbf{M}}$ *the corresponding cubic kernel defined in Algorithm 1. Then under the assumptions above,* $\hat{\mathbf{M}}$ *is the* $\sqrt{n}$*-consistent estimate of* $\boldsymbol{\alpha}$*.*

Proof. Please refer to the Appendix for the detailed proof. □

3.6 Simulations and Applications

In this section, we compare the performance of cubic kernel CKQ with qMAVE from Kong and Xia (2014). For the accuracy comparison, we used the distance of Li, Zha and Chairomonte (2005) between two subspaces of $\mathbb{R}^p$: Let $\mathcal{S}_1$ and $\mathcal{S}_2$ be two subspaces of $\mathbb{R}^p$, and $\mathbf{P}_{\mathcal{S}_1}$ and $\mathbf{P}_{\mathcal{S}_2}$ be the corresponding orthogonal projection, respectively. Let

$$\Delta_F = \|\mathbf{P}_{\mathcal{S}_1} - \mathbf{P}_{\mathcal{S}_2}\|_F$$

where $\|\cdot\|_F$ is the Frobenius norm. The smaller the Δ_F is, the closer the two subspaces are to each other.

Comparison of the Model accuracy

In this section, we consider the following five models to illustrate the efficacy of our proposed methods. Let β_1 and β_2 be p dimensional vectors with $(1, 0, 0, \cdots, 0)'$ and $(1, 0.5, 1, 0, \cdots, 0)'$, respectively. Let β_3, β_4 and β_5 be p dimensional vectors with

$(1, 0.1, 0, \cdots, 0)'$, $(1, 1, 1, 1, 0, \cdots, 0)'$ and $(0, \cdots, 0, 1, 1, 1, 1)'$. Let ϵ be a random variable that is independent of $\mathbf{X} \sim N(0, \mathbf{I}_p)$. The simulation was conducted to make comparison for combinations of the following configurations: $n = 100, 200, 400$, $p = 10, 20$ and $\tau = 0.25, 0.50, 0.75$ using $N = 200$ replicated data. On each generated data, we applied different methods and computed the distance, Δ_F, between these estimates and the true τ-th CQS. Then we took average and standard error from the resulting N distances. The simulated 5 models are:

(A) $Y = \beta_1'\mathbf{X} + 0.1 \exp(1 - \beta_1'\mathbf{X})\epsilon$

(B) $Y = 0.1\beta_2'\mathbf{X} + \exp(\beta_2'\mathbf{X})\epsilon$

(C) $Y = \beta_1'\mathbf{X} + \epsilon$

(D) $Y = 1 + \beta_3'\mathbf{X} + (1 + 0.4\beta_1'\mathbf{X})\epsilon$

(E) $Y = \beta_4'\mathbf{X} + (\beta_5'\mathbf{X})^3 + \epsilon$

Model (A) is the first model considered in Takeuchi (2006). For all $\tau \in (0, 1)$, τ-th CQS is $\mathcal{S}(\beta_1)$. There is only one term involved, which makes it seems unnecessary to apply the quantile esitmator, methods for CS or CMS would work instead. However, error term changes dramatically when $\beta_1'\mathbf{X}$ is negative, masking the true signal, this leads the distribution of the error to have a heavy tail on the negative side. Under such a heavy tail scenario, methods for CS and CMS may not work as desired. So we are going to investigate the performance of quantile based estimator for this case.

Model (B) is Model (C) in Chapter 2. Similar to Model (A), this model also involves a heavy tail, but on the positive side. And the τ-th CQS is also spanned by a single direction, $\mathcal{S}(\beta_2)$, which consists of the τ-th CQS for all τ.

Model (C) has been discussed in many papers, such as Example 4 from Zou and Yuan (2008), where the error term ϵ is taken as t-distribution with 3 degrees of freedom, different from previous examples, this distribution has heavy tail on both

Table 3.1: Quantile Accuracy Comparison for Model (A)

	$p = 10$			$p = 20$		
	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$
$n = 100$						
qMAVE	0.209 (0.060)	0.210 (0.051)	0.261 (0.068)	0.350 (0.029)	0.336 (0.039)	0.363 (0.054)
CKQ	0.201 (0.052)	0.197 (0.049)	0.211 (0.048)	0.334 (0.031)	0.325 (0.037)	0.340 (0.042)
$n = 200$						
qMAVE	0.136 (0.032)	0.142 (0.027)	0.167 (0.054)	0.208 (0.045)	0.223 (0.037)	0.245 (0.021)
CKQ	0.140 (0.028)	0.139 (0.025)	0.138 (0.013)	0.195 (0.032)	0.195 (0.043)	0.205 (0.038)
$n = 400$						
qMAVE	0.097 (0.022)	0.102 (0.020)	0.116 (0.024)	0.136 (0.024)	0.151 (0.029)	0.181 (0.035)
CKQ	0.098 (0.022)	0.096 (0.022)	0.098 (0.015)	0.154 (0.028)	0.153 (0.028)	0.162 (0.030)

said. This is a classical example that demonstrating the central median subspace is superior than the CMS to find the truth, as median is more robust to outliers than mean. Here we will also investigate this classical model. In this model, the τ-th CQS is $\mathcal{S}(\beta_1)$.

Model (D) is Model VII in Luo, Li and Yin (2014), where τ-th CQS is $\mathcal{S}(\beta_1 + 0.4\Phi^{-1}(\tau)\beta_1)$, which changes based on different value of τ. For this model, we are going to see if our proposed estimator could provide a consistent results or not, in terms of different τs.

Model (E) is constructed so that only first and third order term is involved, which is to the favor of cubic kernel method. Here τ-th CS is $\mathcal{S}(\beta_1, \beta_2)$.

For the CKQ algorithm, we used $d_{CS} = 3$ as the starting dimension in the first step as none of the simulating models goes beyond 2 dimensions. And using dimension of 3 would make the estimator for the conditional quantile overfit the simulated data. For the estimator of CS, we used Slice Regression (Wang and Xia 2008).

We also compared our results with qMAVE (Kong and Xia 2014). qMAVE was first designed to recover the CS by stacking information in different τ-th CQS. To make comparison, we retrieved the estimate from qMAVE for a fixed τ to obtain the estimator for τ-th CQS.

From Table 3.1 Table 3.2 and Table 3.3, we can see that with the heavy tailed

Table 3.2: Quantile Accuracy Comparison for Model (B)

	$p = 10$			$p = 20$		
	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$
$n = 100$						
qMAVE	0.652 (0.183)	0.815 (0.160)	0.388 (0.083)	0.832 (0.192)	0.840 (0.141)	0.674 (0.166)
CKQ	0.494 (0.202)	0.669 (0.241)	0.368 (0.073)	0.720 (0.125)	0.773 (0.136)	0.555 (0.184)
$n = 200$						
qMAVE	0.413 (0.207)	0.750 (0.139)	0.300 (0.029)	0.555 (0.129)	0.912 (0.080)	0.430 (0.065)
CKQ	0.276 (0.080)	0.440 (0.187)	0.225 (0.042)	0.476 (0.162)	0.717 (0.229)	0.352 (0.058)
$n = 400$						
qMAVE	0.205 (0.047)	0.567 (0.138)	0.209 (0.051)	0.425 (0.071)	0.809 (0.169)	0.328 (0.061)
CKQ	0.148 (0.046)	0.327 (0.121)	0.157 (0.062)	0.291 (0.062)	0.463 (0.180)	0.241 (0.063)

Table 3.3: Quantile Accuracy Comparison for Model (C)

	$p = 10$			$p = 20$		
	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$
$n = 100$						
qMAVE	0.173 (0.056)	0.134 (0.021)	0.142 (0.028)	0.324 (0.233)	0.233 (0.058)	0.369 (0.223)
CKQ	0.168 (0.031)	0.167 (0.034)	0.164 (0.032)	0.252 (0.049)	0.260 (0.048)	0.267 (0.048)
$n = 200$						
qMAVE	0.108 (0.026)	0.078 (0.024)	0.108 (0.018)	0.168 (0.036)	0.155 (0.035)	0.183 (0.055)
CKQ	0.123 (0.030)	0.114 (0.027)	0.116 (0.028)	0.176 (0.036)	0.177 (0.033)	0.186 (0.036)
$n = 400$						
qMAVE	0.085 (0.024)	0.069 (0.017)	0.073 (0.012)	0.124 (0.023)	0.102 (0.026)	0.121 (0.017)
CKQ	0.085 (0.020)	0.084 (0.022)	0.081 (0.020)	0.115 (0.020)	0.116 (0.022)	0.117 (0.023)

Table 3.4: Quantile Accuracy Comparison for Model (D)

	$p = 10$			$p = 20$		
	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$
$n = 100$						
qMAVE	0.420 (0.085)	0.342 (0.118)	0.313 (0.117)	0.673 (0.155)	0.514 (0.117)	0.505 (0.091)
CKQ	0.373 (0.087)	0.350 (0.082)	0.353 (0.103)	0.592 (0.147)	0.555 (0.114)	0.562 (0.132)
$n = 200$						
qMAVE	0.308 (0.038)	0.209 (0.067)	0.223 (0.022)	0.537 (0.165)	0.379 (0.071)	0.339 (0.049)
CKQ	0.277 (0.060)	0.226 (0.048)	0.228 (0.042)	0.406 (0.067)	0.391 (0.055)	0.384 (0.064)
$n = 400$						
qMAVE	0.235 (0.058)	0.162 (0.037)	0.149 (0.031)	0.369 (0.046)	0.258 (0.029)	0.244 (0.026)
CKQ	0.208 (0.052)	0.162 (0.028)	0.166 (0.025)	0.299 (0.028)	0.266 (0.027)	0.263 (0.026)

Table 3.5: Quantile Accuracy Comparison for Model (E)

	$p = 10$			$p = 20$		
	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$
$n = 100$						
qMAVE	0.728 (0.204)	0.659 (0.236)	0.590 (0.207)	0.929 (0.071)	0.878 (0.139)	0.892 (0.142)
CKQ	0.527 (0.207)	0.521 (0.214)	0.530 (0.231)	0.806 (0.154)	0.794 (0.154)	0.786 (0.156)
$n = 200$						
qMAVE	0.379 (0.160)	0.292 (0.107)	0.443 (0.220)	0.760 (0.180)	0.597 (0.232)	0.677 (0.178)
CKQ	0.310 (0.098)	0.298 (0.111)	0.296 (0.113)	0.515 (0.084)	0.510 (0.074)	0.508 (0.072)
$n = 400$						
qMAVE	0.206 (0.058)	0.175 (0.038)	0.179 (0.050)	0.338 (0.074)	0.287 (0.022)	0.328 (0.051)
CKQ	0.150 (0.037)	0.158 (0.036)	0.166 (0.040)	0.334 (0.073)	0.335 (0.073)	0.339 (0.073)

error term, the quantile based method could provide a relatively accurate estimate. Moreover, in these three cases, the models have the same τ-th central quantile subspace no matter what value does τ takes. Table 3.1 and Table 3.2 reveal, however, that qMAVE does not provide an stable estimate based for different τs, which demonstrate that qMAVE may not estimate a certain quatile very accurately. To be more specific, whenever the tail behavior is not stable, the qMAVE would provide very inaccurate results, this might due to the fact that qMAVE is a local based method, therefore, relatively sensitive to erratic local changes. Compare to qMAVE, CKQ is a method that is very stable across different value of τs, and in most cases, CKQ could improve the performance of qMAVE.

We could further demonstrate this point from Table 3.4 where CKQ is still quite stable across different values of τ, even when the τ-th CQS changes with τ. And in this case, since there is also a direction in the error term, qMAVE also suffer from the unstable results. And CKQ could dramatically improve the results in this case.

Table 3.5 summarized the results with more than one direction in the desired τ-th central quantile subspace, where we could see that CKQ is still quite stable and accurate compare to qMAVE. This is not a surprise to us, as the cubic structure in this model is favorable to our cubic kernel method.

Model (B), (C) and (D) were also used for testing the performance of cubic kernel

Table 3.6: Expectile Accuracy Comparison for Model (B)

	$p = 10$			$p = 20$		
	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$
$n = 200$						
eMAVE	1.355 (0.108)	1.366 (0.105)	1.361 (0.072)	1.365 (0.064)	1.364 (0.071)	1.366 (0.068)
CKE	0.886 (0.339)	0.906 (0.346)	0.796 (0.342)	0.919 (0.342)	0.947 (0.335)	0.927 (0.331)
CKM		1.139 (0.236)			1.249 (0.158)	
$n = 400$						
eMAVE	1.388 (0.038)	1.389 (0.046)	1.381 (0.042)	1.382 (0.041)	1.384 (0.049)	1.373 (0.065)
CKE	0.709 (0.331)	0.763 (0.338)	0.606 (0.273)	0.839 (0.347)	0.949 (0.307)	0.730 (0.320)
CKM		1.111 (0.250)			1.285 (0.158)	

Table 3.7: Expectile Accuracy Comparison for Model (C)

	$p = 10$			$p = 20$		
	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$
$n = 200$						
eMAVE	1.321 (0.119)	1.339 (0.120)	1.386 (0.046)	1.331 (0.098)	1.303 (0.124)	1.327 (0.128)
CKEcs	0.327 (0.132)	0.323 (0.120)	0.329 (0.121)	0.587 (0.207)	0.580 (0.207)	0.563 (0.172)
CKM		0.775 (0.220)			1.202 (0.135)	
$n = 400$						
eMAVE	1.365 (0.076)	1.392 (0.040)	1.405 (0.026)	1.343 (0.083)	1.321 (0.119)	1.329 (0.131)
CKE	0.237 (0.075)	0.238 (0.076)	0.247 (0.086)	0.277 (0.084)	0.275 (0.077)	0.283 (0.080)
CKM		0.524 (0.119)			0.986 (0.153)	

for expectile algorithm. There is no other literature proposed a method for central expectile subspace. In order to make a comparison, we modified the qMAVE method to handle asymmetric least square problem instead of the asymmetric absolute deviation, and call it as eMAVE. Moreover, as the 0.5th-CES is the central mean subspace, we also added the CKM estimator from Chapter 2 at 0.5th-CES for benchmark comparison.

Table 3.6, 3.7 and 3.8 shows that CKE is a much better estimator than eMAVE for the three simulated models. The CMS estimated by CKE evaluated at $\tau = 0.5$ is also a decent estimator when compared with CKM.

Table 3.8: Expectile Accuracy Comparison for Model (D)

	$p = 10$			$p = 20$		
	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$
$n = 200$						
eMAVE	1.778 (0.150)	1.684 (0.215)	1.711 (0.191)	1.815 (0.117)	1.760 (0.161)	1.855 (0.122)
CKE	1.305 (0.170)	1.301 (0.177)	1.302 (0.174)	1.385 (0.102)	1.385 (0.102)	1.385 (0.102)
CKM		1.415 (0.086)			1.583 (0.102)	
$n = 400$						
eMAVE	1.757 (0.133)	1.818 (0.108)	1.783 (0.176)	1.855 (0.103)	1.762 (0.172)	1.843 (0.146)
CKE	1.229 (0.229)	1.228 (0.229)	1.228 (0.230)	1.371 (0.073)	1.371 (0.073)	1.371 (0.073)
CKM		1.302 (0.150)			1.499 (0.054)	

Table 3.9: Variable Selection Results for Model (A)

	$\tau = 0.25$		$\tau = 0.50$		$\tau = 0.75$	
	sensitivity	specificity	sensitivity	specificity	sensitivity	specificity
$n = 100$						
$p = 10$	1.000 (0.000)	0.986 (0.038)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
$p = 20$	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.998 (0.011)
$n = 200$						
$p = 10$	1.000 (0.000)	0.981 (0.043)	1.000 (0.000)	0.990 (0.032)	1.000 (0.000)	0.990 (0.032)
$p = 20$	1.000 (0.000)	0.998 (0.011)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.998 (0.011)
$n = 400$						
$p = 10$	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
$p = 20$	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)

Comparison of Variable Selection

We now provide results for variable selection. Simulations are conducted using the same models (A)-(C), and same set up. On each sample, we used CISE (Chen, Zou and Cook 2010) for CK methods. We also consider a ultra-high dimension ($p \gg n$) simulation with $n = 200, p = 1000$ to test the performance for the one-stage variable selection performance. We report two statistics: fraction of irrelevant variables that are not being selected, i.e. specificity; fraction of relevant variables that are being selected, i.e. sensitivity (Chen, Zou and Cook 2010). The bigger the indices is, the better the estimate is.

From Table 3.9 and Table 3.11, we could tell CISE-CKQ, works fairly well as the

Table 3.10: Variable Selection Results for Model (B)

	$\tau = 0.25$		$\tau = 0.50$		$\tau = 0.75$	
	sensitivity	specificity	sensitivity	specificity	sensitivity	specificity
$n = 100$						
$p = 10$	1.000 (0.000)	0.118 (0.093)	0.971 (0.096)	0.068 (0.113)	0.928 (0.141)	0.180 (0.144)
$p = 20$	0.710 (0.367)	0.552 (0.249)	0.507 (0.388)	0.604 (0.244)	0.812 (0.221)	0.409 (0.169)
$n = 200$						
$p = 10$	0.739 (0.245)	0.733 (0.184)	0.971 (0.139)	0.404 (0.231)	1.000 (0.000)	0.472 (0.230)
$p = 20$	0.826 (0.263)	0.317 (0.199)	0.986 (0.070)	0.317 (0.163)	1.000 (0.000)	0.105 (0.087)
$n = 400$						
$p = 10$	1.000 (0.000)	0.174 (0.172)	0.812 (0.169)	0.323 (0.263)	1.000 (0.000)	0.255 (0.183)
$p = 20$	0.957 (0.115)	0.246 (0.169)	0.913 (0.206)	0.407 (0.184)	1.000 (0.000)	0.169 (0.107)

Table 3.11: Variable Selection Results for Model (C)

	$\tau = 0.25$		$\tau = 0.50$		$\tau = 0.75$	
	sensitivity	specificity	sensitivity	specificity	sensitivity	specificity
$n = 100$						
$p = 10$	1.000 (0.000)	0.875 (0.141)	1.000 (0.000)	0.929 (0.159)	1.000 (0.000)	0.821 (0.192)
$p = 20$	1.000 (0.000)	0.790 (0.181)	1.000 (0.000)	0.877 (0.149)	1.000 (0.000)	0.742 (0.176)
$n = 200$						
$p = 10$	1.000 (0.000)	0.951 (0.140)	1.000 (0.000)	0.946 (0.130)	1.000 (0.000)	0.967 (0.068)
$p = 20$	1.000 (0.000)	0.831 (0.169)	1.000 (0.000)	0.870 (0.165)	1.000 (0.000)	0.879 (0.159)
$n = 400$						
$p = 10$	1.000 (0.000)	0.870 (0.128)	1.000 (0.000)	0.967 (0.077)	1.000 (0.000)	0.875 (0.229)
$p = 20$	1.000 (0.000)	0.903 (0.150)	1.000 (0.000)	0.964 (0.070)	1.000 (0.000)	0.889 (0.175)

Table 3.12: Variable Selection Results for Model (D)

	$\tau = 0.25$		$\tau = 0.50$		$\tau = 0.75$	
	sensitivity	specificity	sensitivity	specificity	sensitivity	specificity
$n = 100$						
$p = 10$	0.522 (0.104)	0.962 (0.070)	0.543 (0.144)	0.935 (0.106)	0.565 (0.172)	0.957 (0.061)
$p = 20$	0.500 (0.000)	0.976 (0.028)	0.543 (0.144)	0.966 (0.047)	0.522 (0.104)	0.937 (0.097)
$n = 200$						
$p = 10$	0.630 (0.224)	0.995 (0.026)	0.565 (0.172)	0.984 (0.043)	0.543 (0.144)	0.973 (0.053)
$p = 20$	0.739 (0.255)	0.940 (0.110)	0.761 (0.255)	0.935 (0.100)	0.739 (0.255)	0.882 (0.140)
$n = 400$						
$p = 10$	0.543 (0.144)	1.000 (0.000)	0.652 (0.235)	0.946 (0.118)	0.826 (0.243)	0.995 (0.026)
$p = 20$	0.761 (0.255)	0.913 (0.122)	0.739 (0.255)	0.954 (0.096)	0.717 (0.253)	0.947 (0.085)

Table 3.13: Variable Selection Results for Model (E)

	$\tau = 0.25$		$\tau = 0.50$		$\tau = 0.75$	
	sensitivity	specificity	sensitivity	specificity	sensitivity	specificity
$n = 100$						
$p = 10$	0.761 (0.255)	0.962 (0.070)	1.000 (0.000)	0.986 (0.051)	0.609 (0.211)	0.924 (0.118)
$p = 20$	0.543 (0.144)	0.976 (0.040)	1.000 (0.000)	0.989 (0.022)	0.543 (0.144)	0.966 (0.036)
$n = 200$						
$p = 10$	0.826 (0.243)	0.973 (0.106)	1.000 (0.000)	0.986 (0.038)	0.957 (0.144)	0.940 (0.091)
$p = 20$	0.674 (0.243)	0.954 (0.086)	1.000 (0.000)	0.984 (0.029)	0.804 (0.250)	0.940 (0.090)
$n = 400$						
$p = 10$	1.000 (0.000)	0.946 (0.098)	1.000 (0.000)	0.990 (0.046)	1.000 (0.000)	0.967 (0.068)
$p = 20$	0.761 (0.255)	0.944 (0.085)	1.000 (0.000)	0.995 (0.015)	1.000 (0.000)	0.959 (0.077)

Table 3.14: Variable Selection Results for Model (A)-(E) with $n = 200, p = 1000$

	$\tau = 0.25$		$\tau = 0.50$		$\tau = 0.75$	
	sensitivity	specificity	sensitivity	specificity	sensitivity	specificity
Model (A)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
Model (B)	0.826 (0.198)	0.984 (0.002)	0.696 (0.223)	0.984 (0.001)	0.797 (0.241)	0.984 (0.002)
Model (C)	0.500 (0.000)	0.995 (0.005)	0.500 (0.000)	0.992 (0.005)	0.500 (0.000)	0.992 (0.005)
Model (D)	0.587 (0.194)	0.992 (0.006)	1.000 (0.000)	0.995 (0.006)	0.565 (0.172)	0.993 (0.005)
Model (E)	0.957 (0.144)	0.992 (0.005)	0.957 (0.144)	0.993 (0.005)	0.957 (0.144)	0.992 (0.004)

sensitivity and specificity are very close to 1 in both cases. And from Table 3.14, this is still the case when we consider the ultra-high dimension settings. This shows that our variable selection procedure works very well.

According to Table 3.10, Table 3.12 and Table 3.13, CISE-CKQ has a reasonable sensitivity but the specificity is relatively low. This indicates CISE-CKQ may pick some of the non-important variable by mistake. Although this may cause some useless information included in the follow up studies, most of the important information could be captured. The high specificity Table 3.14 is because of the sparsity in the true directions.

3.7 Real Data Analysis

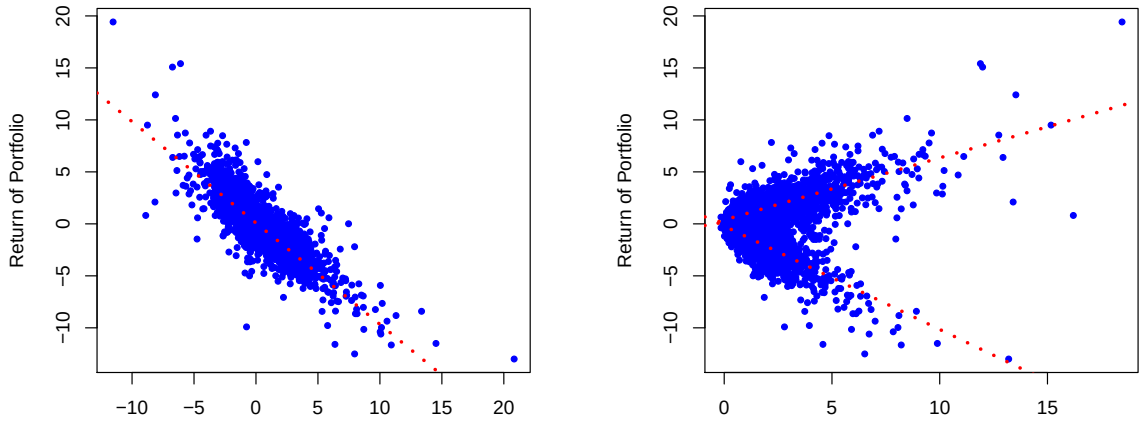
Kong and Xia (2014) studied the factor that affects the volatility of a portfolio by investigating the CS through qMAVE. Here we are going to compare the performance of CKQ estimator and qMAVE for the same data set. Same to Kong and Xia (2014), we pulled the daily return of portfolio, Y, from the following website:

`http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`

To make a fair comparison, we used the same dataset with the same set of variables. To be more specific, we include daily data points from July 1st 1926 to Mar 28th 2013. The response variable Y was taken as the return of intersection of small market equity and low ratio book equity and market equity. The explanatory variable was constructed in the following way: $X_1, \cdots, X_5$ represent the return of portfolio in the previous five days; $X_6, \cdots, X_{10}$ was taken as the absolute value of the past five days return, which could be used as a good indicator of the volatility. $X_{11}, \cdots, X_{15}$ was pulled from market return of the present day and past four days and $X_{16}, \cdots, X_{20}$ are the corresponding absolute value of the market return.

Kong and Xia (2014) suggest to use the structural dimension of 2 which is the structural dimension that we adopt to use here in this section. They have pointed out that the first direction was in the CMS and the second direction was clearly in the central variance subspace. In this section, we consider the τth-CQS for $\tau = 0.25, 0.50, 0.75$, so we are expecting to recover both directions when we traverse through all the quantiles. So for each specific given quantile, we are going to set the structural dimension of the CQS as 2. Similar directions were distracted from the data for different τ, we only presenting the two directions from the 0.25th-CQS.

As we can tell from the graph, we could see that CKQ successfully picked up the first two directions, where the first direction is linear in the CMS and the second



(a) Response vs First Direction (b) Response vs Second Direction

Figure 3.1: 0.25th-CQS for Return of Portfolio

one shows a change in the variation, which is also consistent with our intuition and Kong and Xia (2014). To further access the performance of our estimator, we used bootstrap method for further simulations studies. We bootstrapped the data with replacement and then perform CKQ on each bootstrap sample. Then we record the distance, Δ_F, of each bootstrap estimation and the original estimation. $N = 200$ samples were taken and Table 3.15 shows the results.

Table 3.15: Model Accuracy for Real Data

	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$
qMAVE	1.429 (0.139)	1.437 (0.145)	1.425 (0.147)
CKQ	0.755 (0.182)	0.837 (0.216)	0.884 (0.164)

Table 3.15 shows that our estimator are very stable across different τs. And also it provided a better performance than qMAVE. This might indicate that our CKQ estimator might fit this data better.

Moreover, beyond the qMAVE, our method are also capable of variable selection. Kong and Xia (2014) accessed the variable importance by simply looking at the

Table 3.16: Variable Selection Results for Real Data

	$\tau = 0.25$		$\tau = 0.50$		$\tau = 0.75$	
	sensitivity	specificity	sensitivity	specificity	sensitivity	specificity
CKQ	1.000 (0.000)	0.999 (0.010)	1.000 (0.000)	1.000 (0.000)	0.998 (0.035)	0.999 (0.007)

magnitude of the coefficients. It works fine for this data set, as the features are constructed from the same variables, which indicates that all the variables are of the same scale. However, in a more general frame work, assessing importance of features by magnitude may not lead to desirable results. But for our method, we could adopt our variable selection procedure to find what are some of the important features.

CISE-CKQ on the original data showed that variable X_{11} and X_{16} are important variables when $\tau = 0.25, 0.50, 0.75$. Variable X_{11} represents the market return of the present day and X_{16} represents market volatility of the present day. These findings were inline with Kong and Xia (2014) where they pointed out that capital asset pricing model (CAPM) would suggest return of a portfolio strongly depend on the present day market performance. To better assess the stability of our estimator, we did another simulation study using bootstrap. At each iteration, we bootstrapped data from the original data set, permuting all the variables except for holding X_{11} and X_{16} fixed. By doing so, we manually made X_{11} and X_{16} to be relevant. $N = 200$ iterations were run and our results is summarized in Table 3.16.

From Table 3.16, we can tell that our estimator provided a very stable variable selection procedure for this data as both sensitivity and specificity are close to 1.

3.8 Discussion

In this chapter, we proposed a cubic kernel for estimating I-functional of the T-central subspace, in particularly, we developed details for estimating τ-th CQS, τ-th CES and respective variable selection. Simulation results show the strength and usefulness of

cubic kernel methods: CKQ and CKE.

Chapter 4 Minimum Discrepancy Approach of Moment Kernels with Applications in T-Central Subspace

4.1 Introduction

This chapter will reformulate our SDR methods in Chapter 2 and Chapter 3, so that we have an optimal solutions. This follows the work of Cook (2004), Cook and Ni (2005) and Cook and Zhang (2014) together with Qian,Ding and Cook (2019) for large p small n problem.

This chapter contains the following parts. In Section 4.2, we establish the framework for the minimum discrepancy approach on cubic kernel. Section 4.3 outlines the detailed algorithm for solving the optimization problem. Section 4.4 develops hypothesis tests for estimating the structural dimension and Section 4.5 discusses the future works.

4.2 Proposed Framework

Cubic Kernels on T-Central Subspace.

Chapter 2 introduced *cubic kernel* (CK) to recover the desired T-central subspace. CK was obtained by minimizing an objective function of the form $[T(Y) - a - \beta'_{T\mathbf{Z}}\mathbf{Z} - \mathbf{Z}'\boldsymbol{\Sigma}_{T\mathbf{ZZ}}\mathbf{Z} - (\mathbf{Z}' \otimes \mathbf{Z}')\mathcal{M}_{T\mathbf{ZZZ}}\mathbf{Z}]^2$, where $\beta_{T\mathbf{Z}}, \boldsymbol{\Sigma}_{T\mathbf{ZZ}}, \mathcal{M}_{T\mathbf{ZZZ}}$ are kernel matrices constructed similar to OLS, pHd and third moment matrix. In Proposition 2.2, we showed that under mild conditions, $\mathcal{S}(\beta_{T\mathbf{Z}}, \boldsymbol{\Sigma}_{T\mathbf{ZZ}}, \mathcal{M}'_{T\mathbf{ZZZ}}) \subseteq \mathcal{S}_{T(Y|\mathbf{Z})}$. Under this setup, the T-central subspace, $\mathcal{S}_{T(Y|\mathbf{Z})}$ could be estimated by a basis matrix $\mathbf{M}$ such that the column space of $\mathcal{S}(\mathbf{M}) = \mathcal{S}(\beta_{T\mathbf{Z}}, \boldsymbol{\Sigma}_{T\mathbf{ZZ}}, \mathcal{M}'_{T\mathbf{ZZZ}})$, if the coverage condition is

assumed in addition. And $\mathbf{M}$ was obtained by the following construction

$$\mathbf{M} = \omega_1^2 \beta_{T\mathbf{Z}} \beta'_{T\mathbf{Z}} + \omega_2^2 \mathbf{\Sigma}_{T\mathbf{ZZ}} \mathbf{\Sigma}'_{T\mathbf{ZZ}} + \omega_3^2 \mathcal{M}'_{T\mathbf{ZZZ}} \mathcal{M}_{T\mathbf{ZZZ}},$$

where the $\omega_i, i = 1, 2, 3$ are the appropriate weight correspond to each kernel. Although, this approach combined different kernels with a weight function, the downside is its ignorance on the covariance information between different kernels. In this chapter, we are going to construct $\mathbf{M}$ from a different aspect.

Minimum Discrepancy Approach

We begin this section by introducing some additional notations: Let $\mathbf{B}^{(0)} \in \mathbb{R}^{p \times d}$ denote the basis for desired T-central subspace, i.e. $\mathcal{S}(\mathbf{B}^{(0)}) = \mathcal{S}_{T(Y|\mathbf{Z})}$. Denote $\boldsymbol{\xi}$ as the matrix column binding the kernel matrices,

$$\boldsymbol{\xi} = (\beta_{T\mathbf{Z}}, \mathbf{\Sigma}_{T\mathbf{ZZ}}, \mathcal{M}'_{T\mathbf{ZZZ}}) = (\underbrace{\xi_1}_{\beta_{T\mathbf{Z}}}, \underbrace{\xi_2, \cdots, \xi_{p+1}}_{\mathbf{\Sigma}_{T\mathbf{ZZ}}}, \underbrace{\xi_{p+2}, \cdots, \xi_{p^2+p+1}}_{\mathcal{M}'_{T\mathbf{ZZZ}}}).$$

For each $i = 1, 2, \cdots, p^2 + p + 1$, there exists a vector $\mathbf{c}_i$ such that $\xi_i = \mathbf{B}^{(0)} \mathbf{c}_i^{(0)}$. Let $\mathbf{C}^{(0)} = (\mathbf{c}_1^{(0)}, \mathbf{c}_2^{(0)}, \cdots, \mathbf{c}_{p^2+p+1}^{(0)}) \in \mathbb{R}^{d \times (p^2+p+1)}$, we have

$$\boldsymbol{\xi} = \mathbf{B}^{(0)} \mathbf{C}^{(0)}.$$

The idea of estimating $\mathcal{S}_{T(Y|\mathbf{Z})}$ is then by finding a d dimensional subspace that is "closest" to $\boldsymbol{\xi}$. "Close" is in terms of a quadratic distance with optimality in some sense. Let vec$(\cdot)$ denote the linear operator that converts the columns of a matrix into a column vector. Then, the quadratic distance is defined through the following definition (Cook and Ni 2005):

$$D(\mathbf{B}, \mathbf{C}) = [\text{vec}(\boldsymbol{\xi}\mathbf{R}_n) - \text{vec}(\mathbf{BC})]' \mathbf{V}_n [\text{vec}(\boldsymbol{\xi}\mathbf{R}_n) - \text{vec}(\mathbf{BC})],$$

Here $\mathbf{V}_n \in \mathbb{R}^{p(p^2+p+1)\times p(p^2+p+1)}$ is a positive definite matrix. $\mathbf{R}_n \in \mathbb{R}^{(p^2+p+1)\times(p^2+p+1)}$ is a non-singular matrix which represents how we manipulate the columns of $\boldsymbol{\xi}$. In Chapter 2, we used least square procedure to find weights for each column of $\boldsymbol{\xi}$, $\mathbf{R}_n$ in that case would be a diagonal matrix with weights that given by their methodology filled on the diagonal. $\mathbf{B}^{(0)} \in \mathbb{R}^{p\times d}$ represents a basis and $\mathbf{C}^{(0)} \in \mathbb{R}^{d\times(p^2+p+1)}$ represents the coordinates of $\boldsymbol{\xi}\mathbf{R}_n$ relative to $\mathbf{B}$.

It's clear that this problem is over parameterized: $\mathbf{B}, \mathbf{C}$ are an over-parameterized version of $\boldsymbol{\xi}$. However, this gives us some advantage to get rid of $\mathbf{R}_n$ in the question, the non-uniqueness of this problem make sure that if we took $\tilde{\mathbf{C}} = \mathbf{C}\mathbf{R}_n^{-1}$, then the problem could be re-written as

$$D(\mathbf{B}, \mathbf{C}) = [\mathrm{vec}(\boldsymbol{\xi}) - \mathrm{vec}(\mathbf{B}\mathbf{C}\mathbf{R}_n^{-1})]'(\mathbf{R}_n \otimes I)\mathbf{V}_n(\mathbf{R}_n' \otimes I)[\mathrm{vec}(\boldsymbol{\xi}) - \mathrm{vec}(\mathbf{B}\mathbf{C}\mathbf{R}_n^{-1})]$$
$$D(\mathbf{B}, \tilde{\mathbf{C}}) = [\mathrm{vec}(\boldsymbol{\xi}) - \mathrm{vec}(\mathbf{B}\tilde{\mathbf{C}})]'\tilde{\mathbf{V}}_n[\mathrm{vec}(\boldsymbol{\xi}) - \mathrm{vec}(\mathbf{B}\tilde{\mathbf{C}})]$$

Since we are minimizing over $\mathbf{B}, \mathbf{C}$, it's equivalent to minimize over $\tilde{\mathbf{C}}$ so that $\mathbf{R}_n$ is absorbed in this procedure. In this way, we see that, as long as $\mathbf{R}_n$ is taken as a sequence of matrices that converges to a non-singular non-stochastic matrix, then it's equivalent to rewrite the problem in this way. Because $\mathbf{R}_n$ represents how one manipulates the columns of $\boldsymbol{\xi}$, there is little reason for one to choose $\mathbf{R}_n$ as singular matrix, otherwise we may throw important information away.

As for the choice of $\mathbf{V}_n$, the logic follows from the idea of generalized method of moment (GMM). And it could be shown that if $\mathbf{V}_n$ taken to be the asymptotic covariance matrix, then the minimizer could achieve the optimal in the sense of efficiency. To that end, one may need to show that the asymptotic distribution of our estimator should be normal while satisfying some additional requirements along with some assumptions. Therefore, we will discuss more details in the next section.

Asymptotic Behavior

Denote

$$\hat{\boldsymbol{\xi}} = (\hat{\beta}_{T\mathbf{Z}}, \hat{\boldsymbol{\Sigma}}_{T\mathbf{ZZ}}, \hat{\mathcal{M}}'_{T\mathbf{ZZZ}}).$$

We need to find the asymptotic distribution of $\sqrt{n}\text{vec}(\hat{\boldsymbol{\xi}} - \boldsymbol{\xi})$:

Proposition 4.1. *Assume that the data* $(\mathbf{X}_i, Y), i = 1, 2, \cdots, n$ *are simple random variable with finite sixths moment. Then*

$$\sqrt{n}\text{vec}(\hat{\boldsymbol{\xi}} - \boldsymbol{\xi}) \to N(\mathbf{0}, \boldsymbol{\Gamma}),$$

where $\boldsymbol{\Gamma} = \text{cov}\{[T\mathbf{X}', (T\mathbf{X} \otimes \mathbf{X})', (T\mathbf{X} \otimes \mathbf{X} \otimes \mathbf{X})']'\}$.

Proof. We expect to finish the proof in Appendix in the future. □

By showing the asymptotic normality, we further provide the asymptotic efficiency results for proposed estimator.

Proposition 4.2. *Assume that the data* $(\mathbf{X}_i, Y), i = 1, 2, \cdots, n$ *are simple random variable with finite sixths moment. Let* $d = \dim(\mathcal{S}_{T(Y|\mathbf{X})})$ *and* $(\hat{\mathbf{B}}, \hat{\mathbf{C}}) = \arg\min_{\mathbf{B},\mathbf{C}} D(\mathbf{B}, \mathbf{C})$. *Let*

$$\boldsymbol{\Delta}_{\boldsymbol{\xi}} = (\frac{\partial \text{vec}(\mathbf{BC})}{\partial \text{vec}(\mathbf{B})}, \frac{\partial \text{vec}(\mathbf{BC})}{\partial \text{vec}(\mathbf{C})}) \mid_{(\mathbf{B}^{(0)}, \mathbf{C}^{(0)})} = (\mathbf{C}^{(0)\prime} \otimes \mathbf{I}_p, \mathbf{I}_{p^2+p+1} \otimes \mathbf{B}^{(0)})$$

Then we have the following results:

1. $\text{vec}(\hat{\mathbf{B}}\hat{\mathbf{C}})$ *is asymptotically efficient and* $\sqrt{n}(\text{vec}(\hat{\mathbf{B}}\hat{\mathbf{C}}) - \text{vec}(\mathbf{B}^{(0)}\mathbf{C}^{(0)}))$ *is asymptotically normal with mean* $\mathbf{0}$ *and covariance matrix* $\boldsymbol{\Delta}_{\boldsymbol{\xi}}(\boldsymbol{\Delta}_{\boldsymbol{\xi}}\boldsymbol{\Gamma}^{-1}\boldsymbol{\Delta}_{\boldsymbol{\xi}})^{-}\boldsymbol{\Delta}_{\boldsymbol{\xi}}$,
2. $n\hat{D}(\hat{\mathbf{B}}, \hat{\mathbf{C}})$ *has chi-square distribution with degrees of freedom* $(p-k)(p+p^2-k)$,
3. $\mathcal{S}(\hat{\mathbf{B}})$ *is a consistent estimator of* $\mathcal{S}_{T(Y|\mathbf{X})}$.

Proof. Please refer to Appendix for the detailed proof. □

4.3 Computation Methods

Having establish the framework for the minimum discrepancy approach, we need to provide a practical working algorithm. In addition to constructing the CK kernel matrices, we need to deal with some questions in regards the detail. Minimizing the objective function will be the major task, but in order to deal with the quadratic form, we need to discuss how to find the appropriate sequences of matrices $\mathbf{V}_n$.

$\mathbf{V}_n$ in our approach is chosen as the inverse of the asymptotic covariance matrix. As it was given in theorem, the asymmptotic covariance matrix could be written in close form as a function of the sample, so it's natural to construct $\mathbf{V}_n$ by plug-in the sample.

Given the choice of $\mathbf{V}_n$, now we can talk about how to minimize the objective function. The discrepancy function $\hat{D}$ defined in previous section has two free parameters $\mathbf{B}, \mathbf{C}$. As Cook and Ni (2005) pointed out, although there are some general optimization algorithm for this unconstrained quadratic objective function, the *alternating least squares method* is probably more efficient as it utilized the special structure of the objective function. The idea of this method is to fix one of the parameters and solve for the other one using a least square method: Fixing $\mathbf{B}$, then $\text{vec}(\mathbf{C})$ will be the solution of regressing $\mathbf{V}_n^{1/2}\text{vec}(\hat{\boldsymbol{\xi}})$ on $\mathbf{V}_n(I \otimes \mathbf{B})$; and then fixing $\mathbf{C}$ will result in d least square problems where each one is regressing $\text{vec}(\hat{\boldsymbol{\xi}} - \mathbf{B}_{(-k)}\mathbf{C}_{(-k)})$ on $(C_k' \otimes I)\mathbf{Q}_{\mathbf{B}_{(-k)}}$, where B_k and C_k represent the kth column and row of $\mathbf{B}$ and $\mathbf{C}$ respectively. $\mathbf{B}_{(-k)}$, $\mathbf{C}_{(-k)}$ represent the matrix after removing B_k and C_k from $\mathbf{B}$ and $\mathbf{C}$. $\mathbf{Q}_{(\cdot)}$ represents the projection operator that project the matrix onto column null space of the given argument matrix. At each iteration, the minimization of the quadratic optimization problems became $d + 1$ linear regression problem. We name this algorithm as *Minimum Discrepancy Approach for Cubic Kernel (MDA-CK)*. Now the detailed version of the algorithm is given as follow:

Algorithm for MDA-CK: Let $\{\mathbf{X}_i, Y_i\}, i = 1, 2, \cdots, n$ be an i.i.d sample, and

assume that the structural dimension d for the CMS is known. Then

1. Standardize the predictor $\hat{\mathbf{Z}}_i = \hat{\mathbf{\Sigma}}_{\mathbf{X}}^{-1/2}(\mathbf{X}_i - \bar{\mathbf{X}})$, for $i = 1 \cdots, n$.

2. Construct kernel matrices:

$$\hat{\beta}_{Y\mathbf{Z}} = \frac{1}{n}\sum_{i=1}^{n} \hat{\mathbf{Z}}_i Y_i, \hat{\mathbf{\Sigma}}_{r_1\mathbf{ZZ}} = \frac{1}{n}\sum_{i=1}^{n} \hat{r}_{1i}\hat{\mathbf{Z}}_i\hat{\mathbf{Z}}_i',$$

   where $\hat{r}_{1i}$ is the estimated version of r_1 for ith observation and $r_1 = Y - a_1 - b_1\mathbf{Z}'\boldsymbol{\beta}_{Y\mathbf{Z}}$, with $a_1, b_1 \in \mathbb{R}$ being the OLS coefficients of Y on $(1, \mathbf{Z}'\boldsymbol{\beta}_{Y\mathbf{Z}})$.

3. Construct $\hat{\mathcal{M}}_{r_2\mathbf{ZZZ}}$ and $\hat{\hat{\mathcal{M}}}_{r_2\mathbf{ZZZ}}$. First we construct p^2 vectors for $j, k = 1, \cdots, p$:

$$\begin{aligned}\hat{\mathbf{m}}_{jk} = &\frac{1}{n}\sum_{i=1}^{n} \hat{r}_{2i}\hat{\mathbf{Z}}_i(\mathbf{e}_\mathbf{j}'\hat{\mathbf{Z}}_i)(\mathbf{e}_\mathbf{k}'\hat{\mathbf{Z}}_i) \\ &- \frac{1}{n}\sum_{i=1}^{n} \hat{r}_{2i}(\mathbf{e}_\mathbf{k}'\hat{\mathbf{Z}}_i)\mathbf{e}_\mathbf{j} - \frac{1}{n}\sum_{i=1}^{n} \hat{r}_{2i}(\mathbf{e}_\mathbf{j}'\hat{\mathbf{Z}}_i)\mathbf{e}_\mathbf{k} - \frac{1}{n}\sum_{i=1}^{n} \hat{r}_{2i}(\mathbf{e}_\mathbf{j}'\mathbf{e}_\mathbf{k})\hat{\mathbf{Z}}_\mathbf{i},\end{aligned}$$

   where $\mathbf{e}_i$ is a unit vector with ith element 1 and $\hat{r}_{2i}$ is ith element in the residual vector $\hat{r}_2$, $r_2 = Y - a_2 - b_2\mathbf{Z}'\boldsymbol{\beta}_{Y\mathbf{Z}} - c_2\mathbf{Z}'\mathbf{\Sigma}_{r_1\mathbf{ZZ}}\mathbf{Z}$ with $a_2, b_2, c_2 \in \mathbb{R}$ being the OLS coefficients of Y on $(1, \mathbf{Z}'\boldsymbol{\beta}_{Y\mathbf{Z}}, \mathbf{Z}'\mathbf{\Sigma}_{r_1\mathbf{ZZ}}\mathbf{Z})$ and $\mathbf{\Sigma}_{r_1\mathbf{ZZ}} = \mathrm{E}(r_1\mathbf{ZZ})$. Let $\hat{\mathcal{M}}_{r_2\mathbf{ZZZ}}' = (\hat{\mathbf{m}}_{11}, \cdots, \hat{\mathbf{m}}_{pp})$, which is a $p \times p^2$ matrix.

4. Construct $\hat{\boldsymbol{\xi}} = (\hat{\beta}_{T\mathbf{Z}}, \hat{\mathbf{\Sigma}}_{T\mathbf{ZZ}}, \hat{\mathcal{M}}_{T\mathbf{ZZZ}}')$ and $\hat{\mathbf{V}}_n = \hat{\mathbf{\Gamma}}^{-1}$, where $\hat{\mathbf{\Gamma}} = \hat{\mathrm{cov}}\{[T\mathbf{X}', (T\mathbf{X} \otimes \mathbf{X})', (T\mathbf{X} \otimes \mathbf{X} \otimes \mathbf{X})']'\}$. Choose an initial $\mathbf{B} = (B_1, B_2, \cdots, B_d)$ and calculate least square coefficient

$$\mathrm{vec}(\mathbf{C}) = [(\mathbf{I} \otimes \mathbf{B})'\hat{\mathbf{V}}_n(\mathbf{I} \otimes \mathbf{B})]^{-1}(\mathbf{I} \otimes \mathbf{B})\hat{\mathbf{V}}_n\mathrm{vec}(\hat{\boldsymbol{\xi}})$$

   .

5. For $k = 1, 2, \cdots, d$:

- Assign

$$\alpha_k = \text{vec}(\hat{\boldsymbol{\xi}} - \mathbf{B}_{(-k)}\mathbf{C}_{(-k)})$$

and find a new B_k such that is orthogonal to $\mathbf{B}_{(-k)}$ by

$$\hat{B}_k = \mathbf{Q}_{\mathbf{B}_{(-k)}}[\mathbf{Q}_{\mathbf{B}_{(-k)}}(C_k' \otimes \mathbf{I})\hat{\mathbf{V}}_n(C_k \otimes \mathbf{I})\mathbf{Q}_{\mathbf{B}_{(-k)}}]^{-}\mathbf{Q}_{\mathbf{B}_{(-k)}}(C_k' \otimes \mathbf{I})\hat{\mathbf{V}}_n\alpha_k$$

- Update $\mathbf{B} = (B_1, B_2, \cdots, B_{k-1}, \hat{B}_k/\|\hat{B}_k\|, B_{k+1}, \cdots, B_d)$
- Update $\mathbf{C} = \arg\min_{\mathbf{C}^*} \hat{D}(\mathbf{B}, \mathbf{C}^*)$

6. Repeat step 5 until $\|\hat{D}(\mathbf{B}, \mathbf{C})^{t+1} - \hat{D}(\mathbf{B}, \mathbf{C})^t\| < 10^{-6}$.

Remark 4.1. *The above algorithm has some instant extensions:*

1. *The above algorithm is designed specifically for CMS, central mean subspace. This algorithm could be easily extended to the case with linear functional (L-functional), by replacing the Y_i as $T(Y_i)$, where $T(\cdot)$ stands for the transformation on the response that induced by the given conditional statistical functional. The underlying theorem that support this claim could be find in Chapter 2.*

2. *To extend the results implicit functional (I-functional). An approach that is similar to Chapter 3 could also be adopted here: We first find a basis of the central subspace using some existing method denote as $\hat{\mathbf{B}}_{CS}$. And then based on the reduced dimension, we find a nonparametric estimator of $\hat{T}(Y_i \mid \mathbf{B}_{CS}\mathbf{X}_i)$ depends on which type of statistical functional of interests. Then treat $\hat{T}(Y_i \mid \mathbf{B}_{CS}\mathbf{X}_i)$ as Y_i in above algorithm, we could find the desired T-central subspace. See Chapter 3 for further technical detail.*

4.4 Order Determination

Sections 2.4 and 3.4 used ladle estimator proposed by Luo and Li (2016) and BIC methods (Li Artemiou and Li 2011) to determine the structural dimension for a given T-central subspace. Ladle estimator is great in a sense that it could remove ambiguity by combining information from the variability and magnitude of the eigenvector and eigenvalue of the kernel matrices with high accuracy. However, the ladle estimator itself relies on bootstrap to access the variability information, which leads to a high computational burden. Although similar approach could be taken for MDA-CK method, but with the increasing complexity in algorithm itself, ladle estimator may not be desirable, as the computation time for a single run could be long. Luckily we have established Proposition 4.2, which would be a great setup for the structural dimension hypothesis test. So for MDA-CK, we decide to adopt a similar testing procedure proposed by Cook and Ni (2005)

For order determination, we will be mainly considering the following sequence of marginal dimension hypotheses:

$$H_0^{(k)} : d = k \quad \text{v.s.} \quad H_1^{(k)} : d > k, \text{ for } k = 0, 1, \ldots, p-1.$$

Then the structural dimension d will be determined by the first hypothesis that is not being rejected. For each fixed k, the hypothesis is a marginal dimension test. And deriving the distribution of test statistics under $H_0^{(k)}$ would lead to a test statistic. However, the exact distribution could not be derived easily, so we are using the asymptotic distribution for the development of test statistics.

According to Proposition 4.2, under $H_0^{(k)}$, $n\hat{D}(\hat{\mathbf{B}}, \hat{\mathbf{C}})$ has an asymptotic chi-square distribution with degrees of freedom $(p-k)(p+p^2-k)$. Notice here $n\hat{D}(\hat{\mathbf{B}}, \hat{\mathbf{C}})$ could be seen as a function of the structural dimension k. So for each $H_0^{(k)}$, the test statistic

$\pi^{(k)}$ under level α is given by

$$\phi^{(k)} = \begin{cases} 0 & \text{if} \quad n\hat{D}(\hat{\mathbf{B}}, \hat{\mathbf{C}}) < \chi_{1-\alpha;(p-k)(p+p^2-k)} \\ 1 & \text{if} \quad n\hat{D}(\hat{\mathbf{B}}, \hat{\mathbf{C}}) \geq \chi_{1-\alpha;(p-k)(p+p^2-k)} \end{cases}$$

And the estimated structural dimension $\hat{d}$ is given by:

$$\hat{d} = \min\{k : \phi^{(k)} = 0\}.$$

4.5 Discussion and Future Work

In this chapter, we proposed the minimum discrepancy approach for the cubic kernel (CK) targeting T-central subspace. In particular, we developed details for the L-functionals and its corresponding asymptotic theory and order determination. In addition to the existing results, we will also study arge p small n problem following the approach from Qian, Ding and Cook (2019) in the future. Moreover, simulation studies will be conducted for recovering the CMS: We will start with the choice of $\mathbf{V}_n = \mathbf{I}$, block diagonal matrix and then move to the asymptotic covariance matrix.

MDA-CK could also be extended to I-functionals as it was done in Chapter 3. Asymptotic theory will be developed and simulations will also be conducted to study the central quantile subspace and central expectile subspace.

Appendices

A Chapter 2 Detailed Proofs for Proposition 2.1 and 2.2

In this appendix, we will provide detailed proofs for Proposition 2.1 and Proposition 2.2 in in the main text. To do so, we will prove two lemmas first.

Lemma D1. *Let α be any $p \times q$ matrix where $q \leq p$. Assume that $\mathrm{E}(\mathbf{Z}|\alpha'\mathbf{Z})$ is linear, $\mathrm{Var}(\mathbf{Z}|\alpha'\mathbf{Z})$ is constant, and that $\mathcal{M}^{(3)}(\mathbf{Z}|\alpha'\mathbf{Z}) = 0$. Then for some $p \times 1$ vector $\mathbf{b}_0$, we have*

$$\mathrm{E}((\mathbf{Z}' \otimes \mathbf{Z}')\mathcal{D}\mathbf{Z}|\alpha'\mathbf{Z}) = (\mathbf{Z}' \otimes \mathbf{Z}')(\mathbf{P}_\alpha \otimes \mathbf{P}_\alpha)\mathcal{D}\mathbf{P}_\alpha\mathbf{Z} + (\mathbf{P}_\alpha\mathbf{b}_0)'\mathbf{Z}$$

Proof. Recall that $\mathcal{D}$ is a $p \times p \times p$ array with k-th face $\mathbf{D}_k$ having element d_{ijk} in its i-th row and j-th column. Then all d_{ijk} are the same for any permutation of the index ijk, and $\mathbf{D}_k' = \mathbf{D}_k$, we may also treat $\mathcal{D}$ as a $p^2 \times p$ matrix. $\mathbf{P}_\mathcal{S}$ denotes the projection operator with respect to the standard inner product on $\mathcal{S}$, where $\mathcal{S}$ is represent any linear subspace. And define $\mathbf{Q}_\mathcal{S} = \mathbf{I} - \mathbf{P}_\mathcal{S}$. In this proof, we are taking $\mathcal{S} = \mathcal{S}(\alpha)$, space spanned by column of α.

$$\begin{aligned}
\mathrm{E}((\mathbf{Z}' \otimes \mathbf{Z}')\mathcal{D}\mathbf{Z}|\alpha'\mathbf{Z}) =& tr(\mathrm{E}((\mathbf{Z}' \otimes \mathbf{Z}')\mathcal{D}\mathbf{Z}|\alpha'\mathbf{Z})) \\
=& \mathrm{E}(tr((\mathbf{Z}' \otimes \mathbf{Z}')\mathcal{D}\mathbf{Z}|\alpha'\mathbf{Z})) \\
=& \mathrm{E}(tr(\mathbf{Z}(\mathbf{Z}' \otimes \mathbf{Z}')\mathcal{D}|\alpha'\mathbf{Z})) \\
=& \mathrm{E}(tr(\mathcal{D}'(\mathbf{Z} \otimes \mathbf{Z})\mathbf{Z}'|\alpha'\mathbf{Z})) \\
=& \mathrm{E}(tr(\mathcal{D}'(\mathbf{Z} \otimes \mathbf{Z}\mathbf{Z}')|\alpha'\mathbf{Z})) \\
=& tr(\mathcal{D}'\mathrm{E}((\mathbf{Z} \otimes \mathbf{Z}\mathbf{Z}')|\alpha'\mathbf{Z})) \\
=& tr(\mathcal{D}'\mathcal{M}_1),
\end{aligned}$$

where $\mathcal{M}_1 = \mathrm{E}((\mathbf{Z} \otimes \mathbf{Z}\mathbf{Z}')|\alpha'\mathbf{Z})$. Next, develop a different expression for $\mathcal{M}_1$, Notice

$$\begin{aligned}
\mathcal{M}^{(3)}(\mathbf{Z}|\alpha'\mathbf{Z}) = \mathrm{E}(\mathbf{Z} \otimes \mathbf{Z}\mathbf{Z}'|\alpha'\mathbf{Z}) \\
&- \mathrm{E}(\mathrm{E}(\mathbf{Z}|\alpha'\mathbf{Z}) \otimes (\mathbf{Z} - \mathrm{E}(\mathbf{Z}|\alpha'\mathbf{Z}))(\mathbf{Z} - \mathrm{E}(\mathbf{Z}|\alpha'\mathbf{Z}))'|\alpha'\mathbf{Z}) \\
&- \mathrm{E}((\mathbf{Z} - \mathrm{E}(\mathbf{Z}|\alpha'\mathbf{Z})) \otimes \mathrm{E}(\mathbf{Z}|\alpha'\mathbf{Z})(\mathbf{Z} - \mathrm{E}(\mathbf{Z}|\alpha'\mathbf{Z}))'|\alpha'\mathbf{Z}) \\
&- \mathrm{E}((\mathbf{Z} - \mathrm{E}(\mathbf{Z}|\alpha'\mathbf{Z})) \otimes (\mathbf{Z} - \mathrm{E}(\mathbf{Z}|\alpha'\mathbf{Z}))\mathrm{E}(\mathbf{Z}|\alpha'\mathbf{Z})^T|\alpha'\mathbf{Z}) \\
&- \mathrm{E}(\mathbf{Z}|\alpha'\mathbf{Z}) \otimes \mathrm{E}(\mathbf{Z}|\alpha'\mathbf{Z})\mathrm{E}(\mathbf{Z}|\alpha'\mathbf{Z})' \\
= \mathcal{M}_1 - \mathcal{M}_2 - \mathcal{M}_3 - \mathcal{M}_4 - \mathcal{M}_5,
\end{aligned}$$

where $\mathcal{M}_2, \ldots, \mathcal{M}_5$ are defined implicitly as the corresponding terms in the expression. The no skewness condition $\mathcal{M}^{(3)}(\mathbf{Z}|\alpha'\mathbf{Z}) = 0$ implies $\mathcal{M}_1 = \mathcal{M}_2 + \mathcal{M}_3 + \mathcal{M}_4 + \mathcal{M}_5$ and we claim

$$tr(\mathcal{D}'\mathcal{M}_i) = \begin{cases} (\mathbf{P}_\alpha \mathbf{b}_1)'\mathbf{Z} & \text{for } i = 2, 3, 4 \\ (\mathbf{Z}' \otimes \mathbf{Z}')(\mathbf{P}_\alpha \otimes \mathbf{P}_\alpha)\mathcal{D}\mathbf{P}_\alpha\mathbf{Z} & \text{for } i = 5 \end{cases} \tag{4.1}$$

Therefore, by the no skewness condition and (4.1) we have

$$\begin{aligned}
\mathrm{E}((\mathbf{Z}' \otimes \mathbf{Z}')\mathcal{D}\mathbf{Z}|\alpha'\mathbf{Z}) =& tr(\mathcal{D}'\mathcal{M}_1) = \sum_{i=2}^{5} tr(\mathcal{D}'\mathcal{M}_i) \\
=& (\mathbf{P}_\alpha \mathbf{b}_1)'\mathbf{Z} + (\mathbf{Z}' \otimes \mathbf{Z}')(\mathbf{P}_\alpha \otimes \mathbf{P}_\alpha)
\end{aligned}$$

And the proof is complete if we can show (4.1) holds. To show (4.1) holds, let's denote $\mathbf{W} = \mathrm{E}(\mathbf{Z}|\alpha'\mathbf{Z})$ and $\mathbf{V} = \mathbf{Z} - \mathrm{E}(\mathbf{Z}|\alpha'\mathbf{Z})$ and let $\mathbf{Q}(i, j)$ be a $p^2 \times p^2$ permutation matrix such that permuting the i-th and j-th row. Then we have

$$\begin{aligned}
tr(\mathcal{D}'\mathcal{M}_2) &= \mathrm{E}(tr(\mathcal{D}'(\mathbf{W} \otimes \mathbf{V}\mathbf{V}'))|\alpha'\mathbf{Z}) = \mathrm{E}(tr(\mathbf{V}'\mathcal{D}'(\mathbf{W} \otimes \mathbf{V}))) \\
tr(\mathcal{D}'\mathcal{M}_3) &= \mathrm{E}(tr(\mathcal{D}'(\mathbf{V} \otimes \mathbf{W}\mathbf{V}'))|\alpha'\mathbf{Z}) = \mathrm{E}(tr(\mathbf{V}'\mathcal{D}'(\mathbf{V} \otimes \mathbf{W}))) \\
tr(\mathcal{D}'\mathcal{M}_4) &= \mathrm{E}(tr(\mathcal{D}'(\mathbf{V} \otimes \mathbf{V}\mathbf{W}'))|\alpha'\mathbf{Z}) = \mathrm{E}(tr(\mathbf{W}'\mathcal{D}'(\mathbf{V} \otimes \mathbf{V})))
\end{aligned}$$

$(\mathbf{W} \otimes \mathbf{V})$ is a $p^2 \times 1$ vector whose $[(i-1)p+j]$-th row is $w_i v_j$. And $[(j-1)p+i]$-th row of $(\mathbf{V} \otimes \mathbf{W})$ has the same element $w_i v_j$. Therefore

$$(\mathbf{W} \otimes \mathbf{V}) = \prod_i \prod_j \mathbf{Q}((i-1)p+j, (j-1)p+i)(\mathbf{V} \otimes \mathbf{W})$$

And, by restriction of the array $\mathcal{D}$, $\mathcal{D}'$ is a $p \times p^2$ matrix whose $[(i-1)p+j]$-th column and $[(j-1)p+i]$-th column are the same. Thus

$$\mathcal{D}' \prod_i \prod_j \mathbf{Q}((i-1)p+j, (j-1)p+i) = \mathcal{D}'$$

Therefore, we could show $tr(\mathcal{D}'\mathcal{M}_2) = tr(\mathcal{D}'\mathcal{M}_3)$ by

$$\begin{aligned}
tr(\mathcal{D}'\mathcal{M}_2) &= \mathrm{E}(tr(\mathbf{V}'\mathcal{D}'(\mathbf{W} \otimes \mathbf{V}))) \\
&= \mathrm{E}(tr(\mathbf{V}'\mathcal{D}' \prod_i \prod_j \mathbf{Q}((i-1)p+j, (j-1)p+i)(\mathbf{V} \otimes \mathbf{W}))) \\
&= \mathrm{E}(tr(\mathbf{V}'\mathcal{D}'(\mathbf{V} \otimes \mathbf{W}))) \\
&= tr(\mathcal{D}'\mathcal{M}_3)
\end{aligned}$$

we could also show $tr(\mathcal{D}'\mathcal{M}_3) = tr(\mathcal{D}'\mathcal{M}_4)$ by

$$\begin{aligned}
tr(\mathcal{D}'\mathcal{M}_3) &= \mathrm{E}(tr(\mathbf{V}'\mathcal{D}'(\mathbf{V} \otimes \mathbf{W}))) \\
&= \mathrm{E}(tr(\mathbf{V}'(\sum_i \mathbf{D}_i' v_i) \otimes \mathbf{W})) \\
&= \mathrm{E}(tr(\sum_i \mathbf{W}'(\mathbf{D}_i v_i) \otimes \mathbf{V})) \\
&= \mathrm{E}(tr(\mathbf{W}'(\sum_i \mathbf{D}_i' v_i) \otimes \mathbf{V})) \\
&= \mathrm{E}(tr(\mathbf{W}'\mathcal{D}'(\mathbf{V} \otimes \mathbf{V}))) \\
&= tr(\mathcal{D}'\mathcal{M}_4)
\end{aligned}$$

Then let's show $tr(\mathcal{D}'\mathcal{M}_4) = (\mathbf{P}_\alpha \mathbf{b}_1)'\mathbf{Z}$

$$\begin{aligned} tr(\mathcal{D}'\mathcal{M}_4) &= tr(\mathcal{D}'\text{vec}(\mathbf{Q}_\alpha)\mathbf{Z}'\mathbf{P}_\alpha) \\ &= tr(\mathbf{Z}'\mathbf{P}_\alpha\mathcal{D}'\text{vec}(\mathbf{Q}_\alpha)) \\ &= \text{vec}(\mathbf{Q}_\alpha)'\mathcal{D}\mathbf{P}_\alpha\mathbf{Z} \\ &= (\mathbf{P}_\alpha\mathbf{b}_1)'\mathbf{Z} \end{aligned}$$

where $\mathbf{b}_1 = \mathcal{D}'\text{vec}(\mathbf{Q}_\alpha)$.

To complete the case for $i = 5$ in (4.1), it remains to show $tr(\mathcal{D}'\mathcal{M}_5) = (\mathbf{Z}' \otimes \mathbf{Z}')(\mathbf{P}_\alpha \otimes \mathbf{P}_\alpha)\mathcal{D}\mathbf{P}_\alpha\mathbf{Z}$

$$\begin{aligned} tr(\mathcal{D}'\mathcal{M}_5) &= tr(\mathcal{D}'(\mathbf{P}_\alpha \otimes \mathbf{P}_\alpha)(\mathbf{Z} \otimes \mathbf{Z}\mathbf{Z}')\mathbf{P}_\alpha) \\ &= tr(\mathcal{D}'(\mathbf{P}_\alpha \otimes \mathbf{P}_\alpha)(\mathbf{Z} \otimes \mathbf{Z})\mathbf{Z}'\mathbf{P}_\alpha) \\ &= tr(\mathbf{Z}'\mathbf{P}_\alpha\mathcal{D}'(\mathbf{P}_\alpha \otimes \mathbf{P}_\alpha)(\mathbf{Z} \otimes \mathbf{Z})) \\ &= (\mathbf{Z}' \otimes \mathbf{Z}')(\mathbf{P}_\alpha \otimes \mathbf{P}_\alpha)\mathcal{D}\mathbf{P}_\alpha\mathbf{Z} \\ &= (\mathbf{P}_\alpha\mathbf{b}_1)'\mathbf{Z} \end{aligned}$$

□

Based on Lemma 1, we develop the following lemma.

Lemma D2. *Let the columns of α forms a basis for $\mathcal{S}(\beta_{T\mathbf{Z}}, \mathbf{\Sigma}_{T\mathbf{ZZ}}, \mathcal{M}'_{T\mathbf{ZZZ}})$ then*

$$\begin{aligned} & \text{E}((T_Y - \text{E}(T_Y)))(\mathbf{Z}' \otimes \mathbf{Z}')\mathcal{D}\mathbf{Z}|\alpha'\mathbf{Z}) \\ = \; & \text{E}[(T_Y - \text{E}(T_Y))(\mathbf{Z}' \otimes \mathbf{Z}')(\mathbf{P}_\alpha \otimes \mathbf{P}_\alpha)\mathcal{D}\mathbf{P}_\alpha\mathbf{Z} + (T_Y \\ & - \text{E}(T_Y))(\mathbf{P}_\alpha\mathbf{b}_0)'\mathbf{Z}] \end{aligned}$$

where, $T_Y = T(Y)$ for ease of notation.

Proof. Recall that $\beta_{T\mathbf{Z}}$ represents population OLS slope estimate for regression of $T(Y)$ on $\mathbf{Z}$, $\mathbf{\Sigma}_{T\mathbf{ZZ}}$ represents the population kernel for principal Hessian directions, $\mathcal{M}_{T\mathbf{ZZZ}}$ is the kernel for third moment (Yin and Cook 2004).

By the construction of $\beta_{T\mathbf{Z}}, \mathbf{\Sigma}_{T\mathbf{ZZ}}, \mathcal{M}_{T\mathbf{ZZZ}}$ and the fact that columns of α forms a basis for $\mathcal{S}(\beta_{T\mathbf{Z}}, \mathbf{\Sigma}_{T\mathbf{ZZ}}, \mathcal{M}'_{T\mathbf{ZZZ}})$, we have:

$$\begin{aligned}
\mathcal{M}_{T\mathbf{ZZZ}} &= \mathrm{E}((T_Y - \mathrm{E}(T_Y))(\mathbf{Z} \otimes \mathbf{ZZ}')) - \mathrm{E}(T_Y\mathbf{Z}) \otimes I \\
&\quad - I \otimes \mathrm{E}(T_Y\mathbf{Z}) - (I \otimes I)\mathrm{vec}(I)\mathrm{E}(T_Y\mathrm{Z})' \\
\mathcal{M}_{T\mathbf{ZZZ}} &= (\mathbf{P}_\alpha \otimes \mathbf{P}_\alpha)\mathcal{M}_{T\mathbf{ZZZ}}\mathbf{P}_\alpha \\
&= \mathrm{E}((T_Y - \mathrm{E}(T_Y))(\mathbf{P}_\alpha \otimes \mathbf{P}_\alpha)(\mathbf{Z} \otimes \mathbf{ZZ}')\mathbf{P}_\alpha) - \mathrm{E}(T_Y\mathbf{Z}) \otimes \mathbf{P}_\alpha \\
&\quad - \mathbf{P}_\alpha \otimes \mathrm{E}(T_Y\mathbf{Z}) - (\mathbf{P}_\alpha \otimes \mathbf{P}_\alpha)\mathrm{vec}(I)\mathrm{E}(T_Y\mathrm{Z})'
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
&\mathrm{E}((T_Y - \mathrm{E}(T_Y))(\mathbf{Z} \otimes \mathbf{ZZ}')) \\
&= \mathrm{E}((T_Y - \mathrm{E}(T_Y))(\mathbf{P}_\alpha \otimes \mathbf{P}_\alpha)(\mathbf{Z} \otimes \mathbf{ZZ}')\mathbf{P}_\alpha) \\
&= + \mathrm{E}(T_Y\mathbf{Z}) \otimes \mathbf{Q}_\alpha + \mathbf{Q}_\alpha \otimes \mathrm{E}(T_Y\mathbf{Z}) \\
&\quad + (I \otimes I - \mathbf{P}_\alpha \otimes \mathbf{P}_\alpha)\mathrm{vec}(I)\mathrm{E}(T_Y\mathrm{Z})'
\end{aligned} \tag{4.2}$$

$$\begin{aligned}
&tr(\mathcal{D}'\mathrm{E}((T_Y - \mathrm{E}(T_Y))(\mathbf{P}_\alpha \otimes \mathbf{P}_\alpha)(\mathbf{Z} \otimes \mathbf{ZZ}')\mathbf{P}_\alpha)) \\
&= \mathrm{E}((T_Y - \mathrm{E}(T_Y))tr(\mathcal{D}'(\mathbf{P}_\alpha \otimes \mathbf{P}_\alpha)(\mathbf{Z} \otimes \mathbf{ZZ}')\mathbf{P}_\alpha)) \\
&= \mathrm{E}((T_Y - \mathrm{E}(T_Y))tr(\mathbf{Z}'\mathbf{P}_\alpha\mathcal{D}'(\mathbf{P}_\alpha \otimes \mathbf{P}_\alpha)(\mathbf{Z} \otimes \mathbf{Z}))) \\
&= tr(\mathrm{E}((T_Y - \mathrm{E}(T_Y))(\mathbf{Z}' \otimes \mathbf{Z}')(\mathbf{P}_\alpha \otimes \mathbf{P}_\alpha)\mathcal{D}\mathbf{P}_\alpha\mathbf{Z})) \\
&= \mathrm{E}((T_Y - \mathrm{E}(T_Y))(\mathbf{Z}' \otimes \mathbf{Z}')(\mathbf{P}_\alpha \otimes \mathbf{P}_\alpha)\mathcal{D}\mathbf{P}_\alpha\mathbf{Z})
\end{aligned} \tag{4.3}$$

$$
\begin{aligned}
&(I \otimes I - \mathbf{P}_\alpha \otimes \mathbf{P}_\alpha)\mathrm{vec}(I) \\
&= ((\mathbf{P}_\alpha + \mathbf{Q}_\alpha) \otimes (\mathbf{P}_\alpha + \mathbf{Q}_\alpha) - \mathbf{P}_\alpha \otimes \mathbf{P}_\alpha)\mathrm{vec}(I) \\
&= (\mathbf{Q}_\alpha \otimes \mathbf{Q}_\alpha + \mathbf{P}_\alpha \otimes \mathbf{Q}_\alpha + \mathbf{Q}_\alpha \otimes \mathbf{P}_\alpha)\mathrm{vec}(I) \\
&= \mathrm{vec}(\mathbf{Q}_\alpha)
\end{aligned} \tag{4.4}
$$

Therefore,

$$
\begin{aligned}
&\mathrm{E}((T_Y - \mathrm{E}(T_Y))(\mathbf{Z}' \otimes \mathbf{Z}')\mathcal{D}\mathbf{Z}) \\
&= tr(\mathrm{E}((T_Y - \mathrm{E}(T_Y))(\mathbf{Z}' \otimes \mathbf{Z}')\mathcal{D}\mathbf{Z})) \\
&= tr(\mathrm{E}((T_Y - \mathrm{E}(T_Y))\mathcal{D}'(\mathbf{Z} \otimes \mathbf{Z}\mathbf{Z}'))) \\
&= tr(\mathcal{D}'\mathrm{E}((T_Y - \mathrm{E}(T_Y))(\mathbf{Z} \otimes \mathbf{Z}\mathbf{Z}'))) \\
&= tr(\mathcal{D}'\mathrm{E}((T_Y - \mathrm{E}(T_Y))(\mathbf{P}_\alpha \otimes \mathbf{P}_\alpha)(\mathbf{Z} \otimes \mathbf{Z}\mathbf{Z}')\mathbf{P}_\alpha)) \\
&\quad + tr(\mathcal{D}'\mathrm{E}(T_Y\mathbf{Z}) \otimes \mathbf{Q}_\alpha) + tr(\mathcal{D}'\mathbf{Q}_\alpha \otimes \mathrm{E}(T_Y\mathbf{Z})) \\
&\quad + tr(\mathcal{D}'(I \otimes I - \mathbf{P}_\alpha \otimes \mathbf{P}_\alpha)\mathrm{vec}(I)\mathrm{E}(T_Y\mathrm{Z})') \\
&= \mathrm{E}((T_Y - \mathrm{E}(T_Y))(\mathbf{Z}' \otimes \mathbf{Z}')(\mathbf{P}_\alpha \otimes \mathbf{P}_\alpha)\mathcal{D}\mathbf{P}_\alpha\mathbf{Z}) \\
&\quad + \mathrm{E}[(T_Y - \mathrm{E}(T_Y))tr(\mathcal{D}'(\mathbf{P}_\alpha\mathbf{Z} \otimes \mathbf{Q}_\alpha + \mathbf{Q}_\alpha \otimes \mathbf{P}_\alpha\mathbf{Z} \\
&\quad + (I \otimes I - \mathbf{P}_\alpha \otimes \mathbf{P}_\alpha)\mathrm{vec}(I)\mathbf{Z}'\mathbf{P}_\alpha))] \\
&= \mathrm{E}((T_Y - \mathrm{E}(T_Y))(\mathbf{Z}' \otimes \mathbf{Z}')(\mathbf{P}_\alpha \otimes \mathbf{P}_\alpha)\mathcal{D}\mathbf{P}_\alpha\mathbf{Z}) \\
&\quad + \mathrm{E}[(T_Y - \mathrm{E}(T_Y))tr(\mathcal{D}'(\mathcal{M}_2 + \mathcal{M}_3 + \mathcal{M}_4))]
\end{aligned}
$$

The forth equality is because of (4.2), and the fifth equality is due to (4.3). And in the last equality, $\mathcal{M}_i, i = 2, 3, 4$ is the same notation as the one in Lemma D1. And the equality holds because of (4.4). Hence, we have completed the proof. □

Proof of Proposition 2.1. Recall that $a \in \mathbb{R}^1$, $\mathbf{b} \in \mathbb{R}^p$ and $\mathbf{C}$ is a $p \times p$

symmetric matrix. Since $\phi(\cdot)$ is a convex function, γ is basis for T-central subspace.

$$\begin{aligned} R(a, \mathbf{b}, \mathbf{C}, \mathcal{D}) &= \mathrm{E}(-T_Y K(\mathbf{Z}) + \phi(K(\mathbf{Z}))) \\ &= \mathrm{E}(-\mathrm{E}(T_Y|\gamma'\mathbf{Z})K(\mathbf{Z}) + \phi(K(\mathbf{Z}))) \\ &= \mathrm{E}(-\mathrm{E}(T_Y|\gamma'\mathbf{Z})\mathrm{E}(K(\mathbf{Z})|\gamma'\mathbf{Z}) + \phi(K(\mathbf{Z})|\gamma'\mathbf{Z})) \\ &\geq \mathrm{E}(-\mathrm{E}(T_Y|\gamma'\mathbf{Z})\mathrm{E}(K(\mathbf{Z})|\gamma'\mathbf{Z}) + \phi(\mathrm{E}(K(\mathbf{Z})|\gamma'\mathbf{Z}))) \end{aligned}$$

Now we are going to deal with $\mathrm{E}(K(\mathbf{Z})|\gamma'\mathbf{Z})$. Notice that

$$\begin{aligned} &\mathrm{E}(\mathbf{Z}\mathbf{Z}'|\gamma'\mathbf{Z}) \\ &= \mathrm{E}(\mathbf{Z}|\gamma'\mathbf{Z})\mathrm{E}(\mathbf{Z}|\gamma'\mathbf{Z}) + \mathrm{Var}(\mathbf{Z}|\gamma'\mathbf{Z}) \\ &= \mathbf{P}_\gamma \mathbf{Z}\mathbf{Z}'\mathbf{P}_\gamma + \mathbf{Q}_\gamma \\ &\mathrm{E}(K(\mathbf{Z})|\gamma'\mathbf{Z}) \\ &= a + \mathbf{b}'\mathrm{E}(\mathbf{Z}|\gamma'\mathbf{Z}) + \mathrm{E}(\mathbf{Z}'\mathbf{C}\mathbf{Z}|\gamma'\mathbf{Z}) + \mathrm{E}((\mathbf{Z}'\otimes\mathbf{Z}')\mathcal{D}\mathbf{Z}|\gamma'\mathbf{Z}) \\ &= a + \mathbf{b}'\mathrm{E}(\mathbf{Z}|\gamma'\mathbf{Z}) + tr(\mathrm{E}(\mathbf{Z}\mathbf{Z}'\mathbf{C}|\gamma'\mathbf{Z})) + \mathrm{E}((\mathbf{Z}'\otimes\mathbf{Z}')\mathcal{D}\mathbf{Z}|\gamma'\mathbf{Z}) \\ &= a + \mathbf{b}'\mathbf{P}_\gamma\mathbf{Z} + tr(\mathbf{P}_\gamma\mathbf{Z}\mathbf{Z}'\mathbf{P}_\gamma\mathbf{C} + \mathbf{Q}_\gamma\mathbf{C}) + (\mathbf{P}_\gamma\mathbf{b}_0)'\mathbf{Z} + (\mathbf{Z}'\otimes\mathbf{Z}')(\mathbf{P}_\gamma\otimes\mathbf{P}_\gamma)\mathcal{D}\mathbf{P}_\gamma\mathbf{Z} \\ &= (a + \mathbf{Q}_\gamma\mathbf{C}) + (\mathbf{b}+\mathbf{b}_0)'\mathbf{P}_\gamma\mathbf{Z} + tr(\mathbf{P}_\gamma\mathbf{C}\mathbf{P}_\gamma\mathbf{Z}\mathbf{Z}') + (\mathbf{Z}'\otimes\mathbf{Z}')(\mathbf{P}_\gamma\otimes\mathbf{P}_\gamma)\mathcal{D}\mathbf{P}_\gamma\mathbf{Z} \end{aligned} \tag{4.5}$$

That is

$$R(a, \mathbf{b}, \mathbf{C}, \mathcal{D}) \geq R((a + tr(\mathbf{Q}_\gamma\mathbf{b}\mathbf{b}')), \mathbf{P}_\gamma(\mathbf{b}+\mathbf{b}_0), \mathbf{P}_\gamma\mathbf{C}\mathbf{P}_\gamma, (\mathbf{P}_\gamma\otimes\mathbf{P}_\gamma)\mathcal{D}\mathbf{P}_\gamma)$$

Since we have assume that the $(\alpha, \beta, \Gamma, \Delta)$ is unique, therefore, we have shown that

$$\mathcal{S}(\beta, \Gamma, \Delta') \subseteq \mathcal{S}_{T(Y|\mathbf{Z})}$$

Otherwise, we could always find that $\mathbf{P}_\gamma\beta, \mathbf{P}_\gamma\Gamma\mathbf{P}_\gamma, (\mathbf{P}_\gamma\otimes\mathbf{P}_\gamma)\Delta\mathbf{P}_\gamma$ has a smaller risk

than the risk of using β, Γ, Δ.

Proof of Proposition 2.2. Notice that Since

$$R(a, \mathbf{b}, \mathbf{C}, \mathcal{D}) = \mathrm{E}(-(T_Y - \mathrm{E}(T_Y))K(\mathbf{Z}) - \mathrm{E}(T_Y)K(\mathbf{Z}) + \phi(K(\mathbf{Z}))$$

$\phi(\cdot)$ is convex, $\mathrm{E}(T_Y)K(\mathbf{Z})$ is also a convex function, thus without loss of generality, we can assume $\mathrm{E}(T_Y) = 0$.

Moreover, we have

$$\mathbf{b}'\mathrm{E}(T_Y\mathbf{Z}) = (\mathbf{P}_\alpha\mathbf{b})'\mathrm{E}(T_Y\mathbf{Z})$$

and

$$\mathrm{E}(T_Y\mathbf{Z}'\mathbf{C}\mathbf{Z}) = \mathrm{E}(T_Y\mathbf{Z}'\mathbf{P}_\alpha\mathbf{C}\mathbf{P}_\alpha\mathbf{Z})$$

Combining these, by using Jensen's Inequality, it is similar to show that

$$\begin{aligned}
R(a, \mathbf{b}, \mathbf{C}, \mathcal{D}) =& \mathrm{E}(L(a + \mathbf{b}'\mathbf{Z} + \mathbf{Z}'\mathbf{C}'\mathbf{Z} + (\mathbf{Z}' \otimes \mathbf{Z}')\mathcal{D}\mathbf{Z}, T_Y)) \\
=& \mathrm{E}(-T_Y((\mathbf{P}_\alpha\mathbf{b})'\mathbf{Z} + \mathbf{Z}'\mathbf{P}_\alpha\mathbf{C}\mathbf{P}_\alpha\mathbf{Z} + (\mathbf{P}_\alpha\mathbf{b}_0)'\mathbf{Z} \\
&+ (\mathbf{Z}' \otimes \mathbf{Z}')(\mathbf{P}_\alpha \otimes \mathbf{P}_\alpha)\mathcal{D}\mathbf{P}_\alpha\mathbf{Z})) + \mathrm{E}(\phi(K(\mathbf{Z}))) \\
=& \mathrm{E}(-T_Y(a + tr(Q_\alpha\mathbf{C}) + (\mathbf{P}_\alpha(\mathbf{b} + \mathbf{b}_0))'\mathbf{Z} + \mathbf{Z}'\mathbf{P}_\alpha\mathbf{C}\mathbf{P}_\alpha\mathbf{Z} \\
&+ (\mathbf{Z}' \otimes \mathbf{Z}')(\mathbf{P}_\alpha \otimes \mathbf{P}_\alpha)\mathcal{D}\mathbf{P}_\alpha\mathbf{Z})) + \mathrm{E}(\phi(K(\mathbf{Z})))
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}(\phi(K(\mathbf{Z}))) =& \mathrm{E}(\mathrm{E}(\phi(K(\mathbf{Z}))|\alpha'\mathbf{Z})) \\
\geq& \mathrm{E}(\phi(\mathrm{E}(K(\mathbf{Z})|\alpha'\mathbf{Z}))) \\
=& \mathrm{E}(\phi(a + tr(Q_\alpha\mathbf{C}) + (\mathbf{P}_\alpha(\mathbf{b} + \mathbf{b}_0))'\mathbf{Z} + \mathbf{Z}'\mathbf{P}_\alpha\mathbf{C}\mathbf{P}_\alpha\mathbf{Z} \\
&+ (\mathbf{Z}' \otimes \mathbf{Z}')(\mathbf{P}_\alpha \otimes \mathbf{P}_\alpha)\mathcal{D}\mathbf{P}_\alpha\mathbf{Z}))
\end{aligned}$$

Therefore,

$$R(a, \mathbf{b}, \mathbf{C}, \mathcal{D}) \geq R((a + tr(\mathbf{Q}_\alpha \mathbf{b}\mathbf{b}')), \mathbf{P}_\alpha(\mathbf{b} + \mathbf{b}_0), \mathbf{P}_\alpha \mathbf{C} \mathbf{P}_\alpha, (\mathbf{P}_\alpha \otimes \mathbf{P}_\alpha)\mathcal{D}\mathbf{P}_\alpha)$$

And by the uniqueness of the β and same argument, the result follows.

B Chapter 3 Detailed Proof for Proposition 3.4

In this supplementary files, we will provide detailed proofs for Proposition 3.4 in the main text. To do so, we will present the following lemma from Pollard (1991).

Lemma D3. *Let $b_n(\boldsymbol{\theta})$ be a sequence of random convex function defined on a convex, open subset Θ of $\mathbb{R}^d$. Suppose $b(\cdot)$ is a real-valued convex function on Θ for which $b_n(\boldsymbol{\theta}) \to b(\boldsymbol{\theta})$ in probability, for each θ in Θ. Then for each compact subset K of Θ,*

$$\sup_{\boldsymbol{\theta}\in\Theta} \|b_n(\boldsymbol{\theta}) - b(\boldsymbol{\theta})\| \xrightarrow{P} 0$$

Proof for Proposition 3.4

The idea for this proof is to approximate the target by a quadratic function whose minimizing value has an is $O_p(n^{-1/2})$ and then to show that the sample version estimator lies close enough to that minimizing value to share its asymptotic properties. Let's first define a new quantity to help us proof the $\sqrt{n}$-consistent. Define

$$\begin{aligned}
\hat{b}_n(\boldsymbol{\theta}) =& \sum_{i=1}^{n} (\hat{T}(Y \mid \hat{\mathbf{B}}_{CS}\mathbf{X}_i) - \left(1, \mathbf{X}_i', \mathbf{X}_i' \otimes \mathbf{X}_i', \mathbf{X}_i' \otimes \mathbf{X}_i' \otimes \mathbf{X}_i'\right) \boldsymbol{\theta})^2 - \sum_{i=1}^{n} \hat{T}^2(Y \mid \hat{\mathbf{B}}_{CS}\mathbf{X}_i) \\
=& -2(\sum_{i=1}^{n} \hat{T}(Y \mid \hat{\mathbf{B}}_{CS}\mathbf{X}_i) \left(1, \mathbf{X}_i', \mathbf{X}_i' \otimes \mathbf{X}_i', \mathbf{X}_i' \otimes \mathbf{X}_i' \otimes \mathbf{X}_i'\right))\boldsymbol{\theta} \\
&+ \boldsymbol{\theta}' \sum_{i=1}^{n} \left(1, \mathbf{X}_i', \mathbf{X}_i' \otimes \mathbf{X}_i', \mathbf{X}_i' \otimes \mathbf{X}_i' \otimes \mathbf{X}_i'\right)' \left(1, \mathbf{X}_i', \mathbf{X}_i' \otimes \mathbf{X}_i', \mathbf{X}_i' \otimes \mathbf{X}_i' \otimes \mathbf{X}_i'\right) \boldsymbol{\theta},
\end{aligned}$$

for $\boldsymbol{\theta} \in \mathbb{R}^{1+p+p^2+p^3}$. and define $\Theta = \{\boldsymbol{\theta} = \left(a, \mathbf{b}', \text{vec}(\mathbf{C})', \text{vec}(\mathcal{D})'\right)' : a \in \mathbb{R}, \mathbf{b} \in \mathbb{R}^p, \mathbf{C}$ is a $p \times p$ symmetric matrix., D is a 3 dimensional $p \times p \times p$ array.$\}$ Notice if we take $\boldsymbol{\theta} \in \Theta$, then

$$\left(1, \mathbf{X}_i', \mathbf{X}_i' \otimes \mathbf{X}_i', \mathbf{X}_i' \otimes \mathbf{X}_i' \otimes \mathbf{X}_i'\right) \boldsymbol{\theta} = a + \mathbf{X}_i'\mathbf{b} + \mathbf{X}_i'\mathbf{C}\mathbf{X}_i + \mathbf{X}_i' \otimes \mathbf{X}_i'\mathcal{D}\mathbf{X}_i = K(\mathbf{X}_i).$$

Denote $\boldsymbol{\theta}^* = \left(a^*, \mathbf{b}^{*\prime}, \mathrm{vec}(\mathbf{C}^*)', \mathrm{vec}(\mathcal{D}^*)'\right)'$, where $a^*, \mathbf{b}^*, \mathbf{C}^*, \mathcal{D}^*$ is the population minimizer as we defined in section 3.2, and denote $\mathfrak{X}_i = \left(1, \mathbf{X}_i', \mathbf{X}_i' \otimes \mathbf{X}_i', \mathbf{X}_i' \otimes \mathbf{X}_i' \otimes \mathbf{X}_i'\right)'$ Then for $\boldsymbol{\theta} \in \Theta$, we have

$$\begin{aligned}\hat{b}_n(\boldsymbol{\theta}^* + \frac{\boldsymbol{\theta}}{\sqrt{n}}) =& \hat{b}_n(\boldsymbol{\theta}^*) - \frac{2}{\sqrt{n}} \sum_{i=1}^n \hat{T}(Y \mid \hat{\mathbf{B}}_{CS}\mathbf{X}_i)\mathfrak{X}_i'\boldsymbol{\theta} + \frac{1}{n}\boldsymbol{\theta}' \sum_{i=1}^n \mathfrak{X}_i\mathfrak{X}_i'\boldsymbol{\theta} \\ =& \hat{b}_n(\boldsymbol{\theta}^*) - \frac{2}{\sqrt{n}} \sum_{i=1}^n \hat{T}(Y \mid \hat{\mathbf{B}}_{CS}\mathbf{X}_i)\mathfrak{X}_i'\boldsymbol{\theta} + \boldsymbol{\theta}'\mathrm{E}(\mathfrak{X}\mathfrak{X}')\boldsymbol{\theta} + o_p(1)\end{aligned}$$

Furthermore, if we take

$$b_n(\boldsymbol{\theta}) = \hat{b}_n(\boldsymbol{\theta}^* + \frac{\boldsymbol{\theta}}{\sqrt{n}}) - \hat{b}_n(\boldsymbol{\theta}^*) + \frac{2}{\sqrt{n}} \sum_{i=1}^n \hat{T}(Y \mid \hat{\mathbf{B}}_{CS}\mathbf{X}_i)\mathfrak{X}_i'\boldsymbol{\theta} \xrightarrow{P} \boldsymbol{\theta}'\mathrm{E}(\mathfrak{X}\mathfrak{X}')\boldsymbol{\theta} = b(\boldsymbol{\theta}),$$

we could surely know that $b_n(\boldsymbol{\theta})$ is a sequence of convex function defined on Θ that converge in probability to a convex function $b(\boldsymbol{\theta})$ for each $\boldsymbol{\theta} \in \Theta$, given $\frac{2}{\sqrt{n}} \sum_{i=1}^n \hat{T}(Y \mid \hat{\mathbf{B}}_{CS}\mathbf{X}_i)\mathfrak{X}_i'\boldsymbol{\theta}$ is bounded in probability for each $\boldsymbol{\theta}$. Then we could apply the convexity lemma to improve the result that this convergence is uniform on any compact set $K \in \Theta$.

Now we are going to show $\frac{2}{\sqrt{n}} \sum_{i=1}^n \hat{T}(Y \mid \hat{\mathbf{B}}_{CS}\mathbf{X}_i)\mathfrak{X}_i$ is bounded in probability to complete above claim. It's suffice to show that the second moment is finite. However, this quantity depends on data, so we consider $\frac{2}{\sqrt{n}} \sum_{i=1}^n T(Y \mid \mathbf{B}_{CS}\mathbf{X}_i)\mathfrak{X}_i$ first. Define function $t : \mathbb{R}^{1+p+p^2+p^3} \to \mathbb{R}$ such that $t(y, \mathbf{B}\mathfrak{x}) = T(y|\mathbf{B}\mathfrak{x})$ and let $\mathfrak{T} = \{t : \mathrm{E}\|t^2(Y, \mathbf{B}_{CS}\mathbf{X})\mathfrak{X}\mathfrak{X}'\| < \infty, \|t\|_\infty < \infty\}$, then

$$\sup_{t\in\mathfrak{T}} \mathrm{E}(\frac{2}{\sqrt{n}} \sum_{i=1}^n t(Y, \mathbf{B}_{CS}\mathbf{X}_i)\mathfrak{X}_i)(\frac{2}{\sqrt{n}} \sum_{i=1}^n t(Y, \mathbf{B}_{CS}\mathbf{X}_i)\mathfrak{X}_i)' \leq \sup_{t\in\mathfrak{T}} \frac{4}{n} \sum_{i=1}^n t^2(Y, \mathbf{B}_{CS}\mathbf{X}_i)\|\mathfrak{X}_i\mathfrak{X}_i'\| < \infty$$

This implies $\frac{2}{\sqrt{n}} \sum_{i=1}^n T(Y \mid \mathbf{B}_{CS}\mathbf{X}_i)\mathfrak{X}_i$ a bounded random variable because $T(Y \mid \mathbf{B}_{CS}\mathbf{X}_i)$ is in $\mathfrak{T}$. Also, by previous lemma, we know that $\hat{T}(Y \mid \hat{\mathbf{B}}_{CS}\mathbf{X}_i)$ for n large enough. This implies that $\frac{2}{\sqrt{n}} \sum_{i=1}^n \hat{T}(Y \mid \hat{\mathbf{B}}_{CS}\mathbf{X}_i)\mathfrak{X}_i$ is also bounded in probability.

By far, we are ready to provide the argument for the $\sqrt{n}$-consistency. As it was pointed out previously, we are going to show that $\hat{\mathbf{M}}$ is close to $\boldsymbol{\theta}^*$. It's equivalent to show that $\hat{\mathbf{M}}$ is within a $\boldsymbol{\theta}/\sqrt{n}$ ball of $\boldsymbol{\theta}^*$, $\{\boldsymbol{\theta}^* + \boldsymbol{\theta}/\sqrt{n} : \|\boldsymbol{\theta}\| < C, \boldsymbol{\theta} \in \Theta\}$ with high probability. In another word, outside of the $\boldsymbol{\theta}/\sqrt{n}$ ball of $\boldsymbol{\theta}^*$, we could not find sample minimizer that could beat $\boldsymbol{\theta}^*$ with high probability, i.e.

$$P(\inf_{\boldsymbol{\theta} \geq C} \hat{b}_n(\boldsymbol{\theta}^* + \boldsymbol{\theta}/\sqrt{n}) - \hat{b}_n(\boldsymbol{\theta}^*) > \epsilon) \to 1.$$

To see this holds, we already have

$$\hat{b}_n(\boldsymbol{\theta}^* + \frac{\boldsymbol{\theta}}{\sqrt{n}}) - \hat{b}_n(\boldsymbol{\theta}^*) = -\frac{2}{\sqrt{n}} \sum_{i=1}^{n} \hat{T}(Y \mid \hat{\mathbf{B}}_{CS}\mathbf{X}_i)\mathfrak{X}_i'\boldsymbol{\theta} + \boldsymbol{\theta}'\mathrm{E}(\mathfrak{X}\mathfrak{X}')\boldsymbol{\theta} + o_p(1)$$

on any compact set in Θ. This quantity is quadratic in $\boldsymbol{\theta}$ and hence dominated by $\boldsymbol{\theta}'\mathrm{E}(\mathfrak{X}\mathfrak{X}')\boldsymbol{\theta}$, which means the quantity is bounded away from zero with probability tends to 1.

C Chapter 4 Detailed Proof for 4.1

In this supplementary files, we will provide detailed proofs for Propositions in the main text.

Proof of Proposition 4.1

In this proof, we are going to use the following notation:

$$
\begin{aligned}
\boldsymbol{\mu}_1 &= \mathrm{E}(\mathbf{X}) & \hat{\boldsymbol{\mu}}_1 &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i \\
\boldsymbol{\mu}_2 &= \mathrm{E}(\mathbf{X}\otimes\mathbf{X}) & \hat{\boldsymbol{\mu}}_2 &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\otimes\mathbf{X}_i \\
\boldsymbol{\mu}_3 &= \mathrm{E}(\mathbf{X}\otimes\mathbf{X}\otimes\mathbf{X}) & \hat{\boldsymbol{\mu}}_3 &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\otimes\mathbf{X}_i\otimes\mathbf{X}_i \\
\boldsymbol{\theta}_1 &= \mathrm{E}(T\mathbf{Z}) & \hat{\boldsymbol{\theta}} &= \frac{1}{n}\sum_{i=1}^{n}T_i\mathbf{Z}_i \\
\boldsymbol{\theta}_2 &= \mathrm{E}(T\mathbf{Z}\otimes\mathbf{Z}) & \hat{\boldsymbol{\theta}}_2 &= \frac{1}{n}\sum_{i=1}^{n}T_i\mathbf{Z}_i\otimes\mathbf{Z}_i \\
\boldsymbol{\theta}_3 &= \mathrm{E}(T\mathbf{Z}\otimes\mathbf{Z}\otimes\mathbf{Z}) & \hat{\boldsymbol{\theta}}_3 &= \frac{1}{n}\sum_{i=1}^{n}T_i\mathbf{Z}_i\otimes\mathbf{Z}_i\otimes\mathbf{Z}_i \\
\boldsymbol{\xi}_0 &= \mathrm{E}(T) & \hat{\boldsymbol{\xi}}_0 &= \frac{1}{n}\sum_{i=1}^{n}T_i \\
\boldsymbol{\xi}_1 &= \mathrm{E}(T\mathbf{X}) & \hat{\boldsymbol{\xi}}_1 &= \frac{1}{n}\sum_{i=1}^{n}T_i\mathbf{X}_i \\
\boldsymbol{\xi}_2 &= \mathrm{E}(T\mathbf{X}\otimes\mathbf{X}) & \hat{\boldsymbol{\xi}}_2 &= \frac{1}{n}\sum_{i=1}^{n}T_i\mathbf{X}_i\otimes\mathbf{X}_i \\
\boldsymbol{\xi}_3 &= \mathrm{E}(T\mathbf{X}\otimes\mathbf{X}\otimes\mathbf{X}) & \hat{\boldsymbol{\xi}}_3 &= \frac{1}{n}\sum_{i=1}^{n}T_i\mathbf{X}_i\otimes\mathbf{X}_i\otimes\mathbf{X}_i.
\end{aligned}
$$

Now we first notice that

$$\begin{aligned}
\boldsymbol{\theta}_1 =& \mathrm{E}(\Sigma^{-1/2}T(\mathbf{X}-\boldsymbol{\mu}_1)) = \Sigma^{-1/2}[\mathrm{E}(T\mathbf{X}) - \boldsymbol{\mu}_1\boldsymbol{\xi}_0] = \Sigma^{-1/2}(\boldsymbol{\xi}_1 - \boldsymbol{\mu}_1\boldsymbol{\xi}_0)\\
\boldsymbol{\theta}_2 =& \mathrm{E}(T\Sigma^{-1/2}(\mathbf{X}-\boldsymbol{\mu}_1)\otimes\Sigma^{-1/2}(\mathbf{X}-\boldsymbol{\mu}_1)) = (\Sigma^{-1/2}\otimes\Sigma^{-1/2})[\boldsymbol{\xi}_2 - \boldsymbol{\mu}_2\boldsymbol{\xi}_0\\
& - \boldsymbol{\mu}\otimes\boldsymbol{\xi}_1 - \boldsymbol{\xi}_1\otimes\boldsymbol{\mu} + 2\boldsymbol{\xi}_0\boldsymbol{\mu}\otimes\boldsymbol{\mu}]\\
\boldsymbol{\theta}_3 =& \mathrm{E}(T\Sigma^{-1/2}(\mathbf{X}-\boldsymbol{\mu}_1)\otimes\Sigma^{-1/2}(\mathbf{X}-\boldsymbol{\mu}_1)\otimes\Sigma^{-1/2}(\mathbf{X}-\boldsymbol{\mu}_1))\\
=& (\Sigma^{-1/2}\otimes\Sigma^{-1/2}\otimes\Sigma^{-1/2})[\boldsymbol{\xi}_3 - \boldsymbol{\xi}_2\otimes\boldsymbol{\mu}_1 - \mathrm{E}(Y\mathbf{X}\otimes\mu\mathbf{X}) + \boldsymbol{\xi}_1\otimes\boldsymbol{\mu}_1\otimes\boldsymbol{\mu}_1\\
& - \boldsymbol{\mu}_1\otimes\boldsymbol{\xi}_2 + \boldsymbol{\mu}_1\otimes\boldsymbol{\xi}_1\otimes\boldsymbol{\mu}_1 + \boldsymbol{\mu}_1\otimes\boldsymbol{\mu}_1\otimes\boldsymbol{\xi}_1 - 3\boldsymbol{\xi}_0\boldsymbol{\mu}_1\otimes\boldsymbol{\mu}_1\otimes\boldsymbol{\mu}_1\\
& - \boldsymbol{\xi}_0\boldsymbol{\mu}_3 + \boldsymbol{\xi}_0\boldsymbol{\mu}_2\otimes\boldsymbol{\mu}_1 + \boldsymbol{\xi}_0\mathrm{E}(\mathbf{X}\otimes\mathbf{X}\boldsymbol{\mu}_1\mathbf{X}) + \boldsymbol{\xi}_0\boldsymbol{\mu}_1\otimes\boldsymbol{\mu}2].
\end{aligned}$$

Then we know that for

$$\begin{aligned}
\sqrt{n}(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1) &= \hat{\Sigma}^{-1/2}(\hat{\boldsymbol{\xi}}_1 - \hat{\boldsymbol{\mu}}_1\hat{\boldsymbol{\xi}}_0) - \Sigma^{-1/2}(\boldsymbol{\xi}_1 - \boldsymbol{\mu}_1\boldsymbol{\xi}_0)\\
&= \hat{\Sigma}^{-1/2}\Sigma^{1/2}\Sigma^{-1/2}(\hat{\boldsymbol{\xi}}_1 - \hat{\boldsymbol{\mu}}_1\hat{\boldsymbol{\xi}}_0) - \Sigma^{-1/2}(\boldsymbol{\xi}_1 - \boldsymbol{\mu}_1\boldsymbol{\xi}_0)\\
&= (\hat{\Sigma}^{-1/2}\Sigma^{1/2} - I + I)\Sigma^{-1/2}(\hat{\boldsymbol{\xi}}_1 - \hat{\boldsymbol{\mu}}_1\hat{\boldsymbol{\xi}}_0) - \Sigma^{-1/2}(\boldsymbol{\xi}_1 - \boldsymbol{\mu}_1\boldsymbol{\xi}_0)\\
&= (\hat{\Sigma}^{-1/2}\Sigma^{1/2} - I + I)\Sigma^{-1/2}(\hat{\boldsymbol{\xi}}_1 - \boldsymbol{\xi}_1 + \boldsymbol{\xi}_1 - \boldsymbol{\mu}_1\boldsymbol{\xi}_0 + \boldsymbol{\mu}_1\boldsymbol{\xi}_0 - \boldsymbol{\mu}_1\hat{\boldsymbol{\xi}}_0 + \boldsymbol{\mu}_1\hat{\boldsymbol{\xi}}_0 - \hat{\boldsymbol{\mu}}_1\hat{\boldsymbol{\xi}}_0)\\
&-\Sigma^{-1/2}(\boldsymbol{\xi}_1 - \boldsymbol{\mu}_1\boldsymbol{\xi}_0)\\
&= (\hat{\Sigma}^{-1/2}\Sigma^{1/2} - I + I)\Sigma^{-1/2}(\hat{\boldsymbol{\xi}}_1 - \boldsymbol{\xi}_1 + (\boldsymbol{\xi}_1 - \boldsymbol{\mu}_1\boldsymbol{\xi}_0) + \boldsymbol{\mu}_1(\boldsymbol{\xi}_0 - \hat{\boldsymbol{\xi}}_0) + (\boldsymbol{\mu}_1 - \hat{\boldsymbol{\mu}}_1)\hat{\boldsymbol{\xi}}_0)\\
&-\Sigma^{-1/2}(\boldsymbol{\xi}_1 - \boldsymbol{\mu}_1\boldsymbol{\xi}_0)\\
&= (\hat{\Sigma}^{-1/2}\Sigma^{1/2} - I)\Sigma^{-1/2}(\hat{\boldsymbol{\xi}}_1 - \boldsymbol{\xi}_1 + (\boldsymbol{\xi}_1 - \boldsymbol{\mu}_1\boldsymbol{\xi}_0) + \boldsymbol{\mu}_1(\boldsymbol{\xi}_0 - \hat{\boldsymbol{\xi}}_0) + (\boldsymbol{\mu}_1 - \hat{\boldsymbol{\mu}}_1)\hat{\boldsymbol{\xi}}_0)\\
&+\Sigma^{-1/2}(\hat{\boldsymbol{\xi}}_1 - \boldsymbol{\xi}_1 + \boldsymbol{\mu}_1(\boldsymbol{\xi}_0 - \hat{\boldsymbol{\xi}}_0) + (\boldsymbol{\mu}_1 - \hat{\boldsymbol{\mu}}_1)\hat{\boldsymbol{\xi}}_0)\\
&= \Sigma^{-1/2}(\hat{\boldsymbol{\xi}}_1 - \boldsymbol{\xi}_1 + \boldsymbol{\mu}_1(\boldsymbol{\xi}_0 - \hat{\boldsymbol{\xi}}_0) + (\boldsymbol{\mu}_1 - \hat{\boldsymbol{\mu}}_1)\hat{\boldsymbol{\xi}}_0) + O_p(\frac{1}{n}).
\end{aligned}$$

The last equality holds because we have $(\hat{\Sigma}^{-1/2}\Sigma^{1/2} - I)\Sigma^{-1/2}(\hat{\boldsymbol{\xi}}_1 - \boldsymbol{\xi}_1) = O_p(\frac{1}{n})$; $(\hat{\Sigma}^{-1/2}\Sigma^{1/2} - I)\Sigma^{-1/2}\boldsymbol{\mu}_1(\boldsymbol{\xi}_0 - \hat{\boldsymbol{\xi}}_0) = O_p(\frac{1}{n})$; $(\hat{\Sigma}^{-1/2}\Sigma^{1/2} - I)\Sigma^{-1/2}(\boldsymbol{\mu}_1 - \hat{\boldsymbol{\mu}}_1)\hat{\boldsymbol{\xi}}_0 = O_p(\frac{1}{n})$; $(\hat{\Sigma}^{-1/2}\Sigma^{1/2} - I)\Sigma^{-1/2}(\boldsymbol{\xi}_1 - \boldsymbol{\mu}_1\boldsymbol{\xi}_0) = 0$;

It's clear to see that in the last line, under our assumption, a multivariate central limit theorem would hold which gives us

$$\sqrt{n}[(\hat{\boldsymbol{\xi}}_1', \hat{\boldsymbol{\mu}}_1', \hat{\boldsymbol{\xi}}_0)' - (\boldsymbol{\xi}_1', \boldsymbol{\mu}_1', \boldsymbol{\xi}_0)] \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\Gamma}_1^*),$$

and

$$\sqrt{n}(\hat{\boldsymbol{\xi}}_1 - \boldsymbol{\xi}_1) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\Gamma}_1),$$

with $\boldsymbol{\Gamma}_1 = \text{cov}(T\mathbf{X})$.

Take $\mathbf{A}_2 = (\hat{\Sigma}^{-1/2}\Sigma^{1/2}) \otimes (\hat{\Sigma}^{-1/2}\Sigma^{1/2})$. Then for $\sqrt{n}(\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2)$ we have

$$\begin{aligned}
\sqrt{n}(\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2) =& (\hat{\Sigma}^{-1/2} \otimes \hat{\Sigma}^{-1/2})[\hat{\boldsymbol{\xi}}_2 - \hat{\boldsymbol{\mu}}_2\hat{\boldsymbol{\xi}}_0] - (\Sigma^{-1/2} \otimes \Sigma^{-1/2})[\boldsymbol{\xi}_2 - \boldsymbol{\mu}_2\boldsymbol{\xi}_0] + O_p(\frac{1}{n}) \\
=& (\mathbf{A}_2 - I + I)(\Sigma^{-1/2} \otimes \Sigma^{-1/2})[\hat{\boldsymbol{\xi}}_2 - \boldsymbol{\xi}_2 + (\boldsymbol{\xi}_2 - \boldsymbol{\mu}_2\boldsymbol{\xi}_0) + \boldsymbol{\mu}_2(\boldsymbol{\xi}_0 - \hat{\boldsymbol{\xi}}_0) \\
& + (\boldsymbol{\mu}_2 - \hat{\boldsymbol{\mu}}_2)\hat{\boldsymbol{\xi}}_0] - (\Sigma^{1/2} \otimes \Sigma^{-1/2})[\boldsymbol{\xi}_2 - \boldsymbol{\mu}_2\boldsymbol{\xi}_0] + O_p(\frac{1}{n}) \\
=& (\mathbf{A}_2 - I)(\Sigma^{-1/2} \otimes \Sigma^{-1/2})[\hat{\boldsymbol{\xi}}_2 - \boldsymbol{\xi}_2 + (\boldsymbol{\xi}_2 - \boldsymbol{\mu}_2\boldsymbol{\xi}_0) + \boldsymbol{\mu}_2(\boldsymbol{\xi}_0 - \hat{\boldsymbol{\xi}}_0) \\
& + (\boldsymbol{\mu}_2 - \hat{\boldsymbol{\mu}}_2)\hat{\boldsymbol{\xi}}_0] + (\Sigma^{-1/2} \otimes \Sigma^{-1/2})[\hat{\boldsymbol{\xi}}_2 - \boldsymbol{\xi}_2 + \boldsymbol{\mu}_2(\boldsymbol{\xi}_0 - \hat{\boldsymbol{\xi}}_0) \\
& + (\boldsymbol{\mu}_2 - \hat{\boldsymbol{\mu}}_2)\hat{\boldsymbol{\xi}}_0] + O_p(\frac{1}{n}) \\
=& (\Sigma^{-1/2} \otimes \Sigma^{-1/2})[\hat{\boldsymbol{\xi}}_2 - \boldsymbol{\xi}_2 + \boldsymbol{\mu}_2(\boldsymbol{\xi}_0 - \hat{\boldsymbol{\xi}}_0) + (\boldsymbol{\mu}_2 - \hat{\boldsymbol{\mu}}_2)\hat{\boldsymbol{\xi}}_0] + O_p(\frac{1}{n}).
\end{aligned}$$

The last equality holds because we have $(\mathbf{A}_2 - I)(\Sigma^{-1/2} \otimes \Sigma^{-1/2})[\hat{\boldsymbol{\xi}}_2 - \boldsymbol{\xi}_2] = O_p(\frac{1}{n})$; $(\mathbf{A}_2 - I)(\Sigma^{-1/2} \otimes \Sigma^{-1/2})[\boldsymbol{\xi}_2 - \boldsymbol{\mu}_2\boldsymbol{\xi}_0] = 0$; $(\mathbf{A}_2 - I)(\Sigma^{-1/2} \otimes \Sigma^{-1/2})[\boldsymbol{\mu}_2(\boldsymbol{\xi}_0 - \hat{\boldsymbol{\xi}}_0)] = O_p(\frac{1}{n})$; $(\mathbf{A}_2 - I)(\Sigma^{-1/2} \otimes \Sigma^{-1/2})[(\boldsymbol{\mu}_2 - \hat{\boldsymbol{\mu}}_2)\hat{\boldsymbol{\xi}}_0] = O_p(\frac{1}{n})$.

It's clear to see that in the last line, under our assumption, a multivariate central limit theorem would hold which gives us

$$\sqrt{n}[(\hat{\boldsymbol{\xi}}_2', \hat{\boldsymbol{\mu}}_2', \hat{\boldsymbol{\xi}}_0)' - (\boldsymbol{\xi}_2', \boldsymbol{\mu}_2', \boldsymbol{\xi}_0)] \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\Gamma}_2^*),$$

and

$$\sqrt{n}(\hat{\boldsymbol{\xi}}_2 - \boldsymbol{\xi}_2) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\Gamma}_2),$$

with $\boldsymbol{\Gamma}_2 = \text{cov}(T\mathbf{X} \otimes \mathbf{X})$.

Take $\mathbf{A}_3 = (\hat{\Sigma}^{-1/2}\Sigma^{1/2}) \otimes (-\hat{\Sigma}^{1/2}\Sigma^{1/2} \otimes \hat{\Sigma}^{-1/2}\Sigma^{1/2}))$. Then for $\sqrt{n}(\hat{\boldsymbol{\theta}}_3 - \boldsymbol{\theta}_3)$ we have

$$\begin{aligned}
\sqrt{n}(\hat{\boldsymbol{\theta}}_3 - \boldsymbol{\theta}_3) =& (\hat{\Sigma}^{-1/2} \otimes \hat{\Sigma}^{-1/2} \otimes \hat{\Sigma}^{-1/2})[\hat{\boldsymbol{\xi}}_3 - \hat{\boldsymbol{\mu}}_3\hat{\boldsymbol{\xi}}_0] - \\
& (\Sigma^{-1/2} \otimes \Sigma^{-1/2} \otimes \Sigma^{-1/2})[\boldsymbol{\xi}_3 - \boldsymbol{\mu}_3\boldsymbol{\xi}_0] + O_p(\frac{1}{n}) \\
=& (\mathbf{A}_3 - I + I)(\Sigma^{-1/2} \otimes \Sigma^{-1/2} \otimes \Sigma^{-1/2})[\hat{\boldsymbol{\xi}}_3 - \boldsymbol{\xi}_3 + (\boldsymbol{\xi}_3 - \boldsymbol{\mu}_3\boldsymbol{\xi}_0) + \boldsymbol{\mu}_3(\boldsymbol{\xi}_0 - \hat{\boldsymbol{\xi}}_0) \\
& + (\boldsymbol{\mu}_3 - \hat{\boldsymbol{\mu}}_3)\hat{\boldsymbol{\xi}}_0] - (\Sigma^{-1/2} \otimes \Sigma^{-1/2} \otimes \Sigma^{-1/2})[\boldsymbol{\xi}_3 - \boldsymbol{\mu}_3\boldsymbol{\xi}_0] + O_p(\frac{1}{n}) \\
=& (\mathbf{A}_3 - I)(\Sigma^{-1/2} \otimes \Sigma^{-1/2} \otimes \Sigma^{-1/2})[\hat{\boldsymbol{\xi}}_3 - \boldsymbol{\xi}_3 + (\boldsymbol{\xi}_3 - \boldsymbol{\mu}_3\boldsymbol{\xi}_0) + \boldsymbol{\mu}_3(\boldsymbol{\xi}_0 - \hat{\boldsymbol{\xi}}_0) \\
& + (\boldsymbol{\mu}_3 - \hat{\boldsymbol{\mu}}_3)\hat{\boldsymbol{\xi}}_0] + (\Sigma^{-1/2} \otimes \Sigma^{-1/2} \otimes \Sigma^{-1/2})[\hat{\boldsymbol{\xi}}_3 - \boldsymbol{\xi}_3 + \boldsymbol{\mu}_3(\boldsymbol{\xi}_0 - \hat{\boldsymbol{\xi}}_0) \\
& + (\boldsymbol{\mu}_3 - \hat{\boldsymbol{\mu}}_3)\hat{\boldsymbol{\xi}}_0] + O_p(\frac{1}{n}) \\
=& (\Sigma^{-1/2} \otimes \Sigma^{-1/2} \otimes \Sigma^{-1/2})[\hat{\boldsymbol{\xi}}_3 - \boldsymbol{\xi}_3 + \boldsymbol{\mu}_3(\boldsymbol{\xi}_0 - \hat{\boldsymbol{\xi}}_0) + (\boldsymbol{\mu}_3 - \hat{\boldsymbol{\mu}}_3)\hat{\boldsymbol{\xi}}_0] + O_p(\frac{1}{n}).
\end{aligned}$$

The last equality holds because we have $(\mathbf{A}_3 - I)(\Sigma^{-1/2} \otimes \Sigma^{-1/2} \otimes \Sigma^{-1/2})[\hat{\boldsymbol{\xi}}_3 - \boldsymbol{\xi}_3] = O_p(\frac{1}{n})$; $(\mathbf{A}_3 - I)(\Sigma^{-1/2} \otimes \Sigma^{-1/2} \otimes \Sigma^{-1/2})[\boldsymbol{\xi}_3 - \boldsymbol{\mu}_3\boldsymbol{\xi}_0] = 0$; $(\mathbf{A}_3 - I)(\Sigma^{-1/2} \otimes \Sigma^{-1/2} \otimes \Sigma^{-1/2})[\boldsymbol{\mu}_3(\boldsymbol{\xi}_0 - \hat{\boldsymbol{\xi}}_0)] = O_p(\frac{1}{n})$; $(\mathbf{A}_2 - I)(\Sigma^{-1/2} \otimes \Sigma^{-1/2} \otimes \Sigma^{-1/2})[(\boldsymbol{\mu}_3 - \hat{\boldsymbol{\mu}}_3)\hat{\boldsymbol{\xi}}_0] = O_p(\frac{1}{n})$.

It's clear to see that in the last line, under our assumption, a multivariate central limit theorem would hold which gives us

$$\sqrt{n}[(\hat{\boldsymbol{\xi}}_3', \hat{\boldsymbol{\mu}}_3', \hat{\boldsymbol{\xi}}_0)' - (\boldsymbol{\xi}_3', \boldsymbol{\mu}_3', \boldsymbol{\xi}_0)] \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\Gamma}_3^*),$$

and

$$\sqrt{n}(\hat{\boldsymbol{\xi}}_3 - \boldsymbol{\xi}_3) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\Gamma}_3),$$

with $\boldsymbol{\Gamma}_3 = \text{cov}(T\mathbf{X} \otimes \mathbf{X} \otimes \mathbf{X})$.

Proof of Proposition 4.2

Recall that

$$\hat{D}(\mathbf{B},\mathbf{C}) = [\mathrm{vec}(\hat{\boldsymbol{\xi}}) - \mathrm{vec}(\mathbf{BC})]'\mathbf{V}_n[\mathrm{vec}(\hat{\boldsymbol{\xi}}) - \mathrm{vec}(\mathbf{BC})]$$

and notice that $\mathbf{V}_n$ is a sequence of random matrices that converge in probability to a positive definite matrix $\mathbf{V}$. First step in this proof will be showing that replacing the sequence of random matrix by their limit will not change the limiting distribution of $n\hat{D}$. i.e. let

$$F(\mathbf{B},\mathbf{C}) = [\mathrm{vec}(\hat{\boldsymbol{\xi}}) - \mathrm{vec}(\mathbf{BC})]'\mathbf{V}_n[\mathrm{vec}(\hat{\boldsymbol{\xi}}) - \mathrm{vec}(\mathbf{BC})],$$

then we need to show that $\arg\min_{\mathbf{B},\mathbf{C}} \hat{D} \stackrel{d}{=} \arg\min_{\mathbf{B},\mathbf{C}} F$, where $\stackrel{d}{=}$ represent equally distributed. First we will show that $\min_{\mathbf{B},\mathbf{C}} n\hat{D} \stackrel{d}{=} \min_{\mathbf{B},\mathbf{C}} nF$

Since $\mathbf{V}_n \stackrel{P}{\to} \mathbf{V}$, for $\forall\epsilon > 0$, $\lim_n P[(1-\epsilon)\mathbf{V} < \mathbf{V}_n < (1+\epsilon)\mathbf{V}] = 1$ and also

$$\begin{aligned}
&P[(1-\epsilon)\mathbf{V} < \mathbf{V}_n < (1+\epsilon)\mathbf{V}] < \\
&P[\{\mathrm{vec}(\hat{\boldsymbol{\xi}}) - \mathrm{vec}(\mathbf{BC})\}'(1-\epsilon)\mathbf{V}\{\mathrm{vec}(\hat{\boldsymbol{\xi}}) - \mathrm{vec}(\mathbf{BC})\} \\
&< \{\mathrm{vec}(\hat{\boldsymbol{\xi}}) - \mathrm{vec}(\mathbf{BC})\}'\mathbf{V}_n\{\mathrm{vec}(\hat{\boldsymbol{\xi}}) - \mathrm{vec}(\mathbf{BC})\} \\
&< \{\mathrm{vec}(\hat{\boldsymbol{\xi}}) - \mathrm{vec}(\mathbf{BC})\}'(1+\epsilon)\mathbf{V}\{\mathrm{vec}(\hat{\boldsymbol{\xi}}) - \mathrm{vec}(\mathbf{BC})\}].
\end{aligned}$$

So we have

$$\begin{aligned}
1 = &\lim_n P[(1-\epsilon)\mathbf{V} < \mathbf{V}_n < (1+\epsilon)\mathbf{V}] < \\
&\liminf_n P[(1-\epsilon)\hat{D} < F < (1+\epsilon)\hat{D}] \leq 1.
\end{aligned}$$

The minimum of $(1-\epsilon)\hat{D}, F$ and $(1+\epsilon)\hat{D}$ have the same ordering, so we have

$$1 \leq \liminf P(|\frac{\min_{\mathbf{B},\mathbf{C}} \hat{D}}{\min_{\mathbf{B},\mathbf{C}} F} - 1| < \epsilon) = 1$$

That is $\min_{\mathbf{B},\mathbf{C}} \hat{D} \xrightarrow{P} \min_{\mathbf{B},\mathbf{C}} F$ and this implies $n\hat{D}$ and nF will yield the same limiting distribution by Slutsky theorem.

Furthermore, we also want to show $\arg\min_{\mathbf{B},\mathbf{C}} \hat{D} \stackrel{d}{=} \arg\min_{\mathbf{B},\mathbf{C}} F$ does not depend on the choice of $\mathbf{V}_n$ as long as the sequence converge in probability to some $\mathbf{V}$. To illustrate this point, we will give out exact asymptotic distribution.

Let $\boldsymbol{\theta} = (\text{vec}(\mathbf{B})', \text{vec}(\mathbf{C})')' \in \mathbb{R}^p$; $\boldsymbol{\theta}_0 = (\text{vec}(\mathbf{B}_0)', \text{vec}(\mathbf{C}_0)')'$ be vectorized version of the true value that minimize the generalized quadratic form on population level and $\hat{\boldsymbol{\theta}} = (\text{vec}(\hat{\mathbf{B}})', \text{vec}(\hat{\mathbf{C}})')'$ be the sample minimizer that minimize $\hat{D}$. Let $\mathbf{g}(\boldsymbol{\theta}) = \text{vec}(\mathbf{BC})$ and set $\mathbf{D}(\boldsymbol{\theta}) = \frac{\partial \mathbf{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ and $\mathbf{D}_0 = \mathbf{D}(\boldsymbol{\theta}_0)$. Define $\mathbf{G}(\boldsymbol{\theta}) = \mathbf{D}'(\hat{\boldsymbol{\theta}})\mathbf{V}_n(\hat{\boldsymbol{\xi}} - \mathbf{g}(\boldsymbol{\theta}))$ and $\Delta\mathbf{G}(\boldsymbol{\theta}) = -\mathbf{D}'(\hat{\boldsymbol{\theta}})\mathbf{V}_n\mathbf{D}'(\hat{\boldsymbol{\theta}})$ be the partial derivative of $\mathbf{G}$ with respect to $\boldsymbol{\theta}$.

Now expanding $\mathbf{G}$ around the point $\hat{\boldsymbol{\theta}}$:

$$\mathbf{G}(\boldsymbol{\theta}_0) = \mathbf{G}(\hat{\boldsymbol{\theta}}_0) + [\int_0^1 \Delta\mathbf{G}(\hat{\boldsymbol{\theta}} + t(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}))\, dt](\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}).$$

Also, because $\hat{\boldsymbol{\theta}}$ is minimizer of $\hat{D}$, by taking first derivative of $\hat{D}$ with respect to $\boldsymbol{\theta}$ we have

$$-2\mathbf{D}'(\hat{\boldsymbol{\theta}})\mathbf{V}_n(\hat{\boldsymbol{\xi}} - \mathbf{g}(\hat{\boldsymbol{\theta}})) = \mathbf{0}.$$

i.e. $\mathbf{G}(\hat{\boldsymbol{\theta}}) = \mathbf{0}$.

Therefore, we have

$$\begin{aligned}\sqrt{n}\mathbf{D}'(\hat{\boldsymbol{\theta}})\mathbf{V}_n(\hat{\boldsymbol{\xi}} - \mathbf{g}(\boldsymbol{\theta}_0)) &= \sqrt{n}\mathbf{G}(\boldsymbol{\theta}_0) \\ &= -\sqrt{n}[\int_0^1 \Delta\mathbf{G}(\hat{\boldsymbol{\theta}} + t(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}))\ dt](\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}) \\ &= -\sqrt{n}[\int_0^1 -\mathbf{D}'(\hat{\boldsymbol{\theta}})\mathbf{V}_n\mathbf{D}'(\hat{\boldsymbol{\theta}} + t(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}))\ dt](\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}) \\ &\xrightarrow{P} -\sqrt{n}\mathbf{D}_0\mathbf{V}\mathbf{D}_0'(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}).\end{aligned}$$

The third equal sign follows from the fact that $\hat{\boldsymbol{\theta}} \xrightarrow{P} \boldsymbol{\theta}_0$ and the last convergence follows from Bounded Convergence Theorem.

This implies that

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{D} N(\mathbf{0}, (\mathbf{D}_0'\mathbf{V}\mathbf{D}_0)^{-1}\mathbf{D}_0'\mathbf{V}\boldsymbol{\Gamma}\mathbf{V}\mathbf{D}_0(\mathbf{D}_0'\mathbf{V}\mathbf{D}_0)^{-1}),$$

$$\sqrt{n}(\mathrm{vec}(\hat{\mathbf{B}}\hat{\mathbf{C}} - \mathbf{B}_0\mathbf{C}_0) \xrightarrow{D} N(\mathbf{0}, \mathbf{D}_0(\mathbf{D}_0'\mathbf{V}\mathbf{D}_0)^{-1}\mathbf{D}_0'\mathbf{V}\boldsymbol{\Gamma}\mathbf{V}\mathbf{D}_0(\mathbf{D}_0'\mathbf{V}\mathbf{D}_0)^{-1}\mathbf{D}_0).$$

Above arguments shows that the asymptotic distribution of $\mathrm{vec}(\hat{\mathbf{B}}\hat{\mathbf{C}})$ is independent of the choice of $\{\mathbf{V}_n\}$ as long as $\mathbf{V}_n \xrightarrow{P} \mathbf{V}$. However, it does have some implicit assumptions such as the smoothness of $\mathbf{g}(\boldsymbol{\theta})$ and the topological property about the point θ_0. However, this is easily satisfied, as the proof itself is also indicates that the reparametrization of $\mathbf{B}, \mathbf{C}$ does not affect the results so we can use the following construction to met the requirement:

Let's consider the following reparametrization of $\hat{\mathbf{B}}$ by proper manipulation the order of column vectors: let $\hat{\mathbf{B}} = (\hat{\mathbf{B}}_1', \hat{\mathbf{B}}_2')'$, with $\hat{\mathbf{B}}_1 \in \mathbb{R}^{d\times d}$, which is non-singular and $\hat{\mathbf{B}}_2 \in \mathbb{R}^{(p-d)\times d}$, then

$$\hat{\mathbf{B}}\hat{\mathbf{C}} = \begin{pmatrix}\hat{\mathbf{B}}_1 \\ \hat{\mathbf{B}}_2\end{pmatrix}\hat{\mathbf{C}} = \begin{pmatrix}\mathbf{I}_d \\ \hat{\mathbf{B}}_2\hat{\mathbf{B}}_1^{-1}\end{pmatrix}\hat{\mathbf{B}}_1\hat{\mathbf{C}}.$$

Under this specific reparametrization, we have defined a full-rank transformation which could easily make the topological requirement satisfied given the fact that vec$(\cdot)$ operator is analytic.

The remaining part in this proposition will be using the F to establish the asymptotic properties of the quadratic forms $\hat{D}$. The following development will require using a lemma from Shapiro (1986) which was named as Shapiro's discrepancy function in Cook and Ni (2005). The lemma was stated as follow without proof:

Lemma D4. *(Shapiro 1986; Cook and Ni 2005) Suppose that $\boldsymbol{\theta}$ is a q-dimensional parameter vector that lies in an open and connected parameter space $\Theta \subseteq \mathbb{R}^q$. Let $\boldsymbol{\theta}_0$ denote the true value of $\boldsymbol{\theta}$. Define $\mathbf{g}(\boldsymbol{\theta}) = (g_1(\boldsymbol{\theta}), \ldots, g_m(\boldsymbol{\theta}))' : \Theta \to \mathbb{R}^m$, where $g_i(\boldsymbol{\theta})$ is twice continuously differentiable on Θ, $i = 1, \ldots, m$. The Jacobian matrix $\Delta = \frac{\partial \mathbf{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mid_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$ need not be of full rank, so $\mathbf{g}$ can be overparameterized. Also assume the following:*

1. *τ_n is an asymptotically normal estimate of the population value $\mathbf{g}(\boldsymbol{\theta}_0) : \sqrt{n}(\tau_n - \mathbf{g}(\boldsymbol{\theta}_0)) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\Gamma})$, where n is the sample size.*

2. *For a known inner-product matrix $\mathbf{V}$, the discrepancy function*

$$H(\tau_n, \mathbf{g}(\boldsymbol{\theta})) = (\tau_n - \mathbf{g}(\boldsymbol{\theta}))'\mathbf{V}(\tau_n - \mathbf{g}(\boldsymbol{\theta}))$$

   satisfies the following properties:

   a) $H(\mathbf{a}, \mathbf{b}) \geq 0 \forall \mathbf{a}, \mathbf{b} \in \mathbb{R}^m$.

   b) $H(\mathbf{a}, \mathbf{b}) = 0$ *iff* $\mathbf{a} = \mathbf{b}$.

   c) *H is at least twice continuously differentiable in $\mathbf{a}, \mathbf{b}$.*

   d) *There are positive constants δ and ϵ such that $H(\mathbf{a}, \mathbf{b}) \geq \epsilon$ whenever $\|\mathbf{a} - \mathbf{b}\| \geq \delta$, where $\|\cdot\|$ represents ordinary Euclidean distance.*

3. *The point* $\boldsymbol{\theta}_0$ *is regular.*

4. $\text{rank}(\Delta) = \text{rank}(\Delta'\mathbf{V}\Delta)$

Then the following holds:

1. *Letting* $\hat{H} = H(\tau, \mathbf{g}(\hat{\boldsymbol{\theta}}))$ *denote the value of the discrepancy function minimized over* Θ*, the asymptotic distribution of* $n\hat{H}$ *is the same as the distribution of the quadratic form* $\mathbf{W}'\mathbf{U}\mathbf{W}$*, where* $\mathbf{W} \sim N(\mathbf{0}, \boldsymbol{\Gamma})$*,* $\mathbf{U} = \mathbf{V} - \mathbf{V}\Delta(\Delta'\mathbf{V}\Delta)^{-}\Delta'\mathbf{V} = \mathbf{V}^{1/2}\mathbf{Q}_{\Phi}\mathbf{V}^{1/2}$*, and* $\Phi = \mathbf{V}\Delta$*.*

2. *If* $\boldsymbol{\Gamma}\mathbf{U}\boldsymbol{\Gamma}\mathbf{U}\boldsymbol{\Gamma} = \boldsymbol{\Gamma}\mathbf{U}\boldsymbol{\Gamma}$*, then* $n\hat{H} \xrightarrow{D} \chi^2_{\text{trace}(\mathbf{U}\boldsymbol{\Gamma})}$*.*

3. *The estimate* $\mathbf{g}(\hat{\boldsymbol{\theta}})$ *that minimizes the discrepancy function is a consistent estimator of* $\mathbf{g}(\boldsymbol{\theta}_0)$ *and* $\sqrt{n}(\mathbf{g}(\hat{\boldsymbol{\theta}}) - \mathbf{g}(\boldsymbol{\theta}_0))$ *has an asymptotically normal distribution with mean* $\mathbf{0}$ *and covariance matrix* $\mathbf{V}^{-1/2}\mathbf{P}_{\Phi}\mathbf{V}^{1/2}\boldsymbol{\Gamma}\mathbf{V}^{1/2}\mathbf{P}_{\Phi}\mathbf{V}^{-1/2}$*.*

4. *When* $\boldsymbol{\Gamma}$ *is nonsingular,* $g(\hat{\boldsymbol{\theta}})$ *is asymptotically efficient and* $n\hat{H} \xrightarrow{D} \chi^2_{m-\text{rank}(\Delta)}$*, if and only if* $\mathbf{V} = (\boldsymbol{\Gamma} + \Delta\mathbf{D}\Delta')^{-1}$*, where* $\mathbf{D}$ *is an arbitrary symmetric matrix.*

It's not hard to verify that $\hat{D}$ met the assumptions for this lemma, so we have finished the proof by taking $\tau_n = \text{vec}(\hat{\boldsymbol{\xi}})$.

Bibliography

Bura, E., & Cook, R. D. (2001). Estimating the structural dimension of regressions via parametric inverse regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2), 393-410.

Chaudhuri, P. (1991). Nonparametric estimates of regression quantiles and their local Bahadur representation. *The Annals of statistics*, 19(2), 760-777.

Chen, X., Zou, C., & Cook, R. D. (2010). Coordinate-independent sparse sufficient dimension reduction and variable selection. *The Annals of Statistics*, 38(6), 3696-3723.

Christou, E., & Akritas, M. G. (2016). Single index quantile regression for heteroscedastic data. *Journal of Multivariate Analysis*, 150, 169-182.

Cook, R. D. (1996). Graphics for regressions with a binary response. *Journal of the American Statistical Association*, 91(435), 983-992.

Cook, R. D. (1998). Principal Hessian directions revisited. *Journal of the American Statistical Association*, 93(441), 84-94.

Cook, R. D. (2004). Testing predictor contributions in sufficient dimension reduction. *The Annals of Statistics*, 32(3), 1062-1092.

Cook, R. D., & Li, B. (2002). Dimension reduction for conditional mean in regression. *The Annals of Statistics*, 30(2), 455-474.

Cook, R. D., & Ni, L. (2005). Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *Journal of the American Statistical Association*, 100(470), 410-428.

Cook, R. D., & Setodji, C. M. (2003). A model-free test for reduced rank in multivariate regression. *Journal of the American Statistical Association*, 98(462), 340-351.

Cook, R. D. & Weisberg, S. (1991). Sliced Inverse Regression for Dimension Reduction: Comment. *Journal of the American Statistical Association*, 86(414) 328-332.

Cook, R. D., & Zhang, X. (2014). Fused estimators of the central subspace in sufficient dimension reduction. *Journal of the American Statistical Association*, 109(506), 815-827.

Durrett, R. (2010). Probability: theory and examples. *Cambridge university press.*

Efron, B. (1991). Regression percentiles using asymmetric squared error loss. *Statistica Sinica*, 93-125.

Fan, J., & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5), 849-911.

Fukumizu, K., Bach, F. R., & Jordan, M. I. (2004). Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research*, 5(Jan), 73-99.

Fukumizu, K., Bach, F. R., & Jordan, M. I. (2009). Kernel dimension reduction in regression. *The Annals of Statistics*, 37(4), 1871-1905.

Guerre, E., & Sabbah, C. (2012). Uniform bias study and Bahadur representation for local polynomial estimators of the conditional quantile function. *Econometric Theory*, 28(1), 87-129.

Hong, S. Y. (2003). Bahadur representation and its applications for local polynomial estimates in nonparametric M-regression. *Journal of Nonparametric Statistics*, 15(2), 237-251.

Jones, M. C. (1994). Expectiles and M-quantiles are quantiles. *Statistics & Probability Letters*, 20(2), 149-153.

Koenker, R., & Bassett Jr, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, 33-50.

Kong, E., & Xia, Y. (2012). A single-index quantile regression model and its estimation. *Econometric Theory*, 28(4), 730-768.

Kong, E., & Xia, Y. (2014). An adaptive composite quantile approach to dimension reduction. *The Annals of Statistics*, 42(4), 1657-1688.

Li, B., Artemiou, A., & Li, L. (2011). Principal support vector machines for linear and nonlinear sufficient dimension reduction. *The Annals of Statistics*, 39(6), 3182-3210.

Li, B., & Wang, S. (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association*, 102(479), 997-1008.

Li, B., Wen, S., & Zhu, L. (2008). On a projective resampling method for dimension reduction with multivariate responses. *Journal of the American Statistical Association*, 103(483), 1177-1186.

Li, B., Zha, H., & Chiaromonte, F. (2005). Contour regression: a general approach to dimension reduction. *The Annals of Statistics*, 33(4), 1580-1616.

Li, K. C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414), 316-327.

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.

Li, K. C., & Duan, N. (1989). Regression analysis under link violation. *The Annals of Statistics*, 17(3), 1009-1052.

Li, L. (2007). Sparse sufficient dimension reduction. *Biometrika*, 94(3), 603-613.

Li, R., Zhong, W., & Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association*, 107(499), 1129-1139.

Luo, W., Li, B., & Yin, X. (2014). On efficient dimension reduction with respect to a statistical functional of interest. *The Annals of Statistics*, 42(1), 382-412.

Luo, W., & Li, B. (2016). Combining eigenvalues and variation of eigenvectors for order determination. *Biometrika*, 103(4), 875-887.

Ma, Y., & Zhu, L. (2012). A semiparametric approach to dimension reduction. *Journal of the American Statistical Association*, 107(497), 168-179.

Ma, Y., & Zhu, L. (2013a). Efficient estimation in sufficient dimension reduction. *The Annals of Statistics*, 41(1), 250.

Ma, Y., & Zhu, L. (2013b). Efficiency loss and the linearity condition in dimension reduction. *Biometrika*, 100(2), 371-383.

Newey, W. K., & Powell, J. L. (1987). Asymmetric least squares estimation and testing. *Econometrica: Journal of the Econometric Society*, 819-847.

Pollard, D. (1991). Asymptotics for least absolute deviation regression estimators. *Econometric Theory*, 7(2), 186-199.

Qian, W., Ding, S., & Cook, R. D. (2019). Sparse minimum discrepancy approach to sufficient dimension reduction with simultaneous variable selection in ultrahigh dimension. *Journal of the American Statistical Association*, 114(527), 1277-1290.

Shapiro, A. (1986). Asymptotic theory of overparameterized structural models. *Journal of the American Statistical Association*, 81(393), 142-149.

Sheng, W., & Yin, X. (2016). Sufficient dimension reduction via distance covariance. *Journal of Computational and Graphical Statistics*, 25(1), 91-104.

Székely, G. J., Rizzo, M. L., & Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *The annals of statistics*, 35(6), 2769-2794.

Takeuchi, I., Le, Q. V., Sears, T. D., & Smola, A. J. (2006). Nonparametric quantile estimation. *Journal of Machine Learning Research*, 7(Jul), 1231-1264.

Wang, H., & Xia, Y. (2008). Sliced regression for dimension reduction. *Journal of the American Statistical Association*, 103(482), 811-821.

Wang, Q., & Yin, X. (2008). A nonlinear multi-dimensional variable selection method for high dimensional data: Sparse MAVE. *Computational Statistics & Data Analysis*, 52(9), 4512-4520.

Weng, J., & Yin, X. (2018). Fourier transform approach for inverse dimension reduction method. *Journal of Nonparametric Statistics*, 30(4), 1049-1071.

Xia, Y. (2007). A constructive approach to the estimation of dimension reduction directions. *The Annals of Statistics*, 35(6), 2654-2690.

Xia, Y., Tong, H., Li, W. K., & Zhu, L. X. (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3), 363-410.

Yang, B., Yin, X. & Zhang, N. (2019). Sufficient Variable Selection using Independence Measures for continuous Response. *Journal of Multivariate Analysis*, 173, 480-493.

Yang, Y., & Zou, H. (2015). Nonparametric multiple expectile regression via ER-Boost. *Journal of Statistical Computation and Simulation*, 85(7), 1442-1458.

Ye, Z., & Weiss, R. E. (2003). Using the bootstrap to select one of a new class of dimension reduction methods. *Journal of the American Statistical Association*, 98(464), 968-979.

Ye, Z., & Weiss, R. E. (2003). Using the bootstrap to select one of a new class of dimension reduction methods. *Journal of the American Statistical Association*, 98(464), 968-979.

Yin, X., & Cook, R. D. (2002). Dimension reduction for the conditional kth moment in regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(2), 159-175.

Yin, X., & Cook, R. D. (2003). Estimating central subspaces via inverse third moments. *Biometrika*, 90(1), 113-125.

Yin, X., & Cook, R. D. (2004). Dimension reduction via marginal fourth moments in regression. *Journal of Computational and Graphical Statistics*, 13(3), 554-570.

Yin, X., Li, B., & Cook, R. D. (2008). Successive direction extraction for estimating the central subspace in a multiple-index regression. *Journal of Multivariate Analysis*, 99(8), 1733-1757.

Yin, X., & Li, B. (2011). Sufficient dimension reduction based on an ensemble of minimum average variance estimators. *The Annals of Statistics*, 39(6), 3392-3416.

Yin, X., & Hilafu, H. (2015). Sequential sufficient dimension reduction for large p, small n problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(4), 879-892.

Zeng, P. (2008). Determining the dimension of the central subspace and central mean subspace. *Biometrika*, 95(2), 469-479.

Zhu, L., Miao, B., & Peng, H. (2006). On sliced inverse regression with high-dimensional covariates. *Journal of the American Statistical Association*, 101(474), 630-643.

Zhu, L. P., & Zhu, L. X. (2009). Dimension reduction for conditional variance in regressions. *Statistica Sinica*, 19(2), 869.

Zhu, L. P., Zhu, L. X., & Wen, S. Q. (2010). On dimension reduction in regressions with multivariate responses. *Statistica Sinica*, 1291-1307.

Zhu, Y., & 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.

Zou, H., & Yuan, M. (2008). Composite quantile regression and the oracle model selection theory. *The Annals of Statistics*, 36(3), 1108-1126.

Vita

Education

- University of Kentucky, Ph.D. in Statistics, expected May 2020, Lexington, KY
- Sichuan University, B.S. in Mathematical Statistics, June 2015, Chengdu, China