



University of Kentucky
UKnowledge

Theses and Dissertations--Statistics

Statistics


2017

INFORMATIONAL INDEX AND ITS APPLICATIONS IN HIGH DIMENSIONAL DATA

Qingcong Yuan

University of Kentucky, qingcong.yuan@uky.edu

Author ORCID Identifier:

 <https://orcid.org/0000-0002-2835-3276>

Digital Object Identifier: <https://doi.org/10.13023/ETD.2017.255>

[Right click to open a feedback form in a new tab to let us know how this document benefits you.](#)

Recommended Citation

Yuan, Qingcong, "INFORMATIONAL INDEX AND ITS APPLICATIONS IN HIGH DIMENSIONAL DATA" (2017). *Theses and Dissertations--Statistics*. 28.
https://uknowledge.uky.edu/statistics_etds/28

This Doctoral Dissertation is brought to you for free and open access by the Statistics at UKnowledge. It has been accepted for inclusion in Theses and Dissertations--Statistics by an authorized administrator of UKnowledge. For more information, please contact UKnowledge@lsv.uky.edu.

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.

Qingcong Yuan, Student

Dr. Xiangrong Yin, Major Professor

Dr. Constance Wood, Director of Graduate Studies

INFORMATIONAL INDEX AND ITS APPLICATIONS IN HIGH
DIMENSIONAL DATA

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

Qingcong Yuan

Lexington, Kentucky

Director: Dr. Xiangrong Yin, Professor of Statistics

Lexington, Kentucky

2017

Copyright© Qingcong Yuan 2017

ABSTRACT OF DISSERTATION

INFORMATIONAL INDEX AND ITS APPLICATIONS IN HIGH DIMENSIONAL DATA

We introduce a new class of measures for testing independence between two random vectors, which uses expected difference of conditional and marginal characteristic functions. By choosing a particular weight function in the class, we propose a new index for measuring independence and study its property. Two empirical versions are developed, their properties, asymptotics, connection with existing measures and applications are discussed. Implementation and Monte Carlo results are also presented.

We propose a two-stage sufficient variable selections method based on the new index to deal with large p small n data. The method does not require model specification and especially focuses on categorical response. Our approach always improves other typical screening approaches which only use marginal relation. Numerical studies are provided to demonstrate the advantages of the method.

We introduce a novel approach to sufficient dimension reduction problems using the new measure. The proposed method requires very mild conditions on the predictors, estimates the central subspace effectively and is especially useful when response is categorical. It keeps the model-free advantage without estimating link function. Under regularity conditions, root- n consistency and asymptotic normality are established. The proposed method is very competitive and robust comparing to existing dimension reduction methods through simulations results.

KEYWORDS: Categorical variable, Distance, Independence, Sufficient variable selection, Sufficient dimension reduction

Author's signature: Qingcong Yuan

Date: July 3, 2017

INFORMATIONAL INDEX AND ITS APPLICATIONS IN HIGH
DIMENSIONAL DATA

By
Qingcong Yuan

Director of Dissertation: Dr. Xiangrong Yin

Director of Graduate Studies: Dr. Constance Wood

Date: July 3, 2017

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my advisor, Dr. Xiangrong Yin. I have been very fortunate to be one of his students. His passion, diligence and enthusiasm for research really motivate me and would always encourage me for my future career. His persistent guidance, support and encouragement helped me overcome many difficulties throughout my learning process.

I greatly appreciate the help of all my dissertation committee members: Dr. Arnold Stromberg, Dr. William Griffith, Dr. Solomon Harrar and Dr. Philip Westgate and the outside examiner Dr. David Fardo. Each of them have provided lots of insights, help and support for my research.

I would like to thank Dr. Heather Bush and Dr. Kristen McQuerry, the precious experience at the Applied Statistics Lab will have a significant impact on my career.

I am thankful to Department of Statistics for the consistent support during my years at UKY, and to all the faculty, staff, peers and friends, who are always ready to help and are pleasant to work with.

Finally, special thanks to my parents Gongren Yuan and Yalan Wang and my husband Zhiheng Xie, for their unconditional love and endless support. Thanks to Xuepu Liu, my mentor since childhood.

TABLE OF CONTENTS

Acknowledgments	iii
List of Tables	vi
List of Figures	vii
Chapter 1 Introduction	1
1.1 Introduction	1
1.2 Overview of the Dissertation	4
Chapter 2 A New Class of Measure for Testing Independence	5
2.1 Introduction	5
2.2 The New Class of Measures	6
2.3 The New Index and Its Properties	9
2.4 Estimation Approaches	11
2.5 Testing Procedure	17
2.6 Simulation Studies	17
2.7 Discussion	24
Chapter 3 Sufficient Variable Selection in High Dimensional Data	26
3.1 Introduction	26
3.2 Methodology	29
3.3 Theoretical Properties	34
3.4 Numerical Studies	37
3.5 Discussion	45
Chapter 4 Sufficient Dimension Reduction in Big Data	46
4.1 Introduction	46
4.2 Methodology	47

4.3	Theoretical Properties	51
4.4	Simulation Studies	53
4.5	Discussion	54
	Appendix	55
	Supplementary Materials for Chapter 2	55
	Supplementary Materials for Chapter 3	79
	Supplementary Materials for Chapter 4	82
	Bibliography	111
	Vita	121

LIST OF TABLES

2.1	Test results using different methods	18
2.2	Empirical type-I error rates for 10,000 tests at nominal significance level 0.1, using B replicates	19
2.3	P-values using different group indicators	20
3.1	Proportions comparison of \mathcal{P}_s and \mathcal{P}_a in example 3.4.1	39
3.2	Proportions comparison of \mathcal{P}_s and \mathcal{P}_a in example 3.4.2	40
3.3	Proportions comparison of \mathcal{P}_s and \mathcal{P}_a in example 3.4.3	42
3.4	Proportions comparison of \mathcal{P}_s and \mathcal{P}_a in example 3.4.4	43
4.1	Comparison of dimension reduction accuracy using dCov and $R_c(\text{slice})$	53
4.2	Comparison of dimension reduction accuracy using dCov and $R_c(\text{slice})$	54
S.4.5.1	Empirical type-I error rates for 10,000 tests at nominal significance level 0.1, using B replicates for models (e) - (g)	77
S.4.5.2	Empirical type-I error rates for 10,000 tests at nominal significance level 0.05, using B replicates for models (a) - (d)	77
S.4.5.3	Empirical type-I error rates for 10,000 tests at nominal significance level 0.05, using B replicates for models (e) - (g)	77

LIST OF FIGURES

2.1	Empirical power comparisons at 0.1 level with different sample size n .	20
2.2	Empirical power for testing independence of \mathbf{X} and Y using five methods, $n = 30$ per group, dimension $p = 10$ and non-centrality parameter δ varies, where group indicator is (a) 1, 2, 3, 4; (b) 1, 8, 0.5, 1.2, except for the purple line, Y is transformed to dummy variables.	22
2.3	Empirical power for testing independence of \mathbf{X} and Y using five methods, $n = 30$ per group, dimension p varies and non-centrality parameter $\delta = 0.2$ where group indicator is (a) 1, 2, 3, 4; (b) 1, 8, 0.5, 1.2.	22
2.4	Empirical power with the change of sample size n	23
3.1	Boxplot by plotting the grouping on the first direction, where 0 is ALL group and 1 is the AML group.	45
S.4.5.1	Empirical power with the change of sample size n for other different parameter combinations.	78

Chapter 1 Introduction

1.1 Introduction

With the fast growing ability of doing computation and the decreasing cost to collect data, nowadays, more and more data with high volume and complexity appear in various fields. For example in microarray gene expression data, there may be thousands of predictor variables. Similar or more complex data appears in financial or network area as well. Traditional methods could not be used directly to deal with data that are of high volume and dimensionality. Facing this challenge, many new methods are developed in Statistics to discover the hidden relationship among data. The way we build models, do estimation and predictions have also changed. In this dissertation, we propose new measures to do independence test and use such measures in two applications of sufficient variable selection and sufficient dimension reduction for high dimensional data.

The Importance to Measure and Test Independence

Measuring and testing independence between variables is important in statistics. Classical Pearson product-moment correlation and covariance measure linear dependence between two random variables. In multivariate normal case, a diagonal covariance matrix implies independence, but not in general case. Likelihood-based methods such as Wilks' Lambda (Wilks, 1935) or Puri and Sen (1993) are not applicable if dimension exceeds sample size, or distributional assumptions do not hold. Multivariate nonparametric approaches are discussed by Taskinen et al. (2005). Rich literature exists on measuring independence. For instance, Blomqvist (1950), Blum et al. (1961) or other methods, see Hollander and Wolfe (1999) and Anderson (2003). A novel distance covariance (dCov, Székely et al., 2007), for testing independence between two random vectors of arbitrary dimensions is very useful, as it is nonparametric but free of tuning parameters. The work of dCov has opened new research such as in Shao

and Zhang (2014), and has been used widely in other areas as well, for instance, in variable selection (Li et al., 2012b) and dimension reduction (Sheng and Yin, 2013, 2016). Huo and Székely (2016) developed a fast algorithm for dCov. Heller et al. (2013) proposed a new method of multivariate test of association effectively deals with continuous and discrete random vectors but may have trouble to deal with nominal random vectors due to its ranking.

Most of the measures for independence treat the two random vectors symmetrically such as aforementioned methods or other informational indexes, say, Kullback-Leibler distance (Kullback, 1959) or more general classes of divergences (Vajda, 1989). These measures involve the ratio of joint density to the product of marginal densities. Although symmetry is important and flexible, especially in the use of correlation analysis, conditional or asymmetry may have wider usage and importance such as in regression analysis where we treat one variable as response conditioning on the other predictors, or vice versa in classification and discriminant analysis. Symmetric measures may be linked to asymmetric measures, for instance, in simple regression, correlation coefficient as a symmetric measure is proportional to the fitted regression coefficient as an asymmetric measure, and they do have different interpretations. Some symmetric measures can be regarded as conditional measures flexibly as in Kullback-Leibler distance or other informational divergences, but not always. For instance, in dCov, both sets are treated equally but they cannot simply be treated as one set conditional on the other one.

Sufficient Variable Selection

Variable screening and variable selection are very popular in modern data analysis. The idea of variable selection is to select a small group of predictors that are related with the response, so that a subset consists of important predictors could be detected. By deleting the irrelevant variables, the accuracy of model fitting and prediction would be greatly improved. Variable screening and selection techniques are particular useful for high dimensional data, where the number of predictors p is much larger than the number of observations n . Those kind of data are very common in daily life, for

example the micro array, image, network, financial data and so on.

The current variable screening and selection procedure could not detect the active predictors which are marginally independent of the response, therefore is not sufficient. However in practice, many predictor variables are correlated, there are many variables that are marginally dependent of the response but are indeed active predictors. And methods based on model assumption may be biased if the assumed model is not reliable, though iterative methods and nonparametric methods could partly solve the problem. Another issue occurs when the response variable is a categorical variable, especially when the categories do not have order relationship, or the response has multiple dimensions. Finding a method with minimum assumptions that could do sufficient variable selection, especially for categorical response is an interesting topic.

Sufficient Dimension Reduction

With the increase of dimensionality, the volume of the space increases so fast that the available data become sparse (Bellman, 1961). The sparsity is a problem to any statistical methods since not enough data is available to do model fitting or make any inference. Therefore, in terms of the situations discussed above, many classical models derived from oversimplified assumptions and nonparametric methods are no longer reliable. High dimensional data would lead to high computational cost to do estimation and inference and it would cause the problem of overfitting.

Sufficient dimension reduction means to find a linear transformation of the predictor matrix, so that if given that transformation, the response and the predictor is independent (Li, 1991; Cook, 1994, 1996). Various ways have been proposed to estimate the dimension reduction subspace (Cook, 2007; Yin, 2010; Ma and Zhu, 2013b). Therefore, dimension reduction that reduces the data dimension but retains (sufficient) important information can play a critical role in high-dimensional data analysis. With dimension reduction as a pre-process, often the number of reduced dimensions is small. Hence, parametric and nonparametric modeling methods can then be readily applied to the reduced data. Our proposed method works especially

well when the response is categorical.

1.2 Overview of the Dissertation

There are three main projects involved in this dissertation. In Chapter 2, we develop a novel class of informational measures to reflect the dependency between two random vectors, especially to deal with categorical data. Simulation studies show that the measure has similar power as the distance covariance measure, which is a very general method currently available for reflecting the dependency of two random vectors. And the newly proposed measure performs best when one of the vectors is a categorical vector. The properties of the measure, asymptotic results and the connection with existing measures are also discussed. In Chapter 3, we propose a two-stage sufficient variable selections algorithm. A nice property is that any independence measure can be adapted to our proposed procedure, thus the procedure does not require particular model specification. This model-free approach makes our method robust against model mis-specification, which is a very appealing property in practice. In addition, our approach always improves over typical screening approach which only uses marginal relation. Sure screening property of the new measure and the two-stage sufficient variable selection algorithm is proved. Simulation examples show that it has superior performance than the Kolmogorov filter (Mai and Zou, 2013), fused Kolmogorov filter (Mai and Zou, 2015) or MV-SIS method (Cui et al., 2015) when the data has two or more classes. The project in chapter 4 introduces a novel approach for the sufficient dimension reduction problem based on the measure in chapter 2. The approach requires very mild conditions on the predictors, and works especially well for categorical response. Under regularity conditions, root- n consistency and asymptotic normality properties are established. These theoretical and methodological developments involve multivariate data and computational techniques, which have wide applications in biostatistics, bioinformatics, business and economics, etc., where high dimensional data sets are often encountered.

Copyright© Qingcong Yuan, 2017.

Chapter 2 A New Class of Measure for Testing Independence

2.1 Introduction

In this chapter, our goal is to establish a new class of measures to test independence between two random vectors. We define it as a conditional class based on characteristic functions, treating one of them as a response, much similar to the idea in classification and discriminant analysis or as in inverse regression. Typical classification and discriminant methods or inverse regression methods only measure the relations in the inverse mean function (or moments), or dependence that involve densities, ours is going to measure the dependence between the two sets of variables without involving densities. The novel class defines a general collection of new measures by choosing different weight functions in the definition, as we will see later in the chapter that the weight function in the class determines the actual measure. For the purpose of illustration, however, in this chapter we use a particular weight function similar to what was used by Székely et al. (2007).

With such a chosen weight function, if a slicing method is chosen, our index is a variant of DISCO (Rizzo and Székely, 2010) method. Such an index has a simple/easy population version and it only needs to calculate Euclidean distance, while keeping the advantage of nonparametric. However, the test defined in DISCO method is only for categorical variable Y and is a type of generalized ANOVA from Two-sample to K-sample extension but using untypical formulation (differences among groups). Ours is defined for both continuous and categorical Y , and in the categorical case of Y , we use typical formulation which is more common and unique (difference between group and overall). Slicing is only one particular approach that we want to show the link to existing approaches. There are many other estimations that can be used, and we provide a smoothing approach (kernel estimation) to demonstrate the advantage.

Our general definition (by choosing a special weight function) is more concise and comparable/parallel with dCov. This index together with dCov forms a class that

is analogous to divergence family (such as Kullback-Leibler (KL)-distance). Such a class (with respect to weight function choices), together with dCov and Hilbert-Schmidt Independence Criterion (HSIC) form a more general class, filling a gap by using characteristic functions to that of densities (of divergences). That is, using the discrepancy between the conditional characteristic function and marginal characteristic function, our class fills the gap for distance-based criterion defined by Sejdinovic et al. (2013) where only discrepancy between joint characteristic function and product of marginal characteristic functions is measured, so that the distance-based class together with our proposed class is comparable with other divergence families, where both joint and conditional discrepancies are measured.

Throughout this chapter $\mathbf{X} \in \mathbb{R}^p$ and $\mathbf{Y} \in \mathbb{R}^q$ are random vectors, where p and q are positive integers. If $p = 1$, we use $\mathbf{X} = X$; if $q = 1$, we use $\mathbf{Y} = Y$. The characteristic functions of \mathbf{X} , $\mathbf{X}|\mathbf{Y}$ and (\mathbf{X}, \mathbf{Y}) are denoted by $f_{\mathbf{X}}$, $f_{\mathbf{X}|\mathbf{Y}}$ and $f_{\mathbf{X},\mathbf{Y}}$, respectively. For complex-valued function $f(\cdot)$, we denote \bar{f} as the complex conjugate of f . Let $|f|^2 = f\bar{f}$, and the Euclidean norm of $\mathbf{X} \in \mathbb{R}^p$ be $|\mathbf{X}|_p$.

The rest of the chapter is organized as follows. We propose the new class in Section 2.2. By choosing a particular weight function, we study the resulting index and its properties in Section 2.3, and obtain special formulas for certain distributions in Section 2.3. An empirical version by slicing on \mathbf{Y} is proposed in Section 2.4, including the establishment of its properties. A smoothing estimation approach using kernel approach is proposed in section 2.4. A permutation test is outlined in section 2.5. Simulations to illustrate its usefulness are presented in Section 4.5. Some concluding remarks are made in Section 2.7. All derivations and proofs are arranged in the appendix.

2.2 The New Class of Measures

The hypothesis test of independence between \mathbf{X} and \mathbf{Y} is as follows:

$$H_0 : f_{\mathbf{X}|\mathbf{Y}} = f_{\mathbf{X}} \text{ vs. } H_1 : f_{\mathbf{X}|\mathbf{Y}} \neq f_{\mathbf{X}}.$$

This is because if \mathbf{X} is independent of \mathbf{Y} , then $f_{\mathbf{X}|\mathbf{Y}} = f_{\mathbf{X}}$; and if $f_{\mathbf{X}|\mathbf{Y}} = f_{\mathbf{X}}$, then $e^{is^T\mathbf{Y}}f_{\mathbf{X}|\mathbf{Y}} = e^{is^T\mathbf{Y}}f_{\mathbf{X}}$ for $s \in \mathbb{R}^q$, by taking expectation over \mathbf{Y} , we obtain $f_{\mathbf{X},\mathbf{Y}} = f_{\mathbf{X}}f_{\mathbf{Y}}$. Suppose that $w(t)$, where $t \in \mathbb{R}^p$, is a nonnegative weight function. We assume that such a weight function ensures the existence of integrals.

Definition 2.2.1. *The nonnegative measure of conditional difference for the characteristic function of $\mathbf{X}|\mathbf{Y}$ is denoted by $\mathcal{C}_{w,\mathbf{Y}}(\mathbf{X}|\mathbf{Y})$, whose squared value is*

$$\mathcal{C}_{w,\mathbf{Y}}^2(\mathbf{X}|\mathbf{Y}) = \|f_{\mathbf{X}|\mathbf{Y}}(t) - f_{\mathbf{X}}(t)\|^2 = \int_{\mathbb{R}^p} |f_{\mathbf{X}|\mathbf{Y}}(t) - f_{\mathbf{X}}(t)|^2 w(t) dt. \quad (2.1)$$

Note that $\mathcal{C}_{w,\mathbf{Y}}^2(\mathbf{X}|\mathbf{Y}) \geq 0$. The term $\mathcal{C}_{w,\mathbf{Y}}^2(\mathbf{X}|\mathbf{Y})$ is a \mathbf{Y} -measurable random variable which depends on w . That is, the subscript w in $\mathcal{C}_{w,\mathbf{Y}}^2(\mathbf{X}|\mathbf{Y})$ indicates that each w may lead to a different index. The expected conditional difference is defined as next:

Definition 2.2.2. *The expectation of the conditional difference (ECD) for the characteristic function of $\mathbf{X}|\mathbf{Y}$ is denoted by $\mathcal{C}_w(\mathbf{X}|\mathbf{Y})$, whose squared value is*

$$\mathcal{C}_w^2(\mathbf{X}|\mathbf{Y}) = \mathbb{E}_{\mathbf{Y}}[\mathcal{C}_{w,\mathbf{Y}}^2(\mathbf{X}|\mathbf{Y})] = \mathbb{E}_{\mathbf{Y}}\left[\int_{\mathbb{R}^p} |f_{\mathbf{X}|\mathbf{Y}}(t) - f_{\mathbf{X}}(t)|^2 w(t) dt\right]. \quad (2.2)$$

Note again that $\mathcal{C}_w^2(\mathbf{X}|\mathbf{Y}) \geq 0$. Although $\mathcal{C}_w^2(\mathbf{X}|\mathbf{Y})$ depends on the choice of w , we omit the subscript w , and write $\mathcal{C}_w^2(\mathbf{X}|\mathbf{Y})$ as $\mathcal{C}^2(\mathbf{X}|\mathbf{Y})$ for simplicity without ambiguity. The next lemma whose proof is in the appendix indicates that $\mathcal{C}^2(\mathbf{X}|\mathbf{Y}) = 0$ is equivalent to the independence of \mathbf{X} and \mathbf{Y} . Thus, $\mathcal{C}^2(\mathbf{X}|\mathbf{Y})$ is a measure of independence.

Lemma 2.2.1. $\mathcal{C}^2(\mathbf{X}|\mathbf{Y}) = 0 \Leftrightarrow \mathcal{C}_{w,\mathbf{Y}}^2(\mathbf{X}|\mathbf{Y}) = 0$ almost surely for $\mathbf{Y} \Leftrightarrow f_{\mathbf{X}|\mathbf{Y}}(t) = f_{\mathbf{X}}(t)$ almost surely for $\mathbf{Y} \in \mathbb{R}^q$ and $t \in \mathbb{R}^p$.

A direct application of (2.2) indicates that

$$\mathcal{C}^2(\mathbf{X}|\mathbf{X}) = \mathbb{E}_{\mathbf{X}}[\mathcal{C}_{w,\mathbf{X}}^2(\mathbf{X}|\mathbf{X})] = \mathbb{E}_{\mathbf{X}}\left[\int_{\mathbb{R}^p} |e^{it\mathbf{X}} - f_{\mathbf{X}}(t)|^2 w(t) dt\right]. \quad (2.3)$$

And thus, a statistic that is similar to correlation type can be defined as

$$R_c = R_c(\mathbf{X}|\mathbf{Y}) = \frac{\mathcal{C}(\mathbf{X}|\mathbf{Y})}{\mathcal{C}(\mathbf{X}|\mathbf{X})}. \quad (2.4)$$

The result below indicates properties of $\mathcal{C}(\mathbf{X}|\mathbf{X})$, $\mathcal{C}(\mathbf{X}|\mathbf{Y})$ and R_c .

Theorem 2.2.2. *The following properties hold:*

1. $\mathcal{C}(\mathbf{X}|\mathbf{X}) = 0$ iff $\mathbf{X} = \mathbf{E}(\mathbf{X})$, almost surely.
2. $\mathcal{C}(\mathbf{W}_1 + \mathbf{W}_2 | \mathbf{V}_1 + \mathbf{V}_2) \leq \mathcal{C}(\mathbf{W}_1 | \mathbf{V}_1) + \mathcal{C}(\mathbf{W}_2 | \mathbf{V}_2)$ for independent random vectors $(\mathbf{W}_1, \mathbf{V}_1)$ and $(\mathbf{W}_2, \mathbf{V}_2)$. Equality holds if and only if \mathbf{W}_1 and \mathbf{V}_1 are both constant, or \mathbf{W}_2 and \mathbf{V}_2 are both constant, or $\mathbf{W}_1, \mathbf{V}_1, \mathbf{W}_2, \mathbf{V}_2$ are mutually independent.
3. $\mathcal{C}(\mathbf{X} + \mathbf{Y} | \mathbf{X} + \mathbf{Y}) \leq \mathcal{C}(\mathbf{X} | \mathbf{X}) + \mathcal{C}(\mathbf{Y} | \mathbf{Y})$ for independent random vectors \mathbf{X} and \mathbf{Y} . Equality holds if and only if at least one of the random vectors \mathbf{X} and \mathbf{Y} is constant.
4. $0 \leq \mathcal{C}(\mathbf{X} | \mathbf{Y}) \leq \mathcal{C}(\mathbf{X} | \mathbf{X})$, and $0 \leq R_c \leq 1$.

Most of the independence measures in literature are symmetric, but ours is asymmetric due to its conditional set up. Certainly, if needed, we can modify it to a symmetric version: $\mathcal{C}_s^2(\mathbf{X}, \mathbf{Y}) = \mathcal{C}^2(\mathbf{X} | \mathbf{Y}) + \mathcal{C}^2(\mathbf{Y} | \mathbf{X})$. Note that the combination of two measures of discrepancies: $\mathcal{C}^2(\mathbf{X} | \mathbf{Y})$, and the discrepancy between the joint characteristic function and the product of two marginal characteristic functions (Sejdinovic et al., 2013) makes a larger class which is comparable with divergence family such as ϕ -divergence (Vajda, 1989) where the discrepancy via joint density over the product of marginal densities is used. Furthermore, in our class, different weight functions can result in different indexes for testing independence. For instance, weight functions used by Sejdinovic et al. (2013) result in Hilbert-Schmidt Information Criterion (HSIC) may be used here. Hence, the choice of weight function is important as the resulting indexes may be very different, and may become a simple one or a complicated one. In this chapter, we consider a particular weight function that is similar to that

was used by Székely et al. (2007). Such a weight function results in a very simple formula of the index.

2.3 The New Index and Its Properties

Let $\tilde{C}(p, \alpha) = \frac{2\pi^{p/2}\Gamma(1-\alpha/2)}{\alpha 2^\alpha \Gamma((p+\alpha)/2)}$ for $0 < \alpha < 2$. In the case, $\alpha = 1$, define $\tilde{c}_p = \tilde{C}(p, 1) = \frac{\pi^{(1+p)/2}}{\Gamma((1+p)/2)}$. Suppose that $t \in \mathbb{R}^p$, let the weight function $w(t) = (\tilde{c}_p |t|_p^{1+p})^{-1}$, which is a positive weight function and is very similar to that was in Székely et al. (2007) and Székely and Rizzo (2009). Hereafter, we use this particular weight function.

Let $(\mathbf{X}', \mathbf{Y}')$ be an iid copy of (\mathbf{X}, \mathbf{Y}) , $\mathbf{X}_{\mathbf{Y}}$ denotes a random variable distributed as $\mathbf{X}|\mathbf{Y}$ (Cook, 2007), $\mathbf{X}'_{\mathbf{Y}'}$ denotes a random variable distributed as $\mathbf{X}'|\mathbf{Y}'$ and $\mathbf{X}'_{\mathbf{Y}}$ denotes a random variable distributed as $\mathbf{X}'|\mathbf{Y}'$ with $\mathbf{Y}' = \mathbf{Y}$. Throughout the manuscript, we assume $E|\mathbf{X}| < \infty$, $E|\mathbf{X}_{\mathbf{Y}}| < \infty$ and $E|\mathbf{X}'_{\mathbf{Y}'}| < \infty$. And these assumptions can guarantee the finiteness of $\mathcal{C}^2(\mathbf{X}|\mathbf{Y})$. Based on these assumptions we can obtain a simpler but equivalent formula of (2.2) and a special case as follows, again the proofs are in the appendix.

Theorem 2.3.1. *An equivalent form of (2.2) can be expressed as follows:*

$$\mathcal{C}^2(\mathbf{X}|\mathbf{Y}) = E|\mathbf{X} - \mathbf{X}'_{\mathbf{Y}}| - E|\mathbf{X}_{\mathbf{Y}} - \mathbf{X}'_{\mathbf{Y}}| = E|\mathbf{X} - \mathbf{X}'| - E|\mathbf{X}_{\mathbf{Y}} - \mathbf{X}'_{\mathbf{Y}}|, \quad (2.5)$$

where the expectation is over all random vectors. For instance, the last expectation is first taking the conditional expectation given \mathbf{Y} , then over \mathbf{Y} .

Note that strictly speaking, $E|\mathbf{X}_{\mathbf{Y}} - \mathbf{X}'_{\mathbf{Y}}| = E_{\mathbf{y}}E[|\mathbf{X} - \mathbf{X}'||\mathbf{Y} = \mathbf{y}, \mathbf{Y}' = \mathbf{y}]$. Also, formula (2.2) is more general than formula (2.5). For instance, conditional Cauchy distribution in section 2.3 can be calculated via (2.2) but not (2.5).

Theorem 2.3.2. 1. $\mathcal{C}^2(\mathbf{X}|\mathbf{X}) = E[\mathcal{C}_{w, \mathbf{X}}^2(\mathbf{X}|\mathbf{X})] = E|\mathbf{X} - \mathbf{X}'|$.

2. $\mathcal{C}^2(a+b\mathbf{B}\mathbf{X}|\mathbf{Y}) = |b|\mathcal{C}^2(\mathbf{X}|\mathbf{Y})$ for all constant vector a , scalar b and orthonormal matrix \mathbf{B} .

3. $R_c = 1$ iff \mathbf{X} is a function of \mathbf{Y} , i.e., $\mathbf{X} = \mathbf{g}(\mathbf{Y})$, where \mathbf{g} is a $p \times 1$ vector function.

Special distributions

In this section, we illustrate the connection between this index and some special distributions including normal, binomial and Cauchy distribution. The derivations of these relations are in the appendix.

Conditional normal distribution. Suppose that $X|Y \sim N(\mu_Y, \sigma_Y^2)$, where $Y \in \{0, 1\}$. For simplicity, we assume that $\sigma_Y^2 = \sigma^2 = 1$, and define that $\Delta = \mu_0 - \mu_1$. Let p_y be the probability for the class $Y = y$, by using the characteristic function of normal distribution, let $\text{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} dt$, the Gaussian Error Function, we have:

$$\mathcal{C}^2(X|Y) = 4p_0p_1 \left[\frac{\Delta}{2} \text{erf}\left(\frac{\Delta}{2}\right) + \frac{e^{-\Delta^2/4} - 1}{\sqrt{\pi}} \right].$$

Note that this equivalence indicates that $\Delta = 0$ iff $\mathcal{C}^2(X|Y) = 0$, as we expected.

Bivariate normal distribution. Suppose that X and Y follow standard normal distribution with correlation coefficient ρ . Then we have that $X|Y \sim N(\rho Y, (1 - \rho^2))$. Our index can be expressed using ρ as follows:

$$\mathcal{C}^2(X|Y) = \frac{2}{\sqrt{\pi}} (1 - \sqrt{1 - \rho^2}).$$

Again, in this case, naturally we have that $\mathcal{C}^2(X|Y) = 0$ iff $\rho = 0$.

Conditional Binomial distribution. Suppose that $X|Y \sim \text{Ber}(n, q_Y)$, where $Y \in \{0, 1\}$. Let p_y be the probability for the class $Y = y$. For $n = 1$ which is Bernoulli distribution, we have that

$$\mathcal{C}^2(X|Y) = 4p_0p_1(q_0 - q_1)^2.$$

For $n = 2$, then we have that $\mathcal{C}^2(X|Y) = 4p_0p_1(q_0 - q_1)^2[1 + (1 - q_0 - q_1)^2]$. It is clear to see that in both cases, $\mathcal{C}^2(X|Y) = 0$ iff $q_0 = q_1$. A general formula of $\mathcal{C}^2(X|Y) = 0$ for conditional Binomial distribution can be found in the appendix.

Conditional Cauchy distribution. Although we do require finiteness of conditional means to develop the equivalence formula for $\mathcal{C}^2(X|Y)$ as in (2.5), the original definition of our index $\mathcal{C}^2(X|Y)$ only requires the existence of its respective characteristic functions. It's well-known that Cauchy has its characteristic function but without finite moments. Nevertheless, we could still do such a calculation. Suppose that Cauchy distribution has density: $p(x|y) = \frac{q_y}{\pi(q_y^2+x^2)}$, where $y \in \{0, 1\}$. Let p_y be the probability for the class $Y = y$, then we have that

$$\mathcal{C}^2(X|Y) = \frac{4p_0p_1}{\pi} \left(q_0 \ln \frac{2q_0}{q_0 + q_1} + q_1 \ln \frac{2q_1}{q_0 + q_1} \right).$$

Again, $q_0 \ln \frac{2q_0}{q_0+q_1} + q_1 \ln \frac{2q_1}{q_0+q_1} \geq 0$, and it is 0 iff $q_0 = q_1$.

2.4 Estimation Approaches

Slicing Estimator

In the previous development of population version, we do not require \mathbf{Y} to be discrete or continuous. We now consider a special sample version of \mathbf{Y} : unless \mathbf{Y} is categorical variable or discrete, otherwise for continuous \mathbf{Y} , we slice it into finite categories. Slicing techniques for continuous variables have been used in many other areas, such as in sufficient dimension reduction, SIR (Li, 1991), SAVE (Cook and Weisberg, 1991), CR (Li et al., 2005), DR (Li and Wang, 2007), SR (Wang and Xia, 2008) and the fused approaches (Cook and Zhang, 2014). The use of slicing in our development is a natural choice because of the last term in the second equation of (2.5) in Theorem 2.3.1, and it is especially for its technical simplicity as well. To facilitate our development, we then further assume that $\mathbf{X} \in \mathbb{R}^p$, Y is a categorical variable with H levels. That is, let $Y = \{1, \dots, H\}$.

The defined measure has no restriction on the dimensions of the random vectors. However, when Y is multivariate with high dimensions, slicing on each element of Y will result in very few observations in each slice, and that may affect the proposed test. Nevertheless, slicing techniques for high dimensional response have been used in

other areas effectively. For instance, one may adapt the slicing schemes developed by Zhu et al. (2010) and Li et al. (2008) to our statistic. In general, effectively dealing with multivariate Y with high dimensions is an interesting but independent topic. We leave a thorough study on such topic as our future research.

Let (\mathbf{X}_k, Y_k) , $k = 1, \dots, n$, be a random sample of (\mathbf{X}, Y) . For the purpose of slicing method, these n observations can be equivalently written as $(\mathbf{X}_{y,k_y}, Y_{y,k_y})$, where $y = 1, \dots, H$, $k_y = 1, \dots, n_y$, where $Y_{y,k_y} = y$ for any k_y .

Definition 2.4.1. *An empirical measure is defined as*

$$\mathcal{C}_n^2(\mathbf{X}|Y) = \sum_{y=1}^H p_y \mathcal{C}_{w,y,n}^2(\mathbf{X}|Y = y) = \sum_{y=1}^H p_y \|f_{\mathbf{X}|y}^n(t) - f_{\mathbf{X}}^n(t)\|^2, \quad (2.6)$$

We establish a different formula for the empirical version which gives us practically simple calculations as follows. Again its proof is in the appendix.

Theorem 2.4.1. *The empirical measure can be written as*

$$\mathcal{C}_n^2(\mathbf{X}|Y) = \frac{1}{n^2} \sum_{y,y'=1}^{H,H} \sum_{k_y,l_{y'}=1}^{n_y,n_{y'}} |\mathbf{X}_{y,k_y} - \mathbf{X}_{y',l_{y'}}| - \frac{1}{n} \sum_{y=1}^H \frac{1}{n_y} \sum_{k_y,l_y=1}^{n_y,n_y} |\mathbf{X}_{y,k_y} - \mathbf{X}_{y,l_y}|. \quad (2.7)$$

Theorem 2.4.1 immediately implies the next result.

Corollary 2.4.2.

$$\mathcal{C}_n^2(\mathbf{X}|Y) = \frac{1}{n^2} \sum_{k,l=1}^{n,n} |\mathbf{X}_k - \mathbf{X}_l| - \frac{1}{n} \sum_{y=1}^H \frac{1}{n_y} \sum_{k_y,l_y=1}^{n_y,n_y} |\mathbf{X}_{y,k_y} - \mathbf{X}_{y,l_y}|. \quad (2.8)$$

$$\mathcal{C}_n^2(\mathbf{X}|Y) \leq \mathcal{C}_n^2(\mathbf{X}|\mathbf{X}) = \frac{1}{n^2} \sum_{k,l=1}^{n,n} |\mathbf{X}_k - \mathbf{X}_l|. \quad (2.9)$$

Based on the empirical measure definition, it is easy to see that the following results hold and thus, we omit the proof.

Lemma 2.4.3. *The following properties hold:*

1. $\mathcal{C}_n^2(\mathbf{X}|Y) \geq 0$

2. $\mathcal{C}_n^2(\mathbf{X}|\mathbf{X}) = 0$ iff every sample observation is identical.

We establish the following result and put the proof in the appendix.

Lemma 2.4.4.

$$\lim_{n \rightarrow \infty} \mathcal{C}_n^2(\mathbf{X}|Y) = \mathcal{C}^2(\mathbf{X}|Y) \text{ almost surely}$$

This lemma indicates that our sample version is properly defined and it is consistent. We now develop asymptotic distribution for the empirical measure. Let $\Gamma(\cdot)$ denote a complex-valued zero-mean Gaussian random process with covariance function $\text{cov}_\Gamma(s, s_0) = [f_{\mathbf{X}}(s - s_0) - f_{\mathbf{X}}(s)\overline{f_{\mathbf{X}}(s_0)}]$, where $s, s_0 \in \mathbb{R}^p$.

Theorem 2.4.5. (*Weak convergence*)

a. Assume that \mathbf{X} and Y are independent, and $E(|\mathbf{X}|) < \infty$, then

$$n\mathcal{C}_n^2(\mathbf{X}|Y) \xrightarrow[n \rightarrow \infty]{D} (H - 1) \|\Gamma(s)\|^2.$$

b. Assume that \mathbf{X} and Y are independent, and $E(|\mathbf{X}|) < \infty$, then

$$n\mathcal{C}_n^2(\mathbf{X}|Y)/S_n \xrightarrow[n \rightarrow \infty]{D} Q,$$

where Q is a nonnegative quadratic form of centered Gaussian random variable with $E(Q) = 1$ and $S_n = (H - 1) \frac{1}{n^2} \sum_{k,l=1}^{n,n} |\mathbf{X}_k - \mathbf{X}_l|$.

c. If \mathbf{X} and Y are dependent, then $n\mathcal{C}_n^2(\mathbf{X}|Y)/S_n \xrightarrow[n \rightarrow \infty]{P} \infty$.

Its proof is in the appendix. We now state the limit distribution. If Q is a quadratic form of centered Gaussian random variable and $E(Q) = 1$, then

$$P\{Q \geq \chi_{1-\alpha_0}^2(1)\} \leq \alpha_0, \text{ for all } 0 < \alpha_0 \leq 0.215,$$

where $\chi_{1-\alpha_0}^2(1)$ is the $(1 - \alpha_0)$ quantile of a chi-square variable with 1 degree of freedom. This result follows from a theorem of Székely and Bakirov (2003, page 189). Thus a test that rejects independence if $n\mathcal{C}_n^2(\mathbf{X}|Y)/S_n \geq \chi_{1-\alpha_0}^2(1)$ has an asymptotic

significance level at most α_0 . The asymptotic test criterion could be quite conservative for many distributions. See Székely et al. (2007), Székely and Rizzo (2009) and Rizzo and Székely (2010), for further comments.

By slicing, the measure is equivalent to DISCO (Rizzo and Székely, 2010) whose definition employed conditional moments directly similar to that of (2.5) and for categorical variable Y only. Hence, in general, DISCO limits certain distributions such as conditional Cauchy distribution in section 2.3. Our theoretical justification differs from theirs but similar to dCov. Both our measure and dCov are defined using characteristic functions, thus theoretical justifications of these two are analogous. For continuous \mathbf{Y} , we change \mathbf{Y} to a class variable by slicing on it. In such a case, our index provides an alternative way to dCov. However, one does not have to use slicing approach, other approaches may be used as well. Thus, our index provides many possible approaches for continuous random vectors which may lead to new research directions. As such, a kernel approach is proposed in the next section.

Kernel Estimator

Note that for continuous \mathbf{Y} , slicing \mathbf{Y} is just one of the approaches. In fact, even for slicing approach, one can improve it by using techniques such as “moving slicing” (Li et al., 2005), and fused approach (Cook and Zhang, 2014). In this section, we propose a nonparametric approach: Kernel method to estimate (2.5), in particular, the last term in (2.5), which differs from DISCO.

For simplicity, let $m = E|\mathbf{X}_{\mathbf{Y}} - \mathbf{X}'_{\mathbf{Y}}|$. Thus, our main goal is to estimate m by using kernel method. Write $m = E_{\mathbf{Y}}E|\mathbf{X}_{\mathbf{Y}} - \mathbf{X}'_{\mathbf{Y}}| = E_{\mathbf{Y}}m(\mathbf{Y})$, then $m(\mathbf{Y}) = E_{(\mathbf{X}, \mathbf{X}')}(|\mathbf{X} - \mathbf{X}'| | \mathbf{Y}) = E_{\mathbf{X}}[m(\mathbf{X}, \mathbf{Y}) | \mathbf{Y}]$, where $m(\mathbf{X}, \mathbf{Y}) = E_{\mathbf{X}'}(|\mathbf{X} - \mathbf{X}'| | \mathbf{Y})$.

For kernel estimation, $K_h(t) = h^{-q}K(t/h)$, $h > 0$ denotes a q -dimensional kernel function. Let $p_0(\mathbf{y})$ be the density function of \mathbf{Y} , then the kernel estimator of $p_0(\mathbf{y})$ is given by $\hat{p}_0(\mathbf{y}) = n^{-1} \sum_{k=1}^n K_h(\mathbf{y}_k - \mathbf{y})$. And thus an estimate of $m(\mathbf{X}, \mathbf{Y})$ is

$$\hat{m}(\mathbf{X}, \mathbf{Y}) = \frac{n^{-1} \sum_{j=1}^n |\mathbf{X} - \mathbf{X}_j| K_h(\mathbf{Y} - \mathbf{Y}_j)}{n^{-1} \sum_{j=1}^n K_h(\mathbf{Y} - \mathbf{Y}_j)}$$

Following this, an estimate of $m(\mathbf{Y})$ is

$$\begin{aligned}\hat{m}(\mathbf{Y}) &= \frac{n^{-1} \sum_{i=1}^n \hat{m}(\mathbf{X}_i, \mathbf{Y}) K_h(\mathbf{Y} - \mathbf{Y}_i)}{n^{-1} \sum_{i=1}^n K_h(\mathbf{Y} - \mathbf{Y}_i)} \\ &= \frac{n^{-2} \sum_{i=1, j=1}^n |\mathbf{X}_i - \mathbf{X}_j| K_h(\mathbf{Y} - \mathbf{Y}_i) K_h(\mathbf{Y} - \mathbf{Y}_j)}{n^{-1} \sum_{j=1}^n K_h(\mathbf{Y} - \mathbf{Y}_j) n^{-1} \sum_{i=1}^n K_h(\mathbf{Y} - \mathbf{Y}_i)}\end{aligned}$$

Finally, an estimate of m is $\hat{m} = \frac{1}{n} \sum_{l=1}^n \hat{m}(\mathbf{Y}_l)$. Hence, the kernel estimator of $C^2(\mathbf{X}|\mathbf{Y})$ is $C_{n,k}^2(\mathbf{X}|\mathbf{Y}) = \frac{1}{n^2} \sum_{i,j} |\mathbf{X}_i - \mathbf{X}_j| - \hat{m}$. We now establish the property for the kernel estimator. For the consistency of the result for Theorem 2.4.7, we need the following regularity conditions (Chen et al., 2015)

Condition A₁: The density functions, $p(\mathbf{x}|\mathbf{y})$ and $p(\mathbf{y})$ are continuous and bounded away from zero. The support of \mathbf{y} is bounded and compact in \mathbb{R}^q .

Condition A₂: The continuous kernel function $K(t)$ is Lipschitz on $[-1, 1]$, and for some $s > q/2$,

$$\int K(t)dt = 1, \int t^i K(t)dt = 0, (1 \leq i \leq s-1), 0 \neq \int t^s K(t)dt < \infty.$$

Condition A₃: As $n \rightarrow \infty$, the bandwidth h satisfies $h \rightarrow 0$, $nh^{2q} \rightarrow \infty$ and $nh^{2s+q/2} \log n \rightarrow 0$.

Condition A₄: We have that $E|\mathbf{X}_y|^4 < \infty$.

Condition A₅: Write $p_1(\mathbf{x}, \mathbf{y}) = p(\mathbf{x}|\mathbf{y})p(\mathbf{y})$, which is s times differentiable with respect to \mathbf{y} , and its s th-order derivative is uniformly bounded by a constant C_0 which does not depend on \mathbf{y} .

Conditions A_1 and A_5 require that the density functions be positive and sufficiently smooth. Condition A_5 facilitates the control of remainder terms in Taylor expansions; one may relax this condition by assuming local Lipschitz properties for the density functions, which are widely imposed in the literature (Li et al. (2011)). Condition A_2 implies that the kernel function is bounded from above, which holds for many well-known kernel functions. Condition A_3 gives conditions on the bandwidth h , which are relatively mild. Condition A_4 requires certain moments to be finite as typical. To prove Theorem 2.4.7, we establish the following lemma which is a direct application

of Lemma S5 of Chen et al. (2015).

Lemma 2.4.6. *Suppose Conditions (A₁)-(A₅) hold, then*

$$\sup_{\mathbf{y} \in \mathbb{R}^q} |\hat{m}(\mathbf{y}) - m(\mathbf{y})| = O(h^s + (nh^q)^{-1/2} \log n), \text{ almost surely.}$$

Here, we directly use the assumptions of Chen et al. (2015) for simplicity. Assumption A seems restrictive, although our simulations show otherwise. However, one can weaken assumption A by using different conditions such as those of Härdle and Stoker (1989) and Samarov (1993), or Wang et al. (2015), which nevertheless, need to modify our estimator with a trim/weight function, respectively, to deal with density near 0 and of large bias. We establish the consistency result below.

Theorem 2.4.7. *Under the assumptions (A₁) – (A₅), we have that $C_{n,k}^2(\mathbf{X}|\mathbf{Y}) \xrightarrow[n \rightarrow \infty]{P} C^2(\mathbf{X}|\mathbf{Y})$.*

Note that the first term in $C_{n,k}^2(\mathbf{X}|\mathbf{Y})$ is a typical U-statistic which is root- n asymptotic normal. By using the technicals in Chen et al. (2015), one can establish the asymptotic normality for the second term in $C_{n,k}^2(\mathbf{X}|\mathbf{Y})$, which however, has rate $nh^{q/2}$. Combining the two terms, we can still manipulate the asymptotic normality at the same rate, however, one of the asymptotic variances in the two terms vanishes in a faster rate. Hence, it is not much useful practically when sample size is large. Even if in the same rate of convergence, for instance, Székely et al. (2007), Székely and Rizzo (2009), Rizzo and Székely (2010), Shao and Zhang (2014) and Wang et al. (2015), asymptotic distributions are not practically used but permutation or bootstrap tests are preferred. We will describe the use of permutation test in the next section. Note that $K_h(t)$ is a q -dimensional kernel function. Therefore, kernel method can be used for Y with any dimensions theoretically. Practically due to the high-dimension issue, kernel method certainly has its own restriction. Nevertheless, there exist kernel estimations in using (conditional) distance covariance as in Wang et al. (2015) and Chen et al. (2015).

2.5 Testing Procedure

To obtain the p -value for the independence test, we used permutation approach (Efron and Tibshirani (1998); Davison and Hinkley (1997)). Based on previous discussions, we use R_c as the illustrative test statistic while calculating the p -value. For example, in slicing method, we use $R_c(\text{slice})$ as the test statistic, and illustrate the procedure as follows: Let π^b represent one permutation of the sample, $b = 1, \dots, B$, where B is the total number of permutations. In our simulations, we set $B = 999$ unless otherwise stated. Let $R_c(\text{slice})^b$ be the test statistic computed corresponding to permuted sample π^b and $R_c(\text{slice})^0$ be the observed test statistic. Compute the p -value using the following formula ($\mathbf{1}(\cdot)$ is the indicator function)

$$\hat{p} = \frac{1 + \sum_{b=1}^B \mathbf{1}(R_c(\text{slice})^b \geq R_c(\text{slice})^0)}{B + 1}.$$

2.6 Simulation Studies

In this section we provide some empirical evidences for the new measure and compare with existing methods, in particular, dCov and DISCO, for both continuous and categorical Y .

Using R_c as the test statistic, three estimation methods are used: slicing [$R_c(\text{slice})$]; Epanechnikov kernel [$R_c(\text{epa})$]; Gaussian kernel, [$R_c(\text{gau})$]. We do not compare our methods to other available testing methods, as Székely et al. (2007) and Rizzo and Székely (2010) have detailed comparisons.

Example 2.6.1. Six characteristics of aircraft designs which appeared during the twentieth century were recorded in the aircraft data (Saviotti (1996)). The data is in `r` package `sm`, the data and example are from Bowman and Azzalini (1997, 2007), also see example 3 in Székely and Rizzo (2009). Two variables wing span(m) and speed (km/h), in period 3 with $n = 230$ designs were considered. We want to test the independence of $\log(\text{Speed})$ and $\log(\text{Span})$.

To apply slicing method, we slice $\log(\text{Span})$ into H groups. The number of observations in each slice is $\lfloor n/H \rfloor$. Table 2.1 reports the corresponding test statistic

and p-value using various number of slices and the two kernel methods. For different numbers of slices, we find that as long as the number of slices is not too small or too big, in other words, the number of data points in each slice is greater than 5 but not close to $n/2$, the test results are very consistent and comparable. In addition, the p-values indicate that all three methods give the same test result as dCov of Székely and Rizzo (2009), which has p-value 0.001.

Table 2.1: Test results using different methods

	$R_c(\text{slice})$						$R_c(\text{epa})$	$R_c(\text{gau})$
	$H = 2$	$H = 5$	$H = 10$	$H = 23$	$H = 46$	$H = 115$		
Test statistic	0.161	0.264	0.328	0.453	0.528	0.752	0.302	0.237
p-value	0.004	0.001	0.001	0.001	0.001	0.007	0.001	0.001

Example 2.6.2. In this example, we study the type-I error for dCov, kernel methods, and slice on continuous variable to apply DISCO and slicing method. We simulate four models. In model 2.6.2 (a), the marginal distributions of \mathbf{X} and that of Y are standard normal, where $p = 5$ and $q = 1$. The elements of \mathbf{X} are independent and are also independent of Y . In models 2.6.2 (b)-(d), the dimensions of \mathbf{X} and Y are the same as in 2.6.2 (a), except that each individual random variable is independently generated from $t(1)$, $\chi^2(1)$ and $\chi^2(3)$ distributions, respectively.

We fix the number of slices at $H = 5$ for DISCO and $R_c(\text{slice})$. The total sample sizes $n = 25, 30, 35, 50, 70, 100$, respectively, and we use the number of replicates $B = \lfloor 200 + 5000/n \rfloor$ as suggest by Székely et al. (2007) to obtain p -value for each test. We use 10,000 tests to obtain the type-I error at nominal significance level 0.1. The empirical type-I error rates for each case are recorded in Table 2.2. It appears that all methods perform similarly, close to the level and none of them can consistently beat the others. Simulation results for additional models and nominal level of 0.05 in the appendix indicate similar conclusion. Some models also appear in example 1 in Székely et al. (2007).

Example 2.6.3. The model is: $(X, Y) = (X, \phi(X))$, where X is standard normal random variable and $\phi(\cdot)$ is the standard normal density (Example 2 in Székely and Rizzo (2009)). Our goal is to make a power comparison. The power is computed as

Table 2.2: Empirical type-I error rates for 10,000 tests at nominal significance level 0.1, using B replicates

		(a) $N(0, 1), p = 5, q = 1$						(b) $t(1), p = 5, q = 1$				
n	B	dCov	DISCO	$R_c(\text{slice})$	$R_c(\text{epa})$	$R_c(\text{gau})$	dCov	DISCO	$R_c(\text{slice})$	$R_c(\text{epa})$	$R_c(\text{gau})$	
25	400	0.094	0.103	0.100	0.096	0.101	0.104	0.097	0.095	0.094	0.103	
30	366	0.102	0.095	0.099	0.100	0.100	0.102	0.100	0.099	0.098	0.097	
35	342	0.105	0.099	0.101	0.102	0.099	0.104	0.100	0.102	0.093	0.095	
50	300	0.103	0.099	0.100	0.097	0.101	0.100	0.106	0.104	0.097	0.103	
70	271	0.103	0.097	0.103	0.100	0.100	0.100	0.098	0.100	0.099	0.098	
100	250	0.101	0.098	0.098	0.104	0.098	0.094	0.105	0.103	0.097	0.102	
		(c) $\chi^2(1), p = 5, q = 1$						(d) $\chi^2(3), p = 5, q = 1$				
n	B	dCov	DISCO	$R_c(\text{slice})$	$R_c(\text{epa})$	$R_c(\text{gau})$	dCov	DISCO	$R_c(\text{slice})$	$R_c(\text{epa})$	$R_c(\text{gau})$	
25	400	0.096	0.099	0.099	0.099	0.098	0.097	0.099	0.098	0.100	0.098	
30	366	0.102	0.094	0.095	0.098	0.098	0.094	0.100	0.100	0.096	0.102	
35	342	0.096	0.102	0.104	0.101	0.098	0.101	0.103	0.104	0.102	0.103	
50	300	0.102	0.097	0.098	0.103	0.099	0.099	0.102	0.102	0.103	0.100	
70	271	0.103	0.099	0.098	0.101	0.100	0.104	0.100	0.102	0.102	0.100	
100	250	0.098	0.101	0.098	0.098	0.102	0.099	0.101	0.102	0.100	0.100	

the proportion of significant tests out of 10,000 at significance level 0.1. Again, we use the number of replicates $B = \lfloor 200 + 5000/n \rfloor$ in each permutation test.

Since Y is continuous, we slice it into several categories for DISCO and slicing methods. Based on example 2.6.1, we use 3, 3 and 4 slices when sample size $n = 10, 15$ and 20, and 5 slices for sample sizes greater than 20. Figure 2.1 plots the power of different methods with the increase of sample size n . We find that for $n \geq 35$, all five methods are equivalently powerful with powers near 1. For $n < 35$, Gaussian kernel is the best, followed by dCov and Epanechnikov kernel. As expected, slicing and DISCO methods may lose certain power for small sample size.

Example 2.6.4. To estimate the acceleration due to gravity at Washington, the dataset (*gravity*) consists of 81 measurements in a series of eight experiments between May, 1934 to July, 1935. The experiments are conducted by the National Bureau of Standards in Washington DC. In each experiment, there are replicated measurements of a reversible pendulum expressed as deviations from $980\text{cm}/\text{sec}^2$. Davison and Hinkley (1997) discussed this data in their example 3.2. The data is also available in r package *boot* (Canty and Ripley (2009)).

Our goal is to show that for categorical variables, changing the values of categorical variable shall not affect the conclusion of a robust method. We test the independence between the original X (*gravity*), then the residuals after fitting a linear model, and

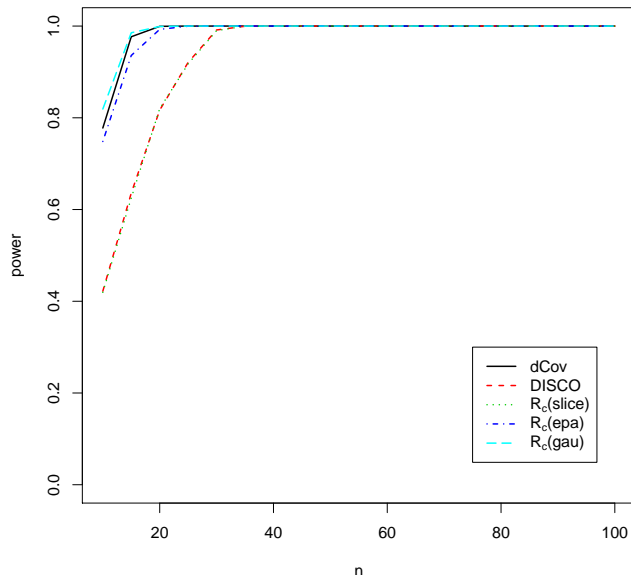


Figure 2.1: Empirical power comparisons at 0.1 level with different sample size n .

the group indicator, respectively, as in Rizzo and Székely (2010). We use three indicators: the original indicator $series0 = (1, 2, 3, 4, 5, 6, 7, 8)$ and two different indicators: $series1 = (1, 10, 15, 20, 45, 70, 200, 500)$ and $series2 = (100, 10, 15, 20, 45, 70, 200, 500)$. Table 2.3 shows if Y indicator changes, only slicing and DISCO methods are robust.

Table 2.3: P-values using different group indicators

Gravity	dCov	DISCO	$R_c(\text{slice})$	$R_c(\text{epa})$	$R_c(\text{gau})$
series0	0.001	0.001	0.001	0.001	0.001
series1	0.133	0.001	0.001	0.401	0.288
series2	0.125	0.001	0.001	0.006	0.019
Residual	dCov	DISCO	$R_c(\text{slice})$	$R_c(\text{epa})$	$R_c(\text{gau})$
series0	0.001	0.046	0.044	0.002	0.001
series1	0.148	0.046	0.046	0.443	0.339
series2	0.264	0.045	0.043	0.008	0.014

Example 2.6.5. In a four group balanced design with common sample size $n = 30$, multivariate observations are generated. The marginal distributions are independent. Group 1 is non-central $t(4)$ with non-centrality parameter δ . Groups 2-4 are all central $t(4)$ distributions. The group indicator is Y . This is example 3 in Rizzo and Székely

(2010). We want to show that for categorical variable, changing values will not change the power of robust methods.

We first look at the empirical power by fixing dimension $p = 10$ and non-centrality parameter δ varies, then we look at the empirical power when p varies and $\delta = 0.2$. Results of the simulations are summarized in Figures 2.2-2.3 at significance level 0.1. We use $B = 199$ in each test and conduct 10,000 tests.

By fixing dimension $p = 10$, and δ varies, Figure 2.2 (a) shows that the empirical power for testing the independence of \mathbf{X} and Y is roughly the same when comparing the five methods with group indicator: 1, 2, 3, and 4. However, when we change the group indicator Y from 1-4 to 1, 8, 0.5, and 1.2, Figure 2.2 (b) shows that the powers of DISCO and slicing methods remain the same. dCov and kernel methods have empirical power much smaller than the others. We also apply dCov for the dummy variables. The purple line in Figure 2.2 (b) shows that, although dCov with dummy variables has higher power compared with treating Y as one dimension, with values (1,0.8,0.5,1.2), it still has less power than $R_c(\text{slice})$ or DISCO method. Figure 2.3 (a) shows that when dimension p varies and non-centrality parameter $\delta = 0.2$, the empirical power for testing the independence of \mathbf{X} and Y is also roughly the same when comparing the five methods with group indicator: 1, 2, 3, and 4. However, when changing the group indicator from 1-4 to 1, 8, 0.5, and 1.2, Figure 2.3 (b) shows only DISCO and $R_c(\text{slice})$ are robust. Therefore, we believe that whether using dummy variable or not, dCov method has less power and not stable comparing with DISCO or $R_c(\text{slice})$.

Example 2.6.6. The model is $Y = a(\beta^T \mathbf{X})^2 \epsilon$, where $\beta = (1, 1, 1, 1, 1, 0, \dots, 0)^T$, $\mathbf{X} \sim N(\mathbf{0}, \Sigma_x)$, Σ_x is a $p \times p$ diagonal matrix with the same diagonal element σ_x^2 , a is a constant and, $\epsilon \sim N(0, \sigma^2)$ is independent of \mathbf{X} . This is an example that \mathbf{X} and Y has non-linear relationship and is similar to Example C in Sheng and Yin (2013).

We use the number of replicates $B = \lfloor 200 + 5000/n \rfloor$ in each permutation test and use 10,000 tests to get the power. We have different combinations for the values of a , p , σ_x^2 and σ^2 . Within each combination, we change the sample size n to see how the power of testing independence of \mathbf{X} and Y will change using different methods.

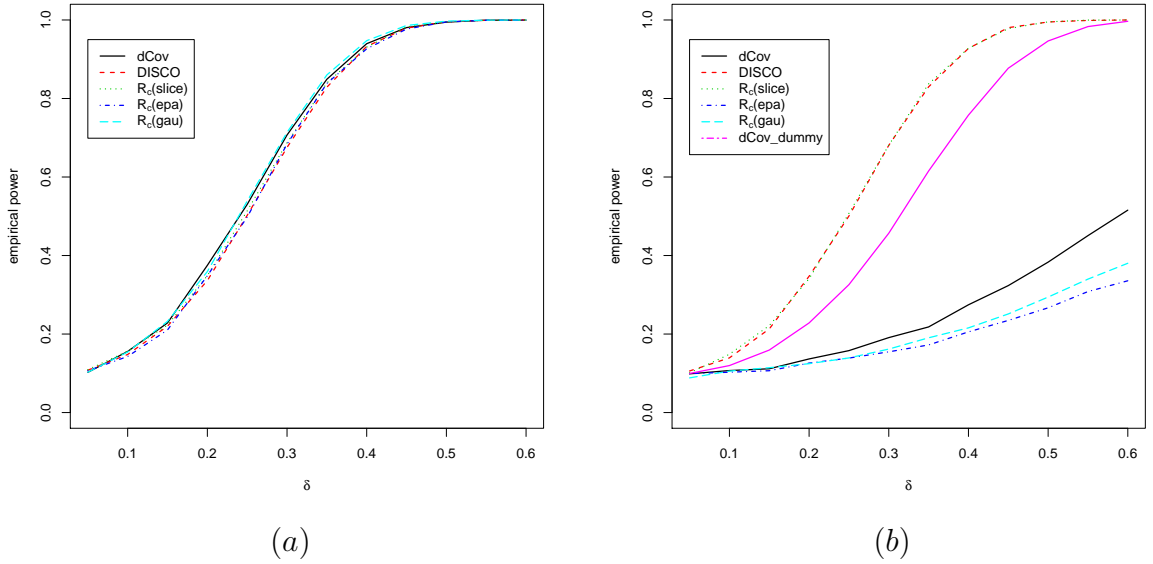


Figure 2.2: Empirical power for testing independence of \mathbf{X} and Y using five methods, $n = 30$ per group, dimension $p = 10$ and non-centrality parameter δ varies, where group indicator is (a) 1, 2, 3, 4; (b) 1, 8, 0.5, 1.2, except for the purple line, Y is transformed to dummy variables.

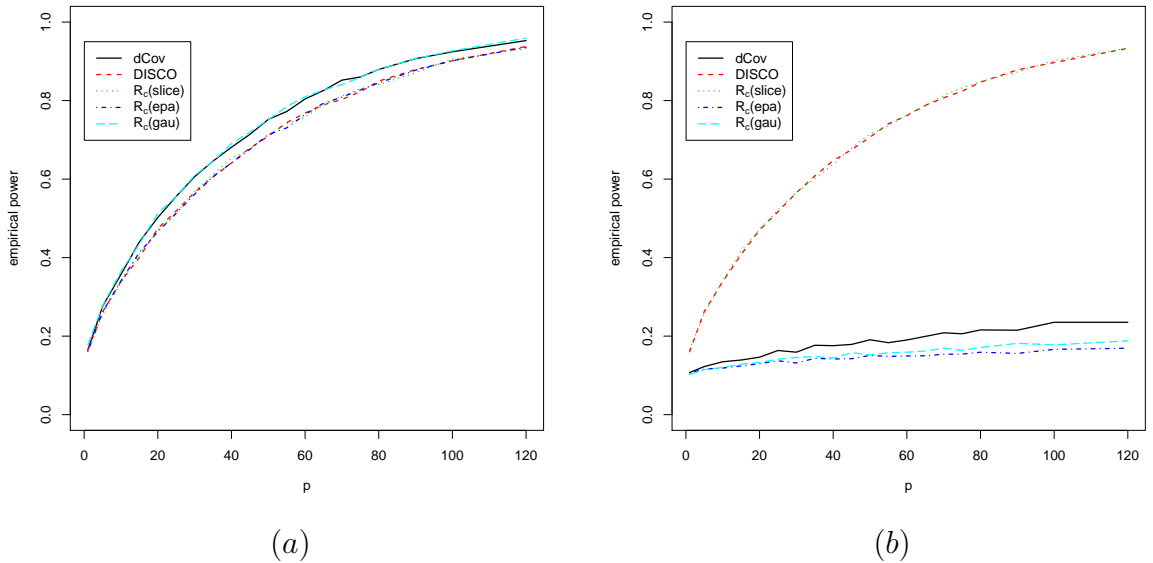
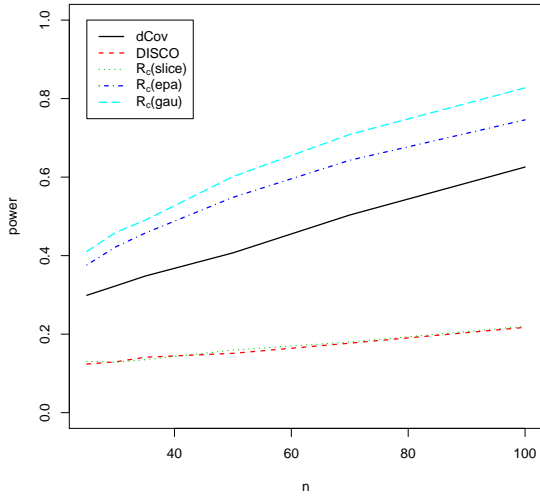
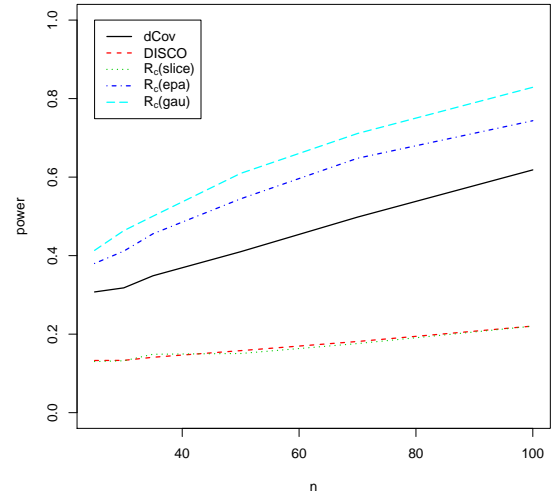


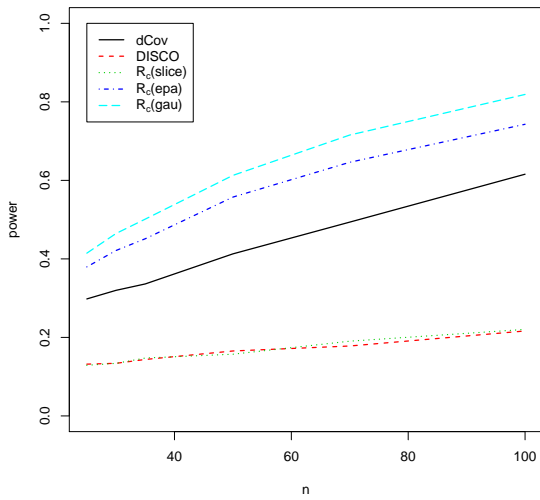
Figure 2.3: Empirical power for testing independence of \mathbf{X} and Y using five methods, $n = 30$ per group, dimension p varies and non-centrality parameter $\delta = 0.2$ where group indicator is (a) 1, 2, 3, 4; (b) 1, 8, 0.5, 1.2.



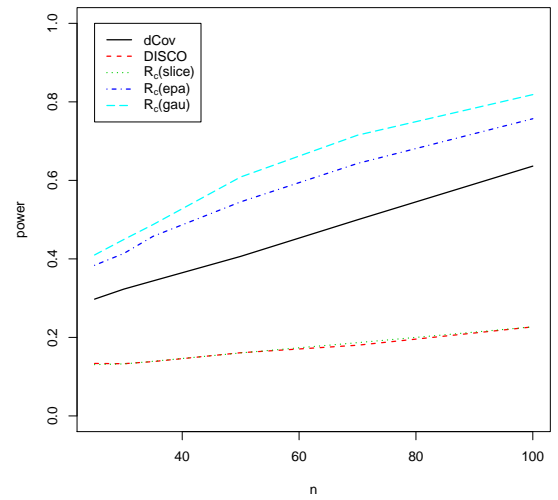
(a) $a = 0.1, p = 10, \sigma_x^2 = 1$ and $\sigma^2 = 1$.



(b) $a = 0.3, p = 10, \sigma_x^2 = 1$ and $\sigma^2 = 1$.



(c) $a = 0.3, p = 10, \sigma_x^2 = 1$ and $\sigma^2 = 4$.



(d) $a = 0.3, p = 10, \sigma_x^2 = 2$ and $\sigma^2 = 1$.

Figure 2.4: Empirical power with the change of sample size n .

Figure 2.4 shows the power change under four different cases. It clearly shows that for such a model with continuous response, discrete methods of DISCO and $R_c(slice)$ are not good, while the two kernel methods are much better than dCov. Additional simulations in the appendix show the same conclusion.

To summarize, for categorical Y , we showed that $R_c(slice)$ is equivalent to DISCO,

which is stable and better than dCov. For continuous Y , we showed that kernel methods with $R_c(gau)$ and $R_c(eps)$ perform better than dCov, and better than the discrete methods (DISCO and slicing) consistently.

2.7 Discussion

We introduce a new class of measures to test independence, which can be used flexibly for continuous and categorical random vectors. We study a particular weight function and its details in the class, however, weight functions used in HSIC or others can be used for developing new independence measures.

Note that dCov calculates Euclidean distance, which does not involve additional “tuning-parameters”, and thus it will lead to a unique value for the same data. Although our measure is defined similarly, it involves conditional distribution, which does require a step of its estimation. However, for continuous Y , simulations indicate that the number of slices does not affect the independence test result much, thus it is not very sensitive to the tuning parameter. We believe this is because, although the value changes when using different tuning parameters for the same data, once the tuning parameter is selected, the effect of such a tuning parameter will be canceled in permutation test. In general, estimating a conditional distribution is more subtle. However, in statistics, when we study the relations between two sets of variables, there are only two ways to do: conditional approaches and correlation-type approaches. Getting rid of directly estimating conditional distributions is important but it does require stronger conditions to do so. Correlation type also explicitly or implicitly requires certain restrictive conditions. Certainly, methods avoiding directly using estimation of conditional distribution could lead us to new interesting research direction.

Székely and Rizzo (2013) discussed the bias of dCov statistic in practice, when the dimensions of the random vectors are large. They constructed an unbiased t -test of independence. Since our measure is defined similarly to theirs, we believe that an analogous calculation will result in a similar unbiased statistic when p tends to infinity while q is fixed. When q tends to infinity as well, such a development seems

straightforward intuitively and theoretically. However, our statistic is different as Y is conditional on.

The index can be used in other areas beyond independence test. In later chapters, we would provide two applications: feature screening and sufficient dimension selection. We believe that it very much worth to investigate it along this direction. The appendix contains the proofs of the theoretical results, additional plots and tables for the numerical studies.

Chapter 3 Sufficient Variable Selection in High Dimensional Data

3.1 Introduction

In this chapter, focusing on categorical response we propose a new sufficient variable selection procedure: a two-stage sufficient variable selections method. Any independence measure can be adapted to our proposed procedure, thus the procedure does not require particular model specification. This model-free approach makes our method robust against model mis-specification, which is a very appealing property in practice. In addition, our approach always improves over typical screening approach which only uses marginal relation. Numerical studies are provided to demonstrate the advantages of the finite sample performances.

Feature screening and variable selection have become increasingly important in various research fields, as data are being collected at a relatively low cost due to modern technology. Many methods have been proposed during the last two decades, penalized approaches such as the least absolute shrinkage and selection operator (LASSO, Tibshirani, 1996), the smoothly clipped absolute deviation (SCAD, Fan and Li, 2001), and the Dantzig selector (Candes and Tao, 2007). These methods have shown promising results in dealing with high dimensional data. However, for the ultrahigh dimensional data, Fan and Lv (2008) pointed out that these aforementioned methods had their limitation due to the challenges of computational cost, statistical accuracy and algorithmic stability. These concerns lead to sure independent screening (SIS, Fan and Lv, 2008) for the ultrahigh dimensional data. SIS is based on the marginal Pearson correlation learning and designs for linear regressions with Gaussian predictors and responses. SIS not only can speed up variable selection drastically but also can improve the estimation accuracy when dimensionality is ultrahigh.

Many existing methods follow SIS with some restrictions on underlying distributions, model specification and structure of the data. Fan and Song (2010) extended

SIS to a generalized linear models using maximum marginal likelihood. Fan et al. (2011) proposed nonparametric independence screening (NIS) in additive models. They used a B-spline basis to do the nonparametric smoothing and ranks the variables according to the strength of marginal nonparametric regression. This method captures the active predictors that have nonlinear relationship with response variable. Chang et al. (2013) proposed marginal empirical likelihood approach for sure independence feature screening in linear and generalized linear models. Fan et al. (2014) discussed the use of nonparametric independence screening in varying coefficient models. Song et al. (2014) proposed varying coefficient independence screening for time-varying coefficient model. Chang et al. (2016) used marginal empirical likelihood to select the variables that locally contribute the response variable in nonparametric additive models, single index and multiple index models and varying coefficient models.

Feature screening methods using more general types of correlations and some model-free screening approaches are also proposed for high-dimensional variable selection. Zhu et al. (2011) proposed a sure independent ranking and screening procedure (SIRS) that does not require a specific model structure on regression functions. Li et al. (2012a) proposed robust rank correlation screening using Kendall τ correlation coefficient instead of the Pearson correlation. Li et al. (2012b) uses distance correlation (DC-SIS, Székely et al., 2007) to do the marginal correlation screening. Mai and Zou (2013) uses Kolmogorov-Smirnov statistic to do variable selection especially for when response variable Y is binary. Mai and Zou (2015) extended it to fused Kolmogorov filter for the cases when Y has more categories or is continuous. Cui et al. (2015) proposed a MV-SIS method based on conditional distribution function that target the marginal sure independence feature screening for ultrahigh dimensional discriminant analysis.

Problems arise when the marginal screening methods fail to identify some important predictors which are marginally independent of the response. For instance, recent methods developed by Zhu et al. (2011), Li et al. (2012b) are only able to detect marginal correlated predictors. As pointed in Zhu et al. (2011), the marginal screen-

ing procedure may miss some active predictors that are marginally independent of the response, thus marginal screening procedure is not sufficient variable selection, and they proposed an iterative feature screening to overcome the problem partly. Many other methods also use the iterative procedure to get a better variable selection result. However, the iterative procedure, making efforts to have sufficiently select variables, is not completely clear. On the other hand, penalized approaches have great impact but may not be sufficiently select variables. One of the difficulties for model-based penalized approaches such as LASSO or model-free based penalized approaches such as sufficient dimension reduction (SDR) sparse solution Li (2007) mainly due to the singularity of sample covariance of the predictors, i.e., large p small n issue.

To overcome the above issue, in this chapter, we propose a new sufficient variable selection procedure. This approach in collaboration with any measure of independence is model-free, thus, it is robust against model mis-specification. In particular, using the newly developed independence measure in chapter 2, we focus on categorical response variable and illustrate the usefulness of the procedure. Feature screening for categorical response and grouped/correlated predictors is of great interest in genome-wide association study (GWAS), discriminant analysis and classification problems.

The rest of this chapter is organized as follows. Section 3.2 develops the sufficient variable selection procedure. Section 3.3 studies the theoretic properties using the independence measure proposed in chapter 2, while Section 3.4 contains simulation studies and real data example, which followed by a short discussion in Section 3.5. Proof of the theorem is in appendix.

Throughout this chapter, we assume that Y is a categorical or continuous response variable, and $\mathbf{X} = (X_1, \dots, X_p)^T$ is a covariate vector. Let (Y_i, \mathbf{X}_i) , $i = 1, \dots, n$, be a random sample from the random vector (Y, \mathbf{X}) . For any random vectors U , V and W , the notation $U \perp\!\!\!\perp V|W$ means that given W , U and V are independent.

3.2 Methodology

Review of sufficient variable selection

Yin and Hilafu (2015) formally defined sufficient variable selection (SVS) and in particular, they discussed the difference between SDR and SVS. Let $\mathbf{X}_{\mathcal{D}} = \{X_k : X_k \in \mathbf{X}\}$ and $\mathbf{X}_{\bar{\mathcal{D}}}$ denotes the complement of $\mathbf{X}_{\mathcal{D}}$. SVS means to find a set $\mathbf{X}_{\mathcal{D}}$ so that $Y \perp\!\!\!\perp \mathbf{X}_{\bar{\mathcal{D}}} | \mathbf{X}_{\mathcal{D}}$, see Cook (2004). That is, given the set $\mathbf{X}_{\mathcal{D}}$, Y is independent of $\mathbf{X}_{\bar{\mathcal{D}}}$. Therefore, the goal of sufficient variable selection is to test the conditional independence of $Y \perp\!\!\!\perp \mathbf{X}_{\bar{\mathcal{D}}}$, given $\mathbf{X}_{\mathcal{D}}$. From this notation, \mathcal{D} and $\bar{\mathcal{D}}$ are the index set of the active and inactive predictors respectively.

While it is relatively easy to make such a statement of sufficient variable selection, it is rather difficult to construct such a test unless p is not too large. For instance, we view that traditional model diagnostic tests are sufficient test procedure for the conditional independence. However, in large p small n case, it is already difficult to build a reasonable model at the first place. Penalized approaches regardless of model-based or model-free may not be sufficient methods due to their ad hoc algorithm and the singularity of sample covariance of predictors, as we mentioned earlier, while SIS type methods are not testing this conditional independence, but using marginal independence tests only. Instead of directly testing such conditional independence, we follow a result of Yin and Hilafu (2015) to test sufficient conditions which then force the the conditional independence. Proposition below is a simplified version of Yin and Hilafu (2015, Proposition 1).

Proposition 3.2.1. *Let \mathbf{X}_1 and \mathbf{X}_2 be random vectors, then the following statement*

(i) or, statement (ii) implies statement (iii):

$$(i) \quad (Y, \mathbf{X}_2) \perp\!\!\!\perp \mathbf{X}_1;$$

$$(ii) \quad \mathbf{X}_1 \perp\!\!\!\perp \mathbf{X}_2 | Y \text{ and } Y \perp\!\!\!\perp \mathbf{X}_1;$$

$$(iii) \quad Y \perp\!\!\!\perp \mathbf{X}_1 | \mathbf{X}_2.$$

Statement (iii) implies that $p(Y|\mathbf{X}_1, \mathbf{X}_2) = p(Y|\mathbf{X}_2)$. Therefore, if statement (iii) holds we can eliminate \mathbf{X}_1 without losing any regression information. Let $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)^T$, after eliminating \mathbf{X}_1 , we treat \mathbf{X}_2 as a new \mathbf{X} , split it, and then do a further test until nothing can be eliminated. Further reduction of similar procedures can be used again on this set, if necessary. Hence, in the end, the final selected set contains $\mathbf{X}_{\mathcal{D}}$. Our procedure is a sufficient variable selection procedure. Thus, statement (iii) is very important. However, directly testing (iii) is impossible as we need (1), a measure of the conditional independence and (2), \mathbf{X}_2 need to contain \mathcal{D} . While testing statistics do exist such as the conditional distance correlation measure by Wang et al. (2015), among others, situation (2) again block the possibility as we discuss early. Nevertheless, Proposition 3.2.1 does provide very nice alternatives to statement (iii), by using statements (i) or (ii), since (iii) can be forced to hold if either statement (i) or statement (ii) holds.

The two statements (i) and (ii) are very general, requiring no particular model or assumptions on Y and \mathbf{X} , but a test index/measure. And one has the flexibility to choose different independence measures, though a chosen measure may bring extra conditions due to the way it is formatted. It is very natural to use statement (i) for continuous Y , since an assigned value of Y is important/meaningful when measuring the dependency between (Y, \mathbf{X}_2) and \mathbf{X}_1 . But statement (i) does not work well for categorical or discrete variable Y , when the value of Y is not meaningful, while on the other hand, statement (ii) can be more useful in such a case. In this chapter, we only focus on statement (ii) to propose a sufficient variable selection procedure, and we shall use the newly developed measure of independence in chapter 2 to illustrate this sufficient procedure, as such a measure has simple sample calculation and without any additional condition.

A measure of independence

In chapter 2, a new measure of independence is proposed: $\mathcal{C}^2(\mathbf{X}|\mathbf{Y}) = E|\mathbf{X} - \mathbf{X}'| - E|\mathbf{X}_{\mathbf{Y}} - \mathbf{X}'_{\mathbf{Y}}|$, where \mathbf{X}' is *i.i.d* copy of \mathbf{X} , and $\mathbf{X}'_{\mathbf{Y}}$ is iid copy of $\mathbf{X}_{\mathbf{Y}}$. Note that notation $\mathbf{X}_{\mathbf{Y}}$ means observations of \mathbf{X} conditioning on \mathbf{Y} . Here $|\cdot|$ is the Euclidean norm in

the respective dimension. An attractive property of $\mathcal{C}^2(\mathbf{X}|\mathbf{Y})$ is that it equals 0 if and only if the two random vectors are independent. This property makes it possible that $\mathcal{C}^2(\mathbf{X}|\mathbf{Y})$ can be used as an independence test statistic. Furthermore, we defined a statistic that is similar to correlation coefficient as follows:

$$R_c(\mathbf{X}|\mathbf{Y}) = \frac{\mathcal{C}(\mathbf{X}|\mathbf{Y})}{\mathcal{C}(\mathbf{X}|\mathbf{X})}, \text{ where } \mathcal{C}^2(\mathbf{X}|\mathbf{X}) = \mathbb{E}|\mathbf{X} - \mathbf{X}'|. \quad (3.1)$$

$0 \leq R_c \leq 1$, the higher value of R_c means a higher dependency between \mathbf{X} and Y . Therefore, we may use the sample version of R_c^2 as the statistic to test independence. Following from chapter 2, there are two ways to estimate $\mathcal{C}^2(\mathbf{X}|\mathbf{Y})$.

- Sclicing estimator. Note that in the definition of population version of $\mathcal{C}^2(\mathbf{X}|\mathbf{Y})$, we do not require Y to be discrete or continuous. We now consider a special sample version of Y : unless Y is categorical variable or discrete, otherwise for continuous Y , we slice it into finite C categories, that is $Y = \{1, \dots, H\}$.

The sample version of $\mathcal{C}^2(\mathbf{X}|Y)$ denoted by $\mathcal{C}_n^2(\mathbf{X}|Y)$, has a very simple form:

$$\mathcal{C}_n^2(\mathbf{X}|Y) = \frac{1}{n^2} \sum_{k,l=1}^{n,n} |\mathbf{X}_k - \mathbf{X}_l| - \frac{1}{n} \sum_{y=1}^H \frac{1}{n_y} \sum_{k,l=1}^{n_y, n_y} |\mathbf{X}_k - \mathbf{X}_l|. \quad (3.2)$$

- Kernel estimator. For continuous Y , usual kernel method can be used to obtain a sample estimate of $\mathcal{C}^2(\mathbf{X}|Y)$, which is denoted by $\mathcal{C}_{n,k}^2(\mathbf{X}|Y)$:

$$\mathcal{C}_{n,k}^2(\mathbf{X}|Y) = \frac{1}{n^2} \sum_{i,j} |\mathbf{X}_i - \mathbf{X}_j| - \hat{m},$$

where $\hat{m} = \frac{1}{n} \sum_{l=1}^n \hat{m}(Y_l)$. $K_h(t) = h^{-1}K(t/h)$, $h > 0$ denotes a 1-dimensional kernel function, and $\hat{m}(\mathbf{Y}) = \frac{n^{-2} \sum_{i=1, j=1}^n |\mathbf{X}_i - \mathbf{X}_j| K_h(Y-Y_i) K_h(Y-Y_j)}{n^{-1} \sum_{j=1}^n K_h(Y-Y_j) n^{-1} \sum_{i=1}^n K_h(Y-Y_i)}$. (Although the original estimation formula is for arbitrary dimensional response, we use a special case when Y is a continuous variable.)

The sample version $\mathcal{C}_n^2(\mathbf{X}|\mathbf{X})$ is the same for the above two methods:

$$\mathcal{C}_n^2(\mathbf{X}|\mathbf{X}) = \frac{1}{n^2} \sum_{k,l=1}^{n,n} |\mathbf{X}_k - \mathbf{X}_l|. \quad (3.3)$$

Different kernels(Gaussian kernel and Epanechnikov kernel) are used for the kernel estimator. Let $R_c^2(\text{slice})$, $R_c^2(\text{gau})$ and $R_c^2(\text{epa})$ be the slicing estimator and kernel estimator with Gaussian and Epanechnikov kernel, respectively in later sections.

Algorithm

In an ultrahigh-dimensional setting, the number of predictors p is usually much larger than the sample size n . Using the notation in earlier sections, sufficient variable selection means to correctly detect $\mathbf{X}_{\mathcal{D}}$, or a set containing such $\mathbf{X}_{\mathcal{D}}$. To achieve this goal, we could use proposition 3.2.1 to test the conditional independence and marginal independence in (ii). Although such an approach is elegant to reach \mathcal{D} , it has a high computational cost and is time consuming. Since our goal of sufficient variable selection is to achieve a set $\mathcal{S} \supseteq \mathcal{D}$, while the size of \mathcal{S} is small enough comparing with data size n , we propose an alternative screening approach as compared with the testing approach for sufficient variable selection.

Based on proposition 3.2.1, the newly proposed two-stage sufficient variable screening method uses both parts in statement (ii). It is different from marginal independence screening that is popular in independence screening area, which only utilizes the second part of statement (ii).

We use the measure $R_c^2(\mathbf{X}|Y)$ to illustrate the method, though it can be replaced by any other appropriate independence measures, for example correlation coefficient. The marginal screening will miss active predictors which are marginally unrelated but jointly related to the response. On the other hand, the two stage approach combines marginal and conditional relationship to fully recover the active predictor set $\mathbf{X}_{\mathcal{D}}$.

- Marginal Screening:

1. Calculate $\mathcal{I}_k^m = R_c^2(X_k|Y)$ for $k = 1, \dots, p$, and sort it by descending order.
 2. For a given model size d , select a set of the X_k 's that correspond to the largest d values of \mathcal{I}_k^m .
- Two-Stage Sufficient Variable Screening: For a given model size d , determine the model size d_m and d_c for the marginal and conditional sequence respectively, $d_m + d_c = d$.
 1. Obtain the first set of active predictors:
Apply the above marginal screening method, obtain the first set of active predictors with size d_m .
 2. Obtain the second set of active predictors:
 - (a) Suppose Y has H groups, or if Y is a continuous variable, slice it into H groups. For observations belong to category y , $y = 1, \dots, H$, calculate $\mathcal{I}_{k,y}^c = R_c^2(\mathbf{X}_{-k}|X_k)$, where \mathbf{X}_{-k} a vector of \mathbf{X} after eliminated X_k , $k = 1, \dots, p$. Compute $\mathcal{I}_k^c = \sum_{y=1}^H p_y \mathcal{I}_{k,y}^c$, where $p_y = n_y/n$ is a weight, n_y is the number of observations in group y , n is the total number of observations.
 - (b) Sort \mathcal{I}_k^c by descending order. The second part of active predictors is a set of X_k 's with the largest d_c values of \mathcal{I}_k^c that have not been selected in the first stage.
 3. An estimate of $\mathbf{X}_{\mathcal{D}}$ is the union of the two sets.

In practice, we use the sample index instead of the population index. Note that d has to be chosen. Typically, $d < p$, in general we can use $d = n - 1$, otherwise, in the simulation examples in section 3.4, we follow existing literature and let d equal to $d_1 = \lfloor n/\log(n) \rfloor$, $d_2 = 2\lfloor n/\log(n) \rfloor$ and $d_3 = 3\lfloor n/\log(n) \rfloor$, respectively. In addition, in screening approach of SVS₂ procedure, we use $d_m = \lfloor 0.95d \rfloor$ since marginal relation is more important in selecting active predictors. while $d_c = d - d_m$ variables are selected in the conditional sequence.

3.3 Theoretical Properties

We now discuss the theoretical properties of the proposed screening approach of the two-stage sufficient variable selection procedure. Two measures are used: $\mathcal{I}_k^m = R_c^2(X_k|Y)$ and $\mathcal{I}_k^c = \sum_{y=1}^H p_y \mathcal{I}_{k,y}^c$, where $\mathcal{I}_{k,y}^c = R_c^2(\mathbf{X}_{-k}|X_k)$ and is computed based on observations in group y . Treat X_k as the variable Y , the measure $\mathcal{I}_{k,y}^c$ is the same as \mathcal{I}_k^m . Thus, we would first study the theoretical properties of marginal screening stage and focus on estimation using slicing method. In particular, we follow what have been studied by Li et al. (2012b) on the theoretical properties of the screening method using distance correlation. The theoretical properties for the conditional stage can be obtained using similar argument, and will be proved to have the similar results as the marginal stage. After that, we will show the sure screening property of the proposed two-stage screening approach.

Note that $R_c(X_k|Y) = 0$ if and only if X_k and Y are independent with $k = 1, \dots, p$, guarantees $R_c^2(X_k|Y)$ ranks the active predictor above the inactive one, i.e. $\max_{i \in \bar{\mathcal{D}}} R_c^2(X_k|Y) < \min_{i \in \mathcal{D}} R_c^2(X_k|Y)$, and separates the active ones from the inactive ones. Hence, the quantity $R_c^2(X_k|Y)$ can be used for variable screening. We use this measure since it is model free and works especially well if the response variable is categorical.

For ease of presentation, let the population and sample version, respectively, be

$$\omega_k = R_c^2(X_k|Y), \hat{\omega}_k = \hat{R}_c^2(X_k|Y) \text{ for } k = 1, \dots, p.$$

While ω_k ranks the importance of X_k at the population level, $\hat{\omega}_k$ helps to select a set of active predictors with large values. Let $\hat{\mathcal{D}}_m$ be the estimated index set of active predictors considering the marginal relationship:

$$\hat{\mathcal{D}}_m = \{k : \hat{\omega}_k \geq cn^{-\tau}, \text{ for } 1 \leq k \leq p\},$$

where c and τ are two pre-specified threshold. And \mathcal{D}_m is the true index set of marginally active predictors. The following three conditions are needed for technical

proofs:

(C_1) Predictor \mathbf{X} satisfies the subexponential tail probability uniformly in p . That is, there exists a positive constant s_0 such that for all $0 < s \leq 2s_0$,

$$\sup_p \max_{1 \leq k \leq p} \mathbb{E}\{\exp(s|X_k|^2)\} < \infty.$$

For some constant $c > 0$ and $0 \leq \tau < \frac{1}{2}$, the dependency measures satisfy:

(C_2)

$$\min_{k \in \mathcal{D}} \omega_k \geq 2cn^{-\tau},$$

(C'_2)

$$\min_{k \in \mathcal{D}} \sum_{y=1}^H p_y \mathcal{I}_{k,y}^c \geq 2cn^{-\tau}.$$

Condition (C_1) is used to facilitate the technical derivations as in Li et al. (2012b). It follows immediately when \mathbf{X} is bounded uniformly, or when it has a multivariate normal distribution, which is widely used in ultrahigh-dimensional data analysis. Condition (C_2) is equivalent to the condition 3 of Fan and Lv (2008) and condition (C_2) in Li et al. (2012b). Condition C'_2 is a similar condition as C_2 for the conditional screening stage. Conditions C_2 and C'_2 reflect the signal strength of individual active predictors, which in turn controls the rate of probability error in selecting the active predictors (Zhu et al., 2011). The following theorem establishes the asymptotic property.

Theorem 3.3.1. *Under condition (C_1), for any $0 < \gamma < \frac{1}{2} - \tau$, there exist positive constants $c_1, c_2 > 0$ such that*

$$Pr(\max_{1 \leq k \leq p} |\hat{\omega}_k - \omega_k| \geq cn^{-\tau}) \leq O(p[\exp(-c_1 n^{1-2(\tau+\gamma)}) + n \exp(-c_2 n^{2\gamma})]) \quad (3.4)$$

Under conditions (C_1) and (C_2), we have that

$$Pr(\mathcal{D}_m \subseteq \hat{\mathcal{D}}_m) \geq 1 - O(s_m[\exp(-c_1 n^{1-2(\tau+\gamma)}) + n \exp(-c_2 n^{2\gamma})]), \quad (3.5)$$

where s_m is the cardinality of \mathcal{D}_m .

The inequality (3.4) in Theorem 3.3.1 shows the rank consistency of $\hat{\omega}_k$, it also indicates that we can handle the non-polynomial (NP) dimensionality of order $\log(p) = o(n^{(1-2\tau)/4})$. If we further assume that X_k is bounded uniformly in p , we can handle the NP dimensionality of order $\log(p) = o(n^{1-2\tau})$. Based on (3.5), the true active predictors survive with probability approaching to one with exponential rate as $n \rightarrow \infty$.

Similarly, define $\hat{\mathcal{D}}_c$ as the estimated index set of active predictors considering the conditional relationship:

$$\hat{\mathcal{D}}_c = \{k : \sum_{y=1}^H p_y \hat{\mathcal{I}}_{k,y}^c \geq cn^{-\tau}, \text{ for } 1 \leq k \leq p\},$$

where $\hat{\mathcal{I}}_{k,y}^c$ is the sample version of $\mathcal{I}_{k,y}^c$, c and τ are two pre-specified threshold. And \mathcal{D}_c is the true index set of conditionally active predictors.

Theorem 3.3.2. *Under condition (C_1) , for any $0 < \gamma < \frac{1}{2} - \tau$, there exist positive constants $c_3, c_4 > 0$ such that*

$$Pr(\max_{1 \leq k \leq p} |\sum_{y=1}^H p_y \hat{\mathcal{I}}_{k,y}^c - \sum_{y=1}^H p_y \mathcal{I}_{k,y}^c| \geq cn^{-\tau}) \leq O(pH[\exp(-c_3 n^{1-2(\tau+\gamma)}) + n \exp(-c_4 n^{2\gamma})]) \quad (3.6)$$

Under conditions (C_1) and (C_2') , we have that

$$Pr(\mathcal{D}_c \subseteq \hat{\mathcal{D}}_c) \geq 1 - O(s_c H[\exp(-c_3 n^{1-2(\tau+\gamma)}) + n \exp(-c_4 n^{2\gamma})]), \quad (3.7)$$

where s_c is the cardinality of \mathcal{D}_c .

Combine the marginal and the conditional procedure together, we would get the following result:

Theorem 3.3.3. *Let $\mathcal{D} = \mathcal{D}_c \cup \mathcal{D}_m$, and $\hat{\mathcal{D}} = \hat{\mathcal{D}}_c \cup \hat{\mathcal{D}}_m$, there exist positive constants*

$c_5, c_6 > 0$ such that

$$Pr(\mathcal{D} \subseteq \hat{\mathcal{D}}) \geq 1 - O[s(\exp(-c_5 n^{1-2(\tau+\gamma)}) + n \exp(-c_6 n^{2\gamma}))],$$

where s is the minimum of s_m and $s_c H$.

We use slicing method as an example to prove the above theorem, however, other methods, for example the kernel method, or dCorr² can lead to similar results.

3.4 Numerical Studies

Simulations

In this section, we assess the performance of the screening approach for the two-stage sufficient variable selection procedure through simulation studies. For each model below, we repeat the experiment 500 times, and report the results in terms of the following criteria:

1. \mathcal{P}_s : the proportion that an individual active predictor is selected for a given model size d in the 500 replications.
2. \mathcal{P}_a : the proportion that all active predictors are selected for a given model size d in the 500 replications.

Note: For a given model size d , we set $\lfloor 0.95d \rfloor$ and $d - \lfloor 0.95d \rfloor$ to be the cutoff point for marginal and conditional sequence, respectively. In a certain replication, for each individual active predictor X_k , if it appears in the estimated set $\mathbf{X}_{\hat{\mathcal{D}}}$, we say this predictor is selected for the given model size. If all the active predictors in the model are selected within the same replication, we say all active predictors are selected for the given model size in this replication.

Note that \mathcal{P}_s measure the probability of an individual active predictors X_s being selected by the variable selection method, while \mathcal{P}_a represents the probability that all active predictors are selected. If \mathcal{P}_s and \mathcal{P}_a are closer to 1, the method is better.

For models with categorical variable Y , we report the results of SVS₂ using $R_c^2(\text{slice})$ as the independence measure in both the marginal and conditional screening stage. In the conditional screening stage, to compute $\mathcal{I}_{k,y}^c = R_c^2(\mathbf{X}_{-k}|X_k)$, for observations belong to category y , we slice X_k into 2, 3 or 4 slices respectively when $n_y \leq 5$, $5 < n_y \leq 15$ or $15 < n_y \leq 20$, and 5 slices when n_y is greater than 20. We compare the results with Kolmogorov filter, fused Kolmogorov filter and MV-SIS methods.

Although we focus on categorical response Y , we also simulate a model with a continuous response. For such a model, we use $R_c^2(\text{gau})$ in the marginal screening stage of SVS₂ method, since it is naturally defined for continuous variables. In the conditional screening stage, we slice the response variable Y into 5 categories (based on Yin and Yuan (2016)), then compute the corresponding value using $R_c^2(\text{gau})$. We compare the result with DC-SIS method.

Example 3.4.1. We generate the following model from example (1.b) in Li et al. (2012b) to compare the finite sample performance of DC-SIS, SVS₂ method with $R_c(\text{gau})$ and SVS₂ method with dCorr:

$$Y = c_1\beta_1X_1X_2 + c_2\beta_2\mathbf{1}(X_{12} < 0) + c_3\beta_3X_{22} + \epsilon,$$

where \mathbf{X} is drawn from a multivariate normal distribution with mean zero and covariance $\Sigma = (\sigma_{ij})_{p \times p}$ with $\sigma_{ij} = 0.5^{|i-j|}$ and $\sigma_{ij} = 0.8^{|i-j|}$ respectively. The error term $\epsilon \sim N(0, 1)$, $\mathbf{1}(\cdot)$ is an indicator function and $(c_1, c_2, c_3) = (2, 3, 2)$. We choose $\beta_i = (-1)^U(a + |Z|)$ for $i = 1, 2, 3$, where $a = 4 \log n / \sqrt{n}$, $U \sim \text{Bernoulli}(0.4)$ and $Z \sim N(0, 1)$. We fix $n = 200$ and vary the dimension p from 2000 to 5000.

From table 3.1, it is clear that SVS₂ procedure using dCorr performs better than the marginal screening procedure via distance correlation method (DC-SIS) in most cases. On the other hand, SVS₂ procedure using $R_c(\text{gau})$ is very comparable with SVS₂ using dCorr. Note that DC-SIS method performs better than SIS and SIRS methods (Li et al., 2012b). Thus we conclude that the SVS₂ procedure does improve the existing marginal screening procedure, even for continuous response variable.

Table 3.1: Proportions comparison of \mathcal{P}_s and \mathcal{P}_a in example 3.4.1

		DC-SIS						SVS ₂ with $R_c^2(\text{gau})$						SVS ₂ with dCorr						
		\mathcal{P}_s			\mathcal{P}_a			\mathcal{P}_s			\mathcal{P}_a			\mathcal{P}_s			\mathcal{P}_a			
X_1	X_2	X_{12}	X_{22}	All	X_1	X_2	X_{12}	X_{22}	All	X_1	X_2	X_{12}	X_{22}	All	X_1	X_2	X_{12}	X_{22}	All	
$p = 2000$ and $\sigma_{ij} = 0.5^{ i-j }$																				
d_1	0.720	0.700	0.990	1.000	0.580	0.906	0.928	0.936	1.000	0.820	0.734	0.712	0.994	1.000	0.604	0.850	0.840	1.000	1.000	0.756
d_2	0.850	0.840	1.000	1.000	0.760	0.940	0.948	0.968	1.000	0.888	0.844	0.828	0.998	1.000	0.756	0.890	0.880	1.000	1.000	0.816
d_3	0.890	0.880	1.000	1.000	0.820	0.948	0.964	0.974	1.000	0.906	0.878	0.876	1.000	1.000	0.816					
$p = 2000$ and $\sigma_{ij} = 0.8^{ i-j }$																				
d_1	0.970	0.980	0.920	1.000	0.880	0.990	0.992	0.876	1.000	0.860	0.972	0.972	0.960	1.000	0.916	0.990	0.990	0.950	1.000	0.958
d_2	0.990	0.990	0.950	1.000	0.940	0.998	0.996	0.906	1.000	0.900	0.990	0.992	0.974	1.000	0.958	1.000	0.990	0.960	1.000	0.974
d_3	1.000	0.990	0.960	1.000	0.960	1.000	0.998	0.922	1.000	0.920	0.998	0.994	0.982	1.000	0.974					
$p = 5000$ and $\sigma_{ij} = 0.5^{ i-j }$																				
d_1	0.590	0.600	0.980	1.000	0.460	0.868	0.866	0.940	1.000	0.746	0.576	0.556	0.988	1.000	0.420	0.720	0.720	0.990	1.000	0.542
d_2	0.720	0.720	0.990	1.000	0.610	0.920	0.892	0.956	1.000	0.810	0.684	0.686	0.990	1.000	0.542	0.790	0.780	0.990	1.000	0.622
d_3	0.790	0.780	0.990	1.000	0.680	0.934	0.926	0.964	1.000	0.854	0.750	0.762	0.996	1.000	0.622					
$p = 5000$ and $\sigma_{ij} = 0.8^{ i-j }$																				
d_1	0.940	0.940	0.900	1.000	0.820	0.976	0.978	0.870	0.998	0.842	0.936	0.926	0.954	1.000	0.856	0.970	0.970	0.962	1.000	0.922
d_2	0.970	0.970	0.930	1.000	0.890	0.988	0.990	0.904	1.000	0.888	0.970	0.974	0.962	1.000	0.922	0.980	0.980	0.950	1.000	0.936
d_3	0.980	0.980	0.950	1.000	0.920	0.996	0.990	0.912	1.000	0.898	0.984	0.982	0.964	1.000	0.936					

Example 3.4.2. Generate $\mathbf{X} = (X_1, X_2, \dots, X_p)^T$ from standard multivariate normal distribution as follows:

$$P(Y = 1|\mathbf{X}) = \exp(g(\beta_4^T \mathbf{X})) / [1 + \exp(g(\beta_4^T \mathbf{X}))]$$

$$g(\beta_4^T \mathbf{X}) = \exp(5\beta_4^T \mathbf{X} - 2) / \{1 + \exp(5\beta_4^T \mathbf{X} - 3)\} - 1.5$$

with $\beta_4 = (1, 1, 0, \dots, 0)^T / \sqrt{2}$, $n = 200$ and vary p from 2000 to 5000. .

This example is a binary classification problem. We compare Kolmogorov filter method (Mai and Zou, 2013), MV-SIS method (Cui et al., 2015) and SVS₂ method with $R_c(\text{slice})$ as the measure for both marginal and conditional sequence. With higher probability to select both individual active predictors and all active predictors, compared with the two existing methods, SVS₂ has superior performance.

Table 3.2: Proportions comparison of \mathcal{P}_s and \mathcal{P}_a in example 3.4.2

	Kolmogorov filter			MV-SIS			SVS ₂ with $R_c^2(\text{slice})$		
	\mathcal{P}_s		\mathcal{P}_a	\mathcal{P}_s		\mathcal{P}_a	\mathcal{P}_s		\mathcal{P}_a
	X_1	X_2	All	X_1	X_2	All	X_1	X_2	All
$n = 200, p = 2000$									
d_1	0.926	0.920	0.850	0.950	0.968	0.924	0.952	0.966	0.918
d_2	0.956	0.964	0.920	0.968	0.976	0.944	0.974	0.986	0.960
d_3	0.968	0.978	0.946	0.976	0.980	0.956	0.982	0.990	0.972
$n = 200, p = 5000$									
d_1	0.820	0.860	0.702	0.900	0.884	0.796	0.916	0.924	0.846
d_2	0.874	0.888	0.772	0.944	0.930	0.880	0.946	0.952	0.900
d_3	0.902	0.916	0.824	0.952	0.938	0.894	0.960	0.962	0.922

Example 3.4.3. Let $Y = \mathbf{1}(\beta^T \mathbf{X} < -1) + 2\mathbf{1}(\beta^T \mathbf{X} > 2)$, where $\mathbf{1}(\cdot)$ is an indicator function and $\beta = (5, 5, 5, -15\rho^{1/2}, 0, \dots, 0)^T$. Generate \mathbf{X} from a multivariate normal distribution with mean zero and covariance $\Sigma = (\sigma_{ij})_{p \times p}$ with $\sigma_{ii} = 1$ for $i = 1, \dots, p$, $\sigma_{i4} = \sigma_{4i} = \rho^{1/2}$ for $i \neq 4$, and $\sigma_{ij} = \rho$, for $i \neq j$, $i \neq 4$, and $j \neq 4$. In this model, $n = 200$ and $p = 2000$.

This example is a classification problem with more than two outcomes. This covariance setup is similar to that in example 4 of Zhu et al. (2011). All predictors are equally correlated with correlation coefficient ρ except for X_4 and X_4 has correlation

$\rho^{1/2}$ with all the other predictors. Note that X_4 is marginally independent of Y , so that the marginal procedure can only pick up X_4 by chance, whereas X_4 is indeed an active predictor when $\rho \neq 0$. The conditional procedure can pick up X_4 correctly. Three variable selection methods are compared: fused Kolmogorov filter (Mai and Zou, 2015), MV-SIS and SVS₂ method. In order to see how the correlation among predictor variables will affect the variable selection result, we use different value of ρ to be 0, 0.1, 0.5 and 0.9.

With the probability to select all the active predictors in the fused Kolmogorov filter or MV-SIS method almost equal to 0, SVS₂ procedure has a very high probability to select all the active predictors. This example demonstrates that SVS₂ is indeed a very powerful tool in picking up active predictors that are marginally independent of the response, compared with marginal screening methods.

Example 3.4.4. Consider model $Y = \mathbf{1}(\boldsymbol{\beta}_1^T \mathbf{X} > 0) + 2\mathbf{1}(\boldsymbol{\beta}_2^T \mathbf{X} > 0)$, where $\mathbf{1}(\cdot)$ is an indicator function. Set $\boldsymbol{\beta}_1 = (1, 1, 0, 0, 0, \dots, 0)^T$, $\boldsymbol{\beta}_2 = (0, 0, 1, 1, 1, -3\rho^{7/10}/(\rho^{1/16} + 1), -3\rho^{7/10}/(\rho^{1/16} + 1), 0, \dots, 0)^T$. Generate \mathbf{X} from a multivariate normal distribution with mean zero and covariance $\boldsymbol{\Sigma} = (\sigma_{ij})_{p \times p}$ with $\sigma_{ii} = 1$ for $i = 1, \dots, p$, $\sigma_{67} = \sigma_{76} = \rho^{1/16}$, $\sigma_{i6} = \sigma_{6i} = \sigma_{7i} = \sigma_{i7} = \rho^{7/10}$ for $i \neq 1, 2, 6, 7$, and $\sigma_{ij} = \rho$, for $i \neq j$, $i \neq 1, 2, 6, 7$, and $j \neq 1, 2, 6, 7$. All other elements in the covariance matrix are zero. In this model, $n = 200$ and $p = 2000$.

We vary the value of ρ to be 0, 0.2, 0.5, 0.8 and 0.9 to see how the correlation among predictor variables will affect the variable selection result. Table 3.4 records the proportions \mathcal{P}_s and \mathcal{P}_a .

This example shows if the active predictors are from two dimensions and all predictors are highly correlated with each other, the marginal selection method will miss some important predictors while the sufficient variable selection method performs significantly better. It also indicates that our procedure will not be affected much by multi-dimensions.

Table 3.3: Proportions comparison of \mathcal{P}_s and \mathcal{P}_a in example 3.4.3

		fused Kolmogorov Filter						MV-SIS						SVS ₂ with R_c^2 (slice)						
		\mathcal{P}_s			\mathcal{P}_a			\mathcal{P}_s			\mathcal{P}_a			\mathcal{P}_s			\mathcal{P}_a			
X_1	X_2	X_3	X_4	All	X_1	X_2	X_3	X_4	All	X_1	X_2	X_3	X_4	All	X_1	X_2	X_3	X_4	All	
$\rho = 0$																				
d_1	0.990	0.984	0.994	NA	0.968	1.000	1.000	1.000	NA	1.000	1.000	1.000	1.000	NA	1.000	1.000	1.000	1.000	NA	1.000
d_2	0.994	0.992	1.000	NA	0.986	1.000	1.000	1.000	NA	1.000	1.000	1.000	1.000	NA	1.000	1.000	1.000	1.000	NA	1.000
d_3	0.998	0.994	1.000	NA	0.992	1.000	1.000	1.000	NA	1.000	1.000	1.000	1.000	NA	1.000	1.000	1.000	1.000	NA	1.000
$\rho = 0.1$																				
d_1	0.984	0.976	0.986	0.006	0.006	1.000	1.000	1.000	0.004	0.006	1.000	1.000	0.004	0.004	1.000	1.000	1.000	1.000	1.000	1.000
d_2	0.998	0.990	0.994	0.030	0.030	1.000	1.000	1.000	0.020	0.030	1.000	1.000	0.020	0.020	1.000	1.000	1.000	1.000	1.000	1.000
d_3	1.000	0.996	1.000	0.050	0.050	1.000	1.000	1.000	0.032	0.050	1.000	1.000	0.032	0.032	1.000	1.000	1.000	1.000	1.000	1.000
$\rho = 0.5$																				
d_1	0.892	0.874	0.896	0.004	0.004	0.976	0.978	0.978	0.000	0.004	0.976	0.978	0.000	0.000	0.984	0.988	0.986	1.000	0.968	0.968
d_2	0.930	0.918	0.930	0.006	0.004	0.984	0.988	0.984	0.004	0.004	0.984	0.984	0.004	0.004	0.992	0.996	0.990	1.000	0.982	0.982
d_3	0.946	0.938	0.952	0.010	0.008	0.990	0.992	0.990	0.008	0.008	0.990	0.992	0.008	0.006	0.994	0.998	0.992	1.000	0.986	0.986
$\rho = 0.9$																				
d_1	0.648	0.658	0.666	0.000	0.000	0.706	0.726	0.728	0.000	0.000	0.710	0.712	0.000	0.000	0.710	0.712	0.712	1.000	0.612	0.612
d_2	0.694	0.712	0.716	0.006	0.004	0.742	0.746	0.766	0.000	0.000	0.746	0.746	0.000	0.000	0.746	0.746	0.756	1.000	0.650	0.650
d_3	0.722	0.748	0.734	0.014	0.010	0.764	0.766	0.778	0.000	0.010	0.764	0.766	0.000	0.000	0.762	0.766	0.776	1.000	0.672	0.672

Leukaemia data analysis

In this section, we apply the screening approach of SVS₂ procedure on a leukaemia data set. The data set contains 72 samples and 7129 genes from high density Affymetrix oligonucleotide arrays. Among the subjects, 25 have acute myeloid leukemia (AML) and 47 have acute lymphoblastic leukemia (ALL).

The data is first analyzed by Golub et al. (1999) and then by Chiaromonte and Martinelli (2002), Dudoit et al. (2002) and Fan and Lv (2008), among others. It is available at <http://portals.broadinstitute.org/cgi-bin/cancer/publications/view/43>.

We treat the grouping of the subjects as the response variable and the explanatory variable has dimension $p = 7129$, which is much larger than sample size $n = 72$. We want to select the genes that can well separate the two groups.

Before applying any method, we preprocess the data following Golub et. al. (1999). Three preprocessing steps were applied as follows and 3194 genes were kept.

- (a) thresholding, gene expression readings of 100 or fewer were set to 100 and expression readings of 16000 or more were set to 16000;
- (b) screening, only genes where $\max_{\min} > 500$ and $\max/\min > 5$ were included, where \max and \min refer to the maximum and minimum readings of a gene expression among the 72 samples respectively;
- (c) transformation, gene expression readings of the genes selected were log-transformed, and were also standardized to have mean 0 and variance 1.

We first select $n - 1 = 71$ variables using SVS₂ procedure, then apply sliced inverse regression (SIR) method (Li, 1991) with sparse solution (Li, 2007) to reduce the dimension of selected $n - 1$ variables, and select active variables. We finally select 10 genes. Figure 1 shows the boxplot of the estimate of first direction by SIR using the 10 selected genes. From the plot, we clearly see the separation of the two groups.

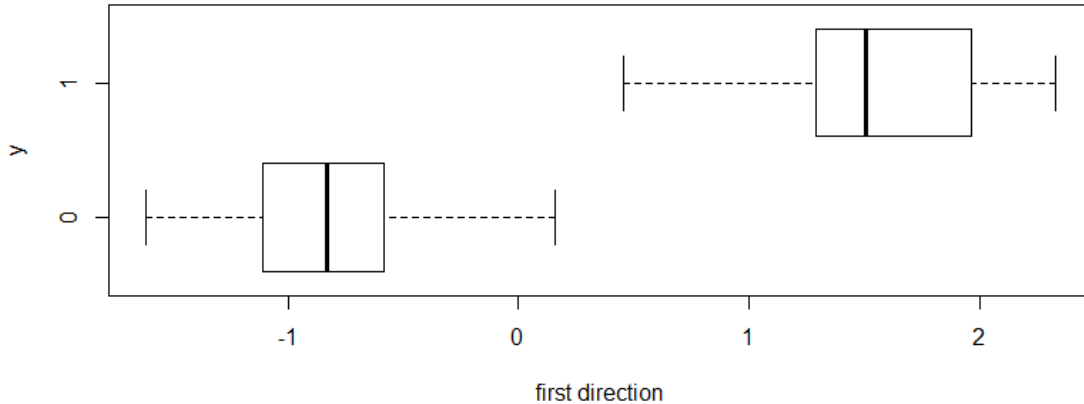


Figure 3.1: Boxplot by plotting the grouping on the first direction, where 0 is ALL group and 1 is the AML group.

3.5 Discussion

In this paper, we propose a novel two-stage sufficient variable selection procedure with screening approach, using a newly developed independence measure. This procedure provides a new aspect that does not rely on model assumption and better than those of SIS approaches, while inherits the advantages of model-free property. It is particularly useful when response is categorical or discrete, such as in classification or high dimensional discriminant analysis. In addition, the procedure can detect the active predictors which are marginally independent of the response, and it has an easier computation and interpretation compared with iterative methods (in marginal screening procedures). Although we do not use testing approach in the procedure, it can be implemented with fine statistical testing methods. We expect that the idea of sufficient variable selection shall lead to new research directions on variable selection.

Chapter 4 Sufficient Dimension Reduction in Big Data

4.1 Introduction

For the past 25 years, sufficient dimension reduction is a hot topic, many methods have been developed to estimate the central subspace (Cook, 1996). These methods can be classified into three classes: inverse, forward and joint regression methods. Inverse regression methods use the regression of $\mathbf{X}|\mathbf{Y}$, and require certain conditions on X , such as linearity condition and/or constant covariance condition. Specific methods include sliced inverse regression (SIR; Li, 1991), sliced average variance estimation (SAVE; Cook and Weisberg, 1991) and directional regression (DR; Li and Wang, 2007). Also see Zhu and Fang (1996), Fung et al. (2002), Yin and Cook (2003), Cook and Ni (2005), Li and Dong (2009), Dong and Li (2010) and Cook and Zhang (2015). The forward regression methods include the minimum average variance estimation (MAVE; Xia et al., 2002) and its variants, Xia (2007) and Wang and Xia (2008), average derivative estimate (Härdle and Stoker, 1989; Powell et al., 1989), Ichimura (1993), Härdle et al. (1993), Horowitz and Härdle (1996), structure adaptive method (Hristache et al., 2001) and Ma and Zhu (2013a). The forward methods require nonparametric approaches such as kernel smoothing. Joint regression methods require the joint distribution of (\mathbf{Y}, \mathbf{X}) , and methods include principal hessian direction (PHD; Li, 1992; Cook, 1998a), the fourier method (Zhu and Zeng, 2006), Zeng and Zhu (2010), Yin and Cook (2005) and Yin et al. (2008). They require either smoothing techniques or stronger conditions.

In this chapter, we develop a new sufficient dimension reduction method based on the measure in chapter 2 to estimate the central subspace. It is similar to the classical inverse approaches, such as SIR and SAVE, but without requiring any linear or constant variance condition and can exhaustively recover the central subspace without smoothing requirement. On the other hand, its algorithm keeps the advantage of Sheng and Yin (2013, 2016) needs no smoothing, and it requires very mild

conditional on the predictors. It is particularly useful when response is categorical, or discrete but its numerical value is not meaningful, compared with Sheng and Yin (2013, 2016).

This chapter is organized as follows: Section 4.2 includes a detailed description of the proposed method. Section 4.3 include some theoretical properties. Section 4.4 presents two simulation examples. The appendix contain the proofs.

4.2 Methodology

A measure of divergence

In chapter 2, we propose a new measure of divergence for independence between two random vectors. Let $\mathbf{X} \in \mathbb{R}^p$ and $\mathbf{Y} \in \mathbb{R}^q$, where p and q are positive integers, then the measure between \mathbf{X} and \mathbf{Y} with finite first moments is a nonnegative number, $\mathcal{C}(\mathbf{X}|\mathbf{Y})$, defined by

$$\mathcal{C}^2(\mathbf{X}|\mathbf{Y}) = \int_{\mathbb{R}^p} |f_{\mathbf{X}|\mathbf{Y}}(t) - f_{\mathbf{X}}(t)|^2 w(t) dt, \quad (4.1)$$

where $f_{\mathbf{X}|\mathbf{Y}}$ and $f_{\mathbf{X}}$ stand for the characteristic functions of $\mathbf{X}|\mathbf{Y}$ and \mathbf{X} , respectively. Let $|f|^2 = f\bar{f}$ for a complex-valued function f , with \bar{f} being the conjugate of f . The weight function $w(t)$ is a specially chosen positive function. More details of $w(t)$ can be found in chapter 2. They also give an equivalent formula as

$$\mathcal{C}^2(\mathbf{X}|\mathbf{Y}) = \mathbb{E}|\mathbf{X} - \mathbf{X}'_{\mathbf{Y}}| - \mathbb{E}|\mathbf{X}_{\mathbf{Y}} - \mathbf{X}'_{\mathbf{Y}}| = \mathbb{E}|\mathbf{X} - \mathbf{X}'| - \mathbb{E}|\mathbf{X}_{\mathbf{Y}} - \mathbf{X}'_{\mathbf{Y}}|, \quad (4.2)$$

where the expectation is over all random vectors. For instance, the last expectation is first taking the conditional expectation given \mathbf{Y} , then over \mathbf{Y} . $(\mathbf{X}', \mathbf{Y}')$ is an iid copy of (\mathbf{X}, \mathbf{Y}) , $\mathbf{X}_{\mathbf{Y}}$ denotes a random variable distributed as $\mathbf{X}|\mathbf{Y}$, $\mathbf{X}'_{\mathbf{Y}'}$ denotes a random variable distributed as $\mathbf{X}'|\mathbf{Y}'$ and $\mathbf{X}'_{\mathbf{Y}}$ denotes a random variable distributed as $\mathbf{X}'|\mathbf{Y}'$ with $\mathbf{Y}' = \mathbf{Y}$.

An attractive property of $\mathcal{C}^2(\mathbf{X}|\mathbf{Y})$ is that it equals 0 if and only if the two random vectors are independent 2). This property makes it possible that $\mathcal{C}^2(\mathbf{X}|\mathbf{Y})$ can be used

as a sufficient dimension reduction tool. What's more, the measure works well for both continuous and categorical \mathbf{Y} . This is particularly useful when the class index of dataset is not meaningful, where other measures do not attain similar power.

Review of sufficient dimension reduction

Let B be a matrix and $\mathcal{S}(B)$ be the subspace spanned by the column vectors of B . $\dim(\mathcal{S}(B))$ is the dimension of $\mathcal{S}(B)$. $P_{B(\Sigma_X)}$ denotes the projection operator which projects onto $\mathcal{S}(B)$ with respect to the inner product $\langle a, b \rangle = a^T \Sigma_X b$, that is, $P_{B(\Sigma_X)} = B(B^T \Sigma_X B)^{-1} B^T \Sigma_X$. Let $Q_{B(\Sigma_X)}$ be the projection of the orthogonal complement of $B(\Sigma_X)$. $Q_{B(\Sigma_X)} = I - P_{B(\Sigma_X)}$, where I is the identity matrix.

Let β be a $p \times q$ matrix with $q \leq p$, and \perp be the independence notation. The following conditional independence leads to the definition of sufficient dimension reduction:

$$\mathbf{Y} \perp \mathbf{X} | \beta^T \mathbf{X}, \quad (4.3)$$

where (4.3) indicates that the regression information of \mathbf{Y} given \mathbf{X} is completely contained in the linear combinations of \mathbf{X} , $\beta^T \mathbf{X}$. The column space of β in (4.3), denoted by $\mathcal{S}(\beta)$, is called a dimension reduction subspace.

If the intersection of all dimension reduction subspace is itself a dimension reduction subspace, then it is called the central subspace (CS), and it is denoted by $\mathcal{S}_{Y|X}$ (Li, 1991; Cook, 1994, 1996). Under mild conditions, CS exists (Yin et al., 2008). Throughout the chapter, we assume CS exists, which is unique. Furthermore, let d denote the structural dimension of the central subspace, and let $\Sigma_{\mathbf{X}}$ be the covariance matrix of \mathbf{X} , which is assumed to be nonsingular. Our primary goal is to identify the central subspace by estimating d and a $p \times d$ basis matrix B of CS.

The new sufficient dimension reduction method

Let β be an $p \times d_0$ arbitrary matrix, where $1 \leq d \leq p$. We will show that under mild conditions, solving (4.4) will yield a basis of the central subspace.

$$\max_{\substack{\beta^T \Sigma_{\mathbf{X}} \beta = I_d \\ 1 \leq d \leq p}} \mathcal{C}^2(\beta^T \mathbf{X} | \mathbf{Y}), \quad (4.4)$$

Here the squared divergence between $\beta^T \mathbf{X}$ and \mathbf{Y} is defined as

$$\mathcal{C}^2(\beta^T \mathbf{X} | \mathbf{Y}) = \int_{\mathbb{R}^{d+1}} |f_{\beta^T \mathbf{X} | \mathbf{Y}}(t) - f_{\beta^T \mathbf{X}}(t)|^2 w(t) dt.$$

The conditions $E|\mathbf{X}| < \infty$ and $E|\mathbf{X}_{\mathbf{Y}}| < \infty$ in chapter 2 guarantee that the $\mathcal{C}^2(\beta^T \mathbf{X} | \mathbf{Y})$ is finite, thus throughout the chapter we assume they hold. In the optimization problem (4.4), we use the constraint $\beta^T \Sigma_{\mathbf{X}} \beta = I_d$. The reason is that $\mathcal{C}^2(c\beta^T \mathbf{X} | \mathbf{Y}) = |c| \mathcal{C}^2(\beta^T \mathbf{X} | \mathbf{Y})$ for any constant c (2), and therefore we can always get a bigger value of $\mathcal{C}^2(\beta^T \mathbf{X} | \mathbf{Y})$ by multiplying β a constant with bigger absolute value, so we need a scale constraint to make the maximization procedure work.

We distinguish two cases, when $d = 1$ and when $d > 1$. For the single index of $d = 1$, we can explicitly have the inference, while for multi-index of $d > 1$, only projection matrix is identifiable, thus its inference may not be meaningful.

Single index

The following propositions ensure that if we maximize $\mathcal{C}^2(\beta^T \mathbf{X} | \mathbf{Y})$ with respect to β under the constraint, the solution indeed spans the CS.

Proposition 4.2.1. *Let η to be a basis of the central subspace $\mathcal{S}_{\mathbf{Y} | \mathbf{X}}$ and $\eta^T \Sigma_{\mathbf{X}} \eta = 1$. If $P_{\eta(\Sigma_{\mathbf{X}})}^T \mathbf{X} \perp Q_{\eta(\Sigma_{\mathbf{X}})}^T \mathbf{X}$, then $\mathcal{C}^2(\eta^T \mathbf{X} | \mathbf{Y}) \geq \mathcal{C}^2(\beta^T \mathbf{X} | \mathbf{Y})$ for any $\beta \in \mathbb{R}^p$ with $\beta^T \Sigma_{\mathbf{X}} \beta = 1$. The equality holds if and only if $\text{Span}(\beta) = \text{Span}(\eta)$.*

Multi-index

The following propositions ensure that if we maximize $\mathcal{C}^2(\boldsymbol{\beta}^T \mathbf{X} | \mathbf{Y})$ with respect to $\boldsymbol{\beta}$ under the constraint and some mild conditions, the solution indeed spans the central subspace.

Proposition 4.2.2. *Let $\boldsymbol{\eta}$ be a basis of the central subspace, $\boldsymbol{\beta}$ be a $p \times d_1$ matrix with $d_1 \leq d$, $\dim(\mathcal{S}(\boldsymbol{\beta})) = d_1$, $\boldsymbol{\eta}^T \Sigma_{\mathbf{X}} \boldsymbol{\eta} = I_d$ and $\boldsymbol{\beta}^T \Sigma_{\mathbf{X}} \boldsymbol{\beta} = I_{d_1}$. Assume $\mathcal{S}(\boldsymbol{\beta}) \subseteq \mathcal{S}(\boldsymbol{\eta})$, then $\mathcal{C}^2(\boldsymbol{\beta}^T \mathbf{X} | \mathbf{Y}) \leq \mathcal{C}^2(\boldsymbol{\eta}^T \mathbf{X} | \mathbf{Y})$. The equality holds if and only if $\mathcal{S}(\boldsymbol{\beta}) = \mathcal{S}(\boldsymbol{\eta})$.*

Proposition 4.2.3. *Let $\boldsymbol{\eta}$ be a basis of the central subspace, $\boldsymbol{\beta}$ be a $p \times d_2$ matrix with $\boldsymbol{\eta}^T \Sigma_{\mathbf{X}} \boldsymbol{\eta} = I_d$ and $\boldsymbol{\beta}^T \Sigma_{\mathbf{X}} \boldsymbol{\beta} = I_{d_2}$. Here d_2 could be bigger, less or equal to d . Suppose $P_{\boldsymbol{\eta}(\Sigma_{\mathbf{X}})}^T \mathbf{X} \perp\!\!\!\perp Q_{\boldsymbol{\eta}(\Sigma_{\mathbf{X}})}^T \mathbf{X}$ and $\mathcal{S}(\boldsymbol{\beta}) \not\subseteq \mathcal{S}(\boldsymbol{\eta})$, then $\mathcal{C}^2(\boldsymbol{\beta}^T \mathbf{X} | \mathbf{Y}) < \mathcal{C}^2(\boldsymbol{\eta}^T \mathbf{X} | \mathbf{Y})$.*

Proposition 4.2.2 indicates that if $\mathcal{S}(\boldsymbol{\beta})$ is a subspace of the central subspace $\mathcal{S}_{\mathbf{Y}|\mathbf{X}} = \mathcal{S}(\boldsymbol{\eta})$, then $\mathcal{C}^2(\boldsymbol{\beta}^T \mathbf{X} | \mathbf{Y})$ is always less or equal to $\mathcal{C}^2(\boldsymbol{\eta}^T \mathbf{X} | \mathbf{Y})$ and the equality holds if and only if $\boldsymbol{\beta}$ is also a basis matrix of the central subspace, i. e., $\mathcal{S}(\boldsymbol{\beta}) = \mathcal{S}(\boldsymbol{\eta})$. Proposition 4.2.3 indicates that if $\mathcal{S}(\boldsymbol{\beta})$ is not a subspace of the central subspace, then under a mild condition $\mathcal{C}^2(\boldsymbol{\beta}^T \mathbf{X} | \mathbf{Y})$ is always less than $\mathcal{C}^2(\boldsymbol{\eta}^T \mathbf{X} | \mathbf{Y})$. The above two propositions indicate that we can always identify the central subspace by maximizing $\mathcal{C}^2(\boldsymbol{\beta}^T \mathbf{X} | \mathbf{Y})$ with respect to $\boldsymbol{\beta}$ under the quadratic constraint. The independence condition, $P_{\boldsymbol{\eta}(\Sigma_{\mathbf{X}})}^T \mathbf{X} \perp\!\!\!\perp Q_{\boldsymbol{\eta}(\Sigma_{\mathbf{X}})}^T \mathbf{X}$, in proposition 4.2.3 is not as strong as it seems to be, and it could be satisfied asymptotically when p is reasonably large. Proofs for proposition 4.2.1, 4.2.2 and 4.2.3 are in the appendix.

Estimating the central subspace when d is specified

In this section, we propose an algorithm for estimating the central subspace when the structural dimension d is known. Let $(\mathbf{X}, \mathbf{Y}) = \{(X_i, Y_i), i = 1, \dots, n\}$ be a random sample from (\mathbf{X}, \mathbf{Y}) and let $\boldsymbol{\beta}$ be a $p \times d$ matrix. The sample version of $\mathcal{C}^2(\boldsymbol{\beta}^T \mathbf{X} | \mathbf{Y})$ denoted by $\mathcal{C}_n^2(\boldsymbol{\beta}^T \mathbf{X} | \mathbf{Y})$, has the following form:

$$\mathcal{C}_n^2(\mathbf{X}|\mathbf{Y}) = \frac{1}{n^2} \sum_{k,l=1}^{n,n} |\mathbf{X}_k - \mathbf{X}_l| - \frac{1}{n} \sum_{y=1}^H \frac{1}{n_y} \sum_{k,l=1}^{n_y, n_y} |\mathbf{X}_k - \mathbf{X}_l|. \quad (4.5)$$

Here $|\cdot|$ is the Euclidean norm in the respective dimension. Let $\hat{\Sigma}_{\mathbf{X}}$ be the sample version of Σ_X , then an estimated basis matrix of the central subspace, say $\boldsymbol{\eta}_n$, is

$$\boldsymbol{\eta}_n = \arg \max_{\boldsymbol{\beta}^T \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\beta} = I_d} \mathcal{C}_n^2(\boldsymbol{\beta}^T \mathbf{X}|\mathbf{Y}).$$

To find such an $\boldsymbol{\eta}_n$, we use Sequential Quadratic Programming method (SQP; Gill et al., 1981, Ch.6) to solve the above nonlinear optimization problem. The SQP procedure incorporated in MATLAB can be directly adopted in our algorithm. In this chapter, we use SIR, SAVE and LAD to estimate the initials and we choose the one, which gives the biggest squared distance covariance, as the final initial value.

Note that by invariance law, we can equivalently work on standardized predictor Z -scale, then transform back to X -scale. Indeed, propositions 4.2.1 and 4.2.2 hold for standardized predictor Z . Thus, we write the algorithm under Z scale, and we transform the estimate back into X scale later. This scheme seems to work well in our simulations. In the next section, we show the estimator $\boldsymbol{\eta}_n$ is consistent and asymptotically normal.

If we don't know the dimension, then it can be estimated by using bootstrap method (Ye and Weiss, 2003; Zhu and Zeng, 2006; Sheng and Yin, 2016).

4.3 Theoretical Properties

Single-index case

Proposition 4.3.1. *Let $\boldsymbol{\eta} \in \mathbb{R}^p$ to be a basis of the central subspace with $\boldsymbol{\eta}^T \Sigma_{\mathbf{X}} \boldsymbol{\eta} = 1$, and $\boldsymbol{\eta}_n = \arg \max_{\boldsymbol{\beta}^T \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\beta} = 1} \mathcal{C}_n^2(\boldsymbol{\beta}^T \mathbf{X}|\mathbf{Y})$. Assume $P_{\boldsymbol{\eta}(\Sigma_{\mathbf{X}})}^T \mathbf{X} \perp\!\!\!\perp Q_{\boldsymbol{\eta}(\Sigma_{\mathbf{X}})}^T \mathbf{X}$ and the support of $\mathbf{X} \in \mathbb{R}^p$, say S , is a compact set, then there exists a constant $c = 1$ or $c = -1$ such that $\boldsymbol{\eta}_n \xrightarrow{P} c\boldsymbol{\eta}$ as $n \rightarrow \infty$.*

Similar to the population level, $\mathcal{C}_n^2(\boldsymbol{\beta}^T \mathbf{X} | \mathbf{Y}) = \mathcal{C}_n^2(-\boldsymbol{\beta}^T \mathbf{X} | \mathbf{Y})$, thus maximizing $\mathcal{C}_n^2(\boldsymbol{\beta}^T \mathbf{X} | \mathbf{Y})$ with respect to $\boldsymbol{\beta}$ under the constraint will have two solutions: $\boldsymbol{\eta}_n$ or $-\boldsymbol{\eta}_n$, which, respectively, spans the same subspace. The purpose of using the constant $c = 1$ or $c = -1$ is to make sure that the first nonzero component of $\boldsymbol{\eta}_n$ and $c\boldsymbol{\eta}$ have the same sign.

In general, the support of \mathbf{X} doesn't have to be compact. However, Yin et al. (2008, proposition 11) showed that as long as compact set S is large enough, then $\mathcal{S}_{\mathbf{Y} | \mathbf{X}_s} = \mathcal{S}_{\mathbf{Y} | \mathbf{X}}$, where \mathbf{X}_s is \mathbf{X} restricted onto S . Hence we can restrict our discussion on a compact set S for simplifying the proof. Under such condition, $E|\mathbf{X}| < \infty$ holds, which together with $E|\mathbf{Y}| < \infty$ satisfy the definition of distance covariance. Proof of proposition 4.3.1 is given in the appendix. Indeed, we can further prove the \sqrt{n} -consistency and asymptotic normality of the estimator as stated below. And the proof of proposition 4.3.2 is again delayed in the appendix.

Proposition 4.3.2. *Let $\boldsymbol{\eta} \in \mathbb{R}^p$ to be a basis of the central subspace with $\boldsymbol{\eta}^T \Sigma_{\mathbf{X}} \boldsymbol{\eta} = 1$, and $\boldsymbol{\eta}_n = \arg \max_{\boldsymbol{\beta}^T \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\beta} = 1} \mathcal{C}_n^2(\boldsymbol{\beta}^T \mathbf{X} | \mathbf{Y})$. Under the same conditions as in proposition 4.2.2 and also the regularity conditions in the appendix, there exist a constant $c = 1$ or $c = -1$ such that $\sqrt{n}(\boldsymbol{\eta}_n - c\boldsymbol{\eta}) \rightarrow N(0, V_{11})$, where V_{11} is covariance matrix defined in the appendix.*

Multi-index case

Proposition 4.3.3. *Assume $\boldsymbol{\eta}$ is a basis matrix of the central subspace $\mathcal{S}_{\mathbf{Y} | \mathbf{X}}$ and $\boldsymbol{\eta}^T \Sigma_{\mathbf{X}} \boldsymbol{\eta} = I_d$. Suppose the support of X , say S , is compact, $E|Y| < \infty$ and $P_{\boldsymbol{\eta}(\Sigma_{\mathbf{X}})}^T \mathbf{X} \perp Q_{\boldsymbol{\eta}(\Sigma_{\mathbf{X}})}^T \mathbf{X}$. Let $\boldsymbol{\eta}_n = \arg \max_{\boldsymbol{\beta}^T \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\beta} = I_d} \mathcal{C}_n^2(\boldsymbol{\beta}^T \mathbf{X} | \mathbf{Y})$, then $\boldsymbol{\eta}_n$ is a consistent estimator of a basis of $\mathcal{S}_{\mathbf{Y} | \mathbf{X}}$, that is, there exists a rotation matrix \mathbf{Q} : $\mathbf{Q}^T \mathbf{Q} = I_d$, such that $\boldsymbol{\eta}_n \xrightarrow{P} \boldsymbol{\eta} \mathbf{Q}$.*

Proposition 4.3.4. *Assume $\boldsymbol{\eta}$ is a basis matrix of the central subspace $\mathcal{S}_{\mathbf{Y} | \mathbf{X}}$ and $\boldsymbol{\eta}^T \Sigma_{\mathbf{X}} \boldsymbol{\eta} = I_d$. Suppose the support of \mathbf{X} is compact, $E|\mathbf{Y}| < \infty$ and $P_{\boldsymbol{\eta}(\Sigma_{\mathbf{X}})}^T \mathbf{X} \perp Q_{\boldsymbol{\eta}(\Sigma_{\mathbf{X}})}^T \mathbf{X}$. Let $\boldsymbol{\eta}_n = \arg \max_{\boldsymbol{\beta}^T \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\beta} = I_d} \mathcal{C}_n^2(\boldsymbol{\beta}^T \mathbf{X} | \mathbf{Y})$, then under the regularity conditions given in the appendix, there exists a rotation matrix \mathbf{Q} : $\mathbf{Q}^T \mathbf{Q} = I_d$ such that $\sqrt{n}[\text{vec}(\boldsymbol{\eta}_n) -$*

$\text{vec}(\boldsymbol{\eta}\mathbf{Q})] \xrightarrow{\mathcal{D}} N(0, V_{11}(\boldsymbol{\eta}_Q))$, where $V_{11}(\boldsymbol{\eta}_Q)$ is the covariance matrix defined in the appendix.

Corollary 4.3.5. *Let $\boldsymbol{\eta}$ be a basis matrix of the central subspace and*

$\boldsymbol{\eta}_n = \arg \max_{\boldsymbol{\beta}^T \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\beta} = I_d} \mathcal{C}_n^2(\boldsymbol{\beta}^T \mathbf{X} | \mathbf{Y})$, then under the same assumptions and conditions in proposition 4.3.1, we have $\sqrt{n}[\text{vec}(\boldsymbol{\eta}_n \boldsymbol{\eta}_n^T \hat{\Sigma}) - \text{vec}(\boldsymbol{\eta} \boldsymbol{\eta}^T \Sigma)] \xrightarrow{\mathcal{D}} N(0, V_{22}(\boldsymbol{\eta}_Q))$, where $V_{22}(\boldsymbol{\eta}_Q)$ is the covariance matrix defined in the appendix.

4.4 Simulation Studies

Estimation accuracy is measured by $\Delta_m(\hat{\mathcal{S}}, \mathcal{S}) = \| \mathbf{P}_{\hat{\mathcal{S}}} - \mathbf{P}_{\mathcal{S}} \|$ (Li et al. (2005)), where \mathcal{S} is the real d -dimensional central subspace of \mathbb{R}^p , $\hat{\mathcal{S}}$ is the estimate, $\mathbf{P}_{\mathcal{S}}$, $\mathbf{P}_{\hat{\mathcal{S}}}$ are the orthogonal projections onto \mathcal{S} and $\hat{\mathcal{S}}$, respectively. And $\| \cdot \|$ is the maximum singular value of a matrix. The smaller the Δ_m is, the better the estimate is. Also a method works better if it has smaller standard error of Δ_m . The following two examples show the nice performance of the proposed method in terms of both continuous and categorical response, assuming we already know the dimension d .

Example 4.4.1. The model (model (A); Sheng and Yin (2016)) is $Y = (\beta_1^T \mathbf{X})^2 + (\beta_2^T \mathbf{X}) + 0.1\epsilon$, where $\mathbf{X} \sim N(\mathbf{0}, \mathbf{I}_p)$, $\epsilon \sim N(0, 1)$ and is independent of \mathbf{X} . $\beta_1 = (1, 0, \dots, 0)^T$, $\beta_2 = (0, 1, \dots, 0)^T$. We compare dCov (Sheng and Yin (2016)) with $R_c(\text{slice})$ (uses 6 slice when $n = 100$ and 10 slices for $n > 100$).

Assume there are two dimensions, table 4.1 shows the average estimation accuracy ($\bar{\Delta}_m$) and its standard error (SE) under different (n, p) combinations and 500 replications. Note that $R_c(\text{slice})$ performs consistently better than dCov, under all the different (n, p) combinations.

Table 4.1: Comparison of dimension reduction accuracy using dCov and $R_c(\text{slice})$

	(100,6)		(200,6)		(300,6)		(400,6)		(500,20)	
	dCov	$R_c(\text{slice})$	dCov	$R_c(\text{slice})$	dCov	$R_c(\text{slice})$	dCov	$R_c(\text{slice})$	dCov	$R_c(\text{slice})$
Δ_m	0.190	0.188	0.130	0.101	0.101	0.075	0.087	0.062	0.162	0.119
SE	0.059	0.078	0.039	0.032	0.029	0.023	0.026	0.019	0.026	0.020

Example 4.4.2. This example is an example for categorical Y . In a four groups balanced design, the total number of observations from all groups is n , \mathbf{X} has dimension $p = 6$, with marginal distributions independent. We set up the following scheme: X_1 follows non-central $t(4)$ distribution with non-centrality parameter $\delta = 5$ in the first group, and it follows central $t(4)$ distribution in the other groups. While, $X_2 \sim N(0, 1)$, $X_3 \sim U(0, 1)$, $X_4 \sim N(0, 1)$, $X_5 \sim \chi^2(1)$ and $X_6 \sim \chi^2(3)$ and each of these elements of X follows the same distribution across different groups.

Table 4.2 shows the average estimation accuracy ($\bar{\Delta}_m$) and its standard error (SE), under different (n, p) combinations and replicate 500 times, assume there is one dimension. The dimension reduction accuracy of $R_c(\text{slice})$ is consistently better than that of dCov.

Table 4.2: Comparison of dimension reduction accuracy using dCov and $R_c(\text{slice})$

	(100,6)		(200,6)		(300,6)		(400,6)	
	dCov	$R_c(\text{slice})$	dCov	$R_c(\text{slice})$	dCov	$R_c(\text{slice})$	dCov	$R_c(\text{slice})$
Δ_m	0.956	0.594	0.948	0.449	0.936	0.397	0.928	0.338
SE	0.076	0.187	0.096	0.189	0.100	0.173	0.105	0.152

4.5 Discussion

In this chapter, we propose a new sufficient dimension reduction method. It's asymptotic properties under single and multiple index cases are discussed. Simulation results show its advantage and it is particularly useful when Y is a categorical variable. Along this line, in the future, we will apply the framework of Yin and Hilafu (2015) for large p and small n problem, and further combine the penalized methods such as LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001) and Dantzig selector (Candes and Tao, 2007) for large p and small n data.

Appendix

Supplementary Materials for Chapter 2

This section provides materials related to the newly proposed index in section 2.3. It includes proofs of propositions and theorems stated in chapter 2, and additional simulation results.

Brownian Motion Approach

We use the discrepancy between the characteristic functions and a particular weight function to lead to our index (2.5). However, in this section, we show that a Brownian motion procedure also can derive our index (2.5).

Let W be a two-sided one-dimensional Brownian motion/Wiener process with expectation zero and covariance function $|s| + |t| - |s - t| = 2 \min(s, t)$, $s, t > 0$ (Székely and Rizzo, 2009, (3.3)).

Definition S. 4.5.1. *The Brownian conditional difference or the Wiener conditional difference of a real-valued random vector \mathbf{X} given \mathbf{Y} with finite second moments is a non-negative number defined by $\mathcal{D}_W^2(\mathbf{X}|\mathbf{Y}) = \mathbb{E}(\mathbf{X}_W \mathbf{X}'_W | \mathbf{Y})$, where W does not depend on $(\mathbf{X}, \mathbf{X}', \mathbf{Y})$.*

With this definition, we then have the following result.

Proposition S. 4.5.1. *If \mathbf{X} is an \mathbb{R}^p valued random vector, \mathbf{Y} is an \mathbb{R}^q valued random vector, and $\mathbb{E}[|\mathbf{X}|^2 + \mathbb{E}(|\mathbf{X}|^2 | \mathbf{Y})] < \infty$, then $\mathbb{E}(\mathbf{X}_W \mathbf{X}'_W | \mathbf{Y})$ is nonnegative and finite. Let \mathbf{X} and \mathbf{X}' be iid, and $\mathbf{X}_\mathbf{Y}$ and $\mathbf{X}'_\mathbf{Y}$ be iid; Expectations are taken over every random vector except conditioning on \mathbf{Y} if it appears. Then, (2.5) holds. That is,*

$$\mathcal{C}^2(\mathbf{X}|\mathbf{Y}) = \mathbb{E}[\mathcal{D}_W^2(\mathbf{X}|\mathbf{Y})].$$

Proof of Proposition S.4.5.1:

$$\begin{aligned}\mathcal{D}_W^2(\mathbf{X}|\mathbf{Y}) &= \mathbb{E}[\mathbb{E}(\mathbf{X}_W\mathbf{X}'_W|\mathbf{Y}, W)|\mathbf{Y}] = \mathbb{E}[\mathbb{E}(\mathbf{X}_W|\mathbf{Y}, W)\mathbb{E}(\mathbf{X}'_W|\mathbf{Y}, W)|\mathbf{Y}] \\ &= \mathbb{E}[\{\mathbb{E}(\mathbf{X}_W|\mathbf{Y}, W)\}^2|\mathbf{Y}],\end{aligned}$$

which is nonnegative. Finiteness can be obtained as Székely and Rizzo (2009, page 1262). Note that $\mathcal{D}_W^2(\mathbf{X}|\mathbf{Y}) = \mathbb{E}[\mathbb{E}(\mathbf{X}_W\mathbf{X}'_W|\mathbf{Y}, \mathbf{X}, \mathbf{X}')|\mathbf{Y}]$. Now using the same argument on page 1263 of Székely and Rizzo (2009), we have that

$$\mathbb{E}(\mathbf{X}_W\mathbf{X}'_W|\mathbf{Y}, \mathbf{X}, \mathbf{X}') = \mathbb{E}'|\mathbf{X}_Y - \mathbf{X}'| + \mathbb{E}|\mathbf{X}'_Y - \mathbf{X}| - |\mathbf{X}_Y - \mathbf{X}'_Y| - \mathbb{E}|\mathbf{X} - \mathbf{X}'|,$$

where the first expectation \mathbb{E}' is over \mathbf{X}' , the second expectation is over \mathbf{X} , and the last one is over both \mathbf{X} , and \mathbf{X}' . Thus, by using the fact that \mathbf{X} and \mathbf{X}' are iid, and \mathbf{X}_Y and \mathbf{X}'_Y are iid,

$$\mathcal{D}_W^2(\mathbf{X}|\mathbf{Y}) = \mathbb{E}[(\mathbb{E}'|\mathbf{X}_Y - \mathbf{X}'|)|\mathbf{Y}] + \mathbb{E}[(\mathbb{E}|\mathbf{X}'_Y - \mathbf{X}|)|\mathbf{Y}] - \mathbb{E}[(|\mathbf{X}_Y - \mathbf{X}'_Y|)|\mathbf{Y}] - \mathbb{E}|\mathbf{X} - \mathbf{X}'|.$$

By taking expectation over \mathbf{Y} , and the fact that the first term and the last term are equal, consequently, we have that $\mathcal{C}^2(\mathbf{X}|\mathbf{Y}) = \mathbb{E}[\mathcal{D}_W^2(\mathbf{X}|\mathbf{Y})]$. That is, again (2.5) holds. □

Relations to DISCO

Our index does not require Y to be discrete. However, if Y is categorical variable, then it is much intuitive and clear that our estimation method provides a close link to ANOVA, MANOVA and, most recently DISCO (Rizzo and Székely (2010)).

To be more specific, we can define the following population within distance and sample within distance, total distance and its sample version, respectively, where if we consider $e^{it^T\mathbf{X}_Y}$ as an observation, $\mathbb{E}(e^{it^T\mathbf{X}_Y})$ as the group mean and $\mathbb{E}(e^{it^T\mathbf{X}})$ as the overall mean.

Definition S. 4.5.2. *The population within distance is defined as:*

$$\mathcal{W}^2(\mathbf{X}|Y) = E[\mathcal{W}_w^2(\mathbf{X}|Y)] = E \int |e^{it^T \mathbf{X}_Y} - E e^{it^T \mathbf{X}_Y}|^2 w(t) dt;$$

The sample within distance is defined as:

$$\mathcal{W}_n^2(\mathbf{X}|Y) = \sum_{y=1}^H p_y \|e^{it^T \mathbf{X}_y} - f_{\mathbf{X}|y}^n(t)\|^2.$$

The population total distance is defined as:

$$\mathcal{T}^2(\mathbf{X}|Y) = E[\mathcal{T}_w^2(\mathbf{X}|Y)] = E \int |e^{it^T \mathbf{X}_Y} - E e^{it^T \mathbf{X}}|^2 w(t) dt;$$

The sample total distance is defined as:

$$\mathcal{T}_n^2(\mathbf{X}|Y) = \sum_{y=1}^H p_y \|e^{it^T \mathbf{X}_y} - f_{\mathbf{X}}^n(t)\|^2.$$

We can have their respective equivalent formulas, stated below.

Proposition S. 4.5.2. *The population within distance can be rewritten as:*

$$\mathcal{W}^2(\mathbf{X}|Y) = E[\mathcal{W}_w^2(\mathbf{X}|Y)] = E|\mathbf{X}_Y - \mathbf{X}'_Y|;$$

The sample within distance can be rewritten as:

$$\mathcal{W}_n^2(\mathbf{X}|Y) = \frac{1}{n} \sum_{y=1}^H \frac{1}{n_y} \sum_{k_y, l_y=1}^{n_y, n_y} |\mathbf{X}_{y, k_y} - \mathbf{X}_{y, l_y}|.$$

The population total distance can be rewritten as:

$$\mathcal{T}^2(\mathbf{X}|Y) = \mathcal{C}^2(\mathbf{X}|\mathbf{X}) = E|\mathbf{X} - \mathbf{X}'|;$$

The sample total distance can be rewritten as:

$$\mathcal{T}_n^2(\mathbf{X}|Y) = \frac{1}{n^2} \sum_{y, y'=1}^{H, H} \sum_{k_y, l_{y'}=1}^{n_y, n_{y'}} |\mathbf{X}_{y, k_y} - \mathbf{X}_{y', l_{y'}}|.$$

The following result is a straightforward calculation, thus we omitted its proof.

Proposition S. 4.5.3. 1. $\mathcal{T}^2(\mathbf{X}|Y) = \mathcal{C}^2(\mathbf{X}|Y) + \mathcal{W}^2(\mathbf{X}|Y);$

2. $\mathcal{T}_n^2(\mathbf{X}|Y) = \mathcal{C}_n^2(\mathbf{X}|Y) + \mathcal{W}_n^2(\mathbf{X}|Y).$

Under the null hypothesis, by SLLN, as $n \rightarrow \infty$, $\mathcal{W}_n^2(\mathbf{X}|Y) \rightarrow E|\mathbf{X} - \mathbf{X}'|$. Or note that $E[\mathcal{T}_n^2(\mathbf{X}|Y)] = E[\mathcal{W}_n^2(\mathbf{X}|Y)]$, thus analogous to ANOVA, we may use test statistic,

$$\frac{\mathcal{C}_n^2(\mathbf{X}|Y)/(H-1)}{\mathcal{W}_n^2(\mathbf{X}|Y)/(n-H)},$$

which is the ratio of between distance over within distance. Note that the previous test statistic in Section 2.4,

$$\frac{n\mathcal{C}_n^2(\mathbf{X}|Y)}{S_n} = \frac{\mathcal{C}_n^2(\mathbf{X}|Y)/(H-1)}{\mathcal{T}_n^2(\mathbf{X}|Y)/n} = \frac{n}{n-1} \frac{\mathcal{C}_n^2(\mathbf{X}|Y)/(H-1)}{\mathcal{T}_n^2(\mathbf{X}|Y)/(n-1)}.$$

With negligible factor $\frac{n}{n-1}$, this is the ratio of between distance over total distance. Note that $\frac{n\mathcal{C}_n^2(\mathbf{X}|Y)}{S_n} \frac{(H-1)}{n} = R_{c,n}^2$, an estimator of R_c^2 .

In particular, one can show that for response with two categories, the energy distance of Rizzo and Székely (2010, page 1038) is proportion to $\mathcal{C}^2(\mathbf{X}|Y)$. Indeed, one also can show that $n\mathcal{C}_n^2(\mathbf{X}|Y) = 2S_\alpha$ and $n\mathcal{W}^2(\mathbf{X}|Y) = 2W_\alpha$, with $\alpha = 1$, where S_α and W_α are defined in Rizzo and Székely (2010).

Classical methods of ANOVA or MANVOA for multi-sample usually require normally distributed error (see, e.g., Cochran and Cox (1957); Hand and Taylor (1987); Mardia et al. (1979)), especially for inference. When such condition fails, one may apply F statistics via permutation test procedure (Efron and Tibshirani (1998); Davison and Hinkley (1997)). Rich literature exists in beyond testing the mean differences but on distributions, for instance, Akritas and Arnold (1994) and Gower and Krzanowski (1999) for structured data, and Anderson (2001), McArdle and Anderson (2001), Excoffier et al. (1992) and Zapala and Schork (2006) with applications in ecology and genetics.

The class of α -divergence

We also can extend our measure (2.5) to a one parameter family of measures indexed with a positive exponent α . Note that in our previous application the exponent $\alpha = 1$.

Suppose that $E|\mathbf{X}_Y|^\alpha < \infty$. Let $\mathcal{C}^{(\alpha)}(\mathbf{X}|Y)$ denote the α -measure which is the nonnegative number defined by

$$\mathcal{C}^{(\alpha)}(\mathbf{X}|Y) = E_Y \|f_{\mathbf{X}|Y}(t) - f_{\mathbf{X}}(t)\|_\alpha^2 = E_Y \int_{\mathbb{R}^p} \frac{|f_{\mathbf{X}|Y}(t) - f_{\mathbf{X}}(t)|^2}{\tilde{C}(p, \alpha)|t|^{\alpha+p}} dt.$$

The α -measure statistics are defined by replacing the exponent 1 with expo-

nent α in the respective formulas (2.5) and (2.7). That is, for instance, in (2.7) replace $|\mathbf{X}_{y,k_y} - \mathbf{X}_{y',l_{y'}}|$ by $|\mathbf{X}_{y,k_y} - \mathbf{X}_{y',l_{y'}}|^\alpha$. Lemma 2.4.4 can be generalized for $\|\cdot\|_\alpha$ -norms, so that almost surely convergence of $\mathcal{C}_n^{2(\alpha)}(\mathbf{X}|Y) \rightarrow \mathcal{C}^{2(\alpha)}(\mathbf{X}|Y)$ follows if the α -moments are finite. Similarly one can prove the weak convergence and statistical consistency for α exponents, $0 < \alpha < 2$, provided that α moments are finite. However, when $\alpha = 2$, it leads to $2\mathbb{E}(\mu_Y - \mu)^2$, where μ_Y is the mean for group Y and μ is the overall mean. Thus in such a case, $\mathcal{C}^{2(2)}(\mathbf{X}|Y) = 0$ iff $\mu_Y = \mu$ for all Y . Furthermore, for $0 < \alpha \leq 2$, $n\mathcal{C}_n^{2(\alpha)}(\mathbf{X}|Y) = 2S_\alpha$ and $n\mathcal{W}_n^{2(\alpha)}(\mathbf{X}|Y) = 2W_\alpha$, where S_α and W_α are defined in Rizzo and Székely (2010).

One can consider the Levy fractional Brownian motion $\{W_H^d(t), t \in \mathbb{R}^d\}$, with Hurst index $H \in (0, 1)$, which is a centered Gaussian random process with covariance function (Herbin and Merzbach, 2007):

$$\mathbb{E}[W_H^d(t)W_H^d(s)] = |t|^{2H} + |s|^{2H} - |t - s|^{2H}, t, s \in \mathbb{R}^d.$$

Using Lemma 1 of Székely and Rizzo (2009), we can show that under $\mathbb{E}|\mathbf{X}|^{2h} < \infty$ and $\mathbb{E}|\mathbf{X}_Y|^{2h} < \infty$, for Hurst parameters $0 < H \leq 1$, and $h = 2H$ ($0 < h \leq 2$),

$$\mathcal{C}_{W_H^d}^2(\mathbf{X}|Y) = \mathbb{E}_Y \int_{\mathbb{R}^p} \frac{|f_{\mathbf{X}|Y}(t) - f_{\mathbf{X}}(t)|^2}{\tilde{C}(p, h)|t|^{h+p}} dt = \mathbb{E}|\mathbf{X} - \mathbf{X}'|^h - \mathbb{E}|\mathbf{X}_Y - \mathbf{X}'_Y|^h.$$

When $h = 1$, that is our Theorem 3.1. Theories for $0 < \alpha < 2$ can be established similarly.

Proofs of results for chapter 2

Proof of Lemma 2.2.1: If $\mathbf{X} \perp\!\!\!\perp \mathbf{Y}$, then $f_{\mathbf{X}|Y}(t) = \mathbb{E}[e^{it^T \mathbf{X}} | Y] = \mathbb{E}[e^{it^T \mathbf{X}}] = f_{\mathbf{X}}(t)$. Thus $\mathcal{C}_{w, Y}^2(\mathbf{X}|Y) = 0$, so does $\mathcal{C}^2(\mathbf{X}|Y)$. On the other hand, if $\mathcal{C}^2(\mathbf{X} | \mathbf{Y}) = 0$, then it implies that $\mathcal{C}_{w, Y}^2(\mathbf{X} | \mathbf{Y}) = 0$ almost surely for \mathbf{Y} . Hence, $f_{\mathbf{X}|Y}(t) = f_{\mathbf{X}}(t)$ almost

surely for t . Let $s \in \mathbb{R}^q$, then $e^{is^T \mathbf{Y}} f_{\mathbf{X}|\mathbf{Y}}(t) = e^{is^T \mathbf{Y}} f_{\mathbf{X}}(t)$. Hence,

$$\begin{aligned} \mathbb{E}(e^{is^T \mathbf{Y}} \mathbb{E}[e^{it^T \mathbf{X}} | \mathbf{Y}]) &= \mathbb{E}(e^{is^T \mathbf{Y}} \mathbb{E}[e^{it^T \mathbf{X}}]) \\ \mathbb{E}[e^{is^T \mathbf{Y}} e^{it^T \mathbf{X}}] &= \mathbb{E}(e^{is^T \mathbf{Y}}) \mathbb{E}[e^{it^T \mathbf{X}}] \\ f_{\mathbf{X}, \mathbf{Y}}(t, s) &= f_{\mathbf{X}}(t) f_{\mathbf{Y}}(s) \end{aligned}$$

That means, $\mathbf{X} \perp\!\!\!\perp \mathbf{Y}$. □

Proof of Theorem 2.2.2:

1. $\mathcal{C}^2(\mathbf{X}|\mathbf{X}) = 0$ iff $e^{it^T \mathbf{X}} = \mathbb{E}[e^{it^T \mathbf{X}}]$ almost surely for \mathbf{X}, t ; Note that the right hand side is constant with regards to \mathbf{X} . Hence, \mathbf{X} must be a constant. And $\mathbf{X} = \mathbb{E}(\mathbf{X})$ almost surely. If $\mathbf{X} = \mathbb{E}(\mathbf{X})$ almost surely, the result is obvious.
2. For simplicity, in the following we omit the term $w(t)dt$ in the integrals. Note that by using the independence of $(\mathbf{W}_1, \mathbf{V}_1)$ and $(\mathbf{W}_2, \mathbf{V}_2)$, suppose $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^p$, $\mathbf{V}_1, \mathbf{V}_2 \in \mathbb{R}^q$, we have:

$$\begin{aligned} &\mathcal{C}^2(\mathbf{W}_1 + \mathbf{W}_2 | \mathbf{V}_1 + \mathbf{V}_2) \\ &= \mathbb{E}_{\mathbf{V}_1 + \mathbf{V}_2} \int |f_{\mathbf{W}_1 + \mathbf{W}_2 | \mathbf{V}_1 + \mathbf{V}_2} - f_{\mathbf{W}_1 + \mathbf{W}_2}|^2 \\ &= \mathbb{E}_{\mathbf{V}_1 + \mathbf{V}_2} \int |\mathbb{E}[(\mathbb{E} e^{it^T \mathbf{W}_1 + it^T \mathbf{W}_2} | \mathbf{V}_1, \mathbf{V}_2) | \mathbf{V}_1 + \mathbf{V}_2] - f_{\mathbf{W}_1} f_{\mathbf{W}_2}|^2. \end{aligned}$$

Now apply Propositions 4.6 and 4.5 of Cook (1998b), we have $\mathbf{W}_1 \perp\!\!\!\perp \mathbf{W}_2 | (\mathbf{V}_1, \mathbf{V}_2)$.

Hence,

$$\dots = \mathbb{E}_{\mathbf{V}_1 + \mathbf{V}_2} \int |\mathbb{E}[(\mathbb{E} e^{it^T \mathbf{W}_1} | \mathbf{V}_1, \mathbf{V}_2) \mathbb{E}(e^{it^T \mathbf{W}_2} | \mathbf{V}_1, \mathbf{V}_2) | \mathbf{V}_1 + \mathbf{V}_2] - f_{\mathbf{W}_1} f_{\mathbf{W}_2}|^2.$$

Use $(\mathbf{W}_1, \mathbf{V}_1) \perp \mathbf{V}_2$, we further have

$$\begin{aligned}
\cdots &= \mathbf{E}_{\mathbf{V}_1 + \mathbf{V}_2} \int |\mathbf{E}[f_{\mathbf{W}_1|\mathbf{V}_1} f_{\mathbf{W}_2|\mathbf{V}_2} | \mathbf{V}_1 + \mathbf{V}_2] - f_{\mathbf{W}_1} f_{\mathbf{W}_2}|^2 \\
&= \mathbf{E}_{\mathbf{V}_1 + \mathbf{V}_2} \int |\mathbf{E}[(f_{\mathbf{W}_1|\mathbf{V}_1} - f_{\mathbf{W}_1}) f_{\mathbf{W}_2|\mathbf{V}_2} + f_{\mathbf{W}_1} f_{\mathbf{W}_2|\mathbf{V}_2} | \mathbf{V}_1 + \mathbf{V}_2] - f_{\mathbf{W}_1} f_{\mathbf{W}_2}|^2 \\
&= \mathbf{E}_{\mathbf{V}_1 + \mathbf{V}_2} \int |\mathbf{E}[(f_{\mathbf{W}_1|\mathbf{V}_1} - f_{\mathbf{W}_1}) f_{\mathbf{W}_2|\mathbf{V}_2} | \mathbf{V}_1 + \mathbf{V}_2] + f_{\mathbf{W}_1} \mathbf{E}[f_{\mathbf{W}_2|\mathbf{V}_2} - f_{\mathbf{W}_2} | \mathbf{V}_1 + \mathbf{V}_2]|^2
\end{aligned}$$

Let $a = \mathbf{E}[(f_{\mathbf{W}_1|\mathbf{V}_1} - f_{\mathbf{W}_1}) f_{\mathbf{W}_2|\mathbf{V}_2} | \mathbf{V}_1 + \mathbf{V}_2]$, $b = f_{\mathbf{W}_1} \mathbf{E}[f_{\mathbf{W}_2|\mathbf{V}_2} - f_{\mathbf{W}_2} | \mathbf{V}_1 + \mathbf{V}_2]$,

$$\cdots = \mathbf{E} \int |a|^2 + 2\mathbf{E} \int |ab| + \mathbf{E} \int |b|^2.$$

By using Cauchy-Schwarz inequality twice $\mathbf{E} \int |ab| \leq (\mathbf{E} \int |a|^2 \mathbf{E} \int |b|^2)^{1/2}$,

$$\cdots \leq (\mathbf{E} \int |a|^2)^{1/2} + (\mathbf{E} \int |b|^2)^{1/2}.$$

That is,

$$\mathcal{C}(\mathbf{W}_1 + \mathbf{W}_2 | \mathbf{V}_1 + \mathbf{V}_2) \leq (\mathbf{E} \int |a|^2)^{1/2} + (\mathbf{E} \int |b|^2)^{1/2}. \quad (\text{S.4.5.6})$$

By applying conditional Hölder's inequality, separately on a and b with power 2, we have

$$\begin{aligned}
&\mathcal{C}(\mathbf{W}_1 + \mathbf{W}_2 | \mathbf{V}_1 + \mathbf{V}_2) \\
&\leq (\mathbf{E} \int |f_{\mathbf{W}_1|\mathbf{V}_1} - f_{\mathbf{W}_1}|^2)^{1/2} + (\mathbf{E} \int |f_{\mathbf{W}_2|\mathbf{V}_2} - f_{\mathbf{W}_2}|^2)^{1/2} \quad (\text{S.4.5.7}) \\
&= \mathcal{C}(\mathbf{W}_1 | \mathbf{V}_1) + \mathcal{C}(\mathbf{W}_2 | \mathbf{V}_2).
\end{aligned}$$

We can see that if (i) \mathbf{W}_1 and \mathbf{V}_1 are both constant, (ii) \mathbf{W}_2 and \mathbf{V}_2 are both constant, or (iii) \mathbf{W}_1 , \mathbf{V}_1 , \mathbf{W}_2 and \mathbf{V}_2 are mutually independent, then we have the equality. On the other hand, if we have the equality, then we must have equality in (S.4.5.6) and (S.4.5.7), which implies (i) or (ii) holds. If none of the (i) and (ii) conditions is satisfied, the equality holds only if \mathbf{W}_1 and \mathbf{V}_1 , and

\mathbf{W}_2 and \mathbf{V}_2 are independent, but $\mathbf{W}_1, \mathbf{V}_1$ and $\mathbf{W}_2, \mathbf{V}_2$ are already independent, so they must be mutually independent. We complete the proof.

3. This follows from item 2. above by choosing $\mathbf{W}_1 = \mathbf{V}_1 = \mathbf{X}$, and $\mathbf{W}_2 = \mathbf{V}_2 = \mathbf{Y}$. And the independence in item 2. means (i) \mathbf{X} is constant; or (ii) \mathbf{Y} is constant; or (iii) both of them are constant, because this is the only case when a random vector can be independent of itself.

4. Note that by definition,

$$\begin{aligned}
\mathcal{C}^2(\mathbf{X}|\mathbf{Y}) &= \mathbb{E}_{\mathbf{Y}} \left[\int_{\mathbb{R}^p} |f_{\mathbf{X}|\mathbf{Y}}(t) - f_{\mathbf{X}}(t)|^2 w(t) dt \right] \\
&= \mathbb{E}_{\mathbf{Y}} \left[\int_{\mathbb{R}^p} (\mathbb{E} e^{it^T \mathbf{X}_{\mathbf{Y}}} - \mathbb{E} e^{it^T \mathbf{X}}) (\mathbb{E} e^{-it^T \mathbf{X}_{\mathbf{Y}}} - \mathbb{E} e^{-it^T \mathbf{X}}) w(t) dt \right] \\
&= \mathbb{E}_{\mathbf{Y}} \left[\int_{\mathbb{R}^p} (\mathbb{E} e^{it^T (\mathbf{X}_{\mathbf{Y}} - \mathbf{X}'_{\mathbf{Y}})} - \mathbb{E} e^{it^T (\mathbf{X} - \mathbf{X}'_{\mathbf{Y}})} - \mathbb{E} e^{it^T (\mathbf{X}_{\mathbf{Y}} - \mathbf{X}')} + \mathbb{E} e^{it^T (\mathbf{X} - \mathbf{X}')}) w(t) dt \right] \\
&= \mathbb{E}_{\mathbf{Y}} \left[\int_{\mathbb{R}^p} -\{1 - \mathbb{E} e^{it^T (\mathbf{X}_{\mathbf{Y}} - \mathbf{X}'_{\mathbf{Y}})}\} + \{1 - \mathbb{E} e^{it^T (\mathbf{X} - \mathbf{X}'_{\mathbf{Y}})}\} \right. \\
&\quad \left. + \{1 - \mathbb{E} e^{it^T (\mathbf{X}_{\mathbf{Y}} - \mathbf{X}')}\} - \{1 - \mathbb{E} e^{it^T (\mathbf{X} - \mathbf{X}')}\} w(t) dt \right] \\
&= \mathbb{E}_{\mathbf{Y}} \left[-\mathbb{E} \int_{\mathbb{R}^p} \{1 - \cos[t^T (\mathbf{X}_{\mathbf{Y}} - \mathbf{X}'_{\mathbf{Y}})]\} w(t) dt \right] \\
&\quad + \mathbb{E}_{\mathbf{Y}} \left[\mathbb{E} \int_{\mathbb{R}^p} \{1 - \cos[t^T (\mathbf{X} - \mathbf{X}'_{\mathbf{Y}})]\} w(t) dt \right] \\
&\quad + \mathbb{E}_{\mathbf{Y}} \left[\mathbb{E} \int_{\mathbb{R}^p} \{1 - \cos[t^T (\mathbf{X}_{\mathbf{Y}} - \mathbf{X}')] \} w(t) dt \right] \\
&\quad - \mathbb{E}_{\mathbf{Y}} \left[\mathbb{E} \int_{\mathbb{R}^p} \{1 - \cos[t^T (\mathbf{X} - \mathbf{X}')] \} w(t) dt \right]
\end{aligned}$$

Note that the last three terms are equal

$$\begin{aligned}
&= \mathbb{E}_{\mathbf{Y}} \left[\mathbb{E} \int_{\mathbb{R}^p} \{1 - \cos[t^T (\mathbf{X} - \mathbf{X}')] \} w(t) dt \right] \\
&\quad - \mathbb{E}_{\mathbf{Y}} \left[\mathbb{E} \int_{\mathbb{R}^p} \{1 - \cos[t^T (\mathbf{X}_{\mathbf{Y}} - \mathbf{X}'_{\mathbf{Y}})] \} w(t) dt \right] \\
&= \mathbb{E} \int_{\mathbb{R}^p} \{1 - \cos[t^T (\mathbf{X} - \mathbf{X}')] \} w(t) dt - \mathbb{E}_{\mathbf{Y}} \left[\mathbb{E} \int_{\mathbb{R}^p} \{1 - \cos[t^T (\mathbf{X}_{\mathbf{Y}} - \mathbf{X}'_{\mathbf{Y}})] \} w(t) dt \right] \\
&\leq \mathbb{E} \int_{\mathbb{R}^p} \{1 - \cos[t^T (\mathbf{X} - \mathbf{X}')] \} w(t) dt.
\end{aligned}$$

However,

$$\begin{aligned}
\mathcal{C}^2(\mathbf{X}|\mathbf{X}) &= \mathbb{E}[\mathcal{C}_{w,\mathbf{X}}^2(\mathbf{X}|\mathbf{X})] = \mathbb{E} \int |e^{it^T\mathbf{X}} - \mathbb{E}e^{it^T\mathbf{X}}|^2 w(t) dt \\
&= \mathbb{E} \int (1 - e^{it^T\mathbf{X}} \mathbb{E}e^{-it^T\mathbf{X}} - e^{-it^T\mathbf{X}} \mathbb{E}e^{it^T\mathbf{X}} + \mathbb{E}e^{it^T\mathbf{X}} \mathbb{E}e^{-it^T\mathbf{X}}) w(t) dt \\
&= \int (1 - \mathbb{E}e^{it^T\mathbf{X}} \mathbb{E}e^{-it^T\mathbf{X}}) w(t) dt = \mathbb{E} \int_{\mathbb{R}^p} \{1 - \cos[t^T(\mathbf{X} - \mathbf{X}')]\} w(t) dt.
\end{aligned}$$

Hence, conclusion follows. Consequently, $0 \leq R_c \leq 1$. \square

Proof of Theorem 2.3.1: By the proof in part 4 of Theorem 2.2.2 and Lemma 1 of Székely et al. (2007), we have

$$\begin{aligned}
\mathcal{C}^2(\mathbf{X}|\mathbf{Y}) &= \mathbb{E} \int_{\mathbb{R}^p} \{1 - \cos[t^T(\mathbf{X} - \mathbf{X}')]\} w(t) dt - \mathbb{E}_{\mathbf{Y}} [\mathbb{E} \int_{\mathbb{R}^p} \{1 - \cos[t^T(\mathbf{X}_{\mathbf{Y}} - \mathbf{X}'_{\mathbf{Y}})]\} w(t) dt] \\
&= \mathbb{E}|\mathbf{X} - \mathbf{X}'| - \mathbb{E}|\mathbf{X}_{\mathbf{Y}} - \mathbf{X}'_{\mathbf{Y}}|.
\end{aligned}$$

The last equality holds. Because $\mathbb{E}|\mathbf{X} - \mathbf{X}'| = \mathbb{E}_{\mathbf{Y}} \mathbb{E}[|\mathbf{X} - \mathbf{X}'| | \mathbf{Y}] = \mathbb{E}|\mathbf{X}_{\mathbf{Y}} - \mathbf{X}'|$, and hence, $\mathbb{E}|\mathbf{X}_{\mathbf{Y}} - \mathbf{X}'| = \mathbb{E}|\mathbf{X} - \mathbf{X}'|$, which immediately indicates that the first equality in (2.5) holds. Thus we complete the proof. \square

Proof of Theorem 2.3.2:

1. This can be proved easily by plugging \mathbf{X} for \mathbf{Y} in the second formula of (2.5). Because, $\mathbb{E}|\mathbf{X}_{\mathbf{Y}} - \mathbf{X}'_{\mathbf{Y}}| = \mathbb{E}_{\mathbf{y}} \mathbb{E}[|\mathbf{X} - \mathbf{X}'| | \mathbf{Y} = \mathbf{y}, \mathbf{Y}' = \mathbf{y}]$. If $\mathbf{X} = \mathbf{Y}$, then $\mathbf{X}' = \mathbf{Y}'$ and $\mathbf{X}' = \mathbf{Y}' = \mathbf{Y} = \mathbf{X}$. Hence, $\mathbb{E}|\mathbf{X}_{\mathbf{Y}} - \mathbf{X}'_{\mathbf{Y}}| = 0$. Or by the proof in part 4 of Theorem 2.2.2, and Lemma 1 in Székely et al. (2007) we have $\mathcal{C}^2(\mathbf{X}|\mathbf{X}) = \int (1 - \mathbb{E}e^{it^T\mathbf{X}} \mathbb{E}e^{-it^T\mathbf{X}}) w(t) dt = \mathbb{E}|\mathbf{X} - \mathbf{X}'|$.
2. By using formula (2.5), and note that $\mathbf{B}^T \mathbf{B} = I_p$, we can prove it easily.
3. If $\mathbf{X} = \mathbf{g}(\mathbf{Y})$, for some function \mathbf{g} , then $\mathbf{X}_{\mathbf{Y}} = \mathbf{X}'_{\mathbf{Y}}$. Thus the second term in $\mathcal{C}^2(\mathbf{X}|\mathbf{Y})$ must be 0. Therefore, $\mathcal{C}^2(\mathbf{X}|\mathbf{Y}) = \mathcal{C}^2(\mathbf{X}|\mathbf{X})$, implying that $R_c = 1$. On the other hand, if $R_c = 1$, then the second term in $\mathcal{C}^2(\mathbf{X}|\mathbf{Y})$ must be 0, which means that almost surely for \mathbf{Y} , there is only one \mathbf{X} corresponding to such a value of \mathbf{Y} . Thus, $\mathbf{X} = \mathbf{g}(\mathbf{Y})$. \square

Section 2.3: Conditional normal distribution:

$$\begin{aligned}
\pi\mathcal{C}^2(X|Y) &= \int \mathbb{E}_Y |\mathbb{E}[e^{isX}|Y] - \mathbb{E}e^{isX}|^2 \frac{ds}{s^2} \\
&= \int \mathbb{E}_Y |e^{is\mu_y - s^2/2} - \mathbb{E}[\mathbb{E}(e^{isX}|Y)]|^2 \frac{ds}{s^2} \\
&= \int \mathbb{E}_Y |e^{is\mu_y - s^2/2} - p_0 e^{is\mu_0 - s^2/2} - p_1 e^{is\mu_1 - s^2/2}|^2 \frac{ds}{s^2} \\
&= \int p_0 p_1 |e^{is\mu_0 - s^2/2} - e^{is\mu_1 - s^2/2}|^2 \frac{ds}{s^2} \\
&= \int p_0 p_1 |e^{is\mu_0} - e^{is\mu_1}|^2 e^{-s^2} \frac{ds}{s^2} \\
&= \int 2p_0 p_1 (1 - \cos(s\Delta)) e^{-s^2} \frac{ds}{s^2} \\
&= 2p_0 p_1 F(\Delta),
\end{aligned}$$

where $F(\Delta) = \int (1 - \cos(s\Delta)) e^{-s^2} \frac{ds}{s^2}$. Note that $F(0) = 0$, and $F'(0) = 0$, but

$$F''(\Delta) = \int \cos(s\Delta) e^{-s^2} ds = \sqrt{\pi} e^{-\Delta^2/4}.$$

Thus

$$F'(Y) = \sqrt{\pi} \int_0^Y e^{-z^2/4} dz.$$

By using the function (error function, or Gaussian Error Function), $\text{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} dt$, we have that $\int_0^Y e^{-z^2/4} dz = \sqrt{\pi} \text{erf}(Y/2)$.

Hence,

$$\begin{aligned}
F(\Delta) &= \sqrt{\pi} \int_0^\Delta \int_0^y e^{-z^2/4} dz dy = \sqrt{\pi} \int_0^\Delta \sqrt{\pi} \text{erf}(y/2) dy \\
&= \pi \int_0^\Delta \text{erf}(y/2) dy = 2\pi \int_0^{\Delta/2} \text{erf}(y) dy \\
&= 2\pi \left[\frac{\Delta}{2} \text{erf}\left(\frac{\Delta}{2}\right) + \frac{e^{-\Delta^2/4} - 1}{\sqrt{\pi}} \right],
\end{aligned}$$

where, we have used the fact that $\int \text{erf}(z) dz = z \text{erf}(z) + \frac{e^{-z^2}}{\sqrt{\pi}}$.

Finally,

$$\mathcal{C}^2(X|Y) = 4p_0p_1 \left[\frac{\Delta}{2} \operatorname{erf}\left(\frac{\Delta}{2}\right) + \frac{e^{-\Delta^2/4} - 1}{\sqrt{\pi}} \right].$$

Section 2.3: Bivariate normal distribution:

Note that if $X \sim N(\mu_x, \sigma_x^2)$, then $E(e^{isX}) = e^{is\mu_x - \frac{s^2}{2}\sigma_x^2}$, and $E(e^{sX}) = e^{s\mu_x + \frac{s^2}{2}\sigma_x^2}$.

Hence, $\mathcal{C}^2(X|Y) = F(\rho)/\pi$, where

$$\begin{aligned} F(\rho) &= \int E_Y |e^{is\rho Y - \frac{s^2}{2}(1-\rho^2)} - e^{-\frac{s^2}{2}}|^2 \frac{ds}{s^2} \\ &= \int E_Y |e^{is\rho Y + \frac{\rho^2 s^2}{2}} - 1|^2 \frac{e^{-s^2}}{s^2} ds \\ &= \int E_Y (e^{\rho^2 s^2} - e^{is\rho Y + \frac{\rho^2 s^2}{2}} - e^{-is\rho Y + \frac{\rho^2 s^2}{2}} + 1) \frac{e^{-s^2}}{s^2} ds \\ &= \int (e^{\rho^2 s^2} - 1) \frac{e^{-s^2}}{s^2} ds. \end{aligned}$$

By Taylor expansion, we have that

$$e^{\rho^2 s^2} - 1 = \sum_{n=1}^{\infty} \frac{(\rho^2 s^2)^n}{n!}.$$

Thus,

$$F(\rho) = \rho^2 \sum_{n=1}^{\infty} \frac{\rho^{2(n-1)}}{n!} \int s^{2(n-1)} e^{-s^2} ds = \rho^2 G(\rho).$$

Note that $G(\rho)$ is an increasing function, then

$$\pi \mathcal{C}^2(X|Y) = F(\rho) \leq F(1) = \pi \mathcal{C}^2(X|X).$$

In addition, $F(0) = 0, F'(0) = 0$. Simple calculation shows that $F'(\rho) = \frac{2\rho\sqrt{\pi}}{\sqrt{1-\rho^2}}$.

Therefore, we have $F(\rho) = \int_0^\rho \frac{2z\sqrt{\pi}}{\sqrt{1-z^2}} dz = 2\sqrt{\pi}(1 - \sqrt{1-\rho^2})$, And we have:

$$\mathcal{C}^2(X|Y) = \frac{2}{\sqrt{\pi}}(1 - \sqrt{1-\rho^2}).$$

Section 2.3: Binomial distribution:

Note that if $X_Y \sim Bin(n, q_Y)$, where $Y \in \{0, 1\}$, then we have that

$$\begin{aligned}
& \mathcal{C}^2(X|Y) \\
&= \int \mathbb{E}_Y |\mathbb{E}[e^{itX}|Y] - \mathbb{E}e^{itX}|^2 w(t) dt \\
&= p_0 p_1 \int |(q_0 e^{it} + 1 - q_0)^n - (q_1 e^{it} + 1 - q_1)^n|^2 w(t) dt \\
&= p_0 p_1 \int \left| \sum_{k=0}^n c_n^k q_0^k e^{ikt} (1 - q_0)^{n-k} - \sum_{k=0}^n c_n^k q_1^k e^{ikt} (1 - q_1)^{n-k} \right|^2 w(t) dt \\
&= p_0 p_1 \int \left\{ \sum_{k=0}^n c_n^k e^{ikt} [q_0^k (1 - q_0)^{n-k} - q_1^k (1 - q_1)^{n-k}] \right\} \\
&\quad \times \left\{ \sum_{l=0}^n c_n^l e^{-ilt} [q_0^l (1 - q_0)^{n-l} - q_1^l (1 - q_1)^{n-l}] \right\} w(t) dt \\
&= p_0 p_1 \int \left\{ \sum_{k,l=0}^n c_n^k c_n^l [q_0^k (1 - q_0)^{n-k} - q_1^k (1 - q_1)^{n-k}] [q_0^l (1 - q_0)^{n-l} - q_1^l (1 - q_1)^{n-l}] \right. \\
&\quad \times \left. [(e^{it(k-l)} - 1) + 1] \right\} w(t) dt \\
&= -p_0 p_1 \left\{ \sum_{k,l=0}^n c_n^k c_n^l [q_0^k (1 - q_0)^{n-k} - q_1^k (1 - q_1)^{n-k}] [q_0^l (1 - q_0)^{n-l} - q_1^l (1 - q_1)^{n-l}] |k - l| \right\}.
\end{aligned}$$

Now consider

$$\begin{aligned}
& q_0^k (1 - q_0)^{n-k} - q_1^k (1 - q_1)^{n-k} \\
&= (q_0 - q_1 + q_1)^k (1 - q_0)^{n-k} - q_1^k (1 - q_1)^{n-k} \\
&= \sum_{i=0}^k c_k^i (q_0 - q_1)^i q_1^{k-i} (1 - q_0)^{n-k} - q_1^k (1 - q_1)^{n-k} \\
&= (q_0 - q_1) \sum_{i=1}^k c_k^i (q_0 - q_1)^{i-1} q_1^{k-i} (1 - q_1)^{n-k} + q_1^k [(1 - q_0)^{n-k} - (1 - q_1)^{n-k}] \\
&= (q_0 - q_1) \left[\sum_{i=1}^k c_k^i (q_0 - q_1)^{i-1} q_1^{k-i} (1 - q_1)^{n-k} - q_1^k \sum_{i=1}^{n-k} (1 - q_0)^{n-k-i} (1 - q_1)^{i-1} \right]
\end{aligned}$$

Therefore,

$$\begin{aligned}
& \mathcal{C}^2(X|Y) \\
&= -p_0 p_1 (q_0 - q_1)^2 \left\{ \sum_{k,l=0}^n c_n^k c_n^l \left[\sum_{i=1}^k c_k^i (q_0 - q_1)^{i-1} q_1^{k-i} (1 - q_1)^{n-k} - q_1^k \sum_{i=1}^{n-k} (1 - q_0)^{n-k-i} (1 - q_1)^{i-1} \right] \right. \\
&\quad \times \left. \left[\sum_{i=1}^l c_l^i (q_0 - q_1)^{i-1} q_1^{l-i} (1 - q_1)^{n-l} - q_1^l \sum_{i=1}^{n-l} (1 - q_0)^{n-l-i} (1 - q_1)^{i-1} \right] |k - l| \right\}.
\end{aligned}$$

When $n = 1$, we simply it to $\mathcal{C}^2(X|Y) = 2p_0 p_1 (q_0 - q_1)^2$; and when $n = 2$, we have $\mathcal{C}^2(X|Y) = 4p_0 p_1 (q_0 - q_1)^2 [1 + (q_0 + q_1 - 1)^2]$.

Section 2.3: Conditional Cauchy distribution:

Note that $q_0, q_1 > 0$, and without loss of generality we assume that $q_1 \geq q_0$. Define a function $E_i(x) = \int_{-\infty}^x \frac{e^s}{s} ds$, and integral is taken in the principal as ϵ to ϵ^{-1} when $\epsilon \rightarrow 0$. We then have,

$$\begin{aligned}
\mathcal{C}^2(X|Y) &= \int \mathbf{E}_Y | \mathbf{E}[e^{itX} | Y] - \mathbf{E}e^{itX} |^2 w(t) dt \\
&= \frac{p_0 p_1}{\pi} \int |e^{-q_0|t|} - e^{-q_1|t|}|^2 \frac{dt}{t^2} \\
&= \frac{2p_0 p_1}{\pi} \int_0^{+\infty} [e^{-2q_0 t} - 2e^{-(q_0+q_1)t} + e^{-2q_1 t}] \frac{dt}{t^2}
\end{aligned}$$

$$\begin{aligned}
\mathcal{C}^2(X|Y; \epsilon) &= \frac{2p_0 p_1}{\pi} \int_{\epsilon}^{\epsilon^{-1}} [e^{-2q_0 t} - 2e^{-(q_0+q_1)t} + e^{-2q_1 t}] \frac{dt}{t^2} \\
&= \frac{2p_0 p_1}{\pi} \int_{\epsilon}^{\epsilon^{-1}} [e^{-2q_0 t} - 2e^{-(q_0+q_1)t} + e^{-2q_1 t}] \frac{dt}{t^2}
\end{aligned}$$

Now by using 1.3.2.20 and 1.3.2.12 of Prudnikov et al. (1986), we have that

$$\begin{aligned}
\mathcal{C}^2(X|Y; \epsilon) &= \frac{2p_0p_1}{\pi} \int_{\epsilon}^{\epsilon^{-1}} [e^{-2q_0t} - 2e^{-(q_0+q_1)t} + e^{-2q_1t}] \frac{dt}{t^2} \\
&= \frac{2p_0p_1}{\pi} \left[-\frac{e^{-2q_0t}}{t} - 2q_0E_i(-2q_0t) - 2\left(-\frac{e^{-(q_0+q_1)t}}{t} - (q_0+q_1)E_i(-(q_0+q_1)t)\right) \right. \\
&\quad \left. - \frac{e^{-2q_1t}}{t} - 2q_1E_i(-2q_1t) \right] \Big|_{\epsilon}^{\epsilon^{-1}} \\
&= \frac{2p_0p_1}{\pi} \left[-\frac{e^{-2q_0t}}{t} + 2\frac{e^{-(q_0+q_1)t}}{t} - \frac{e^{-2q_1t}}{t} \right. \\
&\quad \left. - 2q_0E_i(-2q_0t) + 2(q_0+q_1)E_i(-(q_0+q_1)t) - 2q_1E_i(-2q_1t) \right] \Big|_{\epsilon}^{\epsilon^{-1}}
\end{aligned}$$

But $[-\frac{e^{-2q_0t}}{t} + 2\frac{e^{-(q_0+q_1)t}}{t} - \frac{e^{-2q_1t}}{t}] \Big|_{\epsilon}^{\epsilon^{-1}} \rightarrow 0$ as $\epsilon \rightarrow 0$. Thus, as $\epsilon \rightarrow 0$ we can have

$$\begin{aligned}
\mathcal{C}^2(X|Y; \epsilon) &= \frac{2p_0p_1}{\pi} [-2q_0E_i(-2q_0t) + 2(q_0+q_1)E_i(-(q_0+q_1)t) - 2q_1E_i(-2q_1t)] \Big|_{\epsilon}^{\epsilon^{-1}} \\
&= \frac{2p_0p_1}{\pi} \left[-2q_0 \int_{-2q_0\epsilon}^{-2q_0\epsilon^{-1}} \frac{e^t}{t} dt + 2(q_0+q_1) \int_{-(q_0+q_1)\epsilon}^{-(q_0+q_1)\epsilon^{-1}} \frac{e^t}{t} dt - 2q_1 \int_{-2q_1\epsilon}^{-2q_1\epsilon^{-1}} \frac{e^t}{t} dt \right] \\
&= \frac{2p_0p_1}{\pi} \left[2q_0 \int_{2q_0\epsilon^{-1}}^{(q_0+q_1)\epsilon^{-1}} \frac{e^{-t}}{t} dt - 2q_0 \int_{2q_0\epsilon}^{(q_0+q_1)\epsilon} \frac{e^{-t}}{t} dt \right. \\
&\quad \left. - 2q_1 \int_{(q_0+q_1)\epsilon^{-1}}^{2q_1\epsilon^{-1}} \frac{e^{-t}}{t} dt + 2q_1 \int_{(q_0+q_1)\epsilon}^{2q_1\epsilon} \frac{e^{-t}}{t} dt \right] \\
&= \frac{2p_0p_1}{\pi} [2q_0A_1 - 2q_0B_1 - 2q_1A_2 + 2q_1B_2]
\end{aligned}$$

But $A_1 = \int_{2q_0}^{q_0+q_1} e^{-y\epsilon^{-1}} y^{-1} dy \leq (2q_0)^{-1}(q_1 - q_0) e^{-2q_0\epsilon^{-1}} \rightarrow 0$ as $\epsilon \rightarrow 0$. Similarly,

$A_2 \rightarrow 0$ as $\epsilon \rightarrow 0$. Now by using 1.3.2.13 of Prudnikov et al. (1986), we have

$$\begin{aligned}
B_1 &= \ln[(q_0+q_1)\epsilon] + \sum_{k=1}^{\infty} \frac{(-(q_0+q_1)\epsilon)^k}{k!k} - \ln(2q_0\epsilon) - \sum_{k=1}^{\infty} \frac{(-2q_0\epsilon)^k}{k!k} \\
&= \ln \frac{q_0+q_1}{2q_0} + \sum_{k=1}^{\infty} \frac{(-(q_0+q_1)\epsilon)^k - (-2q_0\epsilon)^k}{k!k} \\
&= \ln \frac{q_0+q_1}{2q_0} + \sum_{k=1}^{\infty} \frac{(-(q_0+q_1))^k - (-2q_0)^k}{k!k} \epsilon^k \\
&= \ln \frac{q_0+q_1}{2q_0} \text{ as } \epsilon \rightarrow 0.
\end{aligned}$$

While by similar argument, we have $B_2 = \ln \frac{2q_1}{q_0+q_1}$ as $\epsilon \rightarrow 0$. Therefore,

$$\mathcal{C}^2(X|Y) = \lim_{\epsilon \rightarrow 0} \mathcal{C}^2(X|Y; \epsilon) = \frac{4p_0p_1}{\pi} \left(q_0 \ln \frac{2q_0}{q_0+q_1} + q_1 \ln \frac{q_1}{q_0+q_1} \right).$$

Note that $\mathcal{C}^2(X|Y) \geq 0$, and it is 0 if $q_1 = q_0$; However, $\mathcal{C}^2(X|Y) \geq 0$ increases as $q_1 > q_0$; decreases as $q_1 < q_0$. Thus $\mathcal{C}^2(X|Y) = 0$ iff $q_1 = q_0$.

Proof of Theorem 2.4.1: Following Székely et al. (2007), we have that

$$\begin{aligned} f_{\mathbf{X}|y}^n(t) \overline{f_{\mathbf{X}|y}^n(t)} &= \frac{1}{n_y^2} \sum_{k_y, l_y=1}^{n_y, n_y} \cos t^T(\mathbf{X}_{y, k_y} - \mathbf{X}_{y, l_y}) + v_1 \\ f_{\mathbf{X}|y}^n(t) \overline{f_{\mathbf{X}}^n(t)} &= \frac{1}{nn_y} \sum_{y'=1}^H \sum_{k_y, l_{y'}=1}^{n_y, n_{y'}} \cos t^T(\mathbf{X}_{y, k_y} - \mathbf{X}_{y', l_{y'}}) + v_2 \\ f_{\mathbf{X}}^n(t) \overline{f_{\mathbf{X}}^n(t)} &= \frac{1}{n^2} \sum_{y, y'=1}^{H, H} \sum_{k_y, l_{y'}=1}^{n_y, n_{y'}} \cos t^T(\mathbf{X}_{y, k_y} - \mathbf{X}_{y', l_{y'}}) + v_3, \end{aligned}$$

where v_1, v_2 and v_3 vanish when integral is evaluated. Since

$$\cos t^T(\mathbf{X}_k - \mathbf{X}_l) = 1 - (1 - \cos t^T(\mathbf{X}_k - \mathbf{X}_l)), \text{ and } \int [1 - \cos t^T(\mathbf{X}_k - \mathbf{X}_l)] w(t) dt = |\mathbf{X}_k - \mathbf{X}_l|,$$

by choosing $k = y, k_y$ and $l = y, l_y$, we have

$$\begin{aligned} \cos t^T(\mathbf{X}_{y, k_y} - \mathbf{X}_{y, l_y}) &= 1 - (1 - \cos t^T(\mathbf{X}_{y, k_y} - \mathbf{X}_{y, l_y})) \\ \text{and } \int [1 - \cos t^T(\mathbf{X}_{y, k_y} - \mathbf{X}_{y, l_y})] w(t) dt &= |\mathbf{X}_{y, k_y} - \mathbf{X}_{y, l_y}|; \end{aligned}$$

by choosing $k = y, k_y$ and $l = y', l_{y'}$, we have

$$\begin{aligned} \cos t^T(\mathbf{X}_{y, k_y} - \mathbf{X}_{y', l_{y'}}) &= 1 - (1 - \cos t^T(\mathbf{X}_{y, k_y} - \mathbf{X}_{y', l_{y'}})) \\ \text{and } \int [1 - \cos t^T(\mathbf{X}_{y, k_y} - \mathbf{X}_{y', l_{y'}})] w(t) dt &= |\mathbf{X}_{y, k_y} - \mathbf{X}_{y', l_{y'}}|. \end{aligned}$$

We also have

$$|f_{\mathbf{X}|y}^n(t) - f_{\mathbf{X}}^n(t)|^2 = f_{\mathbf{X}|y}^n(t)\overline{f_{\mathbf{X}|y}^n(t)} - f_{\mathbf{X}|y}^n(t)\overline{f_{\mathbf{X}}^n(t)} - \overline{f_{\mathbf{X}|y}^n(t)}f_{\mathbf{X}}^n(t) + f_{\mathbf{X}}^n(t)\overline{f_{\mathbf{X}}^n(t)}.$$

Therefore,

$$\begin{aligned} \mathcal{C}_{w,y,n}^2(\mathbf{X}|Y=y) &= \|f_{\mathbf{X}|y}^n(t) - f_{\mathbf{X}}^n(t)\|^2 \\ &= \frac{2}{nn_y} \sum_{y'=1}^H \sum_{k_y, l_{y'}=1}^{n_y, n_{y'}} |\mathbf{X}_{y, k_y} - \mathbf{X}_{y', l_{y'}}| - \frac{1}{n_y^2} \sum_{k_y, l_y=1}^{n_y, n_y} |\mathbf{X}_{y, k_y} - \mathbf{X}_{y, l_y}| \\ &\quad - \frac{1}{n^2} \sum_{y, y'=1}^{H, H} \sum_{k_y, l_{y'}=1}^{n_y, n_{y'}} |\mathbf{X}_{y, k_y} - \mathbf{X}_{y', l_{y'}}|. \end{aligned}$$

And thus, we have

$$\begin{aligned} \mathcal{C}_n^2(\mathbf{X}|Y) &= \sum_{y=1}^H p_y \mathcal{C}_{w,y,n}^2(\mathbf{X}|Y=y) \\ &= \frac{2}{n^2} \sum_{y=1}^H \sum_{y'=1}^H \sum_{k_y, l_{y'}=1}^{n_y, n_{y'}} |\mathbf{X}_{y, k_y} - \mathbf{X}_{y', l_{y'}}| - \frac{1}{n} \sum_{y=1}^H \frac{1}{n_y} \sum_{k_y, l_y=1}^{n_y, n_y} |\mathbf{X}_{y, k_y} - \mathbf{X}_{y, l_y}| \\ &\quad - \frac{1}{n^2} \sum_{y, y'=1}^{H, H} \sum_{k_y, l_{y'}=1}^{n_y, n_{y'}} |\mathbf{X}_{y, k_y} - \mathbf{X}_{y', l_{y'}}| \\ &= \frac{1}{n^2} \sum_{y, y'=1}^{H, H} \sum_{k_y, l_{y'}=1}^{n_y, n_{y'}} |\mathbf{X}_{y, k_y} - \mathbf{X}_{y', l_{y'}}| - \frac{1}{n} \sum_{y=1}^H \frac{1}{n_y} \sum_{k_y, l_y=1}^{n_y, n_y} |\mathbf{X}_{y, k_y} - \mathbf{X}_{y, l_y}|. \end{aligned}$$

Note that the summation in the first and third term after the second equality sign are the same. We complete the proof. \square

Proof of Lemma 2.4.4: This can follow from Theorem 2 of Székely et al. (2007) and Theorem 3 of Shao and Zhang (2014). By applying SLLN of V-statistic to achieve the conclusion. Note that let $\xi_{n,y}(t) = f_{\mathbf{X}|y}^n(t) - f_{\mathbf{X}}^n(t)$, then $\mathcal{C}_{w,y,n}^2(\mathbf{X}|y) = \|\xi_{n,y}(t)\|^2$. Hence, by (2.6), we have $\mathcal{C}_n^2(\mathbf{X}|Y) = \mathbb{E}_Y \mathcal{C}_{w,y,n}^2(\mathbf{X}|Y) = \mathbb{E}_Y \|\xi_{n,Y}\|^2 = \sum_{y=1}^H p_y \|f_{\mathbf{X}|y}^n(t) - f_{\mathbf{X}}^n(t)\|^2$.

Define $\xi_y(t) = f_{\mathbf{X}|y}(t) - f_{\mathbf{X}}(t)$, and let $u_{y,k_y} = \exp(it^T \mathbf{X}_{y,k_y}) - f_{\mathbf{X}|y}(t)$ and $v_{y,k_y} =$

$\exp(it^T \mathbf{X}_{y,k_y}) - f_{\mathbf{X}}(t)$. Then, $\xi_{n,y}(t) = \frac{1}{n_y} \sum_{k_y=1}^{n_y} u_{y,k_y} - \frac{1}{n} \sum_{y=1}^H \sum_{k_y=1}^{n_y} v_{y,k_y} + \xi_y(t)$.

In integrals, we can use the symbol $d\omega$, which is defined by $d\omega = w(t)dt$, where $w(t)$ is defined previously. Define the region $D(\delta) = \{t : \delta \leq |t|_p \leq 1/\delta\}$, for each $\delta > 0$, and the random variables $\mathcal{C}_{w,y,n,\delta}^2(\mathbf{X}|y) = \int_{D(\delta)} |\xi_{n,y}(t)|^2 d\omega$. For any fixed δ , the weight function $w(t)$ is bounded on $D(\delta)$. Hence, $\mathcal{C}_{w,y,n,\delta}^2(\mathbf{X}|y)$ is a combination of V -statistics with finite expectation. By the SLLN for V -statistics, it follows that almost surely

$$\lim_{n \rightarrow \infty} \mathcal{C}_{w,y,n,\delta}^2(\mathbf{X}|y) = \mathcal{C}_{w,y,\cdot,\delta}^2(\mathbf{X}|y) = \int_{D(\delta)} |\xi_y(t)|^2 d\omega.$$

Clearly $\mathcal{C}_{w,y,\cdot,\delta}^2(\mathbf{X}|y)$ converges to $\mathcal{C}_{w,y}^2(\mathbf{X}|y)$ as $\delta \rightarrow 0$. Therefore, it remains to prove that almost surely

$$\limsup_{\delta \rightarrow 0} \limsup_{n \rightarrow \infty} |\mathcal{C}_{w,y,n}^2(\mathbf{X}|y) - \mathcal{C}_{w,y,n,\delta}^2(\mathbf{X}|y)| = 0.$$

For each $\delta > 0$,

$$|\mathcal{C}_{w,y,n}^2(\mathbf{X}|y) - \mathcal{C}_{w,y,n,\delta}^2(\mathbf{X}|y)| = \int_{|t| < \delta} |\xi_{n,y}(t)|^2 d\omega + \int_{|t| > \frac{1}{\delta}} |\xi_{n,y}(t)|^2 d\omega. \quad (\text{S.4.5.8})$$

For $z = (z_1, \dots, z_p)^T \in \mathbb{R}^p$, define the function $G(s) = \int_{|z| < s} \frac{1 - \cos z_1}{|z|^{1+p}} dz$. By Lemma 1 of Székely et al. (2007), clearly $G(s)$ is bounded by \tilde{c}_p and $\lim_{s \rightarrow 0} G(s) = 0$. Using the inequality $|a + b + c|^2 \leq 3(|a|^2 + |b|^2 + |c|^2)$, and applying Cauchy-Schwarz inequality, we have that

$$\begin{aligned} |\xi_{n,y}(t)|^2 &\leq 3 \left(\left| \frac{1}{n_y} \sum_{k_y=1}^{n_y} u_{y,k_y} \right|^2 + \left| \frac{1}{n} \sum_{y=1}^H \sum_{k_y=1}^{n_y} v_{y,k_y} \right|^2 + |\xi_y(t)|^2 \right) \\ &\leq 3 \left(\frac{1}{n_y} \sum_{k_y=1}^{n_y} |u_{y,k_y}|^2 + \frac{1}{n} \sum_{y=1}^H \sum_{k_y=1}^{n_y} |v_{y,k_y}|^2 + |\xi_y(t)|^2 \right). \end{aligned} \quad (\text{S.4.5.9})$$

After a suitable change of variables, we have

$$\begin{aligned}\int_{|t|<\delta} \frac{|u_{y,k_y}|^2}{\tilde{c}_p|t|^{1+p}} dt &\leq 2\mathbb{E}_{\mathbf{X}|y}|\mathbf{X} - \mathbf{X}_{y,k_y}|G_{|y}(|\mathbf{X} - \mathbf{X}_{y,k_y}|\delta) \\ \int_{|t|<\delta} \frac{|v_{y,k_y}|^2}{\tilde{c}_p|t|^{1+p}} dt &\leq 2\mathbb{E}_{\mathbf{X}}|\mathbf{X} - \mathbf{X}_{y,k_y}|G(|\mathbf{X} - \mathbf{X}_{y,k_y}|\delta)\end{aligned}$$

Therefore, we have

$$\begin{aligned}\int_{|t|<\delta} |\xi_{n,y}(t)|^2 d\omega &\leq \frac{6}{n_y} \sum_{k_y=1}^{n_y} \mathbb{E}_{\mathbf{X}|y}|\mathbf{X} - \mathbf{X}_{y,k_y}|G_{|y}(|\mathbf{X} - \mathbf{X}_{y,k_y}|\delta) \\ &\quad + \frac{6}{n} \sum_{y=1}^H \sum_{k_y=1}^{n_y} \mathbb{E}_{\mathbf{X}}|\mathbf{X} - \mathbf{X}_{y,k_y}|G(|\mathbf{X} - \mathbf{X}_{y,k_y}|\delta) + 3 \int_{|t|<\delta} |\xi_y(t)|^2 d\omega\end{aligned}$$

By the SLLN, then

$$\begin{aligned}\limsup_{n \rightarrow \infty} \int_{|t|<\delta} |\xi_{n,y}(t)|^2 d\omega &\leq 6\mathbb{E}_{|y}(|\mathbf{X} - \mathbf{X}'|)G_{|y}(|\mathbf{X} - \mathbf{X}'|\delta) \\ &\quad + 6\mathbb{E}(|\mathbf{X} - \mathbf{X}'|)G(|\mathbf{X} - \mathbf{X}'|\delta) + 3 \int_{|t|<\delta} |\xi_y(t)|^2 d\omega\end{aligned}$$

By the Lebesgue Dominated Convergence theorem, we then have

$$\limsup_{\delta \rightarrow 0} \limsup_{n \rightarrow \infty} \int_{|t|<\delta} |\xi_{n,y}(t)|^2 d\omega = 0, \text{ almost surely.}$$

Now consider the second term in equation (S.4.5.8), since $|u_{y,k_y}|^2, |v_{y,k_y}|^2, |\xi_y(t)|^2 \leq 4$ and the inequality (S.4.5.9) implies that $|\xi_{n,y}(t)|^2 \leq 36$. Hence,

$$\int_{|t|>\frac{1}{\delta}} |\xi_{n,y}(t)|^2 d\omega \leq 36 \int_{|t|>\frac{1}{\delta}} \frac{1}{\tilde{c}_p|t|^{1+p}} dt = 36h(\delta).$$

But $h(\delta)$ goes to zero as $\delta \rightarrow 0$. That means $\mathcal{C}_{w,y,n}^2(\mathbf{X}|y) \rightarrow \mathcal{C}_{w,y}^2(\mathbf{X}|y)$ almost surely, for any given y . And the conclusion then follows. \square

Proof of Theorem 2.4.5: The argument is very similar to that presented in the proofs of Theorem 5 and Corollary 2 of Székely et al. (2007) and that of Theorem 4 of Shao and Zhang (2014). Note that $f_{\mathbf{X}|Y}(s) = \mathbb{E}(e^{is\mathbf{X}}|Y)$, $f_{\mathbf{X}}(s) = \mathbb{E}(e^{is\mathbf{X}})$, $p_y = n_y/n$,

where n_y is the number of observations in $Y \in y$, $y = 1, 2, \dots, H$ and $\sum_{y=1}^H n_y = n$. In addition, $f_{\mathbf{X}}(s) = E_Y f_{\mathbf{X}|Y}(s) = \sum_Y p_Y f_{\mathbf{X}|Y}(s)$, where $p_Y = P(Y \in Y)$.

- a. Define the empirical process $\Gamma_{n,y}(s) = \sqrt{n_y}[f_{\mathbf{X}|y}^n(s) - f_{\mathbf{X}}^n(s)]$. Under independence hypothesis, $E_{\mathbf{X}|y}[\Gamma_{n,y}(s)] = 0$ and $E_{\mathbf{X}|y}[\Gamma_{n,y}(s)\overline{\Gamma_{n,y}(s_0)}] = (1 - \frac{n_y}{n})[f_{\mathbf{X}}(s - s_0) - f_{\mathbf{X}}(s)f_{\mathbf{X}}(s_0)] = (1 - \frac{n_y}{n})\text{cov}_{\Gamma}(s, s_0)$. In particular, $E_{\mathbf{X}|y}|\Gamma_{n,y}(s)|^2 = (1 - \frac{n_y}{n})[1 - |f_{\mathbf{X}}(s)|^2] \leq 1$.

Note that $n\mathcal{C}_n^2(\mathbf{X}|Y) = \sum_{y=1}^H \|\Gamma_{n,y}(s)\|^2$.

For each $\delta > 0$, define the region $D(\delta) = \{s : \delta \leq |s|_p < 1/\delta\}$. For each δ we construct a sequence of random variables $\{Q_{n,y}(\delta)\}$ such that

- (i) $Q_{n,y}(\delta) \xrightarrow{D} Q_y(\delta)$ for each $\delta > 0$;
- (ii) $\limsup_{n \rightarrow \infty} E_{|y}|Q_{n,y}(\delta) - \|\Gamma_{n,y}\|^2| \rightarrow 0$ as $\delta \rightarrow 0$;
- (iii) $E_{|y}|Q_y(\delta) - (1 - p_Y)\|\Gamma\|^2| \rightarrow 0$ as $\delta \rightarrow 0$.

Then the weak convergence of $\|\Gamma_{n,y}\|^2$ to $(1 - p_Y)\|\Gamma\|^2$ follows from Theorem 8.6.2 of Resnick (1999). Therefore,

$$n\mathcal{C}_n^2(\mathbf{X}|Y) = \sum_{y=1}^H \|\Gamma_{n,y}(s)\|^2 \Rightarrow (H - 1)\|\Gamma\|^2.$$

Following the construction in Shao and Zhang (2014) and Székely et al. (2007), we define

$$Q_{n,y}(\delta) = \int_{D(\delta)} |\Gamma_{n,y}(s)|^2 d\omega \text{ and } Q_y(\delta) = (1 - p_Y) \int_{D(\delta)} |\Gamma(s)|^2 d\omega.$$

Given $\epsilon = 1/q > 0$, $q \in N$, choose a partition $\{D_k\}_{k=1}^N$ of $D(\delta)$ into $N = N(\epsilon)$ measurable sets with diameter at most ϵ . Then $Q_{n,y}(\delta) = \sum_{k=1}^N \int_{D_k} |\Gamma_{n,y}(s)|^2 d\omega$ and $Q_y(\delta) = (1 - p_Y) \sum_{k=1}^N \int_{D_k} |\Gamma(s)|^2 d\omega$.

Define $Q_{n,y}^q(\delta) = \sum_{k=1}^N \int_{D_k} |\Gamma_{n,y}(s_0(k))|^2 d\omega$ and $Q_y^q(\delta) = (1 - p_Y) \sum_{k=1}^N \int_{D_k} |\Gamma(s_0(k))|^2 d\omega$, where $\{s_0(k)\}_{k=1}^N$ is a set of distinct points such that $s_0(k) \in D_k$. By multivariate CLT and continuous mapping theorem, $Q_{n,y}^q(\delta) \xrightarrow{D} Q_y^q(\delta)$, for any $q \in N$.

Thus based on Theorem 8.6.2 of Resnick (1999), (i) holds if we can show that

$$\limsup_{q \rightarrow \infty} \mathbb{E}|Q_y^q(\delta) - Q_y(\delta)| = 0, \quad (\text{S.4.5.10})$$

$$\text{and } \limsup_{q \rightarrow \infty} \limsup_{n \rightarrow \infty} \mathbb{E}|Q_{n,y}^q(\delta) - Q_{n,y}(\delta)| = 0. \quad (\text{S.4.5.11})$$

Let $\beta_{n,y}(\epsilon) = \sup_{s,s_0} \mathbb{E}||\Gamma_{n,y}(s)|^2 - |\Gamma_{n,y}(s_0)|^2|$ and $\beta(\epsilon) = \sup_{s,s_0} \mathbb{E}||\Gamma(s)|^2 - |\Gamma(s_0)|^2|$, where the supremum is taken over all s and s_0 , under the restrictions: $\delta < |s|_p, |s_0|_p < 1/\delta$ and $|s - s_0|_p < \epsilon$.

$$\begin{aligned} \beta(\epsilon) &= \sup_{s,s_0} \mathbb{E}||\Gamma(s)|^2 - |\Gamma(s_0)|^2| \\ &= \sup_{s,s_0} \mathbb{E}|(\Gamma(s) - \Gamma(s_0))\overline{\Gamma(s)} + \Gamma(s_0)(\overline{\Gamma(s)} - \overline{\Gamma(s_0)})| \\ &\leq \sup_{s,s_0} \mathbb{E}^{1/2}|\Gamma(s) - \Gamma(s_0)|^2 (\mathbb{E}^{1/2}|\Gamma(s)|^2 + \mathbb{E}^{1/2}|\Gamma(s_0)|^2) \\ &\leq 2 \sup_{s,s_0} \mathbb{E}^{1/2}|\Gamma(s) - \Gamma(s_0)|^2 \\ &= 2 \sup_{s,s_0} |\text{cov}_\Gamma(s, s) - \text{cov}_\Gamma(s, s_0) - \text{cov}_\Gamma(s_0, s) + \text{cov}_\Gamma(s_0, s_0)|^{1/2}. \end{aligned}$$

Since $f_{\mathbf{X}}(s)$ is uniform continuous in $s \in \mathbb{R}^p$, it is clear that $\beta(\epsilon) \rightarrow 0$ as $\epsilon \rightarrow 0$.

To show (S.4.5.10), note that

$$\begin{aligned} \mathbb{E}|Q_y^q(\delta) - Q_y(\delta)| &= (1 - p_Y) \mathbb{E} \left| \sum_{k=1}^N \int_{D_k} |\Gamma(s_0(k))|^2 d\omega - \int_{D(\delta)} |\Gamma(s)|^2 d\omega \right| \\ &= (1 - p_Y) \mathbb{E} \left| \sum_{k=1}^N \int_{D_k} (|\Gamma(s_0(k))|^2 - |\Gamma(s)|^2) d\omega \right| \\ &\leq (1 - p_Y) \beta(1/q) \int_{D(\delta)} w(s) ds \rightarrow 0 \text{ as } q \rightarrow \infty. \end{aligned}$$

Using the same argument, we can show (S.4.5.11) holds, hence (i) is true.

To prove (ii), note that

$$\mathbb{E} \left| \int_{D(\delta)} |\Gamma_{n,y}(s)|^2 d\omega - \int_{\mathbb{R}^p} |\Gamma_{n,y}(s)|^2 d\omega \right| = \int_{|s| < \delta} \mathbb{E} |\Gamma_{n,y}(s)|^2 d\omega + \int_{|s| > 1/\delta} \mathbb{E} |\Gamma_{n,y}(s)|^2 d\omega.$$

By noting that $\mathbb{E}_{\mathbf{X}|Y} |\Gamma_{n,y}(s)|^2 = (1 - \frac{n_y}{n}) [1 - |f_{\mathbf{X}}(s)|^2]$ and following from the proof of Lemma 2.4.4, we have that

$$\int_{|s| < \delta} \mathbb{E} |\Gamma_{n,y}(s)|^2 d\omega \leq (1 - \frac{n_y}{n}) \mathbb{E} |\mathbf{X} - \mathbf{X}'| G(|\mathbf{X} - \mathbf{X}'| \delta).$$

The fact $\mathbb{E}_{\mathbf{X}|Y} |\Gamma_{n,y}(s)|^2 \leq 1$ implies that $\int_{|s| > 1/\delta} \mathbb{E} |\Gamma_{n,y}(s)|^2 d\omega \leq h(\delta)$, where $h(\delta)$ is defined in Lemma 2.4.4 and goes to zero as $\delta \rightarrow 0$. Thus (ii) holds.

Applying a similar argument, (iii) holds. Thus we complete the proof of (a).

- b. This can easily follow from Corollary 2 of Székely et al. (2007) and see Theorem 4 of Shao and Zhang (2014) as well.

Based on (a), $n\mathcal{C}_n^2(\mathbf{X}|Y) \xrightarrow[n \rightarrow \infty]{D} (H - 1) \|\Gamma(s)\|^2$. Note that

$$\mathbb{E} \|\Gamma(s)\|^2 = \int_{\mathbb{R}^p} \text{cov}_{\Gamma}(s, s) = \int_{\mathbb{R}^p} (1 - |f_{\mathbf{X}}(s)|^2) d\omega = \mathbb{E} |\mathbf{X} - \mathbf{X}'|.$$

By the SLLN for V -statistics, as $n \rightarrow \infty$, $S_n \rightarrow (H - 1) \mathbb{E} |\mathbf{X} - \mathbf{X}'|$, almost surely. Therefore,

$$n\mathcal{C}_n^2(\mathbf{X}|Y)/S_n \xrightarrow[n \rightarrow \infty]{D} Q,$$

where $\mathbb{E}(Q) = 1$ and Q is a nonnegative quadratic form of centered Gaussian random variable following the argument in the proof of Corollary 2 of Székely et al. (2007).

- c. If \mathbf{X} and Y are dependent, then $\mathcal{C}^2(\mathbf{X}|Y) > 0$. Lemma 2.4.4 implies that when for large n , $\mathcal{C}_n^2(\mathbf{X}|Y) > 0$, and thus $n\mathcal{C}_n^2(\mathbf{X}|Y) \rightarrow \infty$ as $n \rightarrow \infty$. By the SLLN, S_n converges to a constant and therefore, as $n \rightarrow \infty$, $n\mathcal{C}_n^2(\mathbf{X}|Y)/S_n \rightarrow \infty$.

□

Proof for the Kernel estimator:

Proof of Theorem 2.4.7: Note that $C_{n,k}^2(\mathbf{X}|\mathbf{Y}) = \frac{1}{n^2} \sum_{i,j} |\mathbf{X}_i - \mathbf{X}_j| - \hat{m}$, the first term is a V-statistic, which is root- n consistent to $E|\mathbf{X} - \mathbf{X}'|$. For the second term,

$$\begin{aligned} \hat{m} - E(m(\mathbf{y})) &= \frac{1}{n} \sum_{i=1}^n \hat{m}(\mathbf{y}_i) - E(m(\mathbf{y})) \\ &= \frac{1}{n} \sum_{i=1}^n (\hat{m}(\mathbf{y}_i) - m(\mathbf{y}_i)) + \frac{1}{n} \sum_{i=1}^n m(\mathbf{y}_i) - E(m(\mathbf{y})) \end{aligned}$$

The first part tends to 0 based on Lemma 2.4.6 and the second part tends to 0 by LLN theory, Thus Theorem 2.4.7 holds. \square

Additional simulation studies

In this section, we report additional simulations results in chapter 2.

Example S. 4.5.1. Following example 2.6.2, we construct models 2.6.2 (e)-(g), where the dimensions of \mathbf{X} and Y are the same as the models 2.6.2 (a)-(d), except that each individual random variable is independently generated from $t(2)$, $t(3)$ and $\chi^2(2)$ distributions, respectively. The empirical type-I errors at the nominal level of 0.1 for models 2.6.2 (e)-(g) are shown in table S.4.5.1, while at the nominal significance level of 0.05 are shown in table S.4.5.2 for models 2.6.2 (a)-(d), and in table S.4.5.3 for models 2.6.2 (e)-(g). Again, we have the same conclusion as in the paper.

Example S. 4.5.2. These additional simulations follow from Example 2.6.6 in the paper, but with different combinations of a , p , σ_x^2 and σ^2 . Figure S.4.5.1 shows similar power changes as in the paper. Again, kernel methods are the best.

Table S.4.5.1: Empirical type-I error rates for 10,000 tests at nominal significance level 0.1, using B replicates for models (e) - (g)

(e) $t(2), p = 5, q = 1$							(f) $t(3), p = 5, q = 1$				
n	B	dCov	DISCO	$R_c(\text{slice})$	$R_c(\text{epa})$	$R_c(\text{gau})$	dCov	DISCO	$R_c(\text{slice})$	$R_c(\text{epa})$	$R_c(\text{gau})$
25	400	0.105	0.103	0.105	0.101	0.102	0.101	0.102	0.101	0.105	0.100
30	366	0.097	0.096	0.096	0.093	0.099	0.101	0.099	0.101	0.096	0.099
35	342	0.105	0.103	0.102	0.097	0.105	0.098	0.102	0.102	0.100	0.096
50	300	0.095	0.096	0.095	0.101	0.102	0.096	0.097	0.097	0.102	0.099
70	271	0.100	0.103	0.103	0.100	0.101	0.098	0.096	0.096	0.097	0.097
100	250	0.098	0.095	0.097	0.098	0.100	0.099	0.098	0.099	0.102	0.102

(g) $\chi^2(2), p = 5, q = 1$						
n	B	dCov	DISCO	$R_c(\text{slice})$	$R_c(\text{epa})$	$R_c(\text{gau})$
25	400	0.100	0.097	0.099	0.099	0.099
30	366	0.099	0.097	0.098	0.096	0.097
35	342	0.097	0.098	0.099	0.099	0.098
50	300	0.102	0.102	0.103	0.103	0.104
70	271	0.100	0.097	0.097	0.095	0.101
100	250	0.100	0.100	0.099	0.100	0.096

Table S.4.5.2: Empirical type-I error rates for 10,000 tests at nominal significance level 0.05, using B replicates for models (a) - (d)

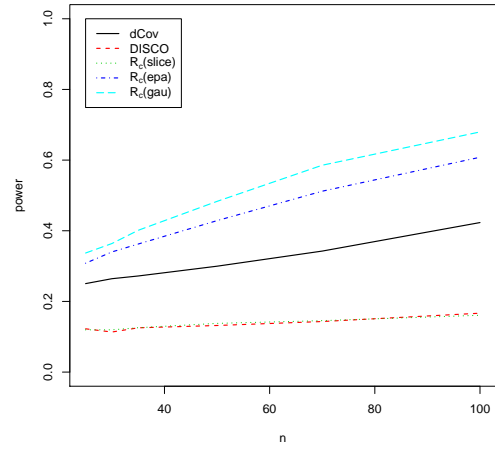
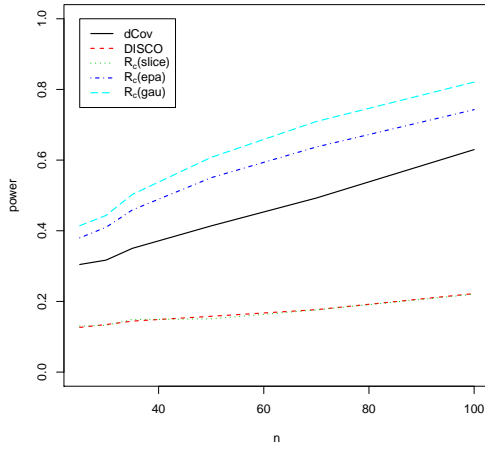
(a) $N(0, 1), p = 5, q = 1$							(b) $t(1), p = 5, q = 1$				
n	B	dCov	DISCO	$R_c(\text{slice})$	$R_c(\text{epa})$	$R_c(\text{gau})$	dCov	DISCO	$R_c(\text{slice})$	$R_c(\text{epa})$	$R_c(\text{gau})$
25	400	0.051	0.054	0.054	0.050	0.051	0.047	0.046	0.048	0.050	0.050
30	366	0.049	0.055	0.055	0.050	0.049	0.050	0.053	0.051	0.049	0.052
35	342	0.049	0.050	0.051	0.049	0.053	0.051	0.047	0.046	0.049	0.050
50	300	0.049	0.051	0.051	0.054	0.054	0.048	0.048	0.048	0.050	0.051
70	271	0.050	0.048	0.048	0.047	0.048	0.045	0.046	0.047	0.051	0.049
100	250	0.047	0.049	0.051	0.049	0.043	0.044	0.046	0.047	0.045	0.046

(c) $\chi^2(1), p = 5, q = 1$							(d) $\chi^2(3), p = 5, q = 1$				
n	B	dCov	DISCO	$R_c(\text{slice})$	$R_c(\text{epa})$	$R_c(\text{gau})$	dCov	DISCO	$R_c(\text{slice})$	$R_c(\text{epa})$	$R_c(\text{gau})$
25	400	0.053	0.054	0.053	0.050	0.050	0.047	0.046	0.046	0.049	0.055
30	366	0.050	0.050	0.050	0.051	0.048	0.048	0.051	0.052	0.050	0.051
35	342	0.052	0.049	0.048	0.052	0.053	0.047	0.052	0.052	0.044	0.048
50	300	0.050	0.050	0.049	0.048	0.049	0.046	0.046	0.048	0.050	0.047
70	271	0.045	0.048	0.047	0.046	0.050	0.046	0.049	0.047	0.049	0.046
100	250	0.051	0.048	0.047	0.046	0.053	0.050	0.048	0.046	0.045	0.051

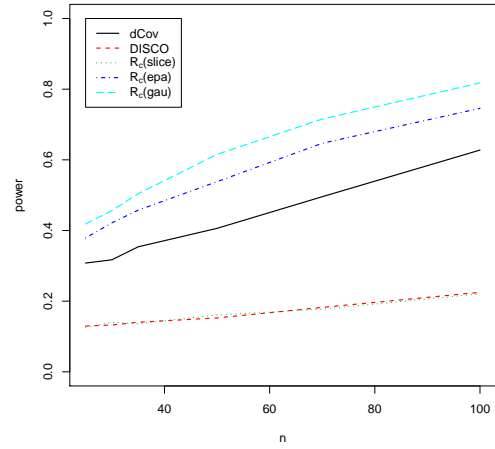
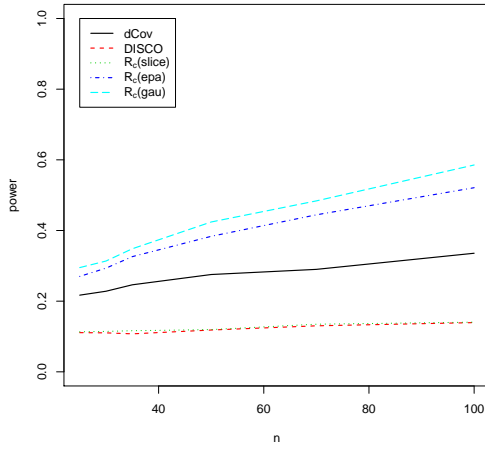
Table S.4.5.3: Empirical type-I error rates for 10,000 tests at nominal significance level 0.05, using B replicates for models (e) - (g)

(e) $t(2), p = 5, q = 1$							(f) $t(3), p = 5, q = 1$				
n	B	dCov	DISCO	$R_c(\text{slice})$	$R_c(\text{epa})$	$R_c(\text{gau})$	dCov	DISCO	$R_c(\text{slice})$	$R_c(\text{epa})$	$R_c(\text{gau})$
25	400	0.051	0.050	0.050	0.054	0.053	0.051	0.050	0.051	0.052	0.049
30	366	0.050	0.049	0.048	0.050	0.051	0.049	0.046	0.045	0.050	0.047
35	342	0.050	0.050	0.049	0.051	0.047	0.050	0.048	0.048	0.055	0.051
50	300	0.052	0.050	0.049	0.049	0.051	0.050	0.050	0.051	0.050	0.048
70	271	0.045	0.047	0.048	0.048	0.045	0.044	0.045	0.046	0.047	0.046
100	250	0.047	0.046	0.045	0.045	0.049	0.046	0.047	0.047	0.047	0.047

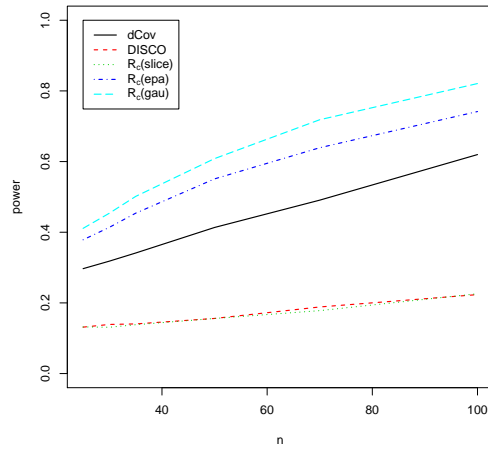
(g) $\chi^2(2), p = 5, q = 1$						
n	B	dCov	DISCO	$R_c(\text{slice})$	$R_c(\text{epa})$	$R_c(\text{gau})$
25	400	0.050	0.048	0.048	0.047	0.050
30	366	0.051	0.052	0.051	0.048	0.050
35	342	0.050	0.050	0.049	0.049	0.050
50	300	0.046	0.050	0.050	0.050	0.048
70	271	0.049	0.049	0.050	0.046	0.048
100	250	0.051	0.052	0.050	0.045	0.049



(a) $a = 0.5, p = 10, \sigma_x^2 = 1$ and $\sigma^2 = 1$. (b) $a = 0.3, p = 15, \sigma_x^2 = 1$ and $\sigma^2 = 1$.



(c) $a = 0.3, p = 20, \sigma_x^2 = 1$ and $\sigma^2 = 1$. (d) $a = 0.3, p = 10, \sigma_x^2 = 1$ and $\sigma^2 = 0.25$.



(e) $a = 0.3, p = 10, \sigma_x^2 = 0.5$ and $\sigma^2 = 1$

Figure S.4.5.1: Empirical power with the change of sample size n for other different parameter combinations.

Supplementary Materials for Chapter 3

This section provides proof of theorem 3.3.1 stated in chapter 3.

Proof of Theorem 3.3.1: We aim to show the uniform consistency of $\hat{\omega}_k$ under regularity condition. We use c as a generic constant, which may take different values at each appearance. Let $\{\tilde{X}_k, \tilde{Y}\}$ be an independent copy of $\{X_k, Y\}$, and \tilde{X}_{kY} be an independent copy of X_{kY} . That, \tilde{X}_{kY} and X_{kY} are \tilde{X}_k and X_k , conditioning on Y respectively.

Define $S_{k1} = E|\tilde{X}_k - X_k|$, $S_{k2} = E|\tilde{X}_{kY} - X_{kY}| = E_{Y=y}E|\tilde{X}_{kY} - X_{kY}| = ES_{k2y}$,

$$\hat{S}_{k1} = \frac{1}{n^2} \sum_{i=1, j=1}^{n, n} |X_{ik} - X_{jk}|$$

$$\hat{S}_{k2y} = \frac{1}{n_y^2} \sum_{i=1, j=1}^{n_y, n_y} |X_{iky} - X_{jky}|.$$

By the definitions, $\mathcal{C}^2(X_k|Y) = S_{k1} - S_{k2}$ and $\mathcal{C}_n^2(X_k|Y) = \hat{S}_{k1} - \sum_{y=1}^H p_y \hat{S}_{k2y} = \hat{S}_{k1} - \hat{S}_{k2}$, where $p_y = n_y/n$.

Note that given $Y = y$, basically term S_{k1} and S_{k2y} , and \hat{S}_{k1} and \hat{S}_{k2y} are of no difference, respectively. Hence, it suffices to prove \hat{S}_{k1} . However, note that $S_{k1} = S_{k2,1}$ and $\hat{S}_{k1} = \hat{S}_{k2,1}$, where $S_{k2,1}$ and $\hat{S}_{k2,1}$ appeared in (Li et al., 2012b, page 1137). And following their proof exactly as in (B.7), we have

$$Pr(|\hat{S}_{k1} - S_{k1}| \geq 4\epsilon) \leq 2 \exp(-\epsilon^2 n^{1-2\gamma}) + 2nc \exp(-sn^{2\gamma}/4),$$

Consequently, for any $y = 1 \dots, H$, we have

$$Pr(|\hat{S}_{k2y} - S_{k2y}| \geq 4\epsilon) \leq 2 \exp(-\epsilon^2 n_y^{1-2\gamma}) + 2n_y c \exp(-sn_y^{2\gamma}/4).$$

By the Bonferroni's inequality, we have that

$$\begin{aligned} Pr(|(\hat{S}_{k1} - \hat{S}_{k2}) - (S_{k1} - S_{k2})| \geq \epsilon) &\leq Pr(|\hat{S}_{k1} - S_{k1}| \geq \frac{\epsilon}{2}) + \sum_{y=1}^H Pr(|\hat{S}_{k2y} - S_{k2y}| \geq \frac{\epsilon}{2}) \\ &= O\{\exp(-c_1\epsilon^2n^{1-2\gamma}) + n \exp(-c_2n^{2\gamma})\}. \end{aligned}$$

In fact, the convergence rate of the denominator of $\hat{\omega}_k$ and itself are also the same as the numerator. Therefore,

$$Pr(|\hat{\omega}_k - \omega_k| > \epsilon) \leq O\{\exp(-c_1\epsilon^2n^{1-2\gamma}) + n \exp(-c_2n^{2\gamma})\}.$$

Let $\epsilon = cn^{-\tau}$, where $0 < \tau + \gamma < 1/2$, we have that

$$\begin{aligned} Pr(\max_{1 \leq k \leq p} |\hat{\omega}_k - \omega_k| \geq cn^{-\tau}) &\leq p \max_{1 \leq k \leq p} Pr(|\hat{\omega}_k - \omega_k| \geq cn^{-\tau}) \\ &\leq O[p\{\exp(-c_1n^{1-2(\tau+\gamma)}) + n \exp(-c_2n^{2\gamma})\}] \end{aligned}$$

Hence, we prove the first part of the Theorem. If $\mathcal{D}_m \not\subseteq \hat{\mathcal{D}}_m$, then there must exist some $k \in \mathcal{D}_m$ such that $\hat{\omega}_k < cn^{-\tau}$. It follows from condition (C2) that $|\hat{\omega}_k - \omega_k| > cn^{-\tau}$ for some $k \in \mathcal{D}_m$, indicating that the events satisfy $\{\mathcal{D}_m \not\subseteq \hat{\mathcal{D}}_m\} \subseteq \{|\hat{\omega}_k - \omega_k| > cn^{-\tau}, \text{ for some } k \in \mathcal{D}_m\}$, and hence, let $\epsilon_n = \{\max_{k \in \mathcal{D}} |\hat{\omega}_k - \omega_k| \leq cn^{-\tau}\} \subseteq \{\mathcal{D}_m \subseteq \hat{\mathcal{D}}_m\}$. Consequently,

$$\begin{aligned} Pr(\mathcal{D}_m \subseteq \hat{\mathcal{D}}_m) &\geq Pr(\epsilon_n) = 1 - Pr(\epsilon_n^c) \\ &= 1 - Pr(\min_{k \in \mathcal{D}} |\hat{\omega}_k - \omega_k| \geq cn^{-\tau}) \\ &= 1 - s_m Pr(|\hat{\omega}_k - \omega_k| \geq cn^{-\tau}) \\ &\geq 1 - O[s_m(\exp(-c_1n^{1-2(\tau+\gamma)}) + n \exp(-c_2n^{2\gamma}))], \end{aligned}$$

where s_m is the cardinality of \mathcal{D}_m . This completes the proof of the second part. \square

Proof of Theorem 3.3.2: Using similar argument, the above theorem also holds

for the marginal screening sequence.

$$\begin{aligned}
Pr\left(\max_{1 \leq k \leq p} \left| \sum_{y=1}^H p_y \hat{\mathcal{I}}_{k,y}^c - \sum_{y=1}^H p_y \mathcal{I}_{k,y}^c \right| \geq cn^{-\tau}\right) &\leq p \max_{1 \leq k \leq p} Pr\left(\left| \sum_{y=1}^H p_y (\hat{\mathcal{I}}_{k,y}^c - \mathcal{I}_{k,y}^c) \right| \geq cn^{-\tau}\right) \\
&= p \max_{1 \leq k \leq p} Pr\left(\left| \sum_{y=1}^H p_y (\hat{\mathcal{I}}_{k,y}^c - \mathcal{I}_{k,y}^c) \right| \geq \sum_{y=1}^H p_y cn^{-\tau}\right) \\
&\leq p \max_{1 \leq k \leq p} \sum_{y=1}^H Pr\left(|p_y (\hat{\mathcal{I}}_{k,y}^c - \mathcal{I}_{k,y}^c)| \geq p_y cn^{-\tau}\right) \\
&= p \max_{1 \leq k \leq p} \sum_{y=1}^H Pr\left(|\hat{\mathcal{I}}_{k,y}^c - \mathcal{I}_{k,y}^c| \geq cn^{-\tau}\right) \\
&\leq O[pH\{\exp(-c_3 n^{1-2(\tau+\gamma)}) + n \exp(-c_4 n^{2\gamma})\}]
\end{aligned}$$

Using similar argument, and denote the true and predictor active predictor set for the marginal dependency to be \mathcal{D}_c and $\hat{\mathcal{D}}_c$, we have similar result as follows:

$$Pr(\mathcal{D}_c \subseteq \hat{\mathcal{D}}_c) \geq 1 - O[s_c H(\exp(-c_3 n^{1-2(\tau+\gamma)}) + n \exp(-c_4 n^{2\gamma}))]. \quad \square$$

Proof of Theorem 3.3.3:

$$\begin{aligned}
Pr(\mathcal{D} \subseteq \hat{\mathcal{D}}) &= Pr((\mathcal{D}_c \cup \mathcal{D}_m) \subseteq (\hat{\mathcal{D}}_c \cup \hat{\mathcal{D}}_m)) \\
&\geq Pr((\mathcal{D}_c \subseteq \hat{\mathcal{D}}_c) \cap (\mathcal{D}_m \subseteq \hat{\mathcal{D}}_m)) \\
&= Pr(\mathcal{D}_c \subseteq \hat{\mathcal{D}}_c) + Pr(\mathcal{D}_m \subseteq \hat{\mathcal{D}}_m) - Pr((\mathcal{D}_c \subseteq \hat{\mathcal{D}}_c) \cup (\mathcal{D}_m \subseteq \hat{\mathcal{D}}_m)) \\
&\geq Pr(\mathcal{D}_c \subseteq \hat{\mathcal{D}}_c) + Pr(\mathcal{D}_m \subseteq \hat{\mathcal{D}}_m) - 1 \\
&\geq 1 - O[s_m(\exp(-c_1 n^{1-2(\tau+\gamma)}) + n \exp(-c_2 n^{2\gamma}))] \\
&\quad + 1 - O[s_c H(\exp(-c_3 n^{1-2(\tau+\gamma)}) + n \exp(-c_4 n^{2\gamma}))] - 1 \\
&\geq 1 - O[s(\exp(-c_5 n^{1-2(\tau+\gamma)}) + n \exp(-c_6 n^{2\gamma}))].
\end{aligned}$$

Where s is the minimum of s_m and $s_c H$. \square

Supplementary Materials for Chapter 4

This section provides proofs of propositions and theorems stated in chapter 4.

Lemma S. 4.5.4. *Suppose $\boldsymbol{\eta}$ is a basis of the central subspace. Let $(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$ be any partition of $\boldsymbol{\eta}$, where $\boldsymbol{\eta}^T \Sigma_X \boldsymbol{\eta} = I_d$. We have $\mathcal{C}^2(\boldsymbol{\eta}_i^T |, Y) < \mathcal{C}^2(\boldsymbol{\eta}^T X | Y)$, $i = 1, 2$.*

Proof: Let $\tilde{X}_1 = \boldsymbol{\eta}_1^T X$, $\tilde{X}_2 = \boldsymbol{\eta}_2^T X$, $F(a, b) = \mathcal{C}^2\left(\begin{pmatrix} a\tilde{X}_1 \\ b\tilde{X}_2 \end{pmatrix} | Y\right)$, $a \in R$ and $b \in R$, and $G_1(a, b) = \partial F(a, b)/\partial a$, $G_2(a, b) = \partial F(a, b)/\partial b$. A simple calculation shows that $aG_1(a, b) + bG_2(a, b) = F(a, b)$

If $(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2) \in \mathcal{S}(\boldsymbol{\eta})$, then $F(0, 1), F(1, 0) > 0$; otherwise, the conclusion automatically holds.

Claim, if $0 \leq \lambda < 1$, then $F(1, \lambda) < F(1, 1)$ and $F(\lambda, 1) < F(1, 1)$.

If not, then there exist a $0 \leq \lambda_0 < 1$ such that $F(1, \lambda_0) \geq F(1, 1)$ or $F(\lambda_0, 1) \geq F(1, 1)$. Without loss of generality, we assume there exist a $0 \leq \lambda_0 < 1$ such that $F(1, \lambda_0) \geq F(1, 1)$.

But $F(1, \lambda) = \lambda F(\frac{1}{\lambda}, 1)$, and as $\lambda \rightarrow \infty$, $F(\frac{1}{\lambda}, 1) \rightarrow F(0, 1) > 0$. Thus $F(1, \lambda) \rightarrow \infty$, as $\lambda \rightarrow \infty$. That means, there exists a $\lambda_1 \in (\lambda_0, \infty)$ such that $F(1, \lambda_1)$ achieves a minimum in (λ_0, ∞) . Hence, $G_2(1, \lambda_1) = 0$. Note that function $F(a, b)$ is a ‘‘ray’’ function, i. e. $F(ca, cb) = cF(a, b)$. Thus using the fact that $F(1, \lambda) = \lambda F(\frac{1}{\lambda}, 1)$, we can have $G_1(\frac{1}{\lambda_1}, 1) = 0$. And it is easy to calculate that $G_1(1, \lambda_1) = G_1(\frac{1}{\lambda_1}, 1) = 0$.

But $0 = 1G_1(1, \lambda_1) + \lambda_1 G_2(1, \lambda_1) = F(1, \lambda_1)$. $F(1, \lambda_1) = 0$ means that $\begin{pmatrix} \tilde{X}_1 \\ \lambda_1 \tilde{X}_2 \end{pmatrix} \perp Y$, which conflicts with our assumption. \square

Proof of Proposition 4.2.1: Let $\boldsymbol{\eta}_0$ be the projection of $\boldsymbol{\beta}$ onto $\boldsymbol{\eta}$, which means $\boldsymbol{\eta}_0 = P_{\boldsymbol{\eta}(\Sigma_X)} \boldsymbol{\beta} = \boldsymbol{\eta}c$, where c is a scalar. Let $\boldsymbol{\eta}_0^\perp = \boldsymbol{\beta} - \boldsymbol{\eta}_0$, where the orthogonality ‘ \perp ’ is the inner product induced by Σ_X , then $1 = \boldsymbol{\beta}^T \Sigma_X \boldsymbol{\beta} = c^2 + \boldsymbol{\eta}_0^{\perp, T} \Sigma_X \boldsymbol{\eta}_0^\perp \geq c^2$.

Now, by (4.1)

$$\begin{aligned}
& \mathcal{C}^2(\boldsymbol{\beta}^T \mathbf{X} | \mathbf{Y}) \\
&= \int |E(e^{i\langle t, \boldsymbol{\beta}^T \mathbf{X} \rangle} | \mathbf{Y}) - Ee^{i\langle t, \boldsymbol{\beta}^T \mathbf{X} \rangle}|^2 dw \\
&= \int |E[E\{e^{i\langle t, (\boldsymbol{\eta}_0^T + \boldsymbol{\eta}_0^{\perp T}) \mathbf{X} \rangle} | Y, \boldsymbol{\eta}^T \mathbf{X}\} | \mathbf{Y}] - Ee^{i\langle t, (\boldsymbol{\eta}_0^T + \boldsymbol{\eta}_0^{\perp T}) \mathbf{X} \rangle}|^2 dw \\
&= \int |E[E\{e^{i\langle t, (\boldsymbol{\eta}_0^T + \boldsymbol{\eta}_0^{\perp T}) \mathbf{X} \rangle} | \boldsymbol{\eta}^T \mathbf{X}\} | \mathbf{Y}] - Ee^{i\langle t, (\boldsymbol{\eta}_0^T + \boldsymbol{\eta}_0^{\perp T}) \mathbf{X} \rangle}|^2 dw \\
&= \int |E[e^{i\langle t, \boldsymbol{\eta}_0^T \mathbf{X} \rangle} E\{e^{i\langle t, \boldsymbol{\eta}_0^{\perp T} \mathbf{X} \rangle} | \boldsymbol{\eta}^T \mathbf{X}\} | \mathbf{Y}] - Ee^{i\langle t, \boldsymbol{\eta}_0^T \mathbf{X} \rangle} Ee^{i\langle t, \boldsymbol{\eta}_0^{\perp T} \mathbf{X} \rangle}|^2 dw \\
&= \int |E[e^{i\langle t, \boldsymbol{\eta}_0^T \mathbf{X} \rangle} | \mathbf{Y}] Ee^{i\langle t, \boldsymbol{\eta}_0^{\perp T} \mathbf{X} \rangle} - Ee^{i\langle t, \boldsymbol{\eta}_0^T \mathbf{X} \rangle} Ee^{i\langle t, \boldsymbol{\eta}_0^{\perp T} \mathbf{X} \rangle}|^2 dw \\
&= \int |Ee^{i\langle t, \boldsymbol{\eta}_0^{\perp T} \mathbf{X} \rangle} \{E[e^{i\langle t, \boldsymbol{\eta}_0^T \mathbf{X} \rangle} | \mathbf{Y}] - Ee^{i\langle t, \boldsymbol{\eta}_0^T \mathbf{X} \rangle}\}|^2 dw \\
&= \int |Ee^{i\langle t, \boldsymbol{\eta}_0^{\perp T} \mathbf{X} \rangle}|^2 |E[e^{i\langle t, \boldsymbol{\eta}_0^T \mathbf{X} \rangle} | \mathbf{Y}] - Ee^{i\langle t, \boldsymbol{\eta}_0^T \mathbf{X} \rangle}|^2 dw \\
&\leq \int |E[e^{i\langle t, \boldsymbol{\eta}_0^T \mathbf{X} \rangle} | \mathbf{Y}] - Ee^{i\langle t, \boldsymbol{\eta}_0^T \mathbf{X} \rangle}|^2 dw \\
&= \mathcal{C}^2(\boldsymbol{\eta}_0^T \mathbf{X} | \mathbf{Y}) \\
&\leq \mathcal{C}^2(\boldsymbol{\eta}^T \mathbf{X}, \mathbf{Y})
\end{aligned}$$

The third equality follows from the assumption $Y \perp\!\!\!\perp X | \boldsymbol{\eta}^T X$, and $\boldsymbol{\eta}_0 = \boldsymbol{\eta}c$. The fourth equality follows from the assumption $P_{\boldsymbol{\eta}(\Sigma_X)}^T X \perp\!\!\!\perp Q_{\boldsymbol{\eta}(\Sigma_X)}^T X$. The last inequality follows from the second property in chapter 2. The maximum is achieved by setting $|c| = 1$, which indicates $\text{Span}(\boldsymbol{\beta}) = \text{Span}(\boldsymbol{\eta})$. \square

Proof of Proposition 4.2.2: Since $\mathcal{S}(\boldsymbol{\beta}) \subseteq \mathcal{S}(\boldsymbol{\eta}) = S_{Y|X}$, $d_1 \leq d$, there exists a matrix A , which satisfies $\boldsymbol{\beta} = \boldsymbol{\eta}A$. Therefore, $\mathcal{C}^2(\boldsymbol{\beta}^T X | Y) = \mathcal{C}^2(A^T \boldsymbol{\eta}^T X | Y)$.

Assume the single value decomposition of A is $U\Sigma V^T$, where U is a $d \times d$ orthogonal matrix, V is a $d_1 \times d_1$ orthogonal matrix and Σ is a $d \times d_1$ diagonal matrix with nonnegative numbers on the diagonal, and it is easy to prove that all nonnegative numbers on the diagonal of Σ are 1. Based on Theorem 2.3.2, part (2) in chapter 2, $\mathcal{C}^2(\boldsymbol{\beta}^T X | Y) = \mathcal{C}^2(V\Sigma^T U^T \boldsymbol{\eta}^T X | Y) = \mathcal{C}^2(\Sigma^T U^T \boldsymbol{\eta}^T X | Y)$.

Let $U^T \boldsymbol{\eta}^T X = (\tilde{X}_1, \dots, \tilde{X}_d)^T$. Since all nonnegative numbers on the diagonal of Σ are 1 and $\Sigma^T U^T \boldsymbol{\eta}^T \mathbf{X} = (\tilde{X}_1, \dots, \tilde{X}_{d_1})^T$, by Lemma S. 4.5.4, we get $\mathcal{C}^2(\Sigma^T U^T \boldsymbol{\eta}^T X | Y) \leq$

$\mathcal{C}^2(U^T \eta^T X|Y)$. The equality holds if and only if $d = d_1$. And again based on Theorem 2.3.2, part (2) in chapter 2, $\mathcal{C}^2(U^T \eta^T X, Y) = \mathcal{C}^2(\eta^T X|Y)$. Thus, $\mathcal{C}^2(\beta^T X|Y) \leq \mathcal{C}^2(\eta^T X, Y)$, and equality holds if and only if $\mathcal{S}(\beta) = \mathcal{S}(\eta)$. \square

Proof of Proposition 4.2.3: For the β and η described in Proposition 2, there exists a rotation matrix \mathbf{Q} such that $\beta \mathbf{Q} = (\eta_a, \eta_b)$, and $\mathcal{S}(\eta_a) \subseteq \mathcal{S}(\eta)$, $\mathcal{S}(\eta_b) \subseteq \mathcal{S}(\eta)^\perp$, where $\mathcal{S}(\eta)^\perp$ is the orthogonal space of $\mathcal{S}(\eta)$.

Since $Y \perp \eta_b^T X | \eta^T X$ and $P_{\eta(\Sigma_X)}^T \mathbf{X} \perp Q_{\eta(\Sigma_X)}^T \mathbf{X}$, therefore $\begin{pmatrix} Y \\ \eta^T X \end{pmatrix} \perp \eta_b^T X$, and according to Proposition 4.3 (Cook, 1998b), $\begin{pmatrix} Y \\ \eta_a^T X \end{pmatrix} \perp \eta_b^T X$. Let $W_1 = \begin{pmatrix} \eta_a^T X \\ \mathbf{0} \end{pmatrix}$, $V_1 = Y$, $W_2 = \begin{pmatrix} \mathbf{0} \\ \eta_b^T X \end{pmatrix}$, and $V_2 = 0$, then $(W_1, V_1) \perp (W_2, V_2)$. According to Theorem 2.3.2, part (3) in chapter 2, $\mathcal{C}(W_1+W_2|V_1+V_2) < \mathcal{C}(W_1|V_1) + \mathcal{C}(W_2|V_2)$, that is $\mathcal{C}^2(\mathbf{Q}^T \beta^T \mathbf{X}|Y) = \mathcal{C}^2(\beta^T \mathbf{X}|Y) < \mathcal{C}^2(\eta_a^T \mathbf{X}|Y) \leq \mathcal{C}^2(\eta^T \mathbf{X}|Y)$. \square

Notations and Conditions

We reconstruct the optimization problem by using the Lagrange multiplier technique, and we introduce the following notations.

Let $(\mathbf{X}, \mathbf{Y}) = \{(X_k, Y_k), k = 1, \dots, n\}$ to be a random sample from the joint distribution of random vector $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}$. Let $\mathcal{L}(\zeta) = \mathcal{C}^2(\beta^T X|Y) + \lambda(\beta^T \Sigma_X \beta - 1)$ and $\mathcal{L}_n(\zeta) = \mathcal{C}_n^2(\beta^T \mathbf{X}|\mathbf{Y}) + \lambda(\beta^T \hat{\Sigma}_X \beta - 1)$.

Here $\zeta = \begin{pmatrix} \beta \\ \lambda \end{pmatrix} \in \mathbb{R}^{p+1}$, $\beta \in \mathbb{R}^p$, $\lambda \in \mathbb{R}$, Σ_X is the covariance matrix of X , and $\hat{\Sigma}_X$ is the sample estimate for Σ_X .

Under the condition $P_{\eta(\Sigma_X)}^T X \perp Q_{\eta(\Sigma_X)}^T X$ and the assumption that CS is unique, let $\eta = \arg \max_{\beta^T \Sigma_X \beta = 1} \mathcal{C}^2(\beta^T X|Y)$ and $\eta_n = \arg \max_{\beta^T \hat{\Sigma}_X \beta = 1} \mathcal{C}_n^2(\beta^T \mathbf{X}|\mathbf{Y})$, then there exist λ_0 and λ_n such that $\begin{pmatrix} \eta \\ \lambda_0 \end{pmatrix}$ is a stationary point for $\mathcal{L}(\zeta)$ and $\begin{pmatrix} \eta_n \\ \lambda_n \end{pmatrix}$ is a stationary point for $\mathcal{L}_n(\zeta)$. On the other hand, since $-\eta$ is another maximizer of $\mathcal{C}^2(\beta^T X|Y)$, and $(-\beta^T) \Sigma_X (-\beta) = \beta^T \Sigma_X \beta$, therefore $\begin{pmatrix} -\eta \\ \lambda_0 \end{pmatrix}$ is also a stationary

point of $\mathcal{L}(\zeta)$. So to speak, $\begin{pmatrix} -\boldsymbol{\eta}_n \\ \lambda_n \end{pmatrix}$ is also stationary point for $\mathcal{L}_n(\zeta)$.

Now let $\theta_n = \begin{pmatrix} \boldsymbol{\eta}_n \\ \lambda_n \end{pmatrix}$, then $\theta_n = \arg \max \mathcal{L}_n(\zeta)$. Note that $\begin{pmatrix} c\boldsymbol{\eta} \\ \lambda_0 \end{pmatrix} = \arg \max \mathcal{L}(\zeta)$, where, $c = \pm 1$. Here $\pm\boldsymbol{\eta}$ and $\boldsymbol{\eta}_n \in \mathbb{R}^p$, λ_0 and $\lambda_n \in \mathbb{R}$.

In order to simplify the proofs, throughout this section in the appendix, without loss of generality, we can assume that the first none zero elements in both of $\boldsymbol{\eta}_n$ and $\boldsymbol{\eta}$ have the same sign by setting $c = 1$, and let $\theta = \begin{pmatrix} \boldsymbol{\eta} \\ \lambda_0 \end{pmatrix}$. Otherwise, set $c = -1$.

Furthermore, we make the following assumption.

Assumption S. 4.5.1. $Var [\phi^{(1)}(X_1, X_2)]$, $Var [\phi^{(2)}(X_{1y}, X_{2y})]$, $Var [\phi^{(4)}(X_1)]$, $Var [\phi^{(5)}(X_1, X_2)]$, $Var [\phi^{(6)}(X_1)]$, $Var [\phi^{(7)}(X_1, X_2)]$ are all $< \infty$, where X_1 and X_2 are iid copies, and X_{1y} and X_{2y} are iid copies, respectively, and

$$\begin{aligned} \phi^{(1)}(X_1, X_2) &= \frac{(X_1 - X_2)(X_1 - X_2)^T \boldsymbol{\eta}}{|\boldsymbol{\eta}^T (X_1 - X_2)|}, \\ \phi^{(2)}(X_{1y}, X_{2y}) &= \frac{(X_{1y} - X_{2y})(X_{1y} - X_{2y})^T \boldsymbol{\eta}}{|\boldsymbol{\eta}^T (X_{1y} - X_{2y})|}, \text{ for } y = 1, \dots, C, \\ \phi^{(4)}(X_1) &= X_1 X_1^T \boldsymbol{\eta}, \\ \phi^{(5)}(X_1, X_2) &= \frac{1}{2}(X_1 X_2^T + X_2 X_1^T) \boldsymbol{\eta}, \\ \phi^{(6)}(X_1) &= \boldsymbol{\eta}^T X_1 X_1^T \boldsymbol{\eta}, \\ \phi^{(7)}(X_1, X_2) &= \frac{1}{2} \boldsymbol{\eta}^T (X_1 X_2^T + X_2 X_1^T) \boldsymbol{\eta}. \end{aligned}$$

Assumption 4.5.1 is needed for Proposition 4.2.3 in the paper and Lemma 4.5.7 in the next Section, which is similar to the assumed conditions of Theorem 6.1.6 (Lehmann, 1999, Ch.6) so that in the spirit of von Mises proposition (Serfling, 1980, Section 6.1), the first nonvanishing term of our Taylor expansion is the linear term. Hence root- n result can be proved. If this term is vanished, then n or higher order-consistency can be proved.

Relevant Lemmas

Lemma S. 4.5.5. *If the support of X , say S , is compact, $E|Y| < \infty$ and $\boldsymbol{\theta}_n \xrightarrow{P} \boldsymbol{\theta}$, then $\mathcal{L}_n(\boldsymbol{\theta}_n) - \mathcal{L}_n(\boldsymbol{\theta}) \xrightarrow{P} 0$.*

Lemma S. 4.5.6. *If the support of X , say S , is compact, $E|Y| < \infty$, then $\boldsymbol{\theta}_n \xrightarrow{P} \boldsymbol{\theta}$.*

Lemma S. 4.5.7. *Under assumption 4.5.1 and the assumptions in Proposition 4.2.3, then $\sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}) \xrightarrow{D} N(0, V)$. The explicit expression for V is in the proof.*

Proofs of the Relevant Lemmas

Proof of Lemma S. 4.5.5:

$$\begin{aligned} \mathcal{L}_n(\boldsymbol{\theta}_n) - \mathcal{L}_n(\boldsymbol{\theta}) &= \mathcal{C}_n^2(\boldsymbol{\eta}_n^T \mathbf{X} | \mathbf{Y}) + \lambda_n(\boldsymbol{\eta}_n^T \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\eta}_n - 1) - \mathcal{C}_n^2(\boldsymbol{\eta}^T \mathbf{X} | \mathbf{Y}) - \lambda_0(\boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\eta} - 1) \\ &= \mathcal{C}_n^2(\boldsymbol{\eta}_n^T \mathbf{X} | \mathbf{Y}) - \mathcal{C}_n^2(\boldsymbol{\eta}^T \mathbf{X} | \mathbf{Y}) + \lambda_n(\boldsymbol{\eta}_n^T \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\eta}_n - 1) - \lambda_0(\boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\eta} - 1). \end{aligned}$$

Since $\boldsymbol{\theta}_n \xrightarrow{P} \boldsymbol{\theta}$, therefore $\boldsymbol{\eta}_n \xrightarrow{P} \boldsymbol{\eta}$ and $\lambda_n \xrightarrow{P} \lambda_0$, and we know $\hat{\Sigma}_{\mathbf{X}} \xrightarrow{a.s.} \Sigma_{\mathbf{X}}$. Hence $\lambda_n \boldsymbol{\eta}_n^T \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\eta}_n \xrightarrow{P} \lambda_0 \boldsymbol{\eta}^T \Sigma_{\mathbf{X}} \boldsymbol{\eta} = \lambda_0$, and $\lambda_0 \boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\eta} \xrightarrow{a.s.} \lambda_0 \boldsymbol{\eta}^T \Sigma_{\mathbf{X}} \boldsymbol{\eta} = \lambda_0$. Therefore $\lambda_n(\boldsymbol{\eta}_n^T \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\eta}_n - 1) - \lambda_0(\boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\eta} - 1) = (\lambda_n \boldsymbol{\eta}_n^T \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\eta}_n - \lambda_0 \boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\eta}) - (\lambda_n - \lambda_0) \xrightarrow{P} 0$. Now in order to prove Lemma 4.5.5, we only need to prove $\mathcal{C}_n^2(\boldsymbol{\eta}_n^T \mathbf{X} | \mathbf{Y}) - \mathcal{C}_n^2(\boldsymbol{\eta}^T \mathbf{X} | \mathbf{Y}) \xrightarrow{P} 0$, which is proved next.

We have that

$$\begin{aligned} a_{kl}(\boldsymbol{\eta}_n) &= |\boldsymbol{\eta}_n^T X_k - \boldsymbol{\eta}_n^T X_l|, \text{ for } k, l = 1, \dots, n, \\ b_{kly}(\boldsymbol{\eta}_n) &= |\boldsymbol{\eta}_n^T X_{ky} - \boldsymbol{\eta}_n^T X_{ly}|, \text{ for } k, l = 1, \dots, n_y, y = 1 \dots, C. \end{aligned}$$

Then we have

$$\mathcal{C}_n^2(\boldsymbol{\eta}_n^T \mathbf{X} | \mathbf{Y}) = \frac{1}{n^2} \sum_{k,l=1}^n a_{kl}(\boldsymbol{\eta}_n) - \frac{1}{n} \sum_{y=1}^C \frac{1}{n_y} \sum_{k,l=1}^{n_y} a_{kly}(\boldsymbol{\eta}_n),$$

and,

$$\mathcal{C}_n^2(\boldsymbol{\eta}^T \mathbf{X} | \mathbf{Y}) = \frac{1}{n^2} \sum_{k,l=1}^n a_{kl}(\boldsymbol{\eta}) - \frac{1}{n} \sum_{y=1}^C \frac{1}{n_y} \sum_{k,l=1}^{n_y} a_{kly}(\boldsymbol{\eta}).$$

Because $\boldsymbol{\eta}_n \rightarrow \boldsymbol{\eta}$ in probability, let $\boldsymbol{\eta}_n = \boldsymbol{\eta} + \boldsymbol{\varepsilon}_n$, then for any $\epsilon > 0$, $|\boldsymbol{\varepsilon}_n| < \epsilon$, when $n \rightarrow \infty$. Hence, by the condition on X , we have that for a positive constant C_x , and large n , $|a_{kl}(\boldsymbol{\eta}_n) - a_{kl}(\boldsymbol{\eta})|$ and $|a_{kly}(\boldsymbol{\eta}_n) - a_{kly}(\boldsymbol{\eta})| \leq \epsilon C_x$. Therefore,

$$\begin{aligned} |\mathcal{C}_n^2(\boldsymbol{\eta}_n^T \mathbf{X} | \mathbf{Y}) - \mathcal{C}_n^2(\boldsymbol{\eta}^T \mathbf{X} | \mathbf{Y})| &\leq \frac{1}{n^2} \sum_{k,l=1}^n |a_{kl}(\boldsymbol{\eta}_n) - a_{kl}(\boldsymbol{\eta})| + \frac{1}{n} \sum_{y=1}^C \frac{1}{n_y} \sum_{k,l=1}^{n_y} |a_{kly}(\boldsymbol{\eta}_n) - a_{kly}(\boldsymbol{\eta})| \\ &\leq 2\epsilon C_x. \end{aligned}$$

Hence, the conclusion follows. \square

Proof of Lemma S. 4.5.6: Suppose $\boldsymbol{\theta}_n$ fails to converge to $\boldsymbol{\theta}$ with probability 1, then there exists a subsequence, still to be indexed by n , and an $\boldsymbol{\theta}^* = \begin{pmatrix} \boldsymbol{\eta}^* \\ \lambda^* \end{pmatrix} \in \mathbb{R}^{p+1}$ satisfying $\boldsymbol{\eta}^{*T} \Sigma_X \boldsymbol{\eta}^* = 1$ and $\boldsymbol{\theta}^* \neq \boldsymbol{\theta}$, such that $\boldsymbol{\theta}_n \xrightarrow{P} \boldsymbol{\theta}^*$. If so, $\boldsymbol{\eta}_n \xrightarrow{P} \boldsymbol{\eta}^*$ and $\lambda_n \xrightarrow{P} \lambda^*$. Note that $\boldsymbol{\eta}^* \neq -\boldsymbol{\eta}$ by setting $c = 1$, previously.

By lemma S. 4.5.5, if $\boldsymbol{\theta}_n \xrightarrow{P} \boldsymbol{\theta}^*$, then $\mathcal{L}_n(\boldsymbol{\theta}_n) - \mathcal{L}_n(\boldsymbol{\theta}^*) \xrightarrow{P} 0$, where $\mathcal{L}_n(\boldsymbol{\theta}^*) = \mathcal{C}_n^2(\boldsymbol{\eta}^{*T} \mathbf{X} | \mathbf{Y}) + \lambda^*(\boldsymbol{\eta}^{*T} \hat{\Sigma}_X \boldsymbol{\eta}^* - 1)$.

We know $\mathcal{C}_n^2(\boldsymbol{\eta}^{*T} \mathbf{X} | \mathbf{Y}) \xrightarrow{a.s.} \mathcal{C}^2(\boldsymbol{\eta}^{*T} X | Y)$. And since $\hat{\Sigma}_X = \frac{1}{n} \sum_{i=1}^n X_i X_i^T - \bar{X} \bar{X}^T \xrightarrow{a.s.} \Sigma_X$, therefore $\lambda^*(\boldsymbol{\eta}^{*T} \hat{\Sigma}_X \boldsymbol{\eta}^* - 1) \xrightarrow{a.s.} \lambda^*(\boldsymbol{\eta}^{*T} \Sigma_X \boldsymbol{\eta}^* - 1)$. Hence $\mathcal{L}_n(\boldsymbol{\theta}^*) \xrightarrow{a.s.} \mathcal{L}(\boldsymbol{\theta}^*)$. With $\mathcal{L}_n(\boldsymbol{\theta}_n) - \mathcal{L}_n(\boldsymbol{\theta}^*) \xrightarrow{P} 0$, we get $\mathcal{L}_n(\boldsymbol{\theta}_n) \xrightarrow{P} \mathcal{L}(\boldsymbol{\theta}^*)$.

On the other hand, since $\boldsymbol{\theta}_n = \arg \max \mathcal{L}_n(\zeta)$, therefore $\mathcal{L}_n(\boldsymbol{\theta}_n) \geq \mathcal{L}_n(\boldsymbol{\theta})$. If we take the limit on both sides of the inequality, we get $\mathcal{L}(\boldsymbol{\theta}^*) \geq \mathcal{L}(\boldsymbol{\theta})$. However, this result conflicts with our assumption that $\boldsymbol{\theta} = \arg \max \mathcal{L}(\zeta)$ and the uniqueness of the CS. Therefore, $\boldsymbol{\theta}_n \xrightarrow{P} \boldsymbol{\theta}$. \square

Proof of Lemma 4.5.7: For simplicity of notation, let $\mathcal{C}_n(\boldsymbol{\eta}) = \mathcal{C}_n^2(\boldsymbol{\eta}^T \mathbf{X} | Y)$. The Taylor expansion of $\mathcal{L}'_n(\boldsymbol{\theta}_n)$ at $\boldsymbol{\theta}$ is $0 = \mathcal{L}'_n(\boldsymbol{\theta}_n) = \mathcal{L}'_n(\boldsymbol{\theta}) + \mathcal{L}''_n(\boldsymbol{\theta})(\boldsymbol{\theta}_n - \boldsymbol{\theta}) + \mathcal{R}_1(\boldsymbol{\theta}_n^*)$, where $|\boldsymbol{\theta}_n^* - \boldsymbol{\theta}| \leq |\boldsymbol{\theta}_n - \boldsymbol{\theta}|$, and $\boldsymbol{\theta}_n^* = \begin{pmatrix} \boldsymbol{\eta}_n^* \\ \lambda_n^* \end{pmatrix}$. Next, we will give explicit expressions of $\mathcal{L}'_n(\boldsymbol{\theta})$, $\mathcal{L}''_n(\boldsymbol{\theta})$ and $\mathcal{R}_1(\boldsymbol{\theta}_n^*)$. With simple calculation,

$$\mathcal{L}'_n(\boldsymbol{\theta}) = \begin{pmatrix} \mathcal{C}'_n(\boldsymbol{\eta}) + 2\lambda_0 \hat{\Sigma}_X \boldsymbol{\eta} \\ \boldsymbol{\eta}^T \hat{\Sigma}_X \boldsymbol{\eta} - 1 \end{pmatrix}; \quad \mathcal{L}''_n(\boldsymbol{\theta}) = \begin{pmatrix} \mathcal{C}''_n(\boldsymbol{\eta}) + 2\lambda_0 \hat{\Sigma}_X & 2\hat{\Sigma}_X \boldsymbol{\eta} \\ 2\boldsymbol{\eta}^T \hat{\Sigma}_X & 0 \end{pmatrix}.$$

Furthermore, we notice that $\mathcal{C}_n''(\boldsymbol{\eta}) = 0$. This is because $\mathcal{C}_n(\boldsymbol{\eta}) = \mathcal{C}_n^2(\boldsymbol{\eta}^T X|Y) = S_1(\boldsymbol{\eta}) - S_2(\boldsymbol{\eta})$, where

$$S_1(\boldsymbol{\eta}) = \frac{1}{n^2} \sum_{k,l=1}^n |\boldsymbol{\eta}^T (X_k - X_l)|,$$

$$S_2(\boldsymbol{\eta}) = \frac{1}{n} \sum_{y=1}^C \frac{1}{n_y} \sum_{k,l=1}^{n_y} |\boldsymbol{\eta}^T (X_{ky} - X_{ly})|.$$

However, a simple calculation shows that

$$S_1''(\boldsymbol{\eta}) = \frac{1}{n^2} \sum_{k,l=1}^n \{[(X_k - X_l)(X_k - X_l)^T][\boldsymbol{\eta}^T (X_k - X_l)(X_k - X_l)^T \boldsymbol{\eta}]^{-\frac{1}{2}} \\ - [(X_k - X_l)(X_k - X_l)^T \boldsymbol{\eta}][\boldsymbol{\eta}^T (X_k - X_l)(X_k - X_l)^T \boldsymbol{\eta}]^{-\frac{3}{2}} \boldsymbol{\eta}^T (X_k - X_l)(X_k - X_l)^T\} = 0.$$

Similarly, $S_2''(\boldsymbol{\eta}) = 0$, therefore $\mathcal{C}_n''(\boldsymbol{\eta}) = 0$.

Thus we obtain that $\begin{pmatrix} \mathcal{C}_n''(\boldsymbol{\eta}) + 2\lambda_0 \hat{\Sigma}_{\mathbf{X}} & 2\hat{\Sigma}_{\mathbf{X}} \boldsymbol{\eta} \\ 2\boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}} & 0 \end{pmatrix} = \begin{pmatrix} 2\lambda_0 \hat{\Sigma}_{\mathbf{X}} & 2\hat{\Sigma}_{\mathbf{X}} \boldsymbol{\eta} \\ 2\boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}} & 0 \end{pmatrix}$, which converges to $\begin{pmatrix} 2\lambda_0 \Sigma_{\mathbf{X}} & 2\Sigma_{\mathbf{X}} \boldsymbol{\eta} \\ 2\boldsymbol{\eta}^T \Sigma_{\mathbf{X}} & 0 \end{pmatrix}$ almost surely.

Since $\begin{vmatrix} 2\lambda_0 \Sigma_{\mathbf{X}} & 2\Sigma_{\mathbf{X}} \boldsymbol{\eta} \\ 2\boldsymbol{\eta}^T \Sigma_{\mathbf{X}} & 0 \end{vmatrix} = -2^{p+1} \lambda_0^{p-1} |\Sigma_{\mathbf{X}}| \neq 0$, thus $\begin{pmatrix} 2\lambda_0 \hat{\Sigma}_{\mathbf{X}} & 2\hat{\Sigma}_{\mathbf{X}} \boldsymbol{\eta} \\ 2\boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}} & 0 \end{pmatrix}$ is invertible when n is large.

As for $\mathcal{R}_1(\boldsymbol{\theta}_n^*)$, let $T_n = \mathcal{L}_n'''(\boldsymbol{\theta}_n^*)$, where T_n is a $(p+1) \times (p+1) \times (p+1)$ array. Each $T_n(j, :, :)$, $j = 1, \dots, p+1$ is a $(p+1) \times (p+1)$ matrix.

Let $\hat{\Sigma}_{\mathbf{X}} = \begin{pmatrix} \hat{\sigma}_{11} & \hat{\sigma}_{12} & \cdots & \hat{\sigma}_{1p} \\ \hat{\sigma}_{21} & \hat{\sigma}_{22} & \cdots & \hat{\sigma}_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\sigma}_{p1} & \hat{\sigma}_{p2} & \cdots & \hat{\sigma}_{pp} \end{pmatrix}$, then we can write

$$T_n(j, :, :) = 2 \begin{pmatrix} 0 & 0 & \cdots & 0 & \hat{\sigma}_{j1} \\ 0 & 0 & \cdots & 0 & \hat{\sigma}_{j2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & \hat{\sigma}_{jp} \\ \hat{\sigma}_{j1} & \hat{\sigma}_{j2} & \cdots & \hat{\sigma}_{jp} & 0 \end{pmatrix}, j = 1, 2, \dots, p \text{ and}$$

$$T_n(p+1, :, :) = 2 \begin{pmatrix} \hat{\sigma}_{11} & \hat{\sigma}_{21} & \cdots & \hat{\sigma}_{p1} & 0 \\ \hat{\sigma}_{12} & \hat{\sigma}_{22} & \cdots & \hat{\sigma}_{p2} & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \hat{\sigma}_{1p} & \hat{\sigma}_{2p} & \cdots & \hat{\sigma}_{pp} & 0 \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix}.$$

The form of $T_n(j, :, :)$, $j = 1, 2, \dots, p+1$ indicates that T_n is not affected by the value of $\boldsymbol{\theta}^*$. The form of $\mathcal{R}_1(\boldsymbol{\theta}_n^*)$ can be written as

$$\mathcal{R}_1(\boldsymbol{\theta}_n^*) = \frac{1}{2} \begin{pmatrix} (\boldsymbol{\theta}_n - \boldsymbol{\theta})^T T_n(1, :, :)(\boldsymbol{\theta}_n - \boldsymbol{\theta}) \\ (\boldsymbol{\theta}_n - \boldsymbol{\theta})^T T_n(2, :, :)(\boldsymbol{\theta}_n - \boldsymbol{\theta}) \\ \vdots \\ (\boldsymbol{\theta}_n - \boldsymbol{\theta})^T T_n(p+1, :, :)(\boldsymbol{\theta}_n - \boldsymbol{\theta}) \end{pmatrix}.$$

Therefore, the Taylor expansion of $\mathcal{L}'_n(\boldsymbol{\theta}_n)$ at $\boldsymbol{\theta}$ can be written as

$$0 = \begin{pmatrix} \mathcal{C}'_n(\boldsymbol{\eta}) + 2\lambda_0 \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\eta} \\ \boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\eta} - 1 \end{pmatrix} + \begin{pmatrix} 2\lambda_0 \hat{\Sigma}_{\mathbf{X}} & 2\hat{\Sigma}_{\mathbf{X}} \boldsymbol{\eta} \\ 2\boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}} & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{\eta}_n - \boldsymbol{\eta} \\ \lambda_n - \lambda_0 \end{pmatrix}$$

$$+ \frac{1}{2} \begin{pmatrix} (\boldsymbol{\theta}_n - \boldsymbol{\theta})^T T_n(1, :, :)(\boldsymbol{\theta}_n - \boldsymbol{\theta}) \\ (\boldsymbol{\theta}_n - \boldsymbol{\theta})^T T_n(2, :, :)(\boldsymbol{\theta}_n - \boldsymbol{\theta}) \\ \vdots \\ (\boldsymbol{\theta}_n - \boldsymbol{\theta})^T T_n(p+1, :, :)(\boldsymbol{\theta}_n - \boldsymbol{\theta}) \end{pmatrix}. \text{ And from the above Taylor expansion of}$$

$\mathcal{L}'_n(\boldsymbol{\theta}_n)$, we obtain that

$$- \begin{pmatrix} 2\lambda_0 \hat{\Sigma}_{\mathbf{X}} & 2\hat{\Sigma}_{\mathbf{X}} \boldsymbol{\eta} \\ 2\boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}} & 0 \end{pmatrix}^{-1} \sqrt{n} \begin{pmatrix} \mathcal{C}'_n(\boldsymbol{\eta}) + 2\lambda_0 \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\eta} \\ \boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\eta} - 1 \end{pmatrix} =$$

$$[I_{p+1} + \frac{1}{2} \begin{pmatrix} 2\lambda_0 \hat{\Sigma}_{\mathbf{X}} & 2\hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta} \\ 2\boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}} & 0 \end{pmatrix}]^{-1} \begin{pmatrix} (\boldsymbol{\theta}_n - \boldsymbol{\theta})^T T_n(1, :, :) \\ (\boldsymbol{\theta}_n - \boldsymbol{\theta})^T T_n(2, :, :) \\ \vdots \\ (\boldsymbol{\theta}_n - \boldsymbol{\theta})^T T_n(p+1, :, :) \end{pmatrix} \sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}).$$

Next, we are going to prove two parts:

$$\text{Part 1: } \begin{pmatrix} 2\lambda_0 \hat{\Sigma}_{\mathbf{X}} & 2\hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta} \\ 2\boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}} & 0 \end{pmatrix}^{-1} \sqrt{n} \begin{pmatrix} \mathcal{C}'_n(\boldsymbol{\eta}) + 2\lambda_0 \hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta} \\ \boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta} - 1 \end{pmatrix} \xrightarrow{\mathcal{D}} N(0, \mathbf{V}).$$

$$\text{Part 2: } \sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}) \stackrel{\mathcal{D}}{=} [I_{p+1} + \frac{1}{2} \begin{pmatrix} 2\lambda_0 \hat{\Sigma}_{\mathbf{X}} & 2\hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta} \\ 2\boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}} & 0 \end{pmatrix}]^{-1} \begin{pmatrix} (\boldsymbol{\theta}_n - \boldsymbol{\theta})^T T_n(1, :, :) \\ (\boldsymbol{\theta}_n - \boldsymbol{\theta})^T T_n(2, :, :) \\ \vdots \\ (\boldsymbol{\theta}_n - \boldsymbol{\theta})^T T_n(p+1, :, :) \end{pmatrix} \sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}).$$

To prove Part 1, we will use the asymptotic properties for U-statistics. We will show that both $\mathcal{C}'_n(\boldsymbol{\eta}) + 2\lambda_0 \hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta}$ and $\boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta}$ are linear combinations of U-statistics.

Based on chapter 2, $\mathcal{C}_n(\boldsymbol{\eta}) = \mathcal{C}_n^2(\boldsymbol{\eta}^T \mathbf{X}, \mathbf{Y}) = S_1(\boldsymbol{\eta}) - S_2(\boldsymbol{\eta})$, where

$$S_1(\boldsymbol{\eta}) = \frac{1}{n^2} \sum_{k,l=1}^n |\boldsymbol{\eta}^T (X_k - X_l)|,$$

$$S_2(\boldsymbol{\eta}) = \frac{1}{n} \sum_{y=1}^C \frac{1}{n_y} \sum_{k,l=1}^{n_y} |\boldsymbol{\eta}^T (X_{ky} - X_{ly})|.$$

Therefore, $\mathcal{C}'_n(\boldsymbol{\eta}) + 2\lambda_0 \hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta} = S'_1(\boldsymbol{\eta}) - S'_2(\boldsymbol{\eta}) + 2\lambda_0 \hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta}$, where

$$S'_1(\boldsymbol{\eta}) = \frac{1}{n^2} \sum_{k,l=1}^n \frac{(X_k - X_l)(X_k - X_l)^T \boldsymbol{\eta}}{|\boldsymbol{\eta}^T (X_k - X_l)|},$$

$$S'_2(\boldsymbol{\eta}) = \frac{1}{n} \sum_{y=1}^C \frac{1}{n_y} \sum_{k,l=1}^{n_y} \frac{(X_{ky} - X_{ly})(X_{ky} - X_{ly})^T \boldsymbol{\eta}}{|\boldsymbol{\eta}^T (X_{ky} - X_{ly})|},$$

$$\hat{\Sigma}_{\mathbf{X}} = \frac{1}{n} \sum_{i=1}^n X_i X_i^T - \bar{X} \bar{X}^T = \frac{1}{n} \sum_{i=1}^n X_i X_i^T - \frac{1}{n^2} \sum_{i,j=1}^n X_i X_j^T.$$

Here $S'_1(\boldsymbol{\eta})$, $S'_2(\boldsymbol{\eta})$, $S'_3(\boldsymbol{\eta})$ and $\hat{\Sigma}_{\mathbf{X}}$ are V-statistics, which can be written as U-

statistics. Let

$$\begin{aligned}
U_{1n} &= \binom{n}{2}^{-1} \sum_{1 \leq k < l \leq n} \frac{(X_k - X_l)(X_k - X_l)^T \boldsymbol{\eta}}{|\boldsymbol{\eta}^T(X_k - X_l)|}, \\
U_{2n_y} &= \binom{n_y}{2}^{-1} \sum_{1 \leq k < l \leq n_y} \frac{(X_{ky} - X_{ly})(X_{ky} - X_{ly})^T \boldsymbol{\eta}}{|\boldsymbol{\eta}^T(X_{ky} - X_{ly})|}, \\
U_{4n} &= \frac{1}{n} \sum_{i=1}^n X_i X_i^T \boldsymbol{\eta}, \\
U_{5n} &= \binom{n}{2}^{-1} \sum_{i < j} \frac{1}{2} (X_i X_j^T + X_j X_i^T) \boldsymbol{\eta}.
\end{aligned}$$

Based on the following calculations, we will write $S'_1(\boldsymbol{\eta})$, $S'_2(\boldsymbol{\eta})$, $S'_3(\boldsymbol{\eta})$ and $\hat{\Sigma}_{\mathbf{X}}$ as linear combinations of these U-statistics.

$$\begin{aligned}
S'_1(\boldsymbol{\eta}) &= \frac{2}{n^2} \binom{n}{2} \left\{ \binom{n}{2}^{-1} \sum_{1 \leq k < l \leq n} \frac{(X_k - X_l)(X_k - X_l)^T \boldsymbol{\eta}}{|\boldsymbol{\eta}^T(X_k - X_l)|} \right\} \\
&= \frac{n-1}{n} U_{1n},
\end{aligned}$$

$$S'_2(\boldsymbol{\eta}) = \frac{1}{n} \sum_{y=1}^C (n_y - 1) U_{2n_y}.$$

And

$$\begin{aligned}
\hat{\Sigma}_{\mathbf{X}} &= \frac{1}{n} \sum_{i=1}^n X_i X_i^T - \bar{X} \bar{X}^T = \frac{1}{n} \sum_{i=1}^n X_i X_i^T - \frac{1}{n^2} \sum_{i,j=1}^n X_i X_j^T \\
&= \frac{1}{n} \sum_{i=1}^n X_i X_i^T - \frac{1}{n^2} \sum_{i=1}^n X_i X_i^T - \frac{1}{n^2} \sum_{i \neq j} X_i X_j^T \\
&= \frac{n-1}{n} \left[\frac{1}{n} \sum_{i=1}^n X_i X_i^T \right] - \frac{2}{n^2} \left(\sum_{i < j} \frac{1}{2} (X_i X_j^T + X_j X_i^T) \right) \\
&= \frac{n-1}{n} \left[\frac{1}{n} \sum_{i=1}^n X_i X_i^T \right] - \frac{n-1}{n} \left[\binom{n}{2}^{-1} \sum_{i < j} \frac{1}{2} (X_i X_j^T + X_j X_i^T) \right].
\end{aligned}$$

That is,

$$\begin{aligned} S'_1(\boldsymbol{\eta}) &= \frac{n-1}{n} U_{1n}, \\ S'_2(\boldsymbol{\eta}) &= \frac{1}{n} \sum_{y=1}^C (n_y - 1) U_{2ny}, \\ \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\eta} &= \frac{n-1}{n} U_{4n} - \frac{n-1}{n} U_{5n}. \end{aligned}$$

Thus $\mathcal{C}'_n(\boldsymbol{\eta}) + 2\lambda_0 \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\eta} = \frac{(n-1)}{n} U_{1n} - \frac{1}{n} \sum_{y=1}^C (n_y - 1) U_{2ny} + 2\lambda_0 \frac{n-1}{n} U_{4n} - 2\lambda_0 \frac{n-1}{n} U_{5n}$.

And $\boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\eta}$ is also a linear combination of U-statistics. Let

$$\begin{aligned} U_{6n} &= \frac{1}{n} \sum_{i=1}^n \boldsymbol{\eta}^T X_i X_i^T \boldsymbol{\eta}, \\ U_{7n} &= \binom{n}{2}^{-1} \sum_{i < j} \frac{1}{2} \boldsymbol{\eta}^T (X_i X_j^T + X_j X_i^T) \boldsymbol{\eta}, \end{aligned}$$

then $\boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\eta} = \frac{n-1}{n} U_{6n} - \frac{n-1}{n} U_{7n}$.

With iid copies of X_k and X_l , and X_{ky} and X_{ly} , respectively, let

$$\begin{aligned} \phi^{(1)}(X_k, X_l) &= \frac{(X_k - X_l)(X_k - X_l)^T \boldsymbol{\eta}}{|\boldsymbol{\eta}^T (X_k - X_l)|}, \\ \phi^{(2)}(X_{ky}, X_{ly}) &= \frac{(X_{ky} - X_{ly})(X_{ky} - X_{ly})^T \boldsymbol{\eta}}{|\boldsymbol{\eta}^T (X_{ky} - X_{ly})|}, \text{ for } y = 1, \dots, C, \\ \phi^{(4)}(X_k) &= X_k X_k^T \boldsymbol{\eta}, \\ \phi^{(5)}(X_k, X_l) &= \frac{1}{2} (X_k X_l^T + X_l X_k^T) \boldsymbol{\eta}, \\ \phi^{(6)}(X_k) &= \boldsymbol{\eta}^T X_k X_k^T \boldsymbol{\eta}, \\ \phi^{(7)}(X_k, X_l) &= \frac{1}{2} \boldsymbol{\eta}^T (X_k X_l^T + X_l X_k^T) \boldsymbol{\eta}. \end{aligned}$$

and let

$$\begin{aligned}\mu_1 &= E \frac{(X - X')(X - X')^T \boldsymbol{\eta}}{|\boldsymbol{\eta}^T(X - X')|}, \\ \mu_{2y} &= E \frac{(X_y - X'_y)(X_y - X'_y)^T \boldsymbol{\eta}}{|\boldsymbol{\eta}^T(X_y - X'_y)|}, \text{ for } y = 1, \dots, C. \\ \mu_4 &= EXX^T \boldsymbol{\eta}, \\ \mu_5 &= (EX)(EX)^T \boldsymbol{\eta}, \\ \mu_6 &= \boldsymbol{\eta}^T (EXX^T) \boldsymbol{\eta}, \\ \mu_7 &= \boldsymbol{\eta}^T (EX)(EX)^T \boldsymbol{\eta}.\end{aligned}$$

Here X and X' are i.i.d copies, and X_Y and X'_Y are i.i.d copies. By Theorem 6.1.6 (Lehmann, 1999, Ch.6), under assumption 4.5.1,

$$\sqrt{n} \begin{pmatrix} U_{1n} - \mu_1 \\ U_{2n1} - \mu_{21} \\ \dots \\ U_{2nC} - \mu_{2C} \\ U_{4n} - \mu_4 \\ U_{5n} - \mu_5 \\ U_{6n} - \mu_6 \\ U_{7n} - \mu_7 \end{pmatrix} \xrightarrow{\mathcal{D}} N(0, \Sigma),$$

$$\text{where } \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12y} & \Sigma_{14} & \Sigma_{15} & \Sigma_{16} & \Sigma_{17} \\ \cdot & \Sigma_{2y2y} & \Sigma_{2y4} & \Sigma_{2y5} & \Sigma_{2y6} & \Sigma_{2y7} \\ \cdot & \cdot & \Sigma_{44} & \Sigma_{45} & \Sigma_{46} & \Sigma_{47} \\ \cdot & \cdot & \cdot & \Sigma_{55} & \Sigma_{56} & \Sigma_{57} \\ \cdot & \cdot & \cdot & \cdot & \Sigma_{66} & \Sigma_{67} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \Sigma_{77} \end{pmatrix}, \text{ and } y = 1, \dots, C \text{ for simplicity}$$

of the expression.

Using Hoeffding's result (Hoeffding, 1948, Section 6), we obtain that

$$\Sigma_{11} = 4\text{cov}(\phi^{(1)}(X_1, X_2), \phi^{(1)}(X_1, X'_2)), \text{ where } X_1, X_2, X'_2 \text{ are i.i.d.}$$

$\Sigma_{12y} = 4\text{cov}(\phi^{(1)}(X_1, X_2), \phi^{(2)}(X_{1y}, X'_{2y}))$, where X_1 and X_2 are iid copies, and X_{1y} and X'_{2y} are i.i.d. copies.

$$\Sigma_{14} = 2\text{cov}(\phi^{(1)}(X_1, X_2), \phi^{(4)}(X_1)), \text{ where } X_1, X_2 \text{ are i.i.d.}$$

$$\Sigma_{15} = 4\text{cov}(\phi^{(1)}(X_1, X_2), \phi^{(5)}(X_1, X'_2)), \text{ where } X_1, X_2, X'_2 \text{ are i.i.d.}$$

$$\Sigma_{16} = 2\text{cov}(\phi^{(1)}(X_1, X_2), \phi^{(6)}(X_1)), \text{ where } X_1, X_2 \text{ are i.i.d.}$$

$$\Sigma_{17} = 4\text{cov}(\phi^{(1)}(X_1, X_2), \phi^{(7)}(X_1, X'_2)), \text{ where } X_1, X_2, X'_2 \text{ are i.i.d.}$$

$$\Sigma_{2y2y} = 4\text{cov}(\phi^{(2)}(X_{1y}, X_{2y}), \phi^{(2)}(X_{1y}, X'_{2y})), \text{ where } X_{1y}, X_{2y}, X'_{2y} \text{ are i.i.d.}$$

$$\Sigma_{2y4} = 2\text{cov}(\phi^{(2)}(X_{1y}, X_{2y}), \phi^{(4)}(X_1)), \text{ where } X_{1y}, X_{2y} \text{ are i.i.d.}$$

$\Sigma_{2y5} = 4\text{cov}(\phi^{(2)}(X_{1y}, X_{2y}), \phi^{(5)}(X_1, X'_2))$, where X_{1y}, X_{2y} are iid, and X_1, X'_2 are i.i.d.

$$\Sigma_{2y6} = 2\text{cov}(\phi^{(2)}(X_{1y}, X_{2y}), \phi^{(6)}(X_1)),$$

where X_{1y}, X_{2y} are i.i.d.

$\Sigma_{2y7} = 4\text{cov}(\phi^{(2)}(X_{1y}, X_{2y}), \phi^{(7)}(X_1, X'_2))$, where X_{1y}, X_{2y} are i.i.d., and X_1, X'_2 are iid copies.

$$\Sigma_{44} = \text{cov}(\phi^{(4)}(X_1), \phi^{(4)}(X_1)).$$

$$\Sigma_{45} = 2\text{cov}(\phi^{(4)}(X_1), \phi^{(5)}(X_1, X'_2)), \text{ where } X_1, X'_2 \text{ are i.i.d.}$$

$$\Sigma_{46} = \text{cov}(\phi^{(4)}(X_1), \phi^{(6)}(X_1)).$$

$$\Sigma_{47} = 2\text{cov}(\phi^{(4)}(X_1), \phi^{(7)}(X_1, X'_2)), \text{ where } X_1, X'_2 \text{ are i.i.d.}$$

$$\Sigma_{55} = 4\text{cov}(\phi^{(5)}(X_1, X_2), \phi^{(5)}(X_1, X'_2)), \text{ where } X_1, X_2, X'_2 \text{ are i.i.d.}$$

$$\Sigma_{56} = 2\text{cov}(\phi^{(5)}(X_1, X_2), \phi^{(6)}(X_1)), \text{ where } X_1, X_2 \text{ are i.i.d.}$$

$$\Sigma_{57} = 4\text{cov}(\phi^{(5)}(X_1, X_2), \phi^{(7)}(X_1, X'_2)), \text{ where } X_1, X_2, X'_2 \text{ are i.i.d.}$$

$$\Sigma_{66} = \text{cov}(\phi^{(6)}(X_1), \phi^{(6)}(X_1)).$$

$$\Sigma_{67} = 2\text{cov}(\phi^{(6)}(X_1), \phi^{(7)}(X_1, X'_2)), \text{ where } X_1, X'_2 \text{ are i.i.d.}$$

$$\Sigma_{77} = 4\text{cov}(\phi^{(7)}(X_1, X_2), \phi^{(7)}(X_1, X'_2)), \text{ where } X_1, X_2, X'_2 \text{ are i.i.d.}$$

Let $\hat{\mathbf{A}} = \begin{pmatrix} 2\lambda_0 \hat{\Sigma}_{\mathbf{X}} & 2\hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta} \\ 2\boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}} & 0 \end{pmatrix}^{-1}$, $\mathbf{A} = \begin{pmatrix} 2\lambda_0 \Sigma_{\mathbf{X}} & 2\Sigma_{\mathbf{X}}\boldsymbol{\eta} \\ 2\boldsymbol{\eta}^T \Sigma_{\mathbf{X}} & 0 \end{pmatrix}^{-1}$ and $\mathbf{B} = \begin{pmatrix} I_p & I_p \otimes (-p_1) & \cdots & I_p \otimes (-p_C) & I_p \otimes 2\lambda_0 & I_p \otimes (-2\lambda_0) & 0_p & 0_p \\ 0_p^T & 0_p^T & \cdots & 0_p^T & 0_p^T & 0_p^T & 1 & -1 \end{pmatrix}$, where 0_p is a $p \times 1$ zero vector, then by the definitions of μ_i , for instance $\mu_6 - \mu_7 = \boldsymbol{\eta}^T \Sigma_{\mathbf{X}} \boldsymbol{\eta} =$

1. In addition, p_1, \dots, p_C are the probabilities in group y for $y = 1, \dots, C$. We have

$$\sqrt{n}\mathbf{B} \begin{pmatrix} U_{1n} - \mu_1 \\ U_{2n1} - \mu_{21} \\ \dots \\ U_{2nC} - \mu_{2C} \\ U_{4n} - \mu_4 \\ U_{5n} - \mu_5 \\ U_{6n} - \mu_6 \\ U_{7n} - \mu_7 \end{pmatrix} = \sqrt{n} \begin{pmatrix} U_{1n} - \sum_{y=1}^C p_y U_{2ny} + 2\lambda_0 U_{4n} - 2\lambda_0 U_{5n} \\ U_{6n} - U_{7n} - 1 \end{pmatrix}.$$

Note that

$$\begin{aligned} & \sqrt{n} \begin{pmatrix} \mathcal{C}'_n(\boldsymbol{\eta}) + 2\lambda_0 \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\eta} \\ \boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\eta} - 1 \end{pmatrix} = \\ & \sqrt{n} \begin{pmatrix} \frac{(n-1)}{n} U_{1n} - \sum_{y=1}^C \frac{n_y-1}{n} U_{2ny} + 2\lambda_0 \frac{n-1}{n} U_{4n} - 2\lambda_0 \frac{n-1}{n} U_{5n} \\ \frac{n-1}{n} U_{6n} - \frac{n-1}{n} U_{7n} - 1 \end{pmatrix}. \end{aligned}$$

Hence,

$$\begin{aligned} & \sqrt{n} \begin{pmatrix} \mathcal{C}'_n(\boldsymbol{\eta}) + 2\lambda_0 \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\eta} \\ \boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\eta} - 1 \end{pmatrix} - \sqrt{n}\mathbf{B} \begin{pmatrix} U_{1n} - \mu_1 \\ U_{2n1} - \mu_{21} \\ \dots \\ U_{2nC} - \mu_{2C} \\ U_{4n} - \mu_4 \\ U_{5n} - \mu_5 \\ U_{6n} - \mu_6 \\ U_{7n} - \mu_7 \end{pmatrix} = \\ & \sqrt{n} \begin{pmatrix} \frac{-1}{n} U_{1n} - \sum_{y=1}^c \frac{n_y - np_y - 1}{n} U_{2ny} + 2\lambda_0 \frac{-1}{n} U_{4n} - 2\lambda_0 \frac{-1}{n} U_{5n} \\ \frac{-1}{n} U_{6n} - \frac{-1}{n} U_{7n} \end{pmatrix} \xrightarrow{P} 0, \text{ by assumption 4.5.1.} \end{aligned}$$

Therefore, by Slutsky's theorem, $\sqrt{n} \begin{pmatrix} \mathcal{C}'_n(\boldsymbol{\eta}) + 2\lambda_0 \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\eta} \\ \boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\eta} - 1 \end{pmatrix} \stackrel{D}{=} \sqrt{n} \mathbf{B} \begin{pmatrix} U_{1n} - \mu_1 \\ U_{2n1} - \mu_{21} \\ \dots \\ U_{2nC} - \mu_{2C} \\ U_{4n} - \mu_4 \\ U_{5n} - \mu_5 \\ U_{6n} - \mu_6 \\ U_{7n} - \mu_7 \end{pmatrix}.$

Hence,

$$\begin{pmatrix} 2\lambda_0 \hat{\Sigma}_{\mathbf{X}} & 2\hat{\Sigma}_{\mathbf{X}} \boldsymbol{\eta} \\ 2\boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}} & 0 \end{pmatrix}^{-1} \sqrt{n} \begin{pmatrix} \mathcal{C}'_n(\boldsymbol{\eta}) + 2\lambda_0 \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\eta} \\ \boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\eta} - 1 \end{pmatrix} \stackrel{D}{=} \sqrt{n} \mathbf{A} \mathbf{B} \begin{pmatrix} U_{1n} - \mu_1 \\ U_{2n1} - \mu_{21} \\ \dots \\ U_{2nC} - \mu_{2C} \\ U_{4n} - \mu_4 \\ U_{5n} - \mu_5 \\ U_{6n} - \mu_6 \\ U_{7n} - \mu_7 \end{pmatrix} \xrightarrow{D} N(0, \mathbf{V}),$$

where $\mathbf{V} = \mathbf{A} \mathbf{B} \Sigma \mathbf{B}^T \mathbf{A}^T$. We complete the proof of Part 1.

Now we prove Part 2: $\sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}) \stackrel{D}{=} [I_{p+1} + \frac{1}{2} \begin{pmatrix} 2\lambda_0 \hat{\Sigma}_{\mathbf{X}} & 2\hat{\Sigma}_{\mathbf{X}} \boldsymbol{\eta} \\ 2\boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}} & 0 \end{pmatrix}^{-1} \begin{pmatrix} (\boldsymbol{\theta}_n - \boldsymbol{\theta})^T T_n(1, :, :) \\ (\boldsymbol{\theta}_n - \boldsymbol{\theta})^T T_n(2, :, :) \\ \vdots \\ (\boldsymbol{\theta}_n - \boldsymbol{\theta})^T T_n(p+1, :, :) \end{pmatrix}] \sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}).$

As shown previously, all elements of $T_n(i, :, :)$, $i = 1, \dots, p+1$ are zero or elements from $\hat{\Sigma}_{\mathbf{X}}$, which are bounded. $\hat{\mathbf{A}}^{-1} \rightarrow \mathbf{A}^{-1}$, thus all elements of $\hat{\mathbf{A}}^{-1}$ are bounded as well. Lemma S. 4.5.6 indicates that $\boldsymbol{\theta}_n \xrightarrow{P} \boldsymbol{\theta}$, we see that

$$[I_{p+1} + \frac{1}{2} \begin{pmatrix} 2\lambda_0 \hat{\Sigma}_{\mathbf{X}} & 2\hat{\Sigma}_{\mathbf{X}} \boldsymbol{\eta} \\ 2\boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}} & 0 \end{pmatrix}^{-1} \begin{pmatrix} (\boldsymbol{\theta}_n - \boldsymbol{\theta})^T T_n(1, :, :) \\ (\boldsymbol{\theta}_n - \boldsymbol{\theta})^T T_n(2, :, :) \\ \vdots \\ (\boldsymbol{\theta}_n - \boldsymbol{\theta})^T T_n(p+1, :, :) \end{pmatrix}] \xrightarrow{P} I_{p+1}.$$

Then by Slutsky's theorem, $\sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}) \stackrel{\mathcal{D}}{=} [I_{p+1} + \frac{1}{2} \begin{pmatrix} 2\lambda\hat{\Sigma}_{\mathbf{X}} & 2\hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta} \\ 2\boldsymbol{\eta}^T\hat{\Sigma}_{\mathbf{X}} & 0 \end{pmatrix}]^{-1} \begin{pmatrix} (\boldsymbol{\theta}_n - \boldsymbol{\theta})^T T_n(1, :, :) \\ (\boldsymbol{\theta}_n - \boldsymbol{\theta})^T T_n(2, :, :) \\ \vdots \\ (\boldsymbol{\theta}_n - \boldsymbol{\theta})^T T_n(p+1, :, :) \end{pmatrix} \sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta})$. Therefore,

$$\sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}) \stackrel{\mathcal{D}}{=} \begin{pmatrix} 2\lambda\hat{\Sigma}_{\mathbf{X}} & 2\hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta} \\ 2\boldsymbol{\eta}^T\hat{\Sigma}_{\mathbf{X}} & 0 \end{pmatrix}^{-1} \sqrt{n} \begin{pmatrix} \mathcal{C}'_n(\boldsymbol{\eta}) + 2\lambda\hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta} \\ \boldsymbol{\eta}^T\hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta} - 1 \end{pmatrix} \stackrel{\mathcal{D}}{\rightarrow} N(0, \mathbf{V}).$$

In other word, $\boldsymbol{\theta}_n$ is \sqrt{n} -consistent estimation of $\boldsymbol{\theta}$. \square

Proof of Consistency

In order to prove the Proposition 4.3.3, we first prove the following Lemma S. 4.5.8.

Proof of Lemma S. 4.5.8

Lemma S. 4.5.8. *If the support of X , say S , is compact, $E|Y| < \infty$ and furthermore, $\boldsymbol{\eta}_n \xrightarrow{P} \boldsymbol{\eta}$, then $\mathcal{C}_n^2(\boldsymbol{\eta}_n^T \mathbf{X} | \mathbf{Y}) - \mathcal{C}_n^2(\boldsymbol{\eta}^T \mathbf{X} | \mathbf{Y}) \xrightarrow{P} 0$.*

Proof of Lemma S. 4.5.8: Based on chapter 2, we have that

$$a_{kl}(\boldsymbol{\eta}_n) = |\boldsymbol{\eta}_n^T X_k - \boldsymbol{\eta}_n^T X_l|, \text{ for } k, l = 1, \dots, n, ,$$

$$b_{kly}(\boldsymbol{\eta}_n) = |\boldsymbol{\eta}_n^T X_{ky} - \boldsymbol{\eta}_n^T X_{ly}|, \text{ for } k, l = 1, \dots, n_y, y = 1 \dots, C .$$

Then we have

$$\mathcal{C}_n^2(\boldsymbol{\eta}_n^T \mathbf{X} | \mathbf{Y}) = \frac{1}{n^2} \sum_{k,l=1}^n a_{kl}(\boldsymbol{\eta}_n) - \frac{1}{n} \sum_{y=1}^C \frac{1}{n_y} \sum_{k,l=1}^{n_y} a_{kly}(\boldsymbol{\eta}_n),$$

and,

$$\mathcal{C}_n^2(\boldsymbol{\eta}^T \mathbf{X} | \mathbf{Y}) = \frac{1}{n^2} \sum_{k,l=1}^n a_{kl}(\boldsymbol{\eta}) - \frac{1}{n} \sum_{y=1}^C \frac{1}{n_y} \sum_{k,l=1}^{n_y} a_{kly}(\boldsymbol{\eta}).$$

Because $\boldsymbol{\eta}_n \rightarrow \boldsymbol{\eta}$ in probability, let $\boldsymbol{\eta}_n = \boldsymbol{\eta} + \boldsymbol{\varepsilon}_n$, then for any $\epsilon > 0$, $|\boldsymbol{\varepsilon}_n| < \epsilon$, when $n \rightarrow \infty$. Hence, by the condition on X , we have that for a positive constant C_x , and large n , $|a_{kl}(\boldsymbol{\eta}_n) - a_{kl}(\boldsymbol{\eta})|$ and $|a_{kly}(\boldsymbol{\eta}_n) - a_{kly}(\boldsymbol{\eta})| \leq \epsilon C_x$.

Therefore,

$$\begin{aligned}
& |\mathcal{C}_n^2(\boldsymbol{\eta}_n^T \mathbf{X} | \mathbf{Y}) - \mathcal{C}_n^2(\boldsymbol{\eta}^T \mathbf{X} | \mathbf{Y})| \\
& \leq \frac{1}{n^2} \sum_{k,l=1}^n |a_{kl}(\boldsymbol{\eta}_n) - a_{kl}(\boldsymbol{\eta})| + \frac{1}{n} \sum_{y=1}^C \frac{1}{n_y} \sum_{k,l=1}^{n_y} |a_{kly}(\boldsymbol{\eta}_n) - a_{kly}(\boldsymbol{\eta})| \\
& \leq 2\epsilon C_x.
\end{aligned}$$

Hence, the conclusion follows. \square

Proof of Proposition 4.3.3

Proof of Proposition 4.3.3: Without loss of generality, we assume $Q = I_d$. Suppose $\boldsymbol{\eta}_n$ is not a consistent estimator of $\mathcal{S}_{Y|X}$, then there exists a subsequence, still to be indexed by n , and an $\boldsymbol{\eta}^*$ satisfying $\boldsymbol{\eta}^{*T} \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\eta}^* = I_d$ such that $\boldsymbol{\eta}_n \xrightarrow{P} \boldsymbol{\eta}^*$ but $\text{Span}(\boldsymbol{\eta}^*) \neq \text{Span}(\boldsymbol{\eta})$.

By Lemma S. 4.5.8, $\mathcal{C}_n^2(\boldsymbol{\eta}_n^T \mathbf{X} | \mathbf{Y}) - \mathcal{C}_n^2(\boldsymbol{\eta}^{*T} \mathbf{X} | \mathbf{Y}) \xrightarrow{P} 0$ and by chapter 2, $\mathcal{C}_n^2(\boldsymbol{\eta}^{*T} \mathbf{X} | \mathbf{Y}) \xrightarrow{a.s.} \mathcal{C}^2(\boldsymbol{\eta}^{*T} X | \mathbf{Y})$, therefore $\mathcal{C}_n^2(\boldsymbol{\eta}_n^T \mathbf{X} | \mathbf{Y}) \xrightarrow{P} \mathcal{C}^2(\boldsymbol{\eta}^{*T} X | \mathbf{Y})$. On the other hand, because $\boldsymbol{\eta}_n = \arg \max_{\boldsymbol{\beta}^T \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\beta} = I_d} \mathcal{C}_n^2(\boldsymbol{\beta}^T \mathbf{X} | \mathbf{Y})$, we have $\mathcal{C}_n^2(\boldsymbol{\eta}_n^T \mathbf{X} | \mathbf{Y}) \geq \mathcal{C}_n^2(\boldsymbol{\eta}^T \mathbf{X} | \mathbf{Y})$. If we take the limit on both sides of the above inequality, we get $\mathcal{C}^2(\boldsymbol{\eta}^{*T} X | \mathbf{Y}) \geq \mathcal{C}^2(\boldsymbol{\eta}^T X | \mathbf{Y})$, however, we have proved that under the assumption $P_{\boldsymbol{\eta}(\Sigma_X)}^T X \perp\!\!\!\perp Q_{\boldsymbol{\eta}(\Sigma_X)}^T X$, $\boldsymbol{\eta} = \arg \max_{\boldsymbol{\beta}^T \Sigma_{\mathbf{X}} \boldsymbol{\beta} = I_d} \mathcal{C}^2(\boldsymbol{\beta}^T \mathbf{X} | \mathbf{Y})$, and we also assume that the central subspace is unique, therefore $\mathcal{C}^2(\boldsymbol{\eta}^{*T} X | \mathbf{Y}) \geq \mathcal{C}^2(\boldsymbol{\eta}^T X | \mathbf{Y})$ conflicts with the above assumption, so $\boldsymbol{\eta}_n$ is a consistent estimator of a basis of the central subspace. \square

Proof of \sqrt{n} -consistency

To prove the \sqrt{n} -consistency of $\text{vec}(\boldsymbol{\eta}_n)$ in Proposition 4.3.4 in chapter 4, we reconstruct the optimization problem by using the Lagrange multiplier technique, and first we introduce the following notations, conditions and we also give a new definition.

Notations and Conditions

For a random sample $(\mathbf{X}, \mathbf{Y}) = \{(X_k, Y_k) : k = 1, \dots, n\}$ from the joint distribution of random vectors X in \mathbb{R}^p and Y in \mathbb{R} .

Let $\mathcal{L}(\zeta) = \mathcal{C}^2(\boldsymbol{\beta}^T X|Y) + \lambda^T(\text{vec}(\boldsymbol{\beta}^T \Sigma_X \boldsymbol{\beta}) - \text{vec}(I_d))$ and $\mathcal{L}_n(\zeta) = \mathcal{C}_n^2(\boldsymbol{\beta}^T \mathbf{X}|\mathbf{Y}) + \lambda^T(\text{vec}(\boldsymbol{\beta}^T \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\beta}) - \text{vec}(I_d))$. Here $\zeta = \begin{pmatrix} \text{vec}(\boldsymbol{\beta}) \\ \lambda \end{pmatrix} \in \mathbb{R}^{pd+d^2}$, $\boldsymbol{\beta} \in \mathbb{R}^{p \times d}$, $\lambda \in \mathbb{R}^{d^2}$, Σ_X is the covariance matrix of X , and $\hat{\Sigma}_{\mathbf{X}}$ is the sample estimate for Σ_X . Let $\boldsymbol{\eta}_n = \arg \max_{\boldsymbol{\beta}^T \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\beta} = I_d} \mathcal{C}_n^2(\boldsymbol{\beta}^T \mathbf{X}|\mathbf{Y})$, then there exists a λ_n such that $\begin{pmatrix} \text{vec}(\boldsymbol{\eta}_n) \\ \lambda_n \end{pmatrix}$ is a stationary point for $\mathcal{L}_n(\zeta)$. Let $\boldsymbol{\theta}_n = \begin{pmatrix} \text{vec}(\boldsymbol{\eta}_n) \\ \lambda_n \end{pmatrix}$, then $\mathcal{L}'_n(\boldsymbol{\theta}_n) = 0$. Let $\boldsymbol{\eta}$ to be a basis of CS, then under the assumption $P_{\boldsymbol{\eta}(\Sigma_X)}^T \mathbf{X} \perp Q_{\boldsymbol{\eta}(\Sigma_X)}^T \mathbf{X}$, there exists a rotation matrix $Q : Q^T Q = I_d$, such that $\boldsymbol{\eta} Q = \arg \max_{\boldsymbol{\beta}^T \Sigma_X \boldsymbol{\beta} = I_d} \mathcal{C}^2(\boldsymbol{\beta}^T X|Y)$. Without loss of generality, we assume $Q = I_d$ here, therefore there exists a λ_0 such that $\begin{pmatrix} \text{vec}(\boldsymbol{\eta}) \\ \lambda_0 \end{pmatrix}$ is a stationary point for $\mathcal{L}(\zeta)$. Let $\boldsymbol{\theta} = \begin{pmatrix} \text{vec}(\boldsymbol{\eta}) \\ \lambda_0 \end{pmatrix}$.

In the proof, we need to take derivatives of $\mathcal{C}^2(\boldsymbol{\eta}^T X|Y)$ and $\mathcal{C}_n^2(\boldsymbol{\eta}^T \mathbf{X}|\mathbf{Y})$ with respect to $\text{vec}(\boldsymbol{\eta})$, so for the simplicity of notation, when we consider the derivatives of $\mathcal{C}^2(\boldsymbol{\eta}^T X|Y)$ and $\mathcal{C}_n^2(\boldsymbol{\eta}^T \mathbf{X}|\mathbf{Y})$, we use $\mathcal{C}(\boldsymbol{\eta})$ and $\mathcal{C}_n(\boldsymbol{\eta})$ to denote $\mathcal{C}^2(\boldsymbol{\eta}^T X|Y)$ and $\mathcal{C}_n^2(\boldsymbol{\eta}^T \mathbf{X}|\mathbf{Y})$, respectively.

Here are additional notations, which will be used later in the following proof. $I_{(d,d)}$ is the vec-permutation matrix. I_m is a identity matrix with rank m , and $I_m(:, i)$ denotes the i th column of I_m . $\mathbf{A} \otimes \mathbf{B}$ denotes Kronecker product between matrix \mathbf{A} and \mathbf{B} . $\text{vec}(\cdot)$ is a vec operator. Furthermore, we give the following definition and assumptions.

Definition S. 4.5.3. Let $\Delta(\boldsymbol{\eta}) = \{\boldsymbol{\alpha} : \|\boldsymbol{\alpha} - \boldsymbol{\eta}\| \leq c\}$, where $\boldsymbol{\alpha}$ is a $p \times d$ matrix and $\boldsymbol{\alpha}^T \Sigma_X \boldsymbol{\alpha} = I_d$, c is a fixed small constant, $\|\cdot\|$ is the Frobenius norm. We define an

indicator function

$$\rho(X, X') = \begin{cases} 0 & \text{if } |\boldsymbol{\alpha}^T(X - X')| \leq \epsilon_0, \text{ for } \boldsymbol{\alpha} \in \Delta(\boldsymbol{\eta}) \\ 1 & \text{if } |\boldsymbol{\alpha}^T(X - X')| > \epsilon_0, \text{ for } \boldsymbol{\alpha} \in \Delta(\boldsymbol{\eta}) \end{cases}$$

where X' is an i.i.d. copy of X and ϵ_0 is a small number. We define the second and third derivative of $\mathcal{C}(\boldsymbol{\eta})$ with respect to $\text{vec}(\boldsymbol{\eta})$ as $\mathcal{C}''(\boldsymbol{\eta})\rho(X, X')$ and $\mathcal{C}'''(\boldsymbol{\eta})\rho(X, X')$. For the simplicity of notation, we will still use $\mathcal{C}''(\boldsymbol{\eta})$ and $\mathcal{C}'''(\boldsymbol{\eta})$ to denote $\mathcal{C}''(\boldsymbol{\eta})\rho(X, X')$ and $\mathcal{C}'''(\boldsymbol{\eta})\rho(X, X')$, respectively.

The reason we use this definition is by definition S. 4.5.3, the second and third derivative of $\mathcal{C}(\boldsymbol{\eta})$ and $\mathcal{C}_n(\boldsymbol{\eta})$ are bounded, near the neighborhood of the central subspace.

Assumption S. 4.5.2. $\text{Var} [\phi^{(1)}(X_1, X_2)], \text{Var} [\phi^{(2)}(X_{1y}, X_{2y})], \text{Var} [\phi^{(4)}(X_1)], \text{Var} [\phi^{(5)}(X_1, X_2)], \text{Var} [\phi^{(6)}(X_1)], \text{Var} [\phi^{(7)}(X_1, X_2)], \text{Var} [\phi^{(8)}(X_1)]$ are all $< \infty$.

Here

$$\begin{aligned} \phi^{(1)}(X_1, X_2) &= \frac{(I_d \otimes (X_1 - X_2))(I_d \otimes (X_1 - X_2)^T)\text{vec}(\boldsymbol{\eta})}{|(I_d \otimes (X_1 - X_2)^T)\text{vec}(\boldsymbol{\eta})|}, \\ \phi^{(2)}(X_{1y}, X_{2y}) &= \frac{(I_d \otimes (X_{1y} - X_{2y}))(I_d \otimes (X_{1y} - X_{2y})^T)\text{vec}(\boldsymbol{\eta})}{|(I_d \otimes (X_{1y} - X_{2y})^T)\text{vec}(\boldsymbol{\eta})|}, \\ \phi^{(4)}(X_1) &= (I_d \otimes X_1 X_1^T \boldsymbol{\eta})(I_{d^2} + I_{d,d}^T)\lambda_0, \\ \phi^{(5)}(X_1, X_2) &= \frac{1}{2}(I_d \otimes (X_1 X_2^T + X_2 X_1^T) \boldsymbol{\eta})(I_{d^2} + I_{d,d}^T)\lambda_0, \\ \phi^{(6)}(X_1) &= \text{vec}(\boldsymbol{\eta}^T X_1 X_1^T \boldsymbol{\eta}), \\ \phi^{(7)}(X_1, X_2) &= \frac{1}{2}\text{vec}(\boldsymbol{\eta}^T (X_1 X_2^T + X_2 X_1^T) \boldsymbol{\eta}), \\ \phi^{(8)}(X_1) &= \text{vec}(X_1 - EX_1)(X_1 - EX_1)^T. \end{aligned}$$

Assumption S. 4.5.3. $\begin{pmatrix} \mathcal{C}''(\boldsymbol{\eta}) + L & (I_d \otimes \Sigma_{\mathbf{X}} \boldsymbol{\eta})(I_{d^2} + I_{(d,d)}) \\ (I_{d^2} + I_{(d,d)}^T)(I_d \otimes \boldsymbol{\eta}^T \Sigma_{\mathbf{X}}) & 0 \end{pmatrix}$ is nonsingular, where L is defined later in the proof.

Assumption 4.5.2 is needed for Proposition 4.3.4 in the main article and Lemma S. 4.5.9 in the next Section, which is similar to the assumed conditions of Theorem 6.1.6 (Lehmann, 1999, Ch.6). This assumption is required by the asymptotic properties of U-statistics.

Assumption 4.5.3 is in the spirit of von Mises proposition (Serfling, 1980, Section 6.1). In this proposition, it claims that if the first nonvanishing term of Taylor expansion is the linear term, then the root-n consistency of the differentiable statistical function can be achieved. In our case, we assume the corresponding matrix is non-singular, which guarantees the root-n consistency. If the matrix is singular, then n or higher order consistency of some parts of our estimates can be proved.

Proof of Lemma S. 4.5.9

In order to prove Proposition 4.3.4, we first prove the following Lemma S. 4.5.9.

Lemma S. 4.5.9. *Under assumptions 4.5.2 and 4.5.3, and the assumptions in Proposition 4.3.4, then $\sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}) \xrightarrow{\mathcal{D}} N(0, V)$. The explicit expression for V is in the proof.*

Proof of Lemma S. 4.5.9: The Taylor expansion of $\mathcal{L}'_n(\boldsymbol{\theta}_n)$ at $\boldsymbol{\theta}$ is $0 = \mathcal{L}'_n(\boldsymbol{\theta}_n) = \mathcal{L}'_n(\boldsymbol{\theta}) + \mathcal{L}''_n(\boldsymbol{\theta})(\boldsymbol{\theta}_n - \boldsymbol{\theta}) + \mathcal{R}_1(\boldsymbol{\theta}_n^*)$, where $\|\boldsymbol{\theta}_n^* - \boldsymbol{\theta}\| \leq \|\boldsymbol{\theta}_n - \boldsymbol{\theta}\|$, where $\|\cdot\|$ is the Frobenius norm and $\boldsymbol{\theta}_n^* = \begin{pmatrix} \text{vec}(\boldsymbol{\eta}_n^*) \\ \lambda_n^* \end{pmatrix}$. Next, we will give explicit expressions of $\mathcal{L}'_n(\boldsymbol{\theta})$, $\mathcal{L}''_n(\boldsymbol{\theta})$

and $\mathcal{R}_1(\boldsymbol{\theta}_n^*)$. With simple calculation, $\mathcal{L}'_n(\boldsymbol{\theta}) = \begin{pmatrix} \mathcal{C}'_n(\boldsymbol{\eta}) + (I_d \otimes \hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta})(I_{d^2} + I_{(d,d)})\lambda_0 \\ \text{vec}(\boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta}) - \text{vec}(I_d) \end{pmatrix}$

$\mathcal{L}''_n(\boldsymbol{\theta}) = \begin{pmatrix} \mathcal{C}''_n(\boldsymbol{\eta}) + \hat{L} & (I_d \otimes \hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta})(I_{d^2} + I_{(d,d)}) \\ (I_{d^2} + I_{(d,d)}^T)(I_d \otimes \boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}}) & 0 \end{pmatrix}$.

where $\hat{L} = (\text{vec}(\hat{L}_{11}), \text{vec}(\hat{L}_{21}), \dots, \text{vec}(\hat{L}_{p1}), \dots, \text{vec}(\hat{L}_{1d}), \text{vec}(\hat{L}_{2d}), \dots, \text{vec}(\hat{L}_{pd}))^T$ and $\hat{L}_{ij} = \hat{\Sigma}_{\mathbf{X}}^T I_p(:, i) \lambda_0^T (I_{d^2} + I_{(d,d)}^T) (I_d(:, j) \otimes I_d)$. It is obvious that $\hat{L} \xrightarrow{a.s.} L$, where $L = (\text{vec}(L_{11}), \text{vec}(L_{21}), \dots, \text{vec}(L_{p1}), \dots, \text{vec}(L_{1d}), \text{vec}(L_{2d}), \dots, \text{vec}(L_{pd}))^T$ and $L_{ij} = \Sigma_{\mathbf{X}}^T I_p(:, i) \lambda_0^T (I_{d^2} + I_{(d,d)}^T) (I_d(:, j) \otimes I_d)$. Here $i = 1, \dots, p$ and $j = 1, \dots, d$.

The remainder term $\mathcal{R}_1(\theta_n^*)$ involves the third derivative of $\mathcal{L}(\zeta)$ at θ_n^* . Let $T_n = \mathcal{L}_n'''(\theta_n^*)$, where T_n is a $(pd + d^2) \times (pd + d^2) \times (pd + d^2)$ array and each $T_n(j, :, :)$, $j = 1, \dots, pd + d^2$, is a $(pd + d^2) \times (pd + d^2)$ matrix. Therefore, the form of $\mathcal{R}_1(\theta_n^*)$ can be written as

$$\mathcal{R}_1(\theta_n^*) = \frac{1}{2} \begin{pmatrix} (\theta_n - \theta)^T T_n(1, :, :)(\theta_n - \theta) \\ (\theta_n - \theta)^T T_n(2, :, :)(\theta_n - \theta) \\ \vdots \\ (\theta_n - \theta)^T T_n(pd + d^2, :, :)(\theta_n - \theta) \end{pmatrix}.$$

Based on the above explicit expression of $\mathcal{L}'_n(\theta)$, $\mathcal{L}''_n(\theta)$ and $\mathcal{R}_1(\theta_n^*)$, the Taylor expansion of $\mathcal{L}'_n(\theta_n)$ at θ can be written as

$$\begin{aligned} 0 &= \begin{pmatrix} \mathcal{C}'_n(\boldsymbol{\eta}) + (I_d \otimes \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\eta})(I_{d^2} + I_{(d,d)})\lambda_0 \\ \text{vec}(\boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\eta}) - \text{vec}(I_d) \end{pmatrix} \\ &+ \begin{pmatrix} \mathcal{C}''_n(\boldsymbol{\eta}) + \hat{L} & (I_d \otimes \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\eta})(I_{d^2} + I_{(d,d)}) \\ (I_{d^2} + I_{(d,d)}^T)(I_d \otimes \boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}}) & 0 \end{pmatrix} \begin{pmatrix} \text{vec}(\boldsymbol{\eta}_n) - \text{vec}(\boldsymbol{\eta}) \\ \lambda_n - \lambda_0 \end{pmatrix} \\ &+ \frac{1}{2} \begin{pmatrix} (\theta_n - \theta)^T T_n(1, :, :)(\theta_n - \theta) \\ (\theta_n - \theta)^T T_n(2, :, :)(\theta_n - \theta) \\ \vdots \\ (\theta_n - \theta)^T T_n(pd + d^2, :, :)(\theta_n - \theta) \end{pmatrix}. \end{aligned}$$

From the above Taylor expansion of $\mathcal{L}'_n(\theta_n)$ at θ , we get

$$\begin{aligned} &- \begin{pmatrix} \mathcal{C}''_n(\boldsymbol{\eta}) + \hat{L} & (I_d \otimes \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\eta})(I_{d^2} + I_{(d,d)}) \\ (I_{d^2} + I_{(d,d)}^T)(I_d \otimes \boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}}) & 0 \end{pmatrix}^{-1} \times \\ &\sqrt{n} \begin{pmatrix} \mathcal{C}'_n(\boldsymbol{\eta}) + (I_d \otimes \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\eta})(I_{d^2} + I_{(d,d)})\lambda_0 \\ \text{vec}(\boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\eta}) - \text{vec}(I_d) \end{pmatrix} = \\ &[I_{pd+d^2} + \frac{1}{2} \begin{pmatrix} \mathcal{C}''_n(\boldsymbol{\eta}) + \hat{L} & (I_d \otimes \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\eta})(I_{d^2} + I_{(d,d)}) \\ (I_{d^2} + I_{(d,d)}^T)(I_d \otimes \boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}}) & 0 \end{pmatrix}^{-1} \times \\ &\begin{pmatrix} (\theta_n - \theta)^T T_n(1, :, :) \\ (\theta_n - \theta)^T T_n(2, :, :) \\ \vdots \\ (\theta_n - \theta)^T T_n(pd + d^2, :, :) \end{pmatrix}] \sqrt{n}(\theta_n - \theta). \end{aligned}$$

Next, we will prove two parts:

$$\text{Part 1: } \begin{pmatrix} \mathcal{C}_n''(\boldsymbol{\eta}) + \hat{L} & (I_d \otimes \hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta})(I_{d^2} + I_{(d,d)}) \\ (I_{d^2} + I_{(d,d)}^T)(I_d \otimes \boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}}) & 0 \end{pmatrix}^{-1} \times \\ \sqrt{n} \begin{pmatrix} \mathcal{C}_n'(\boldsymbol{\eta}) + (I_d \otimes \hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta})(I_{d^2} + I_{(d,d)})\lambda_0 \\ \text{vec}(\boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta}) - \text{vec}(I_d) \end{pmatrix} \rightarrow N(0, V).$$

$$\text{Part 2: } \sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}) \stackrel{\mathcal{D}}{=} \\ [I_{pd+d^2} + \frac{1}{2} \begin{pmatrix} \mathcal{C}_n''(\boldsymbol{\eta}) + \hat{L} & (I_d \otimes \hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta})(I_{d^2} + I_{(d,d)}) \\ (I_{d^2} + I_{(d,d)}^T)(I_d \otimes \boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}}) & 0 \end{pmatrix}^{-1} \times \\ \begin{pmatrix} (\boldsymbol{\theta}_n - \boldsymbol{\theta})^T T_n(1, :, :) \\ (\boldsymbol{\theta}_n - \boldsymbol{\theta})^T T_n(2, :, :) \\ \vdots \\ (\boldsymbol{\theta}_n - \boldsymbol{\theta})^T T_n(pd + d^2, :, :) \end{pmatrix}] \sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}).$$

Proof of part 1: We will show that both $\mathcal{C}_n'(\boldsymbol{\eta}) + (I_d \otimes \hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta})(I_{d^2} + I_{(d,d)})\lambda_0$ and $\text{vec}(\boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta}) - \text{vec}(I_d)$ are linear combinations of U-statistics and the asymptotic distribution can be achieved by the asymptotic property of U-statistics.

Based on chapter 2 $2\mathcal{C}_n(\boldsymbol{\eta}) = S_1(\boldsymbol{\eta}) - S_2(\boldsymbol{\eta})$, where

$$S_1(\boldsymbol{\eta}) = \frac{1}{n^2} \sum_{k,l=1}^n |\boldsymbol{\eta}^T (X_k - X_l)|, \\ S_2(\boldsymbol{\eta}) = \frac{1}{n} \sum_{y=1}^C \frac{1}{n_y} \sum_{k,l=1}^{n_y} |\boldsymbol{\eta}^T (X_{ky} - X_{ly})|.$$

Therefore, $\mathcal{C}_n'(\boldsymbol{\eta}) + (I_d \otimes \hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta})(I_{d^2} + I_{(d,d)})\lambda_0 = S_1'(\boldsymbol{\eta}) - S_2'(\boldsymbol{\eta}) + (I_d \otimes \hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta})(I_{d^2} + I_{(d,d)})\lambda_0$, where

$$S_1'(\boldsymbol{\eta}) = \frac{1}{n^2} \sum_{k,l=1}^n \frac{(I_d \otimes (X_k - X_l))(I_d \otimes (X_k - X_l)^T) \text{vec}(\boldsymbol{\eta})}{|(I_d \otimes (X_k - X_l)^T) \text{vec}(\boldsymbol{\eta})|}, \\ S_2'(\boldsymbol{\eta}) = \frac{1}{n} \sum_{y=1}^C \frac{1}{n_y} \sum_{k,l=1}^{n_y} \frac{(I_d \otimes (X_{ky} - X_{ly}))(I_d \otimes (X_{ky} - X_{ly})^T) \text{vec}(\boldsymbol{\eta})}{|(I_d \otimes (X_{ky} - X_{ly})^T) \text{vec}(\boldsymbol{\eta})|}, \\ \hat{\Sigma}_{\mathbf{X}} = \frac{1}{n} \sum_{i=1}^n X_i X_i^T - \bar{\mathbf{X}} \bar{\mathbf{X}}^T = \frac{1}{n} \sum_{i=1}^n X_i X_i^T - \frac{1}{n^2} \sum_{i,j=1}^n X_i X_j^T.$$

Here $S'_1(\boldsymbol{\eta})$, $S'_2(\boldsymbol{\eta})$ and $\hat{\Sigma}_{\mathbf{X}}$ are V-statistics, which can be written as linear combinations of U-statistics. Let

$$\begin{aligned} U_{1n} &= \binom{n}{2}^{-1} \sum_{1 \leq k < l \leq n} \frac{(I_d \otimes (X_k - X_l))(I_d \otimes (X_k - X_l)^T) \text{vec}(\boldsymbol{\eta})}{|(I_d \otimes (X_k - X_l)^T) \text{vec}(\boldsymbol{\eta})|}, \\ U_{2n_y} &= \binom{n_y}{2}^{-1} \sum_{1 \leq k < l \leq n_y} \left\{ \frac{(I_d \otimes (X_{ky} - X_{ly}))(I_d \otimes (X_{ky} - X_{ly})^T) \text{vec}(\boldsymbol{\eta})}{|(I_d \otimes (X_{ky} - X_{ly})^T) \text{vec}(\boldsymbol{\eta})|} \right\}, \\ U_{4n} &= \frac{1}{n} \sum_{i=1}^n (I_d \otimes X_i X_i^T \boldsymbol{\eta})(I_{d^2} + I_{d,d}^T) \lambda_0, \\ U_{5n} &= \binom{n}{2}^{-1} \sum_{i < j} \frac{1}{2} (I_d \otimes (X_i X_j^T + X_j X_i^T) \boldsymbol{\eta})(I_{d^2} + I_{d,d}^T) \lambda_0. \end{aligned}$$

Through some tedious calculations, we can get $\mathcal{C}'_n(\boldsymbol{\eta}) + (I_d \otimes \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\eta})(I_{d^2} + I_{(d,d)}^T) \lambda_0 = \frac{(n-1)}{n} U_{1n} - \sum_{y=1}^C \frac{n_y-1}{n} U_{2n_y} + \frac{n-1}{n} U_{4n} - \frac{n-1}{n} U_{5n}$.

$\text{vec}(\boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\eta})$ is also a linear combination of U-statistics, let

$$\begin{aligned} U_{6n} &= \frac{1}{n} \sum_{i=1}^n \text{vec}(\boldsymbol{\eta}^T X_i X_i^T \boldsymbol{\eta}) \\ U_{7n} &= \binom{n}{2}^{-1} \sum_{i < j} \frac{1}{2} \text{vec}(\boldsymbol{\eta}^T (X_i X_j^T + X_j X_i^T) \boldsymbol{\eta}), \end{aligned}$$

then $\text{vec}(\boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\eta}) = \frac{n-1}{n} U_{6n} - \frac{n-1}{n} U_{7n}$.

let

$$\begin{aligned} \mu_1 &= E \frac{(I_d \otimes (X - X'))(I_d \otimes (X - X')^T) \text{vec}(\boldsymbol{\eta})}{|(I_d \otimes (X - X')^T) \text{vec}(\boldsymbol{\eta})|}, \\ \mu_{2y} &= E \frac{(I_d \otimes (X_y - X'_y))(I_d \otimes (X_y - X'_y)^T) \text{vec}(\boldsymbol{\eta})}{|(I_d \otimes (X_y - X'_y)^T) \text{vec}(\boldsymbol{\eta})|}, \text{ for } y = 1, \dots, C, \\ \mu_4 &= E(I_d \otimes X X^T \boldsymbol{\eta})(I_{d^2} + I_{(d,d)}^T) \lambda_0, \\ \mu_5 &= (I_d \otimes (EX)(EX)^T \boldsymbol{\eta})(I_{d^2} + I_{(d,d)}^T) \lambda_0, \\ \mu_6 &= \text{vec}(\boldsymbol{\eta}^T (EXX^T) \boldsymbol{\eta}), \\ \mu_7 &= \text{vec}(\boldsymbol{\eta}^T (EX)(EX)^T \boldsymbol{\eta}). \end{aligned}$$

Here \mathbf{X}, \mathbf{X}' are iid copies and $\mathbf{X}_y, \mathbf{X}'_y$ are i.i.d copies.

According to Theorem 6.1.6 (Lehmann, 1999, Ch.6),

$$\sqrt{n} \begin{pmatrix} U_{1n} - \mu_1 \\ U_{2n1} - \mu_{21} \\ \dots \\ U_{2nC} - \mu_{2C} \\ U_{4n} - \mu_4 \\ U_{5n} - \mu_5 \\ U_{6n} - \mu_6 \\ U_{7n} - \mu_7 \end{pmatrix} \xrightarrow{\mathcal{D}} N(0, \Sigma),$$

$$\text{where } \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12y} & \Sigma_{14} & \Sigma_{15} & \Sigma_{16} & \Sigma_{17} \\ \cdot & \Sigma_{2y2y} & \Sigma_{2y4} & \Sigma_{2y5} & \Sigma_{2y6} & \Sigma_{2y7} \\ \cdot & \cdot & \Sigma_{44} & \Sigma_{45} & \Sigma_{46} & \Sigma_{47} \\ \cdot & \cdot & \cdot & \Sigma_{55} & \Sigma_{56} & \Sigma_{57} \\ \cdot & \cdot & \cdot & \cdot & \Sigma_{66} & \Sigma_{67} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \Sigma_{77} \end{pmatrix}, \text{ and } y = 1, \dots, C \text{ for sim-}$$

licity of the expression. And $\Sigma_{ij} = a_{ij} \text{cov}(\phi^{(i)}, \phi^{(j)})$. Here a_{ij} is a constant, which equals to the number of inputs of $\phi^{(i)}$ multiplies the number of inputs of $\phi^{(j)}$.

$$\text{Let } \mathbf{B} = \begin{pmatrix} I_{pd} & (-p_1)I_{pd} & \dots & (-p_C)I_{pd} & I_{pd} & I_{pd} & \mathbf{0} & \mathbf{0} \\ \mathbf{0}^T & \mathbf{0}^T & \dots & \mathbf{0}^T & \mathbf{0}^T & \mathbf{0}^T & I_{d^2 \times d^2} & -I_{d^2 \times d^2} \end{pmatrix}, \text{ where } \mathbf{0}$$

is a $pd \times d^2$ zero matrix, then

$$\sqrt{n} \mathbf{B} \begin{pmatrix} U_{1n} - \mu_1 \\ U_{2n1} - \mu_{21} \\ \dots \\ U_{2nC} - \mu_{2C} \\ U_{4n} - \mu_4 \\ U_{5n} - \mu_5 \\ U_{6n} - \mu_6 \\ U_{7n} - \mu_7 \end{pmatrix} = \sqrt{n} \begin{pmatrix} U_{1n} - \sum_{y=1}^C p_y U_{2ny} + U_{4n} - U_{5n} \\ U_{6n} - U_{7n} - \text{vec}(I_d) \end{pmatrix}.$$

Note that $\sqrt{n} \begin{pmatrix} \mathcal{C}'_n(\boldsymbol{\eta}) + (I_d \otimes \hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta})(I_{d^2} + I_{(d,d)})\lambda_0 \\ \text{vec}(\boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta}) - \text{vec}(I_d) \end{pmatrix} =$

$$\sqrt{n} \begin{pmatrix} \frac{(n-1)}{n}U_{1n} - \sum_{y=1}^C \frac{(n_y-1)}{n_y}U_{2ny} + \frac{n-1}{n}U_{4n} - \frac{n-1}{n}U_{5n} \\ \frac{n-1}{n}U_{6n} - \frac{n-1}{n}U_{7n} - \text{vec}(I_d) \end{pmatrix},$$

under assumption S. 4.5.2,

$$\sqrt{n} \begin{pmatrix} \frac{(n-1)}{n}U_{1n} - \sum_{y=1}^C \frac{(n_y-1)}{n_y}U_{2ny} + \frac{n-1}{n}U_{4n} - \frac{n-1}{n}U_{5n} \\ \frac{n-1}{n}U_{6n} - \frac{n-1}{n}U_{7n} - \text{vec}(I_d) \end{pmatrix} -$$

$$\sqrt{n} \begin{pmatrix} U_{1n} - 2U_{2n} + U_{3n} + U_{4n} - U_{5n} \\ U_{6n} - U_{7n} - \text{vec}(I_d) \end{pmatrix} \xrightarrow{P} 0, \text{ therefore by Slutsky's theorem,}$$

$$\sqrt{n} \begin{pmatrix} \mathcal{C}'_n(\boldsymbol{\eta}) + (I_d \otimes \hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta})(I_{d^2} + I_{(d,d)})\lambda_0 \\ \text{vec}(\boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta}) - \text{vec}(I_d) \end{pmatrix} \stackrel{D}{=} \sqrt{n}\mathbf{B} \begin{pmatrix} U_{1n} - \mu_1 \\ U_{2n1} - \mu_{21} \\ \dots \\ U_{2nC} - \mu_{2C} \\ U_{4n} - \mu_4 \\ U_{5n} - \mu_5 \\ U_{6n} - \mu_6 \\ U_{7n} - \mu_7 \end{pmatrix}.$$

$$\text{Let } A_n = \begin{pmatrix} \mathcal{C}''_n(\boldsymbol{\eta}) + \hat{L} & (I_d \otimes \hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta})(I_{d^2} + I_{(d,d)}) \\ (I_{d^2} + I_{(d,d)}^T)(I_d \otimes \boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}}) & 0 \end{pmatrix}^{-1},$$

$$A = \begin{pmatrix} \mathcal{C}''(\boldsymbol{\eta}) + L & (I_d \otimes \Sigma_{\mathbf{X}}\boldsymbol{\eta})(I_{d^2} + I_{(d,d)}) \\ (I_{d^2} + I_{(d,d)}^T)(I_d \otimes \boldsymbol{\eta}^T \Sigma_{\mathbf{X}}) & 0 \end{pmatrix}^{-1},$$

under assumption S. 4.5.3 and our definition of second derivative of $\mathcal{C}_n(\boldsymbol{\eta})$, by SLLN of

U-statistics, $A_n \xrightarrow{a.s.} A$, therefore $\begin{pmatrix} \mathcal{C}''_n(\boldsymbol{\eta}) + \hat{L} & (I_d \otimes \hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta})(I_{d^2} + I_{(d,d)}) \\ (I_{d^2} + I_{(d,d)}^T)(I_d \otimes \boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}}) & 0 \end{pmatrix}^{-1} \times$

$$\sqrt{n} \begin{pmatrix} \mathcal{C}'_n(\boldsymbol{\eta}) + (I_d \otimes \hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta})(I_{d^2} + I_{(d,d)})\lambda_0 \\ \text{vec}(\boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta}) - \text{vec}(I_d) \end{pmatrix} \stackrel{D}{=} \sqrt{n}\mathbf{A}\mathbf{B} \begin{pmatrix} U_{1n} - \mu_1 \\ U_{2n1} - \mu_{21} \\ \dots \\ U_{2nC} - \mu_{2C} \\ U_{4n} - \mu_4 \\ U_{5n} - \mu_5 \\ U_{6n} - \mu_6 \\ U_{7n} - \mu_7 \end{pmatrix} \longrightarrow N(0, \mathbf{V}),$$

where $\mathbf{V} = \mathbf{A}\mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^T\mathbf{A}^T$.

Proof of part 2:

Under assumption S. 4.5.3 and Definition S. 4.5.3,

$$\begin{aligned} & [I_{pd+d^2} + \frac{1}{2} \begin{pmatrix} \mathcal{C}''_n(\boldsymbol{\eta}) + \hat{L} & (I_d \otimes \hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta})(I_{d^2} + I_{(d,d)}) \\ (I_{d^2} + I_{(d,d)}^T)(I_d \otimes \boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}}) & 0 \end{pmatrix}^{-1} \\ & \times \begin{pmatrix} (\boldsymbol{\theta}_n - \boldsymbol{\theta})^T T_n(1, :, :) \\ (\boldsymbol{\theta}_n - \boldsymbol{\theta})^T T_n(2, :, :) \\ \vdots \\ (\boldsymbol{\theta}_n - \boldsymbol{\theta})^T T_n(pd + d^2, :, :) \end{pmatrix}] \xrightarrow{P} I_{pd+d^2}, \text{ therefore by Slutsky's theorem,} \\ & \sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}) \stackrel{D}{=} \\ & [I_{pd+d^2} + \frac{1}{2} \begin{pmatrix} \mathcal{C}''_n(\boldsymbol{\eta}) + \begin{pmatrix} \text{vec}^T(\hat{L}_{11}) \\ \vdots \\ \text{vec}^T(\hat{L}_{pd}) \end{pmatrix} & (I_d \otimes \hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta})(I_{d^2} + I_{(d,d)}) \\ (I_{d^2} + I_{(d,d)}^T)(I_d \otimes \boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}}) & 0 \end{pmatrix}^{-1} \times \\ & \begin{pmatrix} (\boldsymbol{\theta}_n - \boldsymbol{\theta})^T T_n(1, :, :) \\ (\boldsymbol{\theta}_n - \boldsymbol{\theta})^T T_n(2, :, :) \\ \vdots \\ (\boldsymbol{\theta}_n - \boldsymbol{\theta})^T T_n(pd + d^2, :, :) \end{pmatrix}] \sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}). \end{aligned}$$

Therefore $\sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}) \xrightarrow{D} N(0, \mathbf{V})$, or in other words, $\boldsymbol{\theta}_n$ is \sqrt{n} -consistent estimation of $\boldsymbol{\theta}$.

In the above proof, without loss of generality we assume that $\mathbf{Q} = I_d$. Note that

with an orthogonal matrix \mathbf{Q} , $\mathcal{C}_n^2(\mathbf{Q}^T \boldsymbol{\beta}^T \mathbf{X} | \mathbf{Y}) = \mathcal{C}_n^2(\boldsymbol{\beta}^T \mathbf{X} | \mathbf{Y})$ and $\mathcal{C}^2(\mathbf{Q}^T \boldsymbol{\beta}^T X | Y) = \mathcal{C}^2(\boldsymbol{\beta}^T X | Y)$ (chapter 2). If define $\boldsymbol{\eta}_{\mathbf{Q}} = \boldsymbol{\eta} \mathbf{Q}$, without assuming $\mathbf{Q} = I_d$, then Lemma B holds by using $V(\boldsymbol{\eta}_{\mathbf{Q}})$ which is obtained by replacing every $\boldsymbol{\eta}$ in V with $\boldsymbol{\eta}_{\mathbf{Q}}$. (of course, then $V(\boldsymbol{\eta}_{I_d}) = V$ in the proof). \square

Proof of Proposition 4.3.4

Proof of Proposition 4.3.4: Let $G = (I_{pd}, 0)$ be a $pd \times (pd + d^2)$ matrix, where I_{pd} is a $pd \times pd$ identity matrix. Then $\text{vec}(\boldsymbol{\eta}_n) = G\boldsymbol{\theta}_n$ and $\text{vec}(\boldsymbol{\eta}_{\mathbf{Q}}) = G\boldsymbol{\theta}$. By Lemma 4.5.9, we have $\sqrt{n}(\text{vec}(\boldsymbol{\eta}_n) - \text{vec}(\boldsymbol{\eta}_{\mathbf{Q}})) = \sqrt{n}G(\boldsymbol{\theta}_n - \boldsymbol{\theta}) \xrightarrow{\mathcal{D}} N(0, V_{11}(\boldsymbol{\eta}_{\mathbf{Q}}))$, or in other word, $\sqrt{n}[\text{vec}(\boldsymbol{\eta}_n) - \text{vec}(\boldsymbol{\eta}_{\mathbf{Q}})] \xrightarrow{\mathcal{D}} N(0, V_{11}(\boldsymbol{\eta}_{\mathbf{Q}}))$, where $V_{11}(\boldsymbol{\eta}_{\mathbf{Q}}) = GV(\boldsymbol{\eta}_{\mathbf{Q}})G^T$. \square

Proof of Corollary 4.3.5

Proof of Corollary 4.3.5: In our proof of the \sqrt{n} -consistency, without loss of generality we assume that $\mathbf{Q} = I_d$. Note that with an orthogonal matrix \mathbf{Q} , $\mathcal{C}_n^2(\mathbf{Q}^T \boldsymbol{\beta}^T \mathbf{X} | \mathbf{Y}) = \mathcal{C}_n^2(\boldsymbol{\beta}^T \mathbf{X} | \mathbf{Y})$ and $\mathcal{C}^2(\mathbf{Q}^T \boldsymbol{\beta}^T X | Y) = \mathcal{C}^2(\boldsymbol{\beta}^T X | Y)$ (chapter 2). If define $\boldsymbol{\eta}_{\mathbf{Q}} = \boldsymbol{\eta} \mathbf{Q}$, then Proposition 4 holds by using $V_{11}(\boldsymbol{\eta}_{\mathbf{Q}})$ which is obtained by replacing every $\boldsymbol{\eta}$ in V_{11} with $\boldsymbol{\eta}_{\mathbf{Q}}$. (of course, then $V_{11}(\boldsymbol{\eta}_{I_d}) = V_{11}$ in the proof).

To simplify the proof, here we still use $\boldsymbol{\eta}$ by assuming $\mathbf{Q} = I_d$. Let $A_{11} = \mathcal{C}_n''(\boldsymbol{\eta}) + \hat{L}$, where \hat{L} is given in the proof of Lemma B in the section B.2, $A_{12} = (I_d \otimes \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\eta})(I_{d^2} + I_{(d,d)})$, $A_{21} = A_{12}^T$ and $A_{22} = 0$ (A_{22} is a $d^2 \times d^2$ zero matrix), $A_{22.1} = -A_{21}A_{11}^{-1}A_{12}$ then $D = -(A_{11}^{-1} + A_{11}^{-1}A_{12}A_{22.1}^{-1}A_{21}A_{11}^{-1})$ and $F = A_{11}^{-1}A_{12}A_{22.1}^{-1}$

Without loss of generality, we can expand η_n as $\eta_n = \boldsymbol{\eta} + E_n\{A^*\} + o_p(n^{-1/2})$, and we can expand $\hat{\Sigma}$ as $\hat{\Sigma} = \Sigma + E_n\{\Sigma^*\} + o_p(n^{-1/2})$. Then we can get the asymptotic expansion of $\eta_n \eta_n^T \hat{\Sigma}$ as

$$\eta_n \eta_n^T \hat{\Sigma} = \boldsymbol{\eta} \boldsymbol{\eta}^T \Sigma + E_n\{A^* \boldsymbol{\eta}^T \Sigma\} + E_n\{\boldsymbol{\eta} (A^*)^T \Sigma\} + E_n\{\boldsymbol{\eta} \boldsymbol{\eta}^T \Sigma^*\} + o_p(n^{-1/2}). \quad (\text{S.4.5.12})$$

therefore, $\text{vec}(\eta_n \eta_n^T \hat{\Sigma}) - \text{vec}(\boldsymbol{\eta} \boldsymbol{\eta}^T \Sigma) = [(\Sigma \boldsymbol{\eta} \otimes I_p) + (\Sigma \otimes \boldsymbol{\eta}) I_{(d,p)}] \text{vec}(E_n\{A^*\}) + (I_p \otimes \boldsymbol{\eta} \boldsymbol{\eta}^T) \text{vec}(E_n\{\Sigma^*\}) + o_p(n^{-1/2})$, where $\text{vec}(E_n\{A^*\}) = D[\mathcal{C}'_n(\boldsymbol{\eta}) + (I_d \otimes \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\eta})(I_{d^2} +$

$I_{(d,d)}\lambda_0] + F[\text{vec}(\eta^T \hat{\Sigma}_{\mathbf{X}} \eta) - \text{vec}(I_d)]$ and $\text{vec}(E_n\{\Sigma^*\}) = \frac{1}{n} \sum_{i=1}^n \text{vec}(X_i - \mu)(X_i - \mu)^T - \text{vec}(E(X - \mu)(X - \mu)^T)$, where $\mu = E(X)$.

Let $C = [(\Sigma \eta \otimes I_p) + (\Sigma \otimes \eta)I_{(d,p)}]$, $H = I_p \otimes \eta \eta^T$, $U_{8n} = \frac{1}{n} \sum_{i=1}^n \text{vec}(X_i - \mu)(X_i - \mu)^T$, $\mu_8 = \text{vec}(E(X - \mu)(X - \mu)^T)$, then $\text{vec}(\eta_n \eta_n^T \hat{\Sigma}) - \text{vec}(\eta \eta^T \Sigma) = \frac{(n-1)}{n} CDU_{1n} - \sum_{y=1}^C \frac{(n_y-1)}{n} CDU_{2ny} + \frac{n-1}{n} CDU_{4n} - \frac{n-1}{n} CDU_{5n} + \frac{n-1}{n} CFU_{6n} - \frac{n-1}{n} CFU_{7n} - CF \text{vec}(I_d) + HU_{8n} - H \text{vec}(E(X - \mu)(X - \mu)^T)$,

Let $U_{1n}^* = CDU_{1n}$, $U_{2ny}^* = CDU_{2ny}$, $U_{4n}^* = CDU_{4n}$, $U_{5n}^* = CDU_{5n}$, $U_{6n}^* = CFU_{6n}$, $U_{7n}^* = CFU_{7n}$ and $U_{8n}^* = HU_{8n}$; let $\mu_1^* = CD\mu_1$, $\mu_{2y}^* = CD\mu_{2y}$, $\mu_4^* = CD\mu_4$, $\mu_5^* = CD\mu_5$, $\mu_6^* = CF\mu_6$, $\mu_7^* = CF\mu_7$, $\mu_8^* = H\mu_8$, where U_{1n} , U_{2ny} , U_{4n} , U_{5n} , U_{6n} , U_{7n} , μ_1 , μ_{2y} , μ_4 , μ_5 , μ_6 , μ_7 are defined in the proof of Lemma S. 4.5.9.

According to Theorem 6.1.6 (Lehmann, 1999, Ch.6),

$$\sqrt{n} \begin{pmatrix} U_{1n}^* - \mu_1^* \\ U_{2ny}^* - \mu_{2y}^* \\ \dots \\ U_{2nC}^* - \mu_{2nC}^* \\ U_{4n}^* - \mu_4^* \\ U_{5n}^* - \mu_5^* \\ U_{6n}^* - \mu_6^* \\ U_{7n}^* - \mu_7^* \\ U_{8n}^* - \mu_8^* \end{pmatrix} \xrightarrow{\mathcal{D}} N(0, \Sigma^*),$$

$$\text{where } \Sigma^* = \begin{pmatrix} \Sigma_{11}^* & \Sigma_{12y}^* & \Sigma_{14}^* & \Sigma_{15}^* & \Sigma_{16}^* & \Sigma_{17}^* & \Sigma_{18}^* \\ \cdot & \Sigma_{2y2y}^* & \Sigma_{2y4}^* & \Sigma_{2y5}^* & \Sigma_{2y6}^* & \Sigma_{2y7}^* & \Sigma_{2y8}^* \\ \cdot & \cdot & \Sigma_{44}^* & \Sigma_{45}^* & \Sigma_{46}^* & \Sigma_{47}^* & \Sigma_{48}^* \\ \cdot & \cdot & \cdot & \Sigma_{55}^* & \Sigma_{56}^* & \Sigma_{57}^* & \Sigma_{58}^* \\ \cdot & \cdot & \cdot & \cdot & \Sigma_{66}^* & \Sigma_{67}^* & \Sigma_{68}^* \\ \cdot & \cdot & \cdot & \cdot & \cdot & \Sigma_{77}^* & \Sigma_{78}^* \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \Sigma_{88}^* \end{pmatrix},$$

$\Sigma_{ij}^* = CD\Sigma_{ij}D^T C^T$, $i, j = 1, 2y, 4, 5$, $\Sigma_{ij}^* = CD\Sigma_{ij}F^T C^T$, $i = 1, 2y, 4, 5, j = 6, 7$, $\Sigma_{ij}^* = CF\Sigma_{ij}F^T C^T$, $i, j = 6, 7$, where Σ_{ij} , $i, j = 1, \dots, 7$ are defined in the proof of

lemma S. 4.5.9 in this appendix; $\Sigma_{i8}^* = CD\Sigma_{i8}H^T, i = 1, 2, 4, 5$; $\Sigma_{i8}^* = CF\Sigma_{i8}H^T, i = 6, 7$; $\Sigma_{i8}^* = H\Sigma_{i8}H^T$, where $\Sigma_{i8} = a_{i8}\text{cov}(\phi^{(i)}, \phi^{(8)})$ a_{i8} corresponding to the number of entries in $\phi^{(i)}$.

Let $\mathbf{B}^* = \begin{pmatrix} I_{p^2} & (-p_1)I_{p^2} & \cdots & (-p_C)I_{p^2} & I_{p^2} & -I_{p^2} & I_{p^2} & -I_{p^2} & I_{p^2} \end{pmatrix}$, then by

$$\text{Slutsky's theorem, } \sqrt{n} \begin{pmatrix} \text{vec}(\boldsymbol{\eta}_n \boldsymbol{\eta}_n^T \hat{\Sigma}) - \text{vec}(\boldsymbol{\eta} \boldsymbol{\eta}^T \Sigma) \end{pmatrix} \stackrel{D}{=} \sqrt{n} \mathbf{B}^* \begin{pmatrix} U_{1n}^* - \mu_1^* \\ U_{2n1}^* - \mu_{21}^* \\ \cdots \\ U_{2nC}^* - \mu_{2nC}^* \\ U_{4n}^* - \mu_4^* \\ U_{5n}^* - \mu_5^* \\ U_{6n}^* - \mu_6^* \\ U_{7n}^* - \mu_7^* \\ U_{8n}^* - \mu_8^* \end{pmatrix} \longrightarrow N(0, V_{22}),$$

where $V_{22} = B^* \Sigma^* B^{*T}$. In general, without assuming $\mathbf{Q} = I_d$, we have $V_{22}(\boldsymbol{\eta}_{\mathbf{Q}}) = B^* \Sigma^*(\boldsymbol{\eta}_{\mathbf{Q}}) B^{*T}$, and $\Sigma^*(\boldsymbol{\eta}_{\mathbf{Q}})$ is obtained by replacing every $\boldsymbol{\eta}$ in Σ^* with $\boldsymbol{\eta}_{\mathbf{Q}}$. \square

Bibliography

- Akritis, M. and Arnold, S. (1994). Fully nonparametric hypotheses for factorial designs. i. multivariate repeated measures designs. *Journal of the American Statistical Association*, 89(425):336–343.
- Anderson, M. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, 26(1):32–46.
- Anderson, T. (2003). *An Introduction to Multivariate Statistical Analysis, 3rd Edition*. Wiley, New York.
- Bellman, R. (1961). *Adaptive Control Processes*. Princeton University Press.
- Blomqvist, N. (1950). On a measure of dependence between two random variables. *The Annals of Mathematical Statistics*, pages 593–600.
- Blum, J., Kiefer, j., and Rosenblatt, M. (1961). Distribution free tests of independence based on the sample distribution function. *The Annals of Mathematical Statistics*, 32:485–498.
- Bowman, A. and Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*. Oxford University Press, Oxford.
- Bowman, A. and Azzalini, A. (2007). R package sm: nonparametric smoothing methods (version 2.2).
- Candes, E. and Tao, T. (2007). The dantzig selector: statistical estimation when p is much larger than n (with discussion). *The Annals of Statistics*, pages 2313–2404.
- Canty, A. and Ripley, B. (2009). R package boot: bootstrap r (s-plus) functions (version 1.2-35).

- Chang, J., Tang, C., and Wu, Y. (2013). Marginal empirical likelihood and sure independence feature screening. *Annals of statistics*, 41(4):2123–2148.
- Chang, J., Tang, C., and Wu, Y. (2016). Local independence feature screening for nonparametric and semiparametric models by marginal empirical likelihood. *Annals of statistics*, 44(2):515–539.
- Chen, X., Cook, R., and Zou, C. (2015). Diagnostic studies in sufficient dimension reduction. *Biometrika*, 102(3):545–558.
- Chiaromonte, F. and Martinelli, J. (2002). Dimension reduction strategies for analyzing global gene expression data with a response. *Mathematical Biosciences*, 176(1):123–144.
- Cochran, W. and Cox, G. (1957). *Experimental Designs, 2nd Edition*. Wiley, New York.
- Cook, R. (1994). Using dimension-reduction subspaces to identify important inputs in models of physical systems. In *Proceedings of the section on Physical and Engineering Sciences*, pages 18–25. American Statistical Association Alexandria, VA.
- Cook, R. (1996). Graphics for regressions with a binary response. *Journal of the American Statistical Association*, 91(435):983–992.
- Cook, R. (1998a). Principal hessian directions revisited. *Journal of the American Statistical Association*, 93(441):84–94.
- Cook, R. (1998b). *Regression Graphics: Ideas for Studying Regressions Through Graphics*. Wiley, New York.
- Cook, R. (2004). Testing predictor contributions in sufficient dimension reduction. *The Annals of Statistics*, 32(3):1062–1092.
- Cook, R. (2007). Fisher lecture: dimension reduction in regression. *Statistical Science*, 22:1–26.

- Cook, R. and Ni, L. (2005). Sufficient dimension reduction via inverse regression. *Journal of the American Statistical Association*, 100(470):410–428.
- Cook, R. and Weisberg, S. (1991). Discussion of a paper by k. c. li. *Journal of the American Statistical Association*, 86:328–332.
- Cook, R. and Zhang, X. (2014). Fused estimators of the central subspace in sufficient dimension reduction. *Journal of the American Statistical Association*, 109(506):815–827.
- Cook, R. and Zhang, X. (2015). Foundations for envelope models and methods. *Journal of the American Statistical Association*, 110(510):599–611.
- Cui, H., Li, R., and Zhong, W. (2015). Model-free feature screening for ultrahigh dimensional discriminant analysis. *Journal of the American Statistical Association*, 110(510):630–641.
- Davison, A. and Hinkley, D. (1997). *Bootstrap Methods and Their Application*. Cambridge University Press, Oxford.
- Dong, Y. and Li, B. (2010). Dimension reduction for non-elliptically distributed predictors: second-order methods. *Biometrika*, 97(2):279–294.
- Dudoit, S., Fridlyand, J., and Speed, T. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American statistical association*, 97(457):77–87.
- Efron, B. and Tibshirani, R. (1998). *An Introduction to the Bootstrap*. Chapman and Hall/CRC, Boca Raton, Florida.
- Excoffier, L., Smouse, P., and Quattro, J. (1992). Analysis of molecular variance inferred from metric distances among dna haplotypes: Application to human mitochondrial dna restriction data. *Genetics*, 131(2):479–491.

- Fan, J., Feng, Y., and Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, 106(494):544–557.
- Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.
- Fan, J. and 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.
- Fan, J., Ma, Y., and Dai, W. (2014). Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models. *Journal of the American Statistical Association*, 109(507):1270–1284.
- Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with np-dimensionality. *The Annals of Statistics*, 38(6):3567–3604.
- Fung, W., He, X., Liu, L., and Shi, P. (2002). Dimension reduction based on canonical correlation. *Statistica Sinica*, pages 1093–1113.
- Gill, P., Murray, W., and Wright, M. (1981). *Practical Optimization*. Academic press.
- Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C., and Lander, E. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537.
- Gower, J. and Krzanowski, W. (1999). Analysis of distance for structured multivariate data and extensions to multivariate analysis of variance. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48(4):505–519.
- Hand, D. and Taylor, C. (1987). *Multivariate Analysis of Variance and Repeated Measures*. Chapman and Hall, New York.

- Härdle, W., Hall, P., and Ichimura, H. (1993). Optimal smoothing in single-index models. *The Annals of Statistics*, 21(1):157–178.
- Härdle, W. and Stoker, T. (1989). Investigating smooth multiple regression by the method of average derivatives. *Journal of the American statistical Association*, 84(408):986–995.
- Heller, R., Heller, Y., and Gofine, M. (2013). A consistent multivariate test of association based on ranks of distances. *Biometrika*, 100(2):503–510.
- Herbin, E. and Merzbach, E. (2007). The multiparameter fractional broanian motion. In *Math Everywhere*, pages 93–101. Springer.
- Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, pages 293–325.
- Hollander, M. and Wolfe, D. (1999). *Nonparametric Statistical Methods, 2nd Edition*. Wiley, New York.
- Horowitz, J. and Härdle, W. (1996). Direct semiparametric estimation of single-index models with discrete covariates. *Journal of the American Statistical Association*, 91(436):1632–1640.
- Hristache, M., Juditsky, A., Polzehl, J., and Spokoiny, V. (2001). Structure adaptive approach for dimension reduction. *The Annals of Statistics*, 29(6):1537–1566.
- Huo, X. and Székely, G. (2016). Fast computing for distance covariance. *Technometrics*, 58(4):435–447.
- Ichimura, H. (1993). Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of Econometrics*, 58(1-2):71–120.
- Kullback, S. (1959). *Information Theory and Statistics*. John Wiley & Sons.
- Lehmann, E. (1999). *Elements of large-sample theory*. Springer Verlag, New York.

- Li, B. and Dong, Y. (2009). Dimension reduction for nonelliptically distributed predictors. *The Annals of Statistics*, pages 1272–1298.
- Li, B. and Wang, S. (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association*, 102(479):997–1008.
- Li, B., Wen, S., and Zhu, L. (2008). On a projected resampling method for dimension reduction with multivariate responses. *Journal of the American Statistical Association*, 103(483):1177–1186.
- Li, B., Zha, H., and Chiaromonte, F. (2005). Contour regression: a general approach to dimension reduction. *The Annals of Statistics*, 33(4):1580–1616.
- Li, G., Peng, H., Zhang, J., and Zhu, L. (2012a). Robust rank correlation based screening. *The Annals of Statistics*, pages 1846–1877.
- Li, K. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327.
- Li, K. (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, L. (2007). Sparse sufficient dimension reduction. *Biometrika*, 94:603–613.
- Li, L., Zhu, L., and Zhu, L. (2011). Inference on the primary parameter of interest with the aid of dimension reduction estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(1):59–80.
- Li, R., Zhong, W., and Zhu, L. (2012b). Feature screening via distance correlation learning. *Journal of the American Statistical Association*, 107(499):1129–1139.
- Ma, Y. and Zhu, L. (2013a). Efficient estimation in sufficient dimension reduction. *Annals of statistics*, 41(1):250.
- Ma, Y. and Zhu, L. (2013b). A review on dimension reduction. *International Statistical Review*, 81(1):134–150.

- Mai, Q. and Zou, H. (2013). The kolmogorov filter for variable screening in high-dimensional binary classification. *Biometrika*, 100(1):229–234.
- Mai, Q. and Zou, H. (2015). The fused kolmogorov filter: a nonparametric model-free screening method. *The Annals of Statistics*, 43(4):1471–1497.
- Mardia, K., Kent, J., and Bibby, J. (1979). *Multivariate Analysis*. Academic Press.
- McArdle, B. and Anderson, M. (2001). Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology*, 82(1):290–297.
- Powell, J., Stock, J., and Stoker, T. (1989). Semiparametric estimation of index coefficients. *Econometrica: Journal of the Econometric Society*, 57(6):1403–1430.
- Prudnikov, A., Brychkov, A., and Marichev, O. (1986). *Integrals and Series*. Gordon and Breach Science Publishers.
- Puri, M. and Sen, P. (1993). *Nonparametric Methods in Multivariate Analysis*. Wiley, New York.
- Resnick, S. (1999). *A Probability Path*. Birkhäuser.
- Rizzo, M. and Székely, G. (2010). Disco analysis: A nonparametric extension of analysis of variance. *The Annals of Applied Statistics*, 4(2):1034–1055.
- Samarov, A. (1993). Exploring regression structure using nonparametric function estimation. *Journal of the American Statistical Association*, 88(423):836–847.
- Saviotti, P. (1996). *Technological evolution, variety and the economy*. Edward Elgar, Cheltenham.
- Sejdinovic, D., Sriperumbudur, B., Gretton, A., and Fukumizu, K. (2013). Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5):2263–2291.
- Serfling, R. (1980). *Approximation Theorems of Mathematical Statistics.*, volume 162. John Wiley & Sons.

- Shao, X. and Zhang, J. (2014). Martingale difference correlation and its use in high-dimensional variable screening. *Journal of the American Statistical Association*, 109(507):1302–1318.
- Sheng, W. and Yin, X. (2013). Direction estimation in single-index models via distance covariance. *Journal of Multivariate Analysis*, 122:148–161.
- Sheng, W. and Yin, X. (2016). Sufficient dimension reduction via distance covariance. *Journal of Computational and Graphical Statistics*, 25(1):91–104.
- Song, R., Yi, F., and Zou, H. (2014). On varying-coefficient independence screening for high-dimensional varying-coefficient models. *Statistica Sinica*, 24(4):1735–1752.
- Székely, G. and Bakirov, N. (2003). Extremal probabilities for gaussian quadratic forms. *Probability Theory and Related Fields*, 126(2):184–202.
- Székely, G. and Rizzo, M. (2009). Brownian distance covariance. *The Annals of Applied Statistics*, 3(4):1236–1265.
- Székely, G. and Rizzo, M. (2013). The distance correlation t-test of independence in high dimension. *Journal of Multivariate Analysis*, 117:193–213.
- Székely, G., Rizzo, M., and Bakirov, N. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794.
- Taskinen, S., Oja, H., and Randles, R. (2005). Multivariate nonparametric tests of independence. *Journal of the American Statistical Association*, 100(471):916–925.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.
- Vajda, I. (1989). *Theory of Statistical Inference and Information*. Kluwer Academic Publishers.
- Wang, H. and Xia, Y. (2008). Sliced regression for dimension reduction. *Journal of the American Statistical Association*, 103(482):811–821.

- Wang, X., Pan, W., Hu, W., Tian, Y., and Zhang, H. (2015). Conditional distance correlation. *Journal of the American Statistical Association*, 110(512):1726–1734.
- Wilks, S. (1935). On the independence of k sets of normally distributed statistical variables. *Econometrica*, 3:309–326.
- 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., and Zhu, L. (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):363–410.
- Ye, Z. and Weiss, R. (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. (2010). *Sufficient dimension reduction in regression*. In: Cai, TT.; Shen, X., editors., volume chapter 9. World Scientific.
- Yin, X. and Cook, R. (2003). Estimating central subspaces via inverse third moments. *Biometrika*, 90(1):113–125.
- Yin, X. and Cook, R. (2005). Direction estimation in single-index regressions. *Biometrika*, 92(2):371–384.
- Yin, X. and 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.
- Yin, X., Li, B., and Cook, R. (2008). Successive direction extraction for estimating the central subspace in a multiple-index regression. *Journal of Multivariate Analysis*, 99(8):1733–1757.
- Yin, X. and Yuan, Q. (2016). A new class of measures for testing independence. *Submitted*.

- Zapala, M. and Schork, N. (2006). Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables. *Proceedings of the National Academy of Sciences*, 103(51):19430–19435.
- Zeng, P. and Zhu, Y. (2010). An integral transform method for estimating the central mean and central subspaces. *Journal of Multivariate Analysis*, 101(1):271–290.
- Zhu, L. and Fang, K. (1996). Asymptotics for kernel estimate of sliced inverse regression. *The Annals of Statistics*, 24(3):1053–1068.
- Zhu, L., Li, L., Li, R., and Zhu, L. (2011). Model-free feature screening for ultrahigh dimensional data. *Journal of the American Statistical Association*, 106(496):1464–1475.
- Zhu, L., Wang, T., Zhu, L., and Ferré, L. (2010). Sufficient dimension reduction through discretization-expectation estimation. *Biometrika*, 97(2):295–304.
- Zhu, Y. and Zeng, P. (2006). Fourier methods for estimating the central subspace and the central mean subspace in regression. *Journal of the American Statistical Association*, 101(476):1638–1651.

Vita

Education

- **M.S. in Statistics** University of Kentucky, Lexington, KY, 2012-2014
- **B.S. in Statistics** Shandong University, Jinan, China, 2008-2012
- **B.S. in Finance** Shandong University, Jinan, China, 2009-2012

Experience

- **Research Assistant** University of Kentucky, 2014-2017
- **Teaching Assistant** University of Kentucky, 2012-2014