2019

A NEW INDEPENDENCE MEASURE AND ITS APPLICATIONS IN HIGH DIMENSIONAL DATA ANALYSIS

Chenlu Ke
University of Kentucky, cke237@g.uky.edu
Digital Object Identifier: https://doi.org/10.13023/etd.2019.269

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.

Chenlu Ke, Student

Dr. Xiangrong Yin, Major Professor

Dr. Constance L. Wood, Director of Graduate Studies

A NEW INDEPENDENCE MEASURE AND ITS APPLICATIONS IN HIGH DIMENSIONAL DATA ANALYSIS

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
Chenlu Ke
Lexington, Kentucky

Director: Dr. Xiangrong Yin, Professor of Statistics
Lexington, Kentucky
2019

ABSTRACT OF DISSERTATION

A NEW INDEPENDENCE MEASURE AND ITS APPLICATIONS IN HIGH DIMENSIONAL DATA ANALYSIS

This dissertation has three consecutive topics. First, we propose a novel class of independence measures for testing independence between two random vectors based on the discrepancy between the conditional and the marginal characteristic functions. If one of the variables is categorical, our asymmetric index extends the typical ANOVA to a kernel ANOVA that can test a more general hypothesis of equal distributions among groups. The index is also applicable when both variables are continuous. Second, we develop a sufficient variable selection procedure based on the new measure in a large p small n setting. Our approach incorporates marginal information between each predictor and the response as well as joint information among predictors. As a result, our method is more capable of selecting all truly active variables than marginal selection methods. Furthermore, our procedure can handle both continuous and discrete responses with mixed-type predictors. We establish the sure screening property of the proposed approach under mild conditions. Third, we focus on a model-free sufficient dimension reduction approach using the new measure. Our method does not require strong assumptions on predictors and responses. An algorithm is developed to find dimension reduction directions using sequential quadratic programming. We illustrate the advantages of our new measure and its two applications in high dimensional data analysis by numerical studies across a variety of settings.

KEYWORDS: High dimensional data analysis, Independence, Reproducing Kernel Hilbert Space, Sufficient Dimension Reduction, Sufficient Variable Selection.

Author's signature: Chenlu Ke

Date: June 30, 2019

A NEW INDEPENDENCE MEASURE AND ITS APPLICATIONS IN HIGH DIMENSIONAL DATA ANALYSIS

By
Chenlu Ke

Director of Dissertation: Xiangrong Yin

Director of Graduate Studies: Constance L. Wood

Date: June 30, 2019

ACKNOWLEDGMENTS

First and foremost, 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. He is always passionate, diligent and insightful. His persistent guidance and encouragement helped me overcome many difficulties throughout my graduate study and job search. and will continue inspiring me in my future career.

I would like to thank all my dissertation committee members: Drs. Arnold Stromberg, Solomon Harrar, Derek Young, Chi Wang and Erin Abner. Dr. Stromberg has been a great chair dedicated to support students in all aspects with his knowledge and resource. Dr. Harrar taught me in four courses and his rigorous thinking has greatly impacted me. I finished my first paper with Dr. Young. He opened the door for me in research while I had little experience at the time, and he has encouraged and helped me a lot since then. It is also my honor to have Dr. Wang and Dr. Abner in my committee.

I am grateful for my journey at UK in the past five years. Thanks to all the faculty members. Thanks to peers and friends who I shared the graduate life with - a life that we often complained about but I have already started to miss.

Special thanks to Pinchao for his love and accompany, and for everything that we have experienced together.

Lastly, I would like to thank my parents Jiachun Ke and Qiuhong Chen for their unconditional love and endless support. Even though I live half of the world from home, we are very close as a family. They made me who I am today and I never know what to return. I hope that they are at least a little proud of me for what I have accomplished so far.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1 Introduction

As modern technology allows tremendous data collection at low cost, high dimensional data with complex structure become more and more common in diverse fields of scientific research such as genetics and economics. Traditional statistical methods have been challenged by difficulties in analyzing big data, which stimulates the development of new approaches to discover stories behind data. This dissertation contains three closely related topics in high dimensional data analysis. The first topic introduces a novel measure for testing independence between two random vectors of arbitrary dimensions. The second and the third topics focus on sufficient variable selection and sufficient dimension reduction approaches using the new measure, respectively.

Measuring dependence between random variables/vectors is very important in statistics. However, most classical methods can only detect certain types of dependence or have numerous assumptions that are difficult to assess. More flexible approaches have been developed and successfully used for detecting dependence of variables, such as distance covariance (DCOV, Székely, Rizzo and Bakirov 2007) in statistics literature. A related index for categorical responses termed distance components (DISCO), provides a nonparametric extension of ANOVA (Rizzo et al. 2010). Yin and Yuan (2019) also developed a measure equivalent to DISCO called expectation of conditional difference (ECD). In machine learning literature, Hilbert-Schmidt independence criterion (HSIC, Gretton et al. 2005a, Gretton et al. 2005b, Gretton et al. 2008) has been proposed as a kernel-based counterpart of DCOV. HSIC is built upon mappings of variables into reproducing kernel Hilbert spaces (RKHS) that inherit properties of interest such as independence and homogeneity. With smoothness assumptions, RKHS-based measures yield better power in hypothesis tests than ap-

proaches based on ordinary distances between the unmapped variables. In Chapter 2, we introduce a new measure, expected conditional HSIC (ECHSIC), based on the same idea of ECD, that is, measuring the discrepancy between the conditional and the marginal characteristic function. Whereas most of the independence measures treat two variables symmetrically, we consider one of the variables conditioning on the other, which is a common idea in regression and leads to a kernel ANOVA for testing equal distributions when one of the variable is categorical. Moreover, we establish that ECD is precisely an instance of generalized ECHSIC and both of them belong to the large class of RKHS-based measures.

An important application of our independence measure is in variable selection. Variable selection plays a significant role in modeling modern statistical problems with ultrahigh dimensional data. Although regularization methods such as LASSO (Tibshirani 1996), elastic net (Zou and Hastie 2005), Dantzig selector (Candes and Tao 2004) and many others can deal with cases where the number of predictors p exceeds the sample size n, they may not perform well for large p small n ($p \gg n$) data due to computational cost, statistical accuracy and the stability of algorithms (Fan and Lv 2008). Sure independence screening (Fan and Lv 2008) and similar others (Li, Zhong and Zhu 2012, Cui, Li and Zhong 2015, etc.) have become popular solutions to ultrahigh dimensional variable selection. However, these feature screening approaches only adopt marginal information between each predictor and the response, so significant predictors that are uncorrelated (or even independent) but jointly correlated with the response cannot be picked up. Yin and Hilafu (2015) made a formal definition of sufficient variable selection (SVS), where they employed the idea of sufficient dimension reduction (SDR) as a bridge to tackle the large p small n problem. Enlightened by their work, we propose a SVS method based on ECHSIC and its extension in Chapter 3. While being model-free with sure screening property held under mild conditions, our procedure improves existing methods in

two aspects. First, similar to other SVS procedures like Yang, Yin and Zhang (2019), our method optimizes SIS and related methods by taking joint information among predictors into consideration. Second, our method can handle both continuous and discrete responses with mixed-type predictors.

Another appealing utilization of ECHSIC contributes to sufficient dimension reduction (SDR, Cook 1996). SDR is a powerful tool to extract the key information hidden in the high dimensional data. The extraction of information is based on the notion of sufficiency, which means a set of functions of the predictors provides all the information needed to understand or predict the response. For example, the aim of linear SDR is to find linear combinations of predictors that completely contains the regression information. The number of linear combinations is typically much smaller than the number of predictors and by SDR we are able to downsize the data without loss of information. Various methodologies have been developed including the well-known sliced inverse regression (SIR, Li 1991), sliced average variance estimation (SAVE, Cook and Weisberg 1991), minimum average variance estimator (MAVE, Xia et al. 2002), to name a few. Note that sufficiency is a statistical concept derived from conditional independence, so independence measures and correlation indices can be useful in SDR. In Chapter 4, we propose a new method to achieve SDR via our new measure that is applicable to single-index and multi-index models with either continuous responses or categorical responses. Our method does not require strong assumptions on the predictors compared to other existing dimension reduction approaches. In recent years, SDR has expanded in many directions, for example, from high dimension to ultrahigh dimension, from typical data to functional data, from linearity to non-linearity, etc. Our methods is promising to be adapted to those development in the future.

Chapter 2 Expected Conditional Hilbert-Schmidt Independence Criterion for Testing Independence [1]

2.1 Introduction

Statistical independence/dependence tests have been proposed with a broad variety of measures. However, most classical methods can only detect certain types of dependence or have assumptions that are difficult to assess and meet. For example, the well-known Spearman's correlation can only capture monotonic relationships between the two variables. Likelihood-based methods such as Wilk's Lambda are not applicable if the dimension exceeds the sample size, or when distribution assumptions do not hold. Therefore, testing independence is a challenging task, especially in high dimensional spaces with complicated dependence structures.

More flexible measures have been developed to overcome these difficulties in statistical literature. Wang, Jiang and Liu (2017) proposed a new measure, G^2, to test whether two univariate continuous random variables are dependent and measure the strength of their relationship. The G^2 can be considered as the piecewise R^2 between the sliced variables. And $G^2 = 0$ if and only if $E(X|Y)$, $E(Y|X)$, $Var(X|Y)$ and $Var(Y|X)$ are all constant, which in fact, is not equivalent to the independence of X and Y. The measure can handle nonlinearity and heteroscedastic errors compared to R^2. Its generalization to continuous multivariate variables is intuitive, but it may become difficult and complicated, due to its slicing scheme. Székely, Rizzo and Bakirov (2007) proposed a novel measure for multivariate variables termed distance covariance (DCOV) and related distance correlation (DCOR). Unlike the classical correlation or the G^2, DCOR is zero if and only if the random variables are independent. This

[1] Ke, C., & Yin X. (2019), "Expected Conditional Characteristic Function-based Measures for Testing Independence". *Journal of the American Statistical Association.* Available online at https://doi.org/10.1080/01621459.2019.1604364

measure has led to applications in variable selection (Li, Zhong and Zhu 2012) and dimension reduction (Sheng and Yin 2013; 2016). In addition, developing conditional independence tests has also been attractive since they are essential to statistical inference such as graphical models, Bayesian network analysis and dimension reduction. Su and White (2003, 2007, 2008) proposed a series of difference measures between conditional densities based on smoothing empirical likelihood, conditional characteristic function, and weighted Hellinger distance, respectively. Wang et al. (2015) extended the work of Székely, Rizzo and Bakirov (2007) and developed a conditional independence measure.

Related research exists in machine learning literature as well. Kernel-based methods have been developed and successfully used for detecting dependence of variables (Bach and Jordan 2002a; Gretton et al. 2005a; Gretton et al. 2005b; Sun et al. 2007). Applications of kernel-based approaches can be found in areas including gene selection (Yamanishi, Vert and Kanehisa 2004), fitting graphical models (Bach and Jordan 2002b), dependence detection in fMRI signals (Gretton et al. 2005) and variable selection (Fukumizu, Bach and Jordan 2004). Hilbert Schmidt independence criterion (HSIC) is one of the kernel-based measures for independence that has been proposed (Gretton et al. 2005a; Gretton et al. 2005b; Gretton et al. 2008). HSIC is computed as the Hilbert-Schmidt norm of a cross-covariance operator on mappings of variables into reproducing kernel Hilbert spaces (RKHS). Those mappings inherit properties of interest such as independence and homogeneity. Other than covariance, an RKHS dependence statistic can also rely on distance (Smola et al. 2007) or correlation (Dauxois and Nkiet 1998; Bach and Jordan 2002a; Fukumizu, Bach and Gretton 2007) between the feature mappings. Extensions of HSIC include an associated measure for conditional independence developed by Fukumizu et al. (2008). Sejdinovic et al. (2013) proposed a framework that nicely links HSIC with DCOV; i.e., DCOV is precisely an example of HSIC.

In this chapter, we develop a new class of measures for testing independence of two random vectors based on the discrepancy between the conditional and the marginal characteristic functions. Whereas most of the independence measures treat two variables symmetrically, we consider one of the variables conditioning on the other, which is a common idea in regression, classification, and discriminant analysis. More importantly, when one of the variables is nominal, independence tests based on symmetric measures like HSIC and DCOV still rely on the values of the nominal variable, which is problematic. Hence, there is a lack of appropriate and powerful tests other than the classical (M)ANOVA or nonparametric methods like Kruskal-Wallis. Our work fills this gap. Intuitively, the relation between HSIC/DCOV and our measure is analogous to the relation between a linear regression/correlation and an ANOVA. In fact, our method extends the classical ANOVA to a kernel ANOVA that can test a more general hypothesis of equal distributions among groups. Note that Rizzo and Székely (2010) proposed a measure called distance components (DISCO) that also focuses on multi-sample hypothesis for equal distributions, but our approach generates a much broader class of measures. Essentially, we develop a parallel RKHS framework to HSIC/DCOV that unifies our index and DISCO. Although we are motivated by aforementioned setting, our index is also applicable when both variables are continuous. In addition, if necessary, we can simply obtain a symmetric index by adding a term with switched roles of the two variables.

The rest of the chapter is organized as follows. Section 2 introduces the new measure, a development in a RKHS framework and affiliated properties. Section 3 constructs two empirical estimates and obtains their respective asymptotic distributions. Section 4 provides an algorithm to carry out independence tests using permutations. Section 5 briefly extends the marginal independence measure to a conditional independence measure. Section 6 numerically demonstrates the advantages of our method compared to some existing approaches. Section 7 concludes the chapter with a short

discussion. All proofs are delayed in the appendix.

2.2 Expected Conditional Hilbert-Schmidt Independence Criterion (ECHSIC)

In this section, we introduce a new class of independence measures through two different approaches and discuss related properties.

2.2.1 ECHSIC via Bochner's Theorem

Suppose random vectors $\mathbf{X} \in \mathbb{R}^p$ and $\mathbf{Y} \in \mathbb{R}^q$. Let $(\mathbf{X}', \mathbf{Y}')$ be an i.i.d. copy of $(\mathbf{X}, \mathbf{Y})$, $\varphi_{\mathbf{X}}$ denote the characteristic function of $\mathbf{X}$ and $\varphi_{\mathbf{X}|\mathbf{Y}}$ denote the conditional characteristic function of $\mathbf{X}$ given $\mathbf{Y}$. We use $E_{\mathbf{X}_\mathbf{y}}(\cdot)$ to represent conditional expectation $E(\cdot|\mathbf{Y} = \mathbf{y})$ and $E_{\mathbf{X}_\mathbf{y}, \mathbf{X}'_\mathbf{y}}(\cdot)$ to denote $E(\cdot|\mathbf{Y} = \mathbf{y}, \mathbf{Y}' = \mathbf{y})$. A hypothesis test of independence between $\mathbf{X}$ and $\mathbf{Y}$ is given by $H_0 : \varphi_{\mathbf{X}|\mathbf{Y}} = \varphi_{\mathbf{X}}$ vs. $H_1 : \varphi_{\mathbf{X}|\mathbf{Y}} \neq \varphi_{\mathbf{X}}$. Thus, it is natural to define a measure based on the discrepancy between $\varphi_{\mathbf{X}|\mathbf{Y}}$ and $\varphi_{\mathbf{X}}$. We consider the following distance-like quantity between the two characteristic functions $\varphi_{\mathbf{X}|\mathbf{Y}}$ and $\varphi_{\mathbf{X}}$:

$$\psi^2_\omega(\mathbf{Y}) \equiv \int |\varphi_{\mathbf{X}|\mathbf{Y}}(\mathbf{u}) - \varphi_{\mathbf{X}}(\mathbf{u})|^2 d\omega(\mathbf{u}),$$

where ω is a finite nonnegative Borel measure on $\mathbb{R}^p$. Although $\psi^2_\omega(\mathbf{Y})$ is an intuitive measure of the discrepancy between $\varphi_{\mathbf{X}|\mathbf{Y}}$ and $\varphi_{\mathbf{X}}$, its calculation may be computationally demanding in practice. However, $\psi^2_\omega(\mathbf{Y})$ can also be generated by a positive semi-definite kernel that is induced by ω, which results in an equivalent but simpler representation. The following theorem of Bochner (Wendland 2004, Theorem 6.6) is the key.

Theorem 2.1 (Bochner)**.** *A continuous function $K : \mathbb{R}^p \to \mathbb{C}$ is positive semi-definite if and only if it is the Fourier transform of a finite nonnegative Borel measure ω on*

$\mathbb{R}^p$, *that is,*

$$K(\mathbf{x}) = \int_{\mathbb{R}^p} e^{-i\mathbf{x}^T\mathbf{u}} d\omega(\mathbf{u}), \tag{2.1}$$

for any $\mathbf{x} \in \mathbb{R}^p$.

In the remaining text, we state that a positive semi-definite kernel K is induced by a finite nonnegative Borel measure ω if K is defined by ω according to (2.1). We then obtain an alternative formula for $\psi^2_\omega(\mathbf{Y})$ by applying Bochner's Theorem.

Theorem 2.2. *For a given event* $\mathbf{Y} = \mathbf{Y}' = \mathbf{y}$,

$$\psi^2_\omega(\mathbf{y}) \equiv E_{\mathbf{X}_\mathbf{y}, \mathbf{X}'_\mathbf{y}} K(\mathbf{X} - \mathbf{X}') - 2E_{\mathbf{X}_\mathbf{y}, \mathbf{X}'} K(\mathbf{X} - \mathbf{X}') + E_{\mathbf{X}, \mathbf{X}'} K(\mathbf{X} - \mathbf{X}'), \tag{2.2}$$

where ω *is a finite nonnegative Borel measure on* $\mathbb{R}^p$ *and* K *is a positive semi-definite kernel induced by* ω.

Here we restrict ourselves to a translation-invariant kernel $K : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ that can be written in terms of the difference of their arguments. We henceforth use notation $\psi^2_K(\mathbf{Y})$ instead of $\psi^2_\omega(\mathbf{Y})$ without ambiguity. Note that $\psi^2_K(\mathbf{Y})$ is a $\mathbf{Y}$-measurable random variable. Then a measure of the independence between $\mathbf{X}$ and $\mathbf{Y}$, $\mathcal{I}^2_K(\mathbf{X}|\mathbf{Y})$, is naturally defined by taking the expectation over $\mathbf{Y}$, i.e. $\mathcal{I}^2_K(\mathbf{X}|\mathbf{Y}) \equiv E_\mathbf{Y}\psi^2_K(\mathbf{Y})$. To this end, we formally define our new index.

Definition 2.1. *Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *be two random variables on* $\mathbb{R}^p$ *and* $\mathbb{R}^q$, *respectively. For a given kernel* K *induced by a finite nonnegative Borel measure* ω *on* $\mathbb{R}^p$, *the expected conditional Hilbert-Schmidt independence criterion (ECHSIC) is defined as*

$$\mathcal{I}^2_K(\mathbf{X}|\mathbf{Y}) \equiv E_\mathbf{Y}\left[\int \left|\varphi_{\mathbf{X}|\mathbf{Y}}(\mathbf{u}) - \varphi_\mathbf{X}(\mathbf{u})\right|^2 d\omega(\mathbf{u})\right]. \tag{2.3}$$

Or equivalently,

$$\mathcal{I}^2_K(\mathbf{X}|\mathbf{Y}) \equiv E_\mathbf{Y} E_{\mathbf{X}_\mathbf{Y}, \mathbf{X}'_\mathbf{Y}} K(\mathbf{X} - \mathbf{X}') - E_{\mathbf{X}, \mathbf{X}'} K(\mathbf{X} - \mathbf{X}'). \tag{2.4}$$

Note that the weight function used in DCOV is not integrable and hence, Bochner's theorem does not apply for DCOV. The counterpart of our index using the weight function of DCOV is developed by Yin and Yuan (2019), named as ECD, showing that ECD statistic is actually equivalent to DISCO. In the next two subsections, we introduce a general framework of RKHS that unifies ECD and ECHSIC.

2.2.2 Generalized ECHSIC via MMD

Previously, we proposed ECHSIC in an intuitive way based on the characteristic functions. However, this new class of measures can also be developed via maximum mean discrepancy (MMD, Gretton et al. 2012a) without the requirement for kernels to be translation-invariant. MMD is the largest difference in expectations over functions in the unit ball of a RKHS and can be used to determine if two samples are drawn from different distributions (Gretton et al. 2012a). MMD can be employed to develop HSIC and hence, to measure statistical independence. Now we follow the framework of Gretton et al. (2012a) and Sejdinovic et al. (2013) to generalize ECHSIC via MMD on real spaces. We begin with an introduction to RKHS.

Definition 2.2 (RKHS, Sejdinovic et al. 2013, Definition 8)**.** *Let $\mathbb{H}$ be a Hilbert space of real-valued functions defined on $\mathcal{X}$. A function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a reproducing kernel of $\mathbb{H}$ if:*

1. $\forall \mathbf{x} \in \mathcal{X}$, $K(\cdot, \mathbf{x}) \in \mathbb{H}$;

2. $\forall \mathbf{x} \in \mathcal{X}$, $\forall f \in \mathbb{H}$, $< f, K(\cdot, \mathbf{x}) >_{\mathbb{H}} = f(\mathbf{x})$.

If $\mathbb{H}$ has a reproducing kernel K, it is said to be a reproducing kernel Hilbert space (RKHS) and denoted by $\mathbb{H}_k$.

A reproducing kernel is positive definite. Conversely, Moore-Aronszajn Theorem (Berlinet and Thomas-Agnan 2011, Theorem 3) states that, for every positive definite

function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, there is an associated RKHS $\mathbb{H}_K$ of real-valued functions on $\mathcal{X}$ with the reproducing kernel k. The map $\phi : \mathcal{X} \to \mathbb{H}_K$, $\phi : \mathbf{x} \to K(\cdot, \mathbf{x})$ is called the canonical feature map of K. We say that K is a non-degenerate kernel if its feature map is injective. The concept of feature map can be extended to kernel embeddings of finite signed Borel measures on $\mathcal{X}$. Let $\mathcal{M}(\mathcal{X})$ be the set of all finite signed Borel measures on $\mathcal{X}$ and $\mathcal{M}_+^1(\mathcal{X})$ be the set of all Borel probability measures on $\mathcal{X}$.

Definition 2.3 (Kernel embedding, Sejdinovic et al. 2013, Definition 9)**.** *Let K be a measurable kernel on $\mathcal{X}$, and $\nu \in \mathcal{M}(\mathcal{X})$. The kernel embedding of ν into the RKHS, $\mathbb{H}_K$, is $\mu_K(\nu) \in \mathbb{H}_K$ such that $\langle f, \mu_K(\nu) \rangle_{\mathbb{H}_K} = \int f(\mathbf{x}) d\nu(\mathbf{x})$ for all $f \in \mathbb{H}_K$.*

The kernel embedding can alternatively be defined by the Bochner integral: $\mu_K(\nu) = \int K(\cdot, \mathbf{x}') d\nu(\mathbf{x}')$. To ensure that the kernel embeddings exist, we need to restrict ourselves to a particular class of measures. For a measurable kernel K on $\mathcal{X}$ and $\theta > 0$, denote,

$$\mathcal{M}_K^\theta(\mathcal{X}) \equiv \left\{ \nu \in \mathcal{M}(\mathcal{X}) : \int K^\theta(\mathbf{x}, \mathbf{x}) d|\nu|(\mathbf{x}) < \infty \right\}.$$

The kernel embedding $\mu_K(\nu)$ is well defined for $\nu \in \mathcal{M}_K^{\frac{1}{2}}(\mathcal{X})$ by Riesz representation theorem (Sejdinovic et al. 2013). Therefore, kernel embeddings of Borel probability measures in $\mathcal{M}_+^1(\mathcal{X}) \cap \mathcal{M}_K^{\frac{1}{2}}(\mathcal{X})$ exist, and we can introduce a discrepancy between two Borel probability measures in terms of the Hilbert space distance between their embeddings.

Definition 2.4 (MMD, Sejdinovic et al. 2013, Definition 10)**.** *Let K be a measurable kernel on $\mathcal{X}$, and let probability measures $\mathbf{P}$, $\mathbf{Q} \in \mathcal{M}_+^1(\mathcal{X}) \cap \mathcal{M}_K^{\frac{1}{2}}(\mathcal{X})$. The maximum mean discrepancy (MMD) γ_K between $\mathbf{P}$ and $\mathbf{Q}$ is given by*

$$\gamma_K(\mathbf{P}, \mathbf{Q}) \equiv \|\mu_K(\mathbf{P}) - \mu_K(\mathbf{Q})\|_{\mathbb{H}_K}.$$

Let $\mathbf{P} \equiv \mathbf{P}_{\mathbf{X}|\mathbf{Y}}$ be the probability distribution of $\mathbf{X}$ given $\mathbf{Y}$, $\mathbf{Q} \equiv \mathbf{P}_{\mathbf{X}}$ be the probability distribution of $\mathbf{X}$ and K be a measurable kernel on $\mathbb{R}^p$. Assuming $\mathbf{P}, \mathbf{Q} \in \mathcal{M}_+^1(\mathbb{R}^p) \cap \mathcal{M}_K^{\frac{1}{2}}(\mathbb{R}^p)$, we have

$$\begin{aligned}\gamma_K^2(\mathbf{Y}) &\equiv \gamma_K^2(\mathbf{P}, \mathbf{Q}) \\ &= \langle \mu_K(\mathbf{P}), \mu_K(\mathbf{P}) \rangle_{\mathbb{H}_K} + \langle \mu_K(\mathbf{Q}), \mu_K(\mathbf{Q}) \rangle_{\mathbb{H}_K} - 2 \langle \mu_K(\mathbf{P}), \mu_K(\mathbf{Q}) \rangle_{\mathbb{H}_K} \\ &= E_{\mathbf{X}_\mathbf{Y}} E_{\mathbf{X}'_\mathbf{Y}} K(\mathbf{X}, \mathbf{X}') + E_{\mathbf{X}} E_{\mathbf{X}'} K(\mathbf{X}, \mathbf{X}') - 2 E_{\mathbf{X}_\mathbf{Y}} E_{\mathbf{X}'} K(\mathbf{X}, \mathbf{X}'). \end{aligned} \tag{2.5}$$

Definition 2.5 (Generalized ECHSIC)**.** *Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *be two random variables on* $\mathbb{R}^p$ *and* $\mathbb{R}^q$*, respectively. Assume* $\mathbf{P} \equiv \mathbf{P}_{\mathbf{X}|\mathbf{Y}}, \mathbf{Q} \equiv \mathbf{P}_{\mathbf{X}} \in \mathcal{M}_+^1(\mathbb{R}^p) \cap \mathcal{M}_K^{\frac{1}{2}}(\mathbb{R}^p)$*. For a given measurable kernel* K *on* $\mathbb{R}^p$*, the generalized ECHSIC, is defined as*

$$\begin{aligned}\mathcal{H}_K^2(\mathbf{X}|\mathbf{Y}) &\equiv E_{\mathbf{Y}} \gamma_K^2(\mathbf{Y}) \\ &= E_{\mathbf{Y}} E_{\mathbf{X}_\mathbf{Y}, \mathbf{X}'_\mathbf{Y}} K(\mathbf{X}, \mathbf{X}') - E_{\mathbf{X}, \mathbf{X}'} K(\mathbf{X}, \mathbf{X}'). \end{aligned} \tag{2.6}$$

In particular, $\mathcal{H}_K^2(\mathbf{X}|\mathbf{Y}) = \mathcal{I}_K^2(\mathbf{X}|\mathbf{Y})$ *if* $K(\mathbf{x}, \mathbf{x}') = K(\mathbf{x} - \mathbf{x}')$*.*

2.2.3 A unified framework

We now show that ECD (Yin and Yuan 2019) belongs to the family of generalized ECHSIC measures. The connection between negative type semi-metrics and distance-induced kernels, which are translation-variant positive definite kernels, is the key to our main result. We begin with an introduction to negative type semi-metrics as to define the generalized ECD as well as distance-induced kernels.

Definition 2.6 (Sejdinovic et al. 2013, Definitions 1 and 2)**.** *Let* $\rho : \mathcal{X} \times \mathcal{X} \to [0, \infty)$ *be a function such that* $\forall \mathbf{x}, \mathbf{x}' \in \mathcal{X}$*,*

1. $\rho(\mathbf{x}, \mathbf{x}') = 0$ *if and only if* $\mathbf{x} = \mathbf{x}'$*;*

2. $\rho(\mathbf{x}, \mathbf{x}') = \rho(\mathbf{x}', \mathbf{x})$*;*

3. $\forall n \geq 2$, $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n \in \mathbb{R}^p$, *and* $\alpha_1, \alpha_2, ...\alpha_n \in \mathbb{R}$, *with* $\sum_{i=1}^n \alpha_i = 0$,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j \rho(\mathbf{x}_i, \mathbf{x}_j) \leq 0.$$

Then $(\mathcal{X}, \rho)$ *is said to be a negative type semi-metric and* ρ *is called a semi-metric on* $\mathcal{X}$.

Before we proceed to formally define generalized ECD, we need to introduce a new class of Borel measures

$$\widetilde{\mathcal{M}}_\rho^\theta(\mathcal{X}) \equiv \left\{ \nu \in \mathcal{M}(\mathcal{X}) : \exists \mathbf{x}_0 \in \mathcal{X} \text{ s.t. } \int \rho^\theta(\mathbf{x}, \mathbf{x}_0) d|\nu|(\mathbf{x}) < \infty \right\}.$$

We say that $\nu \in \widetilde{\mathcal{M}}_\rho^\theta(\mathcal{X})$ has a finite θ-moment ($\theta > 0$) with respect to a semi-metric ρ of negative type.

Definition 2.7 (Generalized ECD)**.** *Suppose* $(\mathbb{R}^p, \rho)$ *is a semi-metric space of negative type. Assume* $\mathbf{P} \equiv \mathbf{P}_{\mathbf{X}|\mathbf{Y}}$, $\mathbf{Q} \equiv \mathbf{P}_{\mathbf{X}} \in \mathcal{M}_+^1(\mathbb{R}^p) \cap \widetilde{\mathcal{M}}_\rho^1(\mathbb{R}^p)$, *the generalized ECD is defined as*

$$\mathcal{V}_\rho^2(\mathbf{X}|\mathbf{Y}) \equiv E_{\mathbf{Y}} D_\rho(\mathbf{Y}),$$

where $D_\rho(\mathbf{Y}) \equiv -\int \rho d([\mathbf{P} - \mathbf{Q}] \times [\mathbf{P} - \mathbf{Q}])$.

Note that if $\rho(\mathbf{x}, \mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\|_p$, then $\mathcal{V}_\rho^2(\mathbf{X}|\mathbf{Y})$ is precisely the ECD of Yin and Yuan (2019). Similar to DCOV, we can further extend ECD to a class of α-distance measures by choosing $\rho(\mathbf{x}, \mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\|_p^\alpha$, where $0 < \alpha \leq 2$.

We now introduce distance-induced kernels and illustrate the relation with the semi-metrics of negative type.

Definition 2.8 (Distance-induced kernel, Sejdinovic et al. 2013, Definition 13)**.** *Let* ρ *be a semi-metric of negative type on* $\mathcal{X}$ *and let* $\mathbf{x}_0 \in \mathcal{X}$. *The kernel*

$$K(\mathbf{x}, \mathbf{x}') = \frac{1}{2}[\rho(\mathbf{x}, \mathbf{x}_0) + \rho(\mathbf{x}', \mathbf{x}_0) - \rho(\mathbf{x}, \mathbf{x}')]$$

is said to be the distance-induced kernel induced by ρ *and centered at* $\mathbf{x}_0$.

By varying $\mathbf{x}_0$, ρ induces a family of distance-induced kernels:

$$\mathcal{F}_\rho \equiv \left\{ \frac{1}{2}[\rho(\mathbf{x}, \mathbf{x}_0) + \rho(\mathbf{x}', \mathbf{x}_0) - \rho(\mathbf{x}, \mathbf{x}')] : \mathbf{x}_0 \in \mathcal{X} \right\}.$$

Every $K \in \mathcal{F}_\rho$ is positive definite and non-degenerate, i.e., $\mathbf{x} \mapsto K(\cdot, \mathbf{x})$ is injective. In the opposite, any non-degenerate kernel K on $\mathcal{X}$ can generate a valid semi-metric ρ of negative type on $\mathcal{X}$ by defining

$$\rho(\mathbf{x}, \mathbf{x}') = K(\mathbf{x}, \mathbf{x}) + K(\mathbf{x}', \mathbf{x}') - 2K(\mathbf{x}, \mathbf{x}').$$

It is clear that every distance-induced kernel $K \in \mathcal{F}_\rho$ induced by ρ, also generates ρ. Furthermore, if K generates ρ, then $\mathcal{M}_K^{\frac{n}{2}}(\mathcal{X}) = \widetilde{\mathcal{M}}_\rho^{\frac{n}{2}}(\mathcal{X})$ for all $n \in \mathbb{N}$ (Sejdinovic et al. 2013, Proposition 20). Note that $\mathcal{M}_K^1(\mathcal{X}) \subset \mathcal{M}_K^{\frac{1}{2}}(\mathcal{X})$ and hence, $\mathcal{M}_\rho^1(\mathcal{X}) \subset \mathcal{M}_K^{\frac{1}{2}}(\mathcal{X})$. We are set to build up the connection between $\mathcal{V}_\rho^2(\mathbf{X}|\mathbf{Y})$ and $\mathcal{H}_K^2(\mathbf{X}|\mathbf{Y})$.

Theorem 2.3. *Let $(\mathbb{R}^p, \rho)$ be a semi-metric space of negative type and let K be any kernel that generates ρ. Suppose $\mathbf{P} \equiv \mathbf{P}_{\mathbf{X}|\mathbf{Y}}$ and $\mathbf{Q} \equiv \mathbf{P}_{\mathbf{X}}$ satisfy $\mathbf{P}$, $\mathbf{Q} \in \mathcal{M}_+^1(\mathbb{R}^p) \cap \widetilde{\mathcal{M}}_\rho^1(\mathbb{R}^p)$. Then $\mathcal{V}_\rho^2(\mathbf{X}|\mathbf{Y}) = 2\mathcal{H}_K^2(\mathbf{X}|\mathbf{Y})$.*

2.2.4 Properties of generalized ECHSIC

To make generalized ECHSIC a legitimate measure of independence, the kernel is required to be characteristic, meaning that the feature mapping from the space of probability measures to the RKHS is injective. Conditions under which kernels are characteristic can be found in Fukumizu et al. (2009) and Sriperumbudur et al. (2008, 2010). Examples of characteristic kernels include Gaussian, Laplacian, and inverse multiquadratics. When K is characteristic, $\gamma_K(\mathbf{P}, \mathbf{Q}) = 0$ iff $\mathbf{P} = \mathbf{Q}$, $\forall \mathbf{P}$, $\mathbf{Q} \in \mathcal{M}_+^1(\mathbb{R}^p)$ (Sejdinovic et al. 2013). As a result, the following theorem is trivial but important.

Theorem 2.4. *Let K be a characteristic kernel on $\mathbb{R}^p$, and $\mathbf{P} \equiv \mathbf{P}_{\mathbf{X}|\mathbf{Y}}$, $\mathbf{Q} \equiv \mathbf{P}_{\mathbf{X}} \in \mathcal{M}_+^1(\mathbb{R}^p) \cap \mathcal{M}_K^{\frac{1}{2}}(\mathbb{R}^p)$. Then $\mathcal{H}_K^2(\mathbf{X}|\mathbf{Y}) = 0$ if and only if $\mathbf{X}$ and $\mathbf{Y}$ are independent.*

Note that $\mathcal{H}_K^2(\mathbf{X}|\mathbf{X}) = E_{\mathbf{X}}K(\mathbf{X},\mathbf{X}) - E_{\mathbf{X},\mathbf{X}'}K(\mathbf{X},\mathbf{X}')$. A statistic similar to correlation can be then defined as $\rho_K(\mathbf{X}|\mathbf{Y}) \equiv \dfrac{\mathcal{H}_K^2(\mathbf{X}|\mathbf{Y})}{\mathcal{H}_K^2(\mathbf{X}|\mathbf{X})}$.

Theorem 2.5. *Suppose $E_{\mathbf{X}}k(\mathbf{X},\mathbf{X}) < \infty$. The following properties hold:*

1. $0 \le \mathcal{H}_k^2(\mathbf{X}|\mathbf{Y}) \le \mathcal{H}_k^2(\mathbf{X}|\mathbf{X}) < \infty$, *and thus* $0 \le \rho_k(\mathbf{X}|\mathbf{Y}) \le 1$.

2. $\rho_k(\mathbf{X}|\mathbf{Y}) = 1$ *if and only if* $\mathbf{X}$ *is a function of* $\mathbf{Y}$.

Another critical property of $\mathcal{H}_k^2(\mathbf{X}|\mathbf{Y})$ involves a decompostion analogous to ANOVA. Recall the feature map of a reproducing kernel K, $\phi(\mathbf{x}) = K(\cdot,\mathbf{x})$, and note that $< \phi(\mathbf{x}), \phi(\mathbf{x}') >_{\mathbb{H}_K} = K(\mathbf{x},\mathbf{x}')$ by Definition 2.2. If we treat $(\phi(\mathbf{X}),\mathbf{Y})$ rather than $(\mathbf{X},\mathbf{Y})$ as an individual, then the kernel embedding of $\mathbf{Q} \equiv \mathbf{P}_{\mathbf{X}}$ into the RKHS $\mathbb{H}_K$, $\mu_K(\mathbf{Q})$, can be viewed as the overall mean of $\phi(\mathbf{X})$, while $\mu_K(\mathbf{P})$ for $\mathbf{P} \equiv \mathbf{P}_{\mathbf{X}|\mathbf{Y}}$ can be viewed as the mean of $\phi(\mathbf{X})$ given $\mathbf{Y}$. Let $\mathcal{W}_K^2(\mathbf{X}|\mathbf{Y}) \equiv E_{\mathbf{X},\mathbf{Y}}\|\phi(\mathbf{X}) - \mu_K(\mathbf{P})\|_{\mathbb{H}_K}$, then we have the following decompostion:

$$\begin{aligned}
\mathcal{H}_k^2(\mathbf{X}|\mathbf{X}) &= E_{\mathbf{X}}\|\phi(\mathbf{X}) - \mu_K(\mathbf{Q})\|_{\mathbb{H}_K} \\
&= E_{\mathbf{Y}}\|\mu_K(\mathbf{P}) - \mu_K(\mathbf{Q})\|_{\mathbb{H}_K} + E_{\mathbf{X},\mathbf{Y}}\|\phi(\mathbf{X}) - \mu_K(\mathbf{P})\|_{\mathbb{H}_K} \\
&= \mathcal{H}_k^2(\mathbf{X}|\mathbf{Y}) + \mathcal{W}_K^2(\mathbf{X}|\mathbf{Y}). \qquad (2.7)
\end{aligned}$$

If Y is categorical, equation (2.7) can be regarded as the population decomposition of total variability into between group dispersion and within group dispersion, which is henceforth referred as the kernel ANOVA population decomposition.

One may consider a more general setting as the following

$$\mathcal{H}_{K,a}^2(\mathbf{X}|\mathbf{Y}) \equiv E_{\mathbf{Y}}\left[a(\mathbf{Y})\gamma_K^2(\mathbf{Y})\right], \qquad (2.8)$$

where $a(\cdot)$ is a given nonnegative weight function. Note that under the same conditions of Theorem 2.4, if $a(\cdot)$ has the same support as the probability density function of $\mathbf{Y}$, then $\mathcal{H}_{K,a}^2(\mathbf{X}|\mathbf{Y}) = 0$ if and only if $\mathbf{X}$ and $\mathbf{Y}$ are independent.

2.3 Empirical Estimators and Asymptotic Properties

In this section, we provide two different approaches to estimate $\mathcal{H}_K^2(\mathbf{X}|\mathbf{Y})$: a slicing method that can be applied to both a categorical or a continuous $\mathbf{Y}$, and a kernel regression estimation that is intended for a continuous $\mathbf{Y}$ only. When $\mathbf{Y}$ is categorical or sliced, our method extends the typical ANOVA to a kernel ANOVA that is able to test a more general hypothesis of equal distributions among groups. While if $\mathbf{Y}$ is continuous, our test, based on the kernel regression estimator of generalized ECHSIC, provides an alternative to HSIC.

2.3.1 Slicing on Y: A Kernel ANOVA

Estimating $\mathcal{H}_K^2(\mathbf{X}|\mathbf{Y})$ is straightforward when $\mathbf{Y}$ is categorical. If $\mathbf{Y}$ is continuous, we can make it discrete by slicing.

Let $(\mathbf{X}_t, \mathbf{Y}_t)$, $t = 1, ..., n$ be a random sample of $(\mathbf{X}, \mathbf{Y})$. Assume that $\mathbf{Y}$ has L levels $\{\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, ..., \mathbf{y}^{(L)}\}$ with probability $\{p_1, p_2, ..., p_L\}$ and each level has n_l observations $(\mathbf{X}_t^{(l)}, \mathbf{y}^{(l)})$, $t = 1, ..., n_l$, $l = 1, ..., L$. The empirical generalized ECHSIC is defined as

$$\mathcal{H}_{K,n}^2(\mathbf{X}|\mathbf{Y}) \equiv \frac{1}{n}\sum_{l=1}^{L}\frac{1}{n_l}\sum_{i,j=1}^{n_l} K(\mathbf{X}_i^{(l)}, \mathbf{X}_j^{(l)}) - \frac{1}{n^2}\sum_{i,j=1}^{n} K(\mathbf{X}_i, \mathbf{X}_j) \tag{2.9}$$

$$= \frac{1}{n^2}\text{trace}(\mathbf{KL}), \tag{2.10}$$

where $\mathbf{K}$ is the $n \times n$ Gram matrix with entries $K_{ij} \equiv K(\mathbf{X}_i, \mathbf{X}_j)$, $\mathbf{L}$ is a $n \times n$ matrix with entries $L_{ij} \equiv \sum_{l=1}^{L} \frac{n}{n_l} I\{\mathbf{Y}_i = \mathbf{y}^{(l)}\} I\{\mathbf{Y}_j = \mathbf{y}^{(l)}\} - 1$ and $I\{\cdot\}$ is the indicator function.

Theorem 2.6 (Consistency)**.** *Assuming that $E_{\mathbf{X}} K(\mathbf{X}, \mathbf{X}) < \infty$, we have*

$$\mathcal{H}_{K,n}^2(\mathbf{X}|\mathbf{Y}) \xrightarrow{P} \mathcal{H}_K^2(\mathbf{X}|\mathbf{Y}).$$

Theorem 2.7 (Null Distribution)**.** *Under H_0, if $E_{\mathbf{X}}K(\mathbf{X},\mathbf{X}) < \infty$, then*

$$n\mathcal{H}^2_{K,n}(\mathbf{X}|\mathbf{Y}) \xrightarrow{d} (L-1)\mathcal{H}^2_K(\mathbf{X}|\mathbf{X}) \sum_{i=1}^{\infty} \lambda_i Z_i^2,$$

where $Z_i \overset{i.i.d}{\sim} N(0,1)$ and λ_i are positive constants with $\sum_{i=1}^{\infty} \lambda_i = 1$.

A natural consistent estimator of $\mathcal{H}^2_K(\mathbf{X}|\mathbf{X})$ is given by

$$\mathcal{H}^2_{K,n}(\mathbf{X}|\mathbf{X}) \equiv \frac{1}{n}\sum_{i=1}^{n} K(\mathbf{X}_i,\mathbf{X}_i) - \frac{1}{n^2}\sum_{i,j=1}^{n} K(\mathbf{X}_i,\mathbf{X}_j).$$

Define $\rho_{K,n}(\mathbf{X}|\mathbf{Y}) \equiv \dfrac{\mathcal{H}^2_{K,n}(\mathbf{X}|\mathbf{Y})}{\mathcal{H}^2_{K,n}(\mathbf{X}|\mathbf{X})}$, then we have the following results.

Corollary 2.1. *Assuming that $E_{\mathbf{X}}K(\mathbf{X},\mathbf{X}) < \infty$,*

1. *under H_0, $n\rho_{K,n}(\mathbf{X}|\mathbf{Y}) \xrightarrow{d} (L-1)\sum_{i=1}^{\infty} \lambda_i Z_i^2$, where $Z_i \overset{i.i.d}{\sim} N(0,1)$ and λ_i are positive constants with $\sum_{i=1}^{\infty} \lambda_i = 1$;*
2. *under H_1, then $n\rho_{K,n}(\mathbf{X}|\mathbf{Y}) \xrightarrow{P} \infty$.*

Recall that we introduced a kernel ANOVA population decomposition in the previous section. We now formulate an empirical kernel ANOVA test for equal distributions among groups, where $n\mathcal{H}^2_{K,n}(\mathbf{X}|\mathbf{Y})$ has a nice interpretation as the treatment sum of squares. Considering a random sample $(\phi(\mathbf{X}_t), \mathbf{Y}_t)$, $t = 1, ..., n$ in RKHS $\mathbb{H}_K$, where ϕ denotes the feature map of kernel K, we have the total sum of squares (SST) as

$$SST \equiv \sum_{i=1}^{n} \left\| \phi(\mathbf{X}_i) - \frac{1}{n}\sum_{j=1}^{n} \phi(\mathbf{X}_j) \right\|^2_{\mathbb{H}_k}.$$

Then SST can be decomposed into treatment sum of squares ($SSTr$) and error sum of squares (SSE), i.e. $SST = SSTr + SSE$, where

$$SSTr \equiv \sum_{l=1}^{L} n_l \left\| \frac{1}{n_l}\sum_{i=1}^{n_l} \phi(\mathbf{X}_i^{(l)}) - \frac{1}{n}\sum_{j=1}^{n} \phi(\mathbf{X}_j) \right\|^2_{\mathbb{H}_k}$$

and

$$SSE \equiv \sum_{l=1}^{L} \sum_{i=1}^{n_l} \left\| \phi(\mathbf{X}_i^{(l)}) - \frac{1}{n_l} \sum_{j=1}^{n_l} \phi(\mathbf{X}_j^{(l)}) \right\|_{\mathbb{H}_k}^2 .$$

After some algebra, we can show that, in fact, $SST = n\mathcal{H}_{K,n}^2(\mathbf{X}|\mathbf{X})$ and $SSTr = n\mathcal{H}_{K,n}^2(\mathbf{X}|\mathbf{Y})$. As a consequence, we can propose a test statistic analogous to the F-statistic in ANOVA, namely

$$\begin{aligned} \mathcal{F}_{K,n}(\mathbf{X}|\mathbf{Y}) &\equiv \frac{SSTr/(L-1)}{SSE/(n-L)} \\ &= \frac{\mathcal{H}_{K,n}^2(\mathbf{X}|\mathbf{Y})/(L-1)}{(\mathcal{H}_{K,n}^2(\mathbf{X}|\mathbf{X}) - \mathcal{H}_{K,n}^2(\mathbf{X}|\mathbf{Y}))/(n-L)}. \end{aligned}$$

Under H_0, $\mathcal{F}_{K,n}(\mathbf{X}|\mathbf{Y}) \xrightarrow{d} Q$, where $Q = \sum_{i=1}^{\infty} \lambda_i Z_i^2$, $Z_i \overset{iid}{\sim} N(0,1)$ and $E(Q) = 1$. Székely and Bakirov (2003) proved that

$$P\left(Q \geq \left(\Phi^{-1}(1-\alpha_0/2)\right)^2\right) \leq \alpha_0,$$

for $\alpha_0 \leq 0.215$, where $\Phi(\cdot)$ is the standard normal c.d.f.

2.3.2 Kernel Regression Estimation

In this section, we construct a more sophisticated estimator via kernel estimation for a continuous $\mathbf{Y}$. We start from an alternative formula of $\gamma_K^2(\mathbf{y})$. By formula (2.5),

$$\begin{aligned} \gamma_K^2(\mathbf{y}) &= E\left[K(\mathbf{X}_1, \mathbf{X}_2)|\mathbf{Y}_1 = \mathbf{y}, \mathbf{Y}_2 = \mathbf{y}\right] - E\left[K(\mathbf{X}_1, \mathbf{X}_3)|\mathbf{Y}_1 = \mathbf{y}\right] \\ &\quad -E\left[K(\mathbf{X}_2, \mathbf{X}_4)|\mathbf{Y}_2 = \mathbf{y}\right] + EK(\mathbf{X}_3, \mathbf{X}_4) \\ &= E\left[d_{1234}|\mathbf{Y}_1 = \mathbf{y}, \mathbf{Y}_2 = \mathbf{y}\right], \end{aligned}$$

where $d_{1234} \equiv K_{12} - K_{13} - K_{24} + K_{34}$ and $K_{ts} \equiv K(\mathbf{X}_t, \mathbf{X}_s)$.

For a kernel function $G : \mathbb{R}^q \to \mathbb{R}$ and a bandwidth $h \equiv h(n)$, define $G_h(\mathbf{y}) \equiv h^{-q}G(\mathbf{y}/h)$. The Nadaraya-Watson kernel estimator of the conditional expectation $\gamma_K^2(\mathbf{Y}_{t_1})$ is given by

$$\gamma_{K,n}^2(\mathbf{Y}_{t_1}) = \frac{\frac{1}{n^4}\sum_{t_2t_3t_4t_5} G_{t_1t_2}G_{t_1t_3}d_{t_2t_3t_4t_5}}{\widehat{f}_h^2(\mathbf{Y}_{t_1})},$$

where $G_{ts} \equiv G_h(\mathbf{Y}_t - \mathbf{Y}_s)$ and $\widehat{f}_h(\mathbf{Y}_{t_1}) \equiv \frac{1}{n}\sum_{s=1}^{n} G_{t_1 s}$. A natural estimator immediately follows as

$$\Gamma_{K,G,n}(\mathbf{X}|\mathbf{Y}) \equiv \frac{1}{n^5} \sum_{t_1 t_2 t_3 t_4 t_5} \frac{G_{t_1 t_2} G_{t_1 t_3} d_{t_2 t_3 t_4 t_5}}{\widehat{f}_h^2(\mathbf{Y}_{t_1})}. \tag{2.11}$$

Note that the smoothing kernel G applied on $\mathbf{Y}$ is different from the reproducing kernel K applied on $\mathbf{X}$. We have different requirements on choosing these two kernels, although we use Gaussian for both in our simulations later.

It is known that kernel estimate suffers from random denominator issues, but it can be alleviated by different strategies. Intuitively, we may either assume that the density of $\mathbf{Y}$, $f(\mathbf{y})$, is bounded below by some positive number or impose a trimming function $a_\epsilon(\mathbf{y}) = I\{f(\mathbf{y}) > \epsilon\}$, where $\epsilon > 0$. Another option is to apply a proper weight function $a(\cdot)$ on $\mathbf{Y}$ for the same purpose of dealing with the possible large bias near 0. We adopt the latter approach as in Su and White (2007) and Wang et al. (2015) to deal with the weighted measure $\mathcal{H}^2_{K,a}(\mathbf{X}|\mathbf{Y})$. Consider the following estimator

$$\Gamma_{K,G,a,n}(\mathbf{X}|\mathbf{Y}) \equiv \frac{1}{n^5} \sum_{t_1 t_2 t_3 t_4 t_5} \frac{G_{t_1 t_2} G_{t_1 t_3} d_{t_2 t_3 t_4 t_5} a(\mathbf{Y}_{t_1})}{\widehat{f}_h^2(\mathbf{Y}_{t_1})}. \tag{2.12}$$

We choose $a(\mathbf{Y}_t) = f_h^2(\mathbf{Y}_t)$ ($\hat{a}(\mathbf{Y}_t) = \widehat{f}_h^2(\mathbf{Y}_t)$) so that the denominator in (2.12) vanishes and there is no need for any additional assumption or trimming function. Eventually, we obtain the following statistic (some subscripts are omitted without ambiguity for simplicity)

$$\Gamma_n^V(\mathbf{X}|\mathbf{Y}) \equiv \frac{1}{n^5} \sum_{t_1 t_2 t_3 t_4 t_5} G_{t_1 t_2} G_{t_1 t_3} d_{t_2 t_3 t_4 t_5} \tag{2.13}$$

$$= \frac{1}{n^3} \text{trace}(\mathbf{KHGGH}), \tag{2.14}$$

where $\mathbf{G}$ is a $n \times n$ matrix with entries G_{ij}, $\mathbf{H} \equiv \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T$, $\mathbf{I}$ is the identity matrix and $\mathbf{1}$ is an n vector of ones. A corresponding bias-adjusted statistic is given by

$$\Gamma_n^U(\mathbf{X}|\mathbf{Y}) \equiv \frac{1}{n_5} \sum_{t_1 \neq t_2 \neq t_3 \neq t_4 \neq t_5} G_{t_1 t_2} G_{t_1 t_3} d_{t_2 t_3 t_4 t_5}, \tag{2.15}$$

where $n_5 = n(n-1)(n-2)(n-3)(n-4)$.

Developing the asymptotic properties of $\Gamma_n^U(\mathbf{X}|\mathbf{Y})$ requires some regularity conditions, which are common in literature for kernel estimation (Su and White 2007; Wang et al. 2015). Let $\mathcal{F}_\mu^\alpha$ ($\mu > 0$ and $\alpha > 0$) be the class of functions $f : \mathbb{R}^q \to \mathbb{R}$ satisfying: f is $(m-1)$-times partially differentiable, for $m - 1 < \mu \leq m$; for some $\rho > 0$, $\sup_{\mathbf{y}' \in \phi_{\mathbf{y}\rho}} |f(\mathbf{y}') - f(\mathbf{y}) - Q_f(\mathbf{y}, \mathbf{y}')|/||\mathbf{y}' - \mathbf{y}||^\mu \leq R_f(\mathbf{y})$ for all $\mathbf{y}$, where $\phi_{\mathbf{y}\rho} \equiv \{\mathbf{y}' : ||\mathbf{y}' - \mathbf{y}|| < \rho\}$; $Q_f = 0$ when $m = 1$; Q_f is a $(m-1)$th degree homogeneous polynomial in $\mathbf{y}' - \mathbf{y}$ with coefficients the partial derivatives of f at $\mathbf{y}$ of orders 1 through $m-1$ when $m > 1$; and f, its partial derivatives of order $m-1$ and less, and R_f, have finite αth moments (Robinson 1988). Our conditions are as follows.

(C1) The kernel G is a product of univariate kernel $g : \mathbb{R} \to \mathbb{R}$ such that $\int_{\mathbb{R}} u^i g(u) du = \delta_{i0}$ $(i = 0, 1, ..., \nu - 1)$, and $g(u) = O((1 + |u|^{\nu+1+\epsilon})^{-1})$ for some $\epsilon > 0$, where $\nu \geq 2$ is an integer and δ_{ij} is Kronecker's delta.

(C2) $h^q \to 0$ and $nh^q \to \infty$ as $n \to \infty$.

(C3) The marginal density of $\mathbf{Y}$, $f(\mathbf{y}), \in \mathcal{F}_\nu^\infty$.

(C4) Let $m(\mathbf{y}) \equiv E_{\mathbf{X}_\mathbf{y}} K(\mathbf{X}, \mathbf{X})$, then $m(\mathbf{y}) \in \mathcal{F}_\nu^\infty$.

Condition (C1) characterizes the class of νth order kernel function and it implies $\int u^\nu g(u) du < \infty$. Condition (C2) requires the bandwidth to be chosen appropriately according to n. Conditions (C3) and (C4) imposes smoothness and moment conditions on the marginal and conditional distribution. Similar constraints are used in Su and White (2007).

Theorem 2.8 (Consistency)**.** *Assume that conditions (C1)-(C4) hold and suppose* $E_{\mathbf{X}} K^2(\mathbf{X}, \mathbf{X}) < \infty$*, then*

$$\Gamma_n^U(\mathbf{X}|\mathbf{Y}) \xrightarrow{P} E_{\mathbf{Y}} \left[\gamma_K^2(\mathbf{Y}) f^2(\mathbf{Y})\right].$$

Theorem 2.9 (Asymptotic Null Distribution)**.** *Assume that conditions (C1)-(C3) hold and* $E_{\mathbf{X}}K^2(\mathbf{X},\mathbf{X}) < \infty$*. Under* H_0*, we have*

$$nh^{q/2}\Gamma_n^U(\mathbf{X}|\mathbf{Y}) \xrightarrow{d} N(0, 2\sigma^2),$$

where $\sigma^2 = C^q \left[EK_{12}^2 - 2EK_{12}K_{13} + E^2K_{12}\right] Ef^3(\mathbf{Y})$ *and* $C = \int_{\mathbb{R}}[\int_{\mathbb{R}} g(\mu+\nu)g(\mu)d\mu]^2 dv$.

2.4 An Algorithm via Permutation Procedure

Nonparametric tests that rely on the asymptotic results may have poor power in finite samples (Su and White 2007). An alternative is to use permutation approach (Efron and Tibshirani 1994; Davison and Hinkley 1997) and it has been proved to be successful in area that is related to our measure, such as DCOV tests (Székely, Rizzo and Bakirov 2007; Székely and Rizzo 2009) and DISCO tests (Rizzo and Székely 2010). Davison and Hinkley (1997) suggested that at least 99 and at most 999 random permutations should be sufficient practically. We illustrate the permutation procedure using test statistic $\mathcal{H}_{K,n}^2(\mathbf{X}|\mathbf{Y})$ as follows:

1. Compute $\mathcal{H}_{K,n}^2(\mathbf{X}|\mathbf{Y})$ using formula (2.9) or (2.10) for the data;

2. For each replicate $b = 1, ..., B$, generate a random permutation π^b and compute the statistic of permuted sample $(\mathbf{X}_k, \mathbf{Y}_{\pi^b(k)})$, $k = 1, ..., n$, denoted by $\mathcal{H}_{K,n}^2(\mathbf{X}|\mathbf{Y}_b)$;

3. Compute the empirical p-value by

$$\widehat{p} = \frac{1 + \sum_{b=1}^{B} I\left\{\mathcal{H}_{K,n}^2(\mathbf{X}|\mathbf{Y}_b) \geq \mathcal{H}_{K,n}^2(\mathbf{X}|\mathbf{Y})\right\}}{B+1},$$

   where $I\{\cdot\}$ is the indicator function.

2.5 An Extension to Conditional version

Although we used the conditional characteristic function to develop ECHSIC at the beginning, the measure is still a marginal test statistic. However, we can simply

extend ECHSIC to a conditional independence measure for testing $\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z}$ based on the same idea.

Let $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ be three random vectors in $\mathbb{R}^{p_1}$, $\mathbb{R}^{p_2}$ and $\mathbb{R}^q$, respectively. For a given translation-invariant RKHS kernel K, we consider the following discrepancy between two characteristic functions $\varphi_{\mathbf{X}|\mathbf{Y},\mathbf{Z}}$ and $\varphi_{\mathbf{X}|\mathbf{Z}}$:

$$\widetilde{\psi}_K^2(\mathbf{Y},\mathbf{Z}) \equiv \int |\varphi_{\mathbf{X}|\mathbf{Y},\mathbf{Z}}(\mathbf{u}) - \varphi_{\mathbf{X}|\mathbf{Z}}(\mathbf{u})|^2 \omega(\mathbf{u}) d\mathbf{u},$$

where ω is a finite nonnegative Borel measure on $\mathbb{R}^{p_1}$ that induces K. Then we define the expected conditional Hilbert-Schmidt conditional independence criterion (ECHSCIC) as $\mathcal{I}_K^2(\mathbf{X}|\mathbf{Y};\mathbf{Z}) \equiv E_{(\mathbf{Y},\mathbf{Z})}\widetilde{\psi}_K^2(\mathbf{Y},\mathbf{Z})$. After some algebra, we can show that

$$\begin{aligned}\mathcal{I}_K^2(\mathbf{X}|\mathbf{Y};\mathbf{Z}) &= E_{(\mathbf{Y},\mathbf{Z})}E_{\mathbf{X}_{(\mathbf{Y},\mathbf{Z})},\mathbf{X}'_{(\mathbf{Y},\mathbf{Z})}}K(\mathbf{X}-\mathbf{X}') - E_{\mathbf{Z}}E_{\mathbf{X}_{\mathbf{Z}},\mathbf{X}'_{\mathbf{Z}}}K(\mathbf{X}-\mathbf{X}') \\ &= \mathcal{I}_K^2(\mathbf{X}|(\mathbf{Y},\mathbf{Z})) - \mathcal{I}_K^2(\mathbf{X}|\mathbf{Z}).\end{aligned}$$

That is, ECHSCIC can be written as the difference between two marginal indices. We can also develop a generalized ECHSCIC via MMD, which will result in a more general index as follows:

$$\mathcal{H}_K^2(\mathbf{X}|\mathbf{Y};\mathbf{Z}) \equiv \mathcal{H}_K^2(\mathbf{X}|(\mathbf{Y},\mathbf{Z})) - \mathcal{H}_K^2(\mathbf{X}|\mathbf{Z}).$$

Then we can easily estimate $\mathcal{H}_K^2(\mathbf{X}|\mathbf{Y};\mathbf{Z})$ by

$$\mathcal{H}_{K,n}^2(\mathbf{X}|\mathbf{Y};\mathbf{Z}) \equiv \mathcal{H}_{K,n}^2(\mathbf{X}|(\mathbf{Y},\mathbf{Z})) - \mathcal{H}_{K,n}^2(\mathbf{X}|\mathbf{Z}).$$

We omit the asymptotic properties of $\mathcal{H}_{K,n}^2(\mathbf{X}|\mathbf{Y};\mathbf{Z})$ as well as the algorithm here as they follow straightforward from Sections 3 and 4.

Note that Su and White (2007) proposed a conditional independence measure, denoted as Γ, that is also based on the difference between $\varphi_{\mathbf{X}|\mathbf{Y},\mathbf{Z}}$ and $\varphi_{\mathbf{X}|\mathbf{Z}}$. However, their index and ours are distinct in terms of how to quantify the discrepancy. Su and White (2007) first take a Fourier transformation of the difference $\varphi_{\mathbf{X}|\mathbf{Y},\mathbf{Z}} - \varphi_{\mathbf{X}|\mathbf{Z}}$

(adding an extra parameter), then compute the square norm, while we directly measure the distance between the two characteristic functions, resulting in a simpler formula. In addition, while our index has a clear and intuitive interpretation of kernel ANOVA and RKHS, the counterpart explanation for Γ is not clear. Along the line of ECHSIC, we can estimate ECHSCIC through two paths. One is the slicing method intending for categorical $\mathbf{Y}$ and $\mathbf{Z}$, which is not considered in Su and White (2007). For continuous $\mathbf{Y}$ and $\mathbf{Z}$, we apply the kernel regression estimation approach as in Su and White (2007) as well as Wang et al. (2015) mainly for estimating the conditional expectation terms. Again, the smoothing kernel we apply on $(\mathbf{Y}, \mathbf{Z})$ or $\mathbf{Z}$ is different from the reproducing kernel we apply on $\mathbf{X}$, while Su and White (2007) only use the smoothing kernel.

2.6 Numerical Studies

In this section we provide empirical examples of independence tests using ECHSIC and power comparisons with other existing tests, in particular, the HSIC, I^{NOCCO} (Fukumizu et al. 2008), DCOV and DISCO (or equivalently, ECD). All the tests are implemented using the permutation procedure presented in section 4.

Basic settings are as follows unless otherwise specified. For ECHSIC, HSIC and I^{NOCCO}, Gaussian kernel, which is a translation-invariant characteristic kernel, is applied with parameter σ setting to the heuristic median pairwise distances of the data (Gretton et al. 2008), although more sophisticated methods are available (Fukumizu et al. 2009, Gretton et al. 2012b). For the kernel regression estimation of a continuous $\mathbf{Y}$, we assume $a(\cdot) \equiv 1$ and use Gaussian kernel for G in formula (2.11). The bandwidth h is set to $1.06\tilde{\sigma}n^{-\frac{1}{5}}$ as Silverman (1986) suggested, where n is the sample size and $\tilde{\sigma}$ is estimated by sample standard deviation. One may also use cross-validation, test graph method and other techniques to choose the smoothing parameter. $B = [200 + 5000/n]$ permutation replicates are carried out in each test

(Székely, Rizzo and Bakirov 2007). ϵ_n of I^{NOCCO} is set to be 10^{-6}. Empirical power or Type-I error rate is computed as the proportion of significant tests on 10,000 random samples at significance level of 0.1.

Example 2.1. This example is to examine the Type-I error rates, similarly to Example 1 in Székely, Rizzo and Bakirov (2007). Set $\mathbf{X} \in \mathbb{R}^5$ and $\mathbf{Y} \in \mathbb{R}$ to be independent. In model (a), $\mathbf{X} \sim N(\mathbf{0}, I_5)$, $\mathbf{Y} \sim N(0, 1)$. In model (b)-(d), we repeat the same scheme except that the marginals of $\mathbf{X}$ and $\mathbf{Y}$ are $t(\nu)$, $\nu = 1, 2, 3$. For our slicing method and DISCO, the number of slices is set to 5. Results in **Table 2.1** indicate that the empirical Type-I error rates of all the methods are under reasonable control.

Table 2.1. Example 2.1: empirical Type-I error rates

Model	n	Method					
		ECHSIC (kernel)	HSIC	I^{NOCCO}	DCOV	ECHSIC (slicing)	DISCO
(a)	25	0.1017	0.1018	0.0999	0.1005	0.1009	0.1012
	50	0.1007	0.1047	0.0966	0.1021	0.1065	0.1045
	100	0.0984	0.0993	0.1006	0.0963	0.0959	0.0944
(b)	25	0.0988	0.0989	0.0948	0.0976	0.1000	0.0995
	50	0.0942	0.0967	0.1014	0.1000	0.0978	0.0961
	100	0.0953	0.0991	0.0981	0.1022	0.1001	0.1020
(c)	25	0.1027	0.1018	0.0999	0.0968	0.1017	0.0996
	50	0.1000	0.1005	0.0922	0.0977	0.0974	0.0973
	100	0.1036	0.1019	0.0959	0.1026	0.1025	0.1014
(d)	25	0.1008	0.0952	0.0962	0.1031	0.1018	0.0988
	50	0.0977	0.1004	0.0990	0.0961	0.1019	0.1042
	100	0.1019	0.0998	0.1000	0.1064	0.1024	0.0978

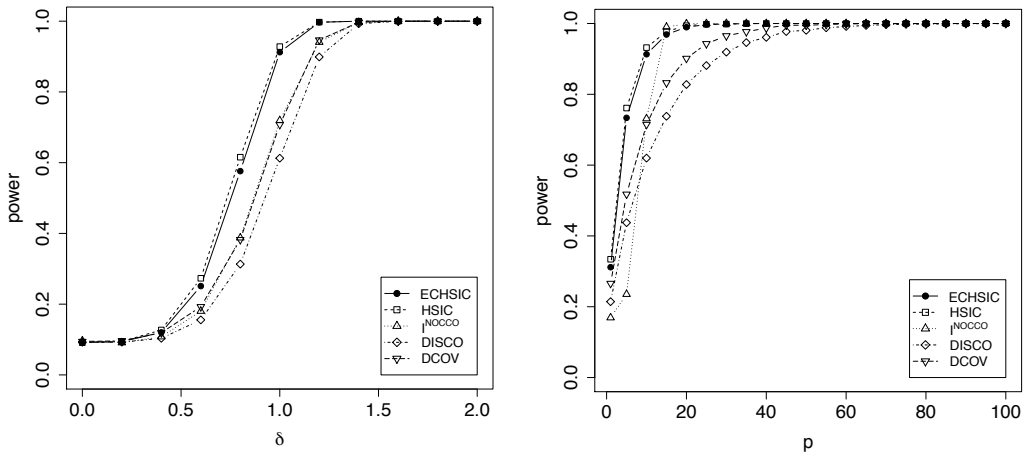
Example 2.2. This example is to examine the performance of ECHSIC when one of the variables is categorical. The setting imitates Example 3 in Rizzo and Székely (2010), a four group balanced design with common sample size 30. Data are generated from distributions with identical independent marginals. Group 1-3 all have central t(4) distributions as marginals. Group 4 has a mixture distribution of two equal-weighted noncentral t(4) with noncentrality parameter δ and $-\delta$, respectively. We treat group indicator as response naturally. Monte Carlo power comparison of our

method with others are summarized in **Figure 2.1**. We use 199 permutations in each test.

In **Figure 2.1a**, power curves with respect to noncentrality parameter δ are presented with dimension fixed at $p = 10$. Each method roughly achieves the nominal significance level 0.1 under the null hypothesis (δ=0). ECHSIC test is generally more powerful than I^{NOCCO}, DCOV and DISCO but slightly less preferred than HSIC. In **Figure 2.1b**, noncentrality parameter is fixed at $\delta = 1$ and power varies with dimension p. We notice that I^{NOCCO} is less capable of detecting dependence than ECHSIC and HSIC when p is small, although it outperforms all the other methods when p gets to 20.

We then replace group indicator 1-4 by 1, 8, 0.2 and 2.5. **Figure 2.1c** and **2.1d** show that the power of HSIC and DCOV decreases dramatically while the performance of our method and DISCO remain the same. I^{NOCCO} is also robust to this variation. When $\mathbf{Y}$ is nominal and its values are not meaningful, ECHSIC tests fix the issue of symmetric measures like HSIC and DCOV because they are only subject to the cohorts of the data but not the values of $\mathbf{Y}$, as seen in (2.9).

Example 2.3. This example aims to show the effect of different kernels on the performance of ECHSIC and HSIC. We imitate the Genome-Wide Association Studies (GWAS) example in Cui, Li and Zhong (2015). In GWAS, we typically have genetic data containing a large number of single-nucleotide polymorphisms (SNPs). The SNPs are categorical predictors with three classes, denoted by {AA, Aa, aa}. We adopt a simple model with only two SNPs and denote Z_{ij} as the indicators of the dominant effect of the jth SNP for ith subject. Z_{ij} is generated in the following way $Z_{ij} = \begin{cases} 1, & \text{if } X_{ij} < q_1 \\ 0, & \text{if } q_1 \le X_{ij} < q_3 \\ -1, & \text{if } X_{ij} \ge q_3 \end{cases}$, where $\mathbf{X}_i = (X_{i1}, X_{i2}) \sim N(0, \Sigma)$,



(a) $p = 10$ (b) $\delta = 1$

(c) $p = 10$ (d) $\delta = 1$

Figure 2.1. Example 2.2: empirical power

$\Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$, and q_1 and q_3 are first and third quartiles of a standard normal distribution, respectively. Then we generate the response by $Y = \beta_1 Z_1 + \beta_2 Z_2 + \epsilon$, where $\beta_j = (-1)^U (a + |Z|)$ for $j = 1, 2$, $a = 2 \log n / \sqrt{n}$, $U \sim$Bernoulli(0.4) and $Z \sim N(0, 1)$. The error term $\epsilon \sim t(1)$, which is largely heavy-tailed. Monte Carlo power comparison of ECHSIC and HSIC with different kernels (Gaussian kernel and kernels induced by semi-metric $\rho(\mathbf{x}, \mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\|_p^\alpha$ with $\alpha = 1/2, 1, 3/2$) are summarized in **Figure 2.2**, assuming a significance level 0.05. In general, ECHSIC is more powerful and less sensitive to the choice of kernel than HSIC.

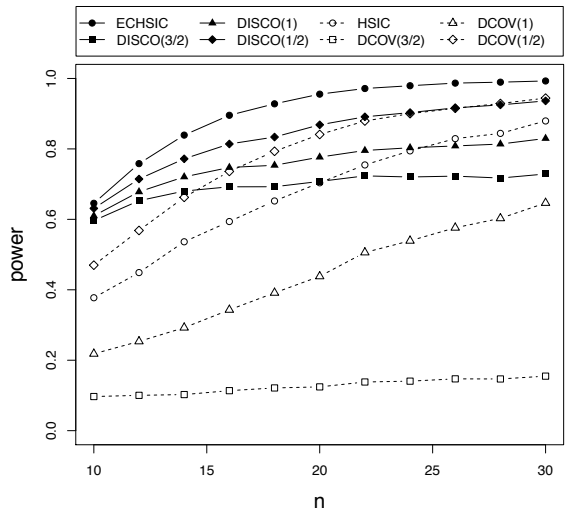


Figure 2.2. Example 2.3: empirical power

Example 2.4. This example is to investigate the effect of number of slices on the performance of the ECHSIC slicing method and DISCO. The Saviotti aircraft data contain six variables of aircraft designs in the twentieth century (Bowman and Azzalini 1997). Two variables, wing span (m) and speed (km/h) for the 230 designs of the third (of three) periods are considered here. As discussed in Example 3 of Székely and Rizzo (2009), the nonlinear relation between speed and wing span is quite evident from the scatter plot and contour plot. Our goal is to test the independence of log(speed) and log(span).

We slice on log(span) to apply ECHSIC slicing method and DISCO. Results are listed in **Table 2.2** with respect to different number of slices. Although our method is not very sensitive to the number of slices, we suggest that each slice should have at least 5 and at most $n/2$ data points.

Example 2.5. This example is to examine the performance of ECHSIC when both variables are univariate continuous. Consider Example 2 in Székely and Rizzo (2009):

Table 2.2. Example 2.4: p-values of ECHSIC and DISCO tests

# of slices	2	5	10	23	46	115
ECHSIC	0.001	0.001	0.001	0.001	0.001	0.001
DISCO	0.005	0.006	0.001	0.002	0.002	0.004

$(X, Y) = (X, \varphi(X))$, where X is standard normal and $\varphi(\cdot)$ is the probability density function of standard normal. The relation between X and Y is deterministic but not monotone. Monte Carlo power comparisons are shown in **Figure 2.3** for varied sample size n. To apply the slicing method and DISCO, we use 2, 3 and 4 slices when $n = 10, 15$ and 20, respectively. While for n greater than 20, we use 5 slices. **Figure 2.3** reveals that the ECHSIC test with the kernel regression estimator has decent performance against the alternative, even with very small sample size. In addition, although slicing on $\mathbf{Y}$ is less preferred in continuous case, we note that the ECHSIC slicing method is still better than DISCO.

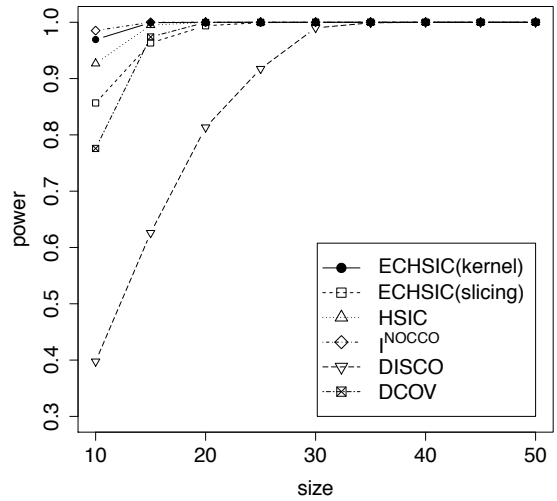


Figure 2.3. Example 2.5: empirical power

Example 2.6. This example (Székely, Rizzo and Bakirov 2007) is to examine the power of ECHSIC when both variables are multivariate continuous. Suppose that $\mathbf{X}$ follows standard multivariate normal with dimension $p = 5$, and $Y_{kj} = X_{kj}\epsilon_{kj}$, $j = 1, ..., p$, where ϵ_{kj} are independent standard normal variables and independent

of X. For multivariate continuous $\mathbf{Y}$, existing slicing techniques in other areas, for example in sufficient dimension reduction such as Zhu, Zhu and Feng (2010), Li, Wen and Zhu (2008) and Cook and Zhang (2014) can be very helpful. However, the kernel regression estimator is still more applicable and accurate. Thus, we only compare with HSIC and DCOV. **Figure 2.4** indicates that ECHSIC with the kernel regression estimator works the best.

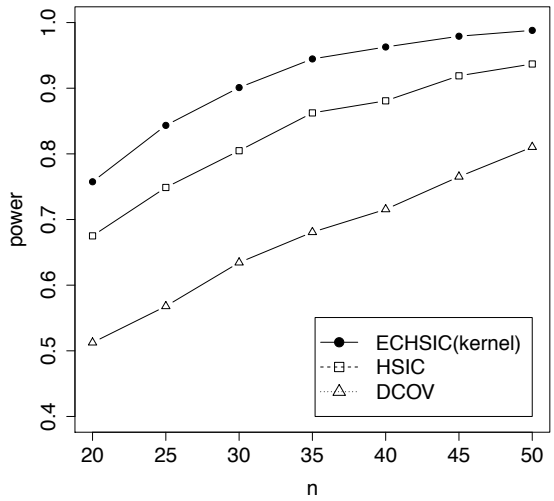


Figure 2.4. Example 2.6: empirical power

Example 2.7. This example evaluates the performance of ECHSIC in regression setups. Two models are generated: (A) $Y = (\boldsymbol{\beta}^T\mathbf{X})^2 + \epsilon$; (B) $Y = 0.2(\boldsymbol{\beta}^T\mathbf{X})^2\epsilon$. Let $\boldsymbol{\beta} = (1,1,1,1,1,0,0,0,0,0)^T$ and $\epsilon \sim N(0,1)$. Predictors are generated based on the following schemes: part (1), standard normal predictors $\mathbf{X} \sim N(\mathbf{0}, \mathbf{I}_{10})$; part (2), non-normal predictors; part (3), discrete predictors. We report the power for sample size $n = 10, 20$ and 50, respectively. To apply the slicing method and DISCO, we slice Y into 2, 4, and 5 levels for n=10, 20 and 50, respectively.

Model A. This is the first model in Sheng and Yin (2013), which has a nonlinear structure in the regression mean function. Predictors for part (2) and part (3) are generated as follows: part (2), $X_1 \sim N(-8,4), X_2 \sim F(4,10), X_3 \sim \chi^2(5), X_4 \sim$

$t(15), X_5 \sim t(3), X_i \sim N(0,1), i = 6, ..., 10$; part (3), $X_i \sim Poisson(1), i = 1, ..., 5, X_i \sim N(0,1), i = 6, ..., 10$.

Model B. This is the third model in Sheng and Yin (2013), which has a nonlinear structure in the regression variance function. Predictors for part (2) and part (3) are generated as follows: part (2), $X_1 \sim N(-8, 4), X_2 \sim t(5), X_3 \sim Gamma(9, 0.5), X_4 \sim F(5, 12), X_5 \sim \chi^2(13), X_i \sim N(0,1), i = 6, ..., 10$; part (3), $X_i \sim Poisson(1), i = 1, ..., 5, X_i \sim N(0,1), i = 6, ..., 10$.

Results in **Table 2.3** indicate that ECHSIC with the kernel regression estimator has the best power in all the models except A(3), in which case ECHSIC is second but very close to the best, DCOV.

Table 2.3. Example 2.7: empirical power

Model	n	Method					
		ECHSIC (kernel)	HSIC	I^{NOCCO}	DCOV	ECHSIC (slicing)	DISCO
A(1)	10	0.2433	0.1775	0.1218	0.2013	0.1220	0.1169
	20	0.3852	0.2290	0.1149	0.2738	0.1462	0.1221
	50	0.6595	0.3480	0.1069	0.4006	0.2115	0.1599
A(2)	10	0.5690	0.4935	0.1760	0.5001	0.3310	0.2401
	20	0.8857	0.8008	0.1268	0.7768	0.5810	0.4019
	50	0.9989	0.9962	0.1037	0.9935	0.9621	0.7810
A(3)	10	0.6528	0.5877	0.1422	0.6945	0.4778	0.4355
	20	0.9433	0.9104	0.1203	0.9783	0.7363	0.7148
	50	1.0000	1.0000	0.1042	1.0000	0.9986	0.9984
B(1)	10	0.2193	0.1786	0.1224	0.2001	0.1062	0.1052
	20	0.3357	0.2173	0.1221	0.2578	0.1168	0.1110
	50	0.5790	0.3054	0.1171	0.3459	0.1459	0.1235
B(2)	10	0.3504	0.3087	0.1275	0.3059	0.1293	0.1376
	20	0.5870	0.5174	0.1171	0.5131	0.3036	0.3415
	50	0.9084	0.8740	0.1002	0.8816	0.6784	0.7645
B(3)	10	0.3548	0.3286	0.1431	0.3042	0.1260	0.1254
	20	0.5723	0.5304	0.1280	0.4739	0.2869	0.2757
	50	0.9131	0.9124	0.1014	0.8362	0.6859	0.6815

Example 2.8. This example is to elaborate the use of the weight function $a(\cdot)$ in the kernel regression estimation of ECHSIC. Although we use $a(\cdot) \equiv 1$ in all the previous examples, our method is actually sensitive to the choice of the weight function, especially when extreme values exist. This is logically, similar to Ordinary Least Squares (OLS) vs Weight Least Squares (WLS). Below we provide an example where the use of a weight function is appropriate. Suppose $Y = \frac{1}{|X|} + \epsilon$, where $X \sim Unif(-3, 3)$

and $\epsilon \sim N(0,1)$. A histogram of $\widehat{f}_h^2(Y_t)$ values is provided in **Figure 2.5**, which shows a heavy tail near 0. Then we compare the power of four methods - our method with $a(\mathbf{Y}_t) \equiv f_h^2(\mathbf{Y}_t)$, our method with $a(\cdot) \equiv 1$, HSIC and DCOV based on 1,000 Monte Carlo simulations.

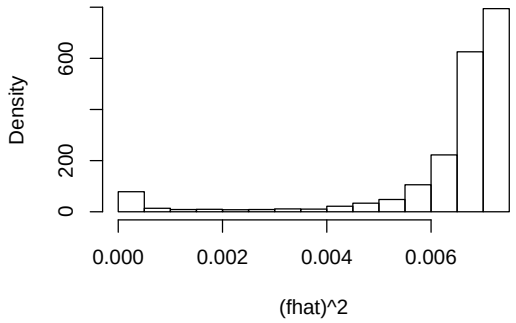


Figure 2.5. Example 2.8: histogram of $\widehat{f}_h^2(Y_t)$

As we can see from **Table 2.4**, the use of the weight function can improve the performance of our method when there are extremely small $\widehat{f}_h^2(\mathbf{Y}_t)$ values. We suggest to check the distribution of $\widehat{f}_h^2(\mathbf{Y}_t)$ before choosing the weight function, which however, could be somewhat subjective.

Table 2.4. Example 2.8: empirical power

n	ECHSIC $a(\mathbf{Y}_t) \equiv f_h^2(\mathbf{Y}_t)$	ECHSIC $a(\cdot) \equiv 1$	HSIC	DCOV
20	0.430	0.225	0.414	0.190
35	0.690	0.357	0.675	0.408
50	0.845	0.443	0.827	0.631

Example 2.9. This example is to compare ECHSCIC with the measure Γ of Su and White (2007), in terms of the type I error rate and the power. Two time series models (Su and White 2007) are generated: (A) $Y_t = 0.3Y_{t-1} + \epsilon_t$, where $\epsilon_t \overset{iid}{\sim} N(0,1)$; (B) $Y_t = e^{-Y_{t-1}^2} + |0.1Y_{t-2}(16 - Y_{t-2})|\epsilon_t$, where $\{\epsilon_t\}$ are i.i.d. sum of 10 uniformly independently distributed random variables each over the range $[-1/7, 1/7]$. We test the null hypothesis $H_0 : f(Y_t|Y_{t-1}, Y_{t-2}) = f(Y_t|Y_{t-1})$, that is, Y_{t-2} has no explanatory

power for Y_t, which is true for Model A but not for Model B. The results based on 200 Monte Carlo simulations are listed in **Table 2.5**, which shows that our measure has a reasonable type I error rate and is more powerful than Γ.

Table 2.5. Example 2.9: empirical type I error rate and power

n	α	ECHSCIC	Γ
100	0.05	0.055	0.050
	0.10	0.100	0.095
200	0.05	0.035	0.070
	0.10	0.060	0.115

(a) Model A, type I error rate

n	α	ECHSCIC	Γ
100	0.05	0.240	0.160
	0.10	0.365	0.240
200	0.05	0.490	0.210
	0.10	0.595	0.385

(b) Model B, power

Example 2.10. This data focuses on comparing our kernel ANOVA with the typical ANOVA and DCOV. The leaf dataset contains a collection of shape and texture features extracted from digital images of leaf specimens originating from a total of 30 different plant species (`http://archive.ics.uci.edu/ml/datasets/Leaf`, Silva, Marcal and da Silva 2013, Silva 2013). Specifically, the relation between elongation and species is studied. The typical ANOVA and the kernel ANOVA decompositions are listed in **Table 2.6**. Both tests indicate that elongation is a significant aspect to distinct different leaf species. However, a residual plot of the fitted ANOVA model in **Figure 2.6** reveals nonconstant variance of the elongation measurements across the species.

Table 2.6. Example 2.10: ANOVA and Kernel ANOVA

		ANOVA				Kernel ANOVA			
Source	Df	Sum	Mean	F	p-value	Sum	Mean	$\mathcal{F}$	p-value
Species	29	11.9107	0.4107	120.46	<0.001	445.1185	15.3489	42.1277	0.001
Error	310	1.0569	0.0034			112.9461	0.3643		
Total	339	12.9676				558.0646			

Therefore, we further examine the assumption of ANOVA by testing the dependence between ANOVA residuals and species. Our method is compared with DCOV. Our kernel ANOVA test in **Table 2.7** also suggests a violation of constant variance. From

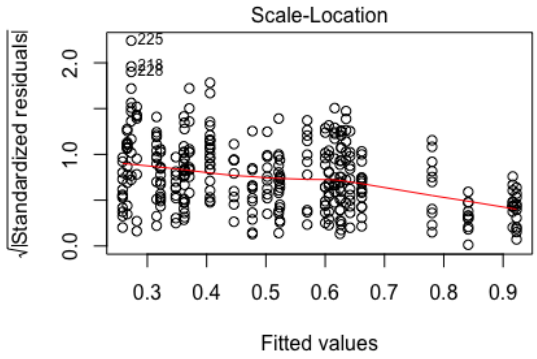


Figure 2.6. Example 2.10: analysis of ANOVA residuals

the above two kernel ANOVA tests, we can conclude that distributions of elongation are different between species in the second moment or higher. As for the DCOV test on ANOVA residuals, since species is a categorical variable, we consider both the original coding as well as a dummy coding (Cui, Li and Zhong 2015), but neither of them detects the heteroscedasticity. Our kernel ANOVA method is more powerful than DCOV.

Table 2.7. Example 2.10: kernel ANOVA and DCOV test on analysis of ANOVA residuals

Kernel ANOVA					
Source	Df	Sum	Mean	$\mathcal{F}$	p-value
Species	29	315.8016	10.8897	1.6937	0.001
Error	310	1993.1540	6.4295		
Total	339	2308.9550			
DCOV					
$nV^2 = 0.0489$, p-value $= 0.326$					
DCOV (dummy coding)					
$nV^2 = 0.0932$, p-value $= 0.12$					

2.7 Discussion

In this chapter, we proposed ECHSIC as a flexible and powerful measure for testing independence between two random vectors, which is especially useful when one of them is categorical. We provided two empirical estimators for the new measure and their associated asymptotic properties. Similar asymptotic results on non-iid samples may also be obtained by using U-statistic (Lee 1990) and those of Su and

White (2007). Another direction of investigating asymptotic distributions is to let the dimension tend to infinity. Székely and Rizzo (2013) indicates that the sample DCOR tends to 1 as the dimension goes to infinity even when $\mathbf{X}$ and $\mathbf{Y}$ are independent. Therefore, they propose a modified DCOR statistic and under independence the distribution of a transformation of the statistic converges to a t-distribution as the dimension tends to infinity. Dueck et al. (2014) studies the limiting theorems of an affinely invariant version of DCOR assuming normal distributions. We can certainly follow these work to develop asymptotic results for our measure when the dimension tends to infinity. Another possible topic for further study is on optimizing over kernels and parameters. However, this is a very challenging problem, although discussion can be found in literature (Fukumizu et al. 2009, Gretton et al. 2012b).

Chapter 3 Sufficient Variable Selection via Expected Conditional Hilbert-Schmidt Independence Criterion

3.1 Introduction

Modern technology allows ultrahigh dimensional data collection at low cost in diverse fields of scientific research. Although regularization methods such as LASSO (Tibshirani 1996), elastic net (Zou and Hastie 2005), Dantzig selector (Candes and Tao 2007) and many others can deal with cases where the number of predictors exceeds the sample size, they may not perform well for ultrahigh dimensional data due to computational cost, statistical accuracy and the stability of algorithms (Fan and Lv 2008). Motivated by these concerns, Fan and Lv (2008) first introduced the concept of sure screening and proposed the sure independence screening (SIS) method, which overcomes the large p small n issue for linear models. However, specifying a correct model for ultrahigh dimensional data can be challenging and their screening procedures may fail in the presence of model mis-specification. Subsequently, Li, Zhong and Zhu (2012) improved SIS using distance correlation (DCOR) instead of Pearson correlation to obtain a model-free procedure. Similar work (Balasubramanian, Sriperumbudur and Lebanon 2013) has also been done by applying a more general measure, Hilbert-Schmidt independence criterion (HSIC). SIS procedures focusing on categorical response models include Kolmogorov sure screening filter (Mai and Zou 2013, 2015) for binary/multi-class classification, a pairwise sure screening procedure for multi-class linear discriminant analysis (Pan, Wang and Li 2016) and MV-SIS based on empirical conditional distribution functions for discriminant analysis with a diverging number of classes (Cui, Li and Zhong 2015).

All the aforementioned feature screening approaches collect only marginal information between the predictors and the response variable, which can result in several

potential issues (Fan and Lv 2008, Zhu et al. 2011). First, some unimportant predictors that are highly correlated with the important predictors can have higher priority for being selected than other important predictors that are relatively weakly related to the response. Second, significant predictors that are uncorrelated (or even independent) but jointly correlated with the response cannot be picked up. Third, the issue of collinearity between predictors intensifies the difficulty. Fan and Lv (2008) proposed an iterative algorithm to overcome these difficulties, which indeed has nice empirical performance. However, the algorithm could be computational demanding and its theoretical justification remains unclear.

Yin and Hilafu (2015) made a formal definition of sufficient variable selection (SVS), where they employed the idea of sufficient dimension reduction (SDR) as a bridge to tackle large p small n problems in variable selection. Their work enlightens an alternative way to optimize SIS and related feature screening procedures by taking both marginal and conditional information into consideration. In fact, the second issue we mentioned above will especially be addressed by their idea without an iterative procedure. Yang, Yin and Zhang (2019) and Yuan and Yin (2017) then developed SVS algorithms based on the paths proposed Yin and Hilafu (2015) using different measures. The former deals with continuous responses while the latter mainly focuses on categorical responses, but both assume that all the predictors are continuous.

SIS and SVS rely heavily on independence measures. Although different measures have been explored in variable selection for various types of response variables, the predictors are mostly restricted to be all continuous. In this chapter, we propose a SVS method based on the independence measure, expected conditional Hilbert-Schmidt independence criterion (ECHSIC), and its extension, expected conditional Hilbert-Schmidt conditional independence criterion (ECHSCIC) that we developed in Chapter 1. Our approach is model-free in a large p small n setting and its sure screening property held under general conditions. We use two paths (Yin and Hilafu 2015)

to improve marginal selection procedures by incorporating information of conditional dependence in extra steps of our algorithm. While sharing advantages with existing SIS and SVS methods, our method inherits the power of the two measures in detecting important variables. More importantly, we can handle either continuous or discrete response with mixed-type predictors. This is achieved because ECHSIC and ECHSCIC can measure independence between variables of different types in a comparable way. Compared to Li, Zhong and Zhu (2012), (Balasubramanian, Sriperumbudur and Lebanon 2013) and Yang, Yin and Zhang (2019), which use DCOR/HSIC that are not appropriate for discrete variables, our method is more legitimate and powerful when we have a discrete response or discrete predictors. In contrast to Yuan and Yin (2017), the measures we adopt are more general and we can cope with a continuous response without slicing.

The rest of this chapter is organized as follows. In section 2, we first review the two independence measures, ECHSIC and ECHSCIC, and the concept of sufficient variable selection. Section 3 introduces a SIS algorithm based on ECHSIC and studies its sure screening property. In Section 4, we develop a novel method to achieve SVS using ECHSIC and ECHSCIC, and establish related theoretical results. We numerically demonstrate the advantages of our method across a variety of settings in Section 5, followed by a short discussion in Section 6. All proofs are deferred in the appendix.

3.2 Preliminaries

In this section, we review marginal independence measure ECHSIC and its extention ECHSCIC for conditional independence test, as well as the concept of sufficient variable selection. The goal of this chapter is to assemble the idea of sufficient variable selection with ECHSIC and ECHSCIC.

3.2.1 ECHSIC

Let $\mathbf{X}$ and $\mathbf{Y}$ be two random variables on $\mathbb{R}^{p_1}$ and $\mathbb{R}^{p_2}$, respectively. The ECHSIC for testing $\mathbf{X} \perp\!\!\!\perp \mathbf{Y}$ is defined as the maximum mean discrepancy (MMD, Gretton et al. 2012a) between the conditional distribution of $\mathbf{X}|\mathbf{Y}$ and the marginal distribution of $\mathbf{X}$, or equivalently,

$$\mathcal{H}_K^2(\mathbf{X}|\mathbf{Y}) \equiv E_{\mathbf{Y}} E_{\mathbf{X}_\mathbf{Y}, \mathbf{X}'_\mathbf{Y}} K(\mathbf{X}, \mathbf{X}') - E_{\mathbf{X},\mathbf{X}'} K(\mathbf{X}, \mathbf{X}')$$

for a characteristic (Fukumizu et al. 2009) positive definite kernel K, where $\mathbf{X}'$ is an i.i.d. copy of $\mathbf{X}$. In Chapter 1, we also introduced a correlation measure, $\rho_K(\mathbf{X}|\mathbf{Y}) \equiv \dfrac{\mathcal{H}_K^2(\mathbf{X}|\mathbf{Y})}{\mathcal{H}_K^2(\mathbf{X}|\mathbf{X})}$. We showed that $0 \leq \rho_K(\mathbf{X}|\mathbf{Y}) \leq 1$, where $\rho_K(\mathbf{X}|\mathbf{Y}) = 0$ if and only if $\mathbf{X}$ and $\mathbf{Y}$ are independent and $\rho_K(\mathbf{X}|\mathbf{Y}) = 1$ if and only if $\mathbf{X}$ is a function of $\mathbf{Y}$. Note that ECHSIC is a parallel framework based on Reproducing Kernel Hilbert Space (RKHS) theory to DCOV/HSIC.

If $\mathbf{Y}$ is finite discrete, given a sample $(\mathbf{X}_t, \mathbf{Y}_t)$, $t = 1, ..., n$, a natural estimator of $\mathcal{H}_K^2(\mathbf{X}|\mathbf{Y})$ is

$$\mathcal{H}_{K,n}^2(\mathbf{X}|\mathbf{Y}) \equiv \frac{1}{n} \sum_{l=1}^{L} \frac{1}{n_l} \sum_{i,j=1}^{n_l} K(\mathbf{X}_i^{(l)}, \mathbf{X}_j^{(l)}) - \frac{1}{n^2} \sum_{i,j=1}^{n} K(\mathbf{X}_i, \mathbf{X}_j), \tag{3.1}$$

where $\mathbf{Y}$ has L levels $\{\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, ..., \mathbf{y}^{(L)}\}$ and within each level we have n_l observations $(\mathbf{X}_t^{(l)}, \mathbf{y}^{(l)})$, $t = 1, ..., n_l$ for $l = 1, ..., L$. This estimator is more appropriate and powerful than DCOV/HSIC when $\mathbf{Y}$ is discrete and leads to a kernel ANOVA test.

If $\mathbf{Y}$ is continuous, one can either slice on $\mathbf{Y}$ to apply the above estimator or use an alternative kernel regression approach. The kernel estimator with a selected smoothing kernel $G : \mathbb{R}^q \to \mathbb{R}$ and a bandwidth $h \equiv h(n)$ is given by

$$\mathcal{H}_{K,G,n}^2(\mathbf{X}|\mathbf{Y}) \equiv \frac{1}{n^3} \sum_{t_1 t_2 t_3 t_4 t_5} \frac{G_{t_1 t_2} G_{t_1 t_3} d_{t_2 t_3 t_4 t_5}}{\sum_{s_1 s_2} G_{t_1 s_1} G_{t_1 s_2}}, \tag{3.2}$$

where $G_{ts} \equiv G_h(\mathbf{Y}_t - \mathbf{Y}_s)$, $G_h(\mathbf{y}) \equiv h^{-p_2} G(\mathbf{y}/h)$, $d_{t_2 t_3 t_4 t_5} \equiv K_{t_2 t_3} - K_{t_2 t_4} - K_{t_3 t_5} + K_{t_4 t_4}$ and $K_{ts} \equiv K(\mathbf{X}_t, \mathbf{X}_s)$. As pointed out in Chapter 1, different approaches can be

employed to address the issue of the random denominator in (3.2). In this chapter, we simply assume that the density of $\mathbf{Y}$ is bounded below by some positive number and hence, no trimming function or weight function needs to be applied.

Note that $\mathcal{H}^2_K(\mathbf{X}|\mathbf{X})$ can be estimated by

$$\mathcal{H}^2_{K,n}(\mathbf{X}|\mathbf{X}) \equiv \frac{1}{n}\sum_{i=1}^{n} K(\mathbf{X}_i, \mathbf{X}_i) - \frac{1}{n^2}\sum_{i,j=1}^{n} K(\mathbf{X}_i, \mathbf{X}_j).$$

Therefore, one can obtain an estimator for $\rho_K(\mathbf{X}|\mathbf{Y})$ as $\rho_{K,n}(\mathbf{X}|\mathbf{Y}) \equiv \dfrac{\mathcal{H}^2_{K,n}(\mathbf{X}|\mathbf{Y})}{\mathcal{H}^2_{K,n}(\mathbf{X}|\mathbf{X})}$ or $\rho_{K,G,n}(\mathbf{X}|\mathbf{Y}) \equiv \dfrac{\mathcal{H}^2_{K,G,n}(\mathbf{X}|\mathbf{Y})}{\mathcal{H}^2_{K,n}(\mathbf{X}|\mathbf{X})}$.

3.2.2 ECHSCIC

Let $\mathbf{Z}$ be a random variables on $\mathbb{R}^q$. ECHSIC can be easily extended to test if $\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z}$. For a given kernel K, the conditional independence measure ECHSCIC is defined as

$$\mathcal{H}^2_K(\mathbf{X}|\mathbf{Y};\mathbf{Z}) \equiv E_{(\mathbf{Y},\mathbf{Z})} E_{\mathbf{X}_{(\mathbf{Y},\mathbf{Z})},\mathbf{X}'_{(\mathbf{Y},\mathbf{Z})}} K(\mathbf{X},\mathbf{X}') - E_{\mathbf{Z}} E_{\mathbf{X}_{\mathbf{Z}},\mathbf{X}'_{\mathbf{Z}}} K(\mathbf{X},\mathbf{X}'), \tag{3.3}$$

where $\mathbf{X}'$ is an i.i.d. copy of $\mathbf{X}$. Similarly, we can introduce a correlation-type index $\rho_K(\mathbf{X}|\mathbf{Y};\mathbf{Z}) \equiv \dfrac{\mathcal{H}^2_K(\mathbf{X}|\mathbf{Y};\mathbf{Z})}{\mathcal{H}^2_K(\mathbf{X}|\mathbf{X};\mathbf{Z})}$ and $0 \le \rho_K(\mathbf{X}|\mathbf{Y};\mathbf{Z}) \le 1$.

If $\mathbf{Z}$ is finite discrete, assume that $\mathbf{Z}$ has L levels and each level contains n_l observations $(\mathbf{X}^{(l)}_t, \mathbf{Y}^{(l)}_t, \mathbf{z}^{(l)})$, $t = 1, ..., n_l$ for $l = 1, ..., L$. $\mathcal{H}^2_K(\mathbf{X}|\mathbf{Y};\mathbf{Z})$ can be estimated straightforwardly by a weight sum of ECHSIC between $\mathbf{X}$ and $\mathbf{Y}$ within each level of $\mathbf{Z}$ as

$$\mathcal{H}^2_{K,n}(\mathbf{X}|\mathbf{Y};\mathbf{Z}) \equiv \sum_{l=1}^{L} \frac{n_l}{n} \mathcal{H}^2_{K,n}(\mathbf{X}^{(l)}|\mathbf{Y}^{(l)}), \tag{3.4}$$

or

$$\mathcal{H}^2_{K,G,n}(\mathbf{X}|\mathbf{Y};\mathbf{Z}) \equiv \sum_{l=1}^{L} \frac{n_l}{n} \mathcal{H}^2_{K,G,n}(\mathbf{X}^{(l)}|\mathbf{Y}^{(l)}), \tag{3.5}$$

depending on whether $\mathbf{Y}$ is discrete or continuous. In addition,

$$\mathcal{H}^2_{K,n}(\mathbf{X}|\mathbf{X};\mathbf{Z}) \equiv \sum_{l=1}^{L} \frac{n_l}{n} \mathcal{H}^2_{K,n}(\mathbf{X}^{(l)}|\mathbf{X}^{(l)}).$$

Correspondingly, $\rho_{K,n}(\mathbf{X}|\mathbf{Y};\mathbf{Z}) \equiv \dfrac{\mathcal{H}^2_{K,n}(\mathbf{X}|\mathbf{Y};\mathbf{Z})}{\mathcal{H}^2_{K,n}(\mathbf{X}|\mathbf{X};\mathbf{Z})}$ if $\mathbf{Y}$ is discrete, $\rho_{K,G,n}(\mathbf{X}|\mathbf{Y};\mathbf{Z}) \equiv \dfrac{\mathcal{H}^2_{K,G,n}(\mathbf{X}|\mathbf{Y};\mathbf{Z})}{\mathcal{H}^2_{K,n}(\mathbf{X}|\mathbf{X};\mathbf{Z})}$ otherwise.

3.2.3 An Introduction to SVS

Yin and Hilafu (2015) propose a new and simple framework for dimension reduction in the large p, small n setting, where they make a formal definition of SVS that is similar to Cook (2004).

Definition 3.1 (SVS, Yin and Hilafu 2015, Definition 1)**.** *Let $\mathbf{X} \in \mathbb{R}^p$. If there is a $p \times q$ matrix B ($q \leq p$), where the columns of B consist of unit vectors of e_j's with jth element 1 and 0 otherwise, such that $\mathbf{Y} \perp\!\!\!\perp \mathbf{X}|B^T\mathbf{X}$, then the column space of B is called the variable selection space. The intersection of all such spaces, if itself satisfies the conditional independence condition above, is called the central variable selection space, denoted by $S^V_{\mathbf{Y}|\mathbf{X}}$, with dimension s.*

Conditions for the existence of $S^V_{\mathbf{Y}|\mathbf{X}}$ are briefly discussed in Yin and Hilafu (2015). Throughout this chapter, we assume that $S^V_{\mathbf{Y}|\mathbf{X}}$ exists and is unique. Let $\mathcal{D} \equiv \{j : e_j \in S^V_{\mathbf{Y}|\mathbf{X}}\}$ and $\bar{\mathcal{D}}$ denote its complement. We write $\mathbf{X}_{\mathcal{D}} \equiv \{X_j : j \in \mathcal{D}\}$ and refer to $\mathbf{X}_{\mathcal{D}}$ as the set of active predictors. Then by definition, $\mathbf{X}_{\mathcal{D}}$ is the smallest subset of the predictors such that

$$\mathbf{Y} \perp\!\!\!\perp \mathbf{X}_{\bar{\mathcal{D}}}|\mathbf{X}_{\mathcal{D}}. \tag{3.6}$$

The goal is to find a reduced set of predictors with a moderate size which can fully cover $\mathbf{X}_{\mathcal{D}}$.

3.3 Sure Independence Screening Using ECHSIC

We start from a SIS algorithm based on ECHSIC in this section before we transit to our SVS method later. Note that SIS is essentially distinct from SVS as SIS looks

for the smallest subset of the predictors, denoted as $\mathbf{X}_{\mathcal{A}}$, such that $\mathbf{Y} \perp\!\!\!\perp \mathbf{X}_{\bar{\mathcal{A}}}$ without conditioning on $\mathbf{X}_{\mathcal{A}}$. Therefore, only marginal information between each predictor and the response is collected in SIS, while the joint effect of predictors on the response is also considered in SVS. In fact, SIS will serve as the first stage of our SVS procedure but extra steps are adopted to ensure that (3.6) is satisfied.

3.3.1 An Algorithm

Let $w_j^M \equiv \rho_K(X_j, \mathbf{Y})$ for $j = 1, ..., p$. We follow SIS and propose the following marginal sure screening procedure:

1. Compute $\widehat{w}_j^M$ for $j = 1, ..., p$ based on **Table 3.1.**;
2. Sort $\widehat{w}_j^M$ in descending order;
3. $\widehat{\mathcal{A}} \equiv \{1 \leq j \leq p : \widehat{w}_j^M$ is among the first d largest of all$\}$.

In practice, we may choose $d = n - 1$ or $d = n/log(n)$.

Table 3.1. Measures for different data types in SIS

$\mathbf{Y}$	X_j	$\widehat{w}_j^M$
Discrete	Continuous	$\rho_{K,n}(X_j\|\mathbf{Y})$
Continuous	Discrete	$\rho_{K,n}(\mathbf{Y}\|X_j)$
	Continuous	$\rho_{K,G,n}(X_j\|\mathbf{Y})$

We refer this procedure to the ECHSIC sure independence screening, or ECH-SIS for short. Notice that when $\mathbf{Y}$ is continuous, $\widehat{w}_j^M$'s are computed differently depending on the type of X_j, but they are still comparable because they are two estimators of the same measure. Therefore, we can rank them together in the second step. If distinct measures are used for continuous predictors and discrete predictors separately, then it is questionable whether they are comparable or not.

3.3.2 Theoretical Properties

Now we study the sure screening property of the proposed ECH-SIS. Define

$$\widehat{\mathcal{A}} \equiv \{1 \leq j \leq p : \widehat{w}_j^M \geq cn^{-\gamma}\},$$

where c and γ are pre-specified threshold values (see condition (C3) below).

Following the literature, conditions below are imposed to facilitate the technical proofs, although they may not be the weakest ones.

(C1) The characteristic positive-definite kernel K and the smoothing kernel G are bounded.

(C2) If $\mathbf{Y}$ is discrete with L levels $\{\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, ..., \mathbf{y}^{(L)}\}$ and $p_l \equiv P(\mathbf{Y} = \mathbf{y}^{(l)})$ for $l = 1, .., L$, then there exists a positive constant c_1 such that $\min_{1 \leq l \leq L} p_l \geq 2c_1 n^{-\tau}$ for $\tau \geq 0$. Furthermore, $L = O(n^{\kappa})$ for $\kappa \geq 0$.

(C3) $\min_{j \in \mathcal{A}} w_j^M \geq 2cn^{-\gamma}$, for some constant $c > 0$ and $0 \leq \gamma < \frac{1}{2}$.

Condition (C1) ensures our measure is finite. The first assumption in condition (C2) requires that the proportion of each level of the response should not be too small. The second assumption allows a diverging number of levels of the response. Condition (C3) assumes that the ECHSIC correlation of active predictors cannot be too small, which is common in the variable selection literature (see condition 3 of Fan and Lv (2008), condition 2 in Li, Zhong and Zhu (2012) and similar others). (C3) reflects the signal strength of individual active predictors, which in turn controls the rate of probability error in selecting the active predictors.

Theorem 3.1 (Sure Screening Property of ECH-SIS)**.** *Under condition (C1) and (C2), for any* $0 \leq \gamma < \frac{1}{2}$, $0 \leq \kappa < \frac{1}{2} - \gamma$ *and* $0 \leq \tau < \frac{1}{4} - \frac{1}{2}\gamma - \frac{1}{2}\kappa$, *there exists a positive constant* $b > 0$ *depending on* c_1 *and* c *such that*

$$P\left(\max_{1 \leq j \leq p} |\widehat{w}_j^M - w_j^M| \geq cn^{-\gamma}\right) \leq O\left(p \exp\{-bn^{1-2\gamma-2\kappa-4\tau} + \kappa \log n\}\right).$$

Furthermore, with condition (C3), we have that

$$P(\max_{j\in\bar{\mathcal{A}}} \widehat{w}_j^M < \min_{j\in\mathcal{A}} \widehat{w}_j^M) \geq 1 - O\left(p\exp\{-bn^{1-2\gamma-2\kappa-4\tau} + \kappa\log n\}\right), \tag{3.7}$$

and

$$P(\mathcal{A} \subset \widehat{\mathcal{A}}) \geq 1 - O\left(s\exp\{-bn^{1-2\gamma-2\kappa-4\tau} + \kappa\log n\}\right), \tag{3.8}$$

where s is the cardinality of $\mathcal{A}$.

The sure screening property holds under mild conditions, allowing categorical responses with a diverging number of levels. If L is fixed, i.e. $\kappa = 0$, then according to Theorem 3.1, we can handle non-polynomial (NP) dimensionality of order $\log(p) = o(n^{1-2\gamma-4\tau})$, that is, if $\log(p) = o(n^{1-2\gamma-4\tau})$, then the probability that ECHSIC ranks active predictors above inactive ones approaches 1 as $n \to \infty$. As a consequence, all truly important predictors can be selected with probability approaching 1 as $n \to \infty$.

3.4 Sufficient Variable Selection Using ECHSIC and ECHSCIC

3.4.1 Methodology

Yin and Hilafu (2015) proposed a framework for dimension reduction and variable selection in the large p small n setting with a sequential implementation. The following proposition is essential.

Proposition 3.1 (Yin and Hilafu 2015, Proposition 1)**.** *Either statement (a) or (b) implies (c) below:*

(a) $\mathbf{X}_1 \perp\!\!\!\perp (\mathbf{X}_2, \mathbf{Y})$*;*

(b) $\mathbf{X}_1 \perp\!\!\!\perp \mathbf{Y}$ *and* $\mathbf{X}_1 \perp\!\!\!\perp \mathbf{X}_2|\mathbf{Y}$*;*

(c) $\mathbf{X}_1 \perp\!\!\!\perp \mathbf{Y}|\mathbf{X}_2$*.*

Assuming that we find a partition of $\mathbf{X}$, $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$, such that statement (c) is true, we can drop $\mathbf{X}_1$ without losing any regression information. Note that for a predictor X_j ($j \in \{1, ..., p\}$), $X_j \in \mathbf{X}_{\bar{\mathcal{D}}}$ iff $X_j \perp\!\!\!\perp \mathbf{Y}|\mathbf{X}_{-j}$, where $\mathbf{X}_{-j} \equiv (X_1, ..., X_{j-1}, X_{j+1}, ..., X_p)$. Therefore, we can exclude variable X_j based on $X_j \perp\!\!\!\perp \mathbf{Y}|\mathbf{X}_{-j}$ for each $j \in \{1, ..., p\}$, which is the condition in statement (c) for a leave-one-out partition of $\mathbf{X}$. Statements (a) and (b) can further optimize this procedure. When $\mathbf{Y}$ is continuous, statement (a) can be simply assessed by a marginal independence measure, which provides us a shortcut to statement (c). When $\mathbf{Y}$ is discrete, the two conditions in statement (b) are easier to verify, so we can take the path from (b) to (c) as (b) also implies (c).

Our bottom line here is that all the truly important variables are included, or equivalently, the excluded ones are truly unimportant. Therefore, it is legitimate to use stronger statements like (a) and (b) rather than (c) itself to exclude variables. Notice that SIS and its family only evaluate the first part of statement (b), which is not sufficient for (c). As a consequence, significant predictors that are uncorrelated (or even independent) but jointly correlated with the response could be left out. In fact, statement (a) and the second part of statement (b) can serve as a remedy to select variables mistakenly eliminated by SIS.

Based on the above discussion, we propose a SVS algorithm built upon two paths (a)→(c) and (b)→(c) for a continuous response and a discrete response, respectively.

3.4.2 An Algorithm

Let $w_j^{S,1} \equiv \rho_K((Y, \mathbf{X}_{-j})|X_j)$ and $w_j^{S,2} \equiv \rho_K(\mathbf{X}_{-j}|X_j; \mathbf{Y})$ for $i = 1, ..., p$. We propose the following 2-stage sufficient variable selection procedure:

1. Compute $\widehat{w}_j^M$ for $j = 1, ..., p$ based on **Table 3.1.**;

2. Sort $\widehat{w}_j^M$ in descending order;

3. $\widehat{\mathcal{D}}_1 \equiv \{1 \le j \le p : \widehat{w}_j^M$ is among the first d_1 largest of all$\}$;

4. Compute $\widehat{w}_j^{S,1}$ if $\mathbf{Y}$ is continuous or $\widehat{w}_j^{S,2}$ if $\mathbf{Y}$ is discrete based on **Table 3.2.**, for $j \in \{1, ..., p\} \setminus \widehat{\mathcal{D}}_1$;

5. Sort $\widehat{w}_j^{S,i}$ in descending order, $i = 1$ or 2;

6. $\widehat{\mathcal{D}}_2 \equiv \{j \in \{1, ..., p\} \setminus \widehat{\mathcal{D}}_1 : \widehat{w}_j^{S,i}$ is among the first d_2 largest of all$\}$, $i = 1$ or 2;

7. $\widehat{\mathcal{D}} = \widehat{\mathcal{D}}_1 \cup \widehat{\mathcal{D}}_2$.

In practice, we may choose $d_1 = [0.9n]$ and $d_2 = n - 1 - d_1$.

Table 3.2. Measures for different data types in SVS

$\mathbf{Y}$	X_j	$\widehat{w}_j^{S,1}$
Continuous	Discrete	$\rho_{K,n}((\mathbf{Y}, \mathbf{X}_{-j})\|X_j)$
	Continuous	$\rho_{K,G,n}((\mathbf{Y}, \mathbf{X}_{-j})\|X_j)$
$\mathbf{Y}$	X_j	$\widehat{w}_j^{S,2}$
Discrete	Discrete	$\rho_{K,n}(\mathbf{X}_{-j}\|X_j; \mathbf{Y})$
	Continuous	$\rho_{K,G,n}(\mathbf{X}_{-j}\|X_j; \mathbf{Y})$

The procedure is referred as the ECHSIC/ECHSCIC-based sufficient variable selection, or ECH-SVS for short. Note that when applying path (a)$\rightarrow$ (c) for continuous response, although statement (a) already implies marginal independence, we still conduct SIS first. The reason is that in practice, marginal relation typically plays an important role and hence, we include a SIS step to secure the ability of ECH-SVS for picking up marginally active variables. Similar to ECH-SIS, we allow both continuous and discrete predictors since their ECHSCIC correlations with the response are commensurate.

3.4.3 Theoretical Properties

We now study the sure screening property of the two ECH-SVS paths. Define

$$\widehat{\mathcal{D}} \equiv \{1 \le j \le p : \widehat{w}_j^{S,1} \ge cn^{-\gamma}\},$$

where c and γ are pre-specified threshold values (see condition (C5) below).

The following conditions are imposed:

(C4) If X_j is discrete with L_j levels $\{x_j^{(1)}, ..., x_j^{(L_j)}\}$ for some j, and $p_l^{(j)} \equiv P(X_j = x_j^{(l)})$ for $l = 1, ..., L_j$, then there exists a positive constant c_1 such that $\min_{j,l} p_l^{(j)} \geq 2c_1 n^{-\tau}$ for $\tau \geq 0$. Furthermore, $\max_j L_j = O(n^{\kappa})$ for $\kappa \geq 0$.

(C5) $\min_{j \in \mathcal{D}} w_j^{S,1} \geq 2cn^{-\gamma}$, for some constant $c > 0$ and $0 \leq \gamma < \frac{1}{2}$.

Theorem 3.2 (Sure Screening Property of ECH-SVS Path 1)**.** *Under condition (C1) and (C4), for any* $0 \leq \gamma < \frac{1}{2}$, $0 \leq \kappa < \frac{1}{2} - \gamma$ *and* $0 \leq \tau < \frac{1}{4} - \frac{1}{2}\gamma - \frac{1}{2}\kappa$, *there exists a positive constant* $b > 0$ *depending on* c_1 *and* c *such that*

$$P\left(\max_{1 \leq j \leq p} |\widehat{w}_j^{S,1} - w_j^{S,1}| \geq cn^{-\gamma}\right) \leq O\left(p \exp\{-bn^{1-2\gamma-2\kappa-4\tau} + \kappa \log n\}\right).$$

Furthermore, with condition (C5), we have that

$$P(\max_{j \in \widehat{\mathcal{D}}} \widehat{w}_j^{S,1} < \min_{j \in \mathcal{D}} \widehat{w}_j^{S,1}) \geq 1 - O\left(p \exp\{-bn^{1-2\gamma-2\kappa-4\tau} + \kappa \log n\}\right), \tag{3.9}$$

and

$$P(\mathcal{D} \subset \widehat{\mathcal{D}}) \geq 1 - O\left(s \exp\{-bn^{1-2\gamma-2\kappa-4\tau} + \kappa \log n\}\right), \tag{3.10}$$

where s *is the cardinality of* $\mathcal{D}$.

Define

$$\widehat{\mathcal{D}}^* \equiv \{1 \leq j \leq p : \widehat{w}_j^{S,2} \geq cn^{-\gamma}\},$$

where c and γ are pre-specified threshold values (see condition (C8) below).

We require several conditions as follows:

(C6) $\mathbf{Y}$ is discrete with L levels $\{\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, ..., \mathbf{y}^{(L)}\}$ and $p_l \equiv P(\mathbf{Y} = \mathbf{y}^{(l)})$ for $l = 1, .., L$. Assume that $L = O(n^{\kappa})$ for $\kappa \geq 0$.

(C7) If X_j is discrete with L_j levels $\{x_j^{(1)}, ..., x_j^{(L_j)}\}$ for some j, and $p_{l_1,l_2}^{(j)} \equiv P(X_j = x_j^{(l_1)} | \mathbf{Y} = \mathbf{y}^{(l_2)})$ for $l_1 = 1, ..., L_j$ and $l_2 = 1, ..., L$, then there exists a positive

constant c_1 such that $\min_{j,l_1,l_2} p^{(j)}_{l_1,l_2} \geq 2c_1 n^{-\tau}$ for $\tau \geq 0$. Furthermore, $\max_j L_j = O(n^{\kappa})$.

(C8) $\min_{j\in\mathcal{D}} w_j^{S,2} \geq 2cn^{-\gamma}$, for some constant $c > 0$ and $0 \leq \gamma < \frac{1}{2}$.

Theorem 3.3 (Sure Screening Property of ECH-SVS Path 2)**.** *Under condition (C1), (C6) and (C7), for any* $0 \leq \gamma < \frac{1}{2}$, $0 \leq \kappa < \frac{1}{2} - \gamma$ *and* $0 \leq \tau < \frac{1}{4} - \frac{1}{2}\gamma - \frac{1}{2}\kappa$, *there exists a positive constant* $b > 0$ *depending on* c_1 *and* c *such that*

$$Pr\left(\max_{1\leq j\leq p} |\widehat{w}_j^{S,2} - w_j^{S,2}| \geq cn^{-\gamma}\right) \leq O\left(p\exp\{-bn^{1-2\gamma-2\kappa-4\tau} + 2\kappa\log n\}\right).$$

Furthermore, with condition (C8), we have that

$$P(\max_{j\in\widehat{\mathcal{D}}} \widehat{w}_j^{S,2} < \min_{j\in\mathcal{D}} \widehat{w}_j^{S,2}) \geq 1 - O\left(p\exp\{-bn^{1-2\gamma-2\kappa-4\tau} + 2\kappa\log n\}\right), \tag{3.11}$$

and

$$P(\mathcal{D} \subset \widehat{\mathcal{D}}^*) \geq 1 - O\left(s\exp\{-bn^{1-2\gamma-2\kappa-4\tau} + 2\kappa\log n\}\right), \tag{3.12}$$

where s *is the cardinality of* $\mathcal{D}$.

The above two theorems show that ECH-SVS can recover $\mathcal{D}$ or the central variable selection space eventually as sample size increases. Notice that $\mathcal{A} \subset \mathcal{D}$, so ECH-SVS captures important variables that are omitted by marginal selection procedures.

3.5 Numerical Studies

For all the numerical studies, if a model size d is given, we report the proportion including a single active predictors X_i, denoted as P_i^s, and the proportion including all active predictors, denoted as P_a (Li, Zhong and Zhu 2012). For SIS results, we also report the median of the minimum model size (MMS) that includes all active predictors, along with a robust standard deviation calculated as RSD=IQR/1.34 (Cui, Li and Zhong 2015). ECH-SIS is mainly compared with MV-SIS (categorical

response only; Cui, Li and Zhong 2015), DC-SIS (Li, Zhong and Zhu 2012) and HR-SIS (Yang, Yin and Zhang 2019). ECH-SVS is mainly compared with DC-SVS (Yang, Yin and Zhang 2019), HR-SVS (Yang, Yin and Zhang 2019) and ECD-SVS (Yuan and Yin 2017). Note that DC-SVS and HR-SVS only handle continuous responses with continuous predictors, while ECD-SVS only allows categorical responses with continuous predictors. Hence, to make them comparable for other cases, we code categorical variables into dummies as in Cui, Li and Zhong (2015) when necessary.

Example 3.1 (Cui, Li and Zhong 2015, Example 3.1)**.** The response Y is generated from two different distributions: (i) balanced, a discrete uniform distribution with L categories where $P(Y = l) = 1/L$ for $l = 1, ..., L$; (ii) unbalanced, the sequence of probabilities $P(Y = l) = 2[1 + (l - 1)/(L - 1)]/3L$ is an arithmetic progression with $\max_{1\leq l\leq L} P(Y = l) = 2\min_{1\leq l\leq L} P(Y = l)$. Given $Y = l$, the predictor $\mathbf{X}$ is generated by letting $\mathbf{X} = \boldsymbol{\mu}_l + \boldsymbol{\epsilon}$, where the mean term $\boldsymbol{\mu}_l = (\mu_{l1}, ..., \mu_{lp})$ is a p-dimensional vector with lth element $\mu_{ll} = 3$ but others are all zero, and $\boldsymbol{\epsilon}$ is a p-dimensional error term. Here, we consider two cases of the error term: (1) $\epsilon_i \sim N(0, 1)$; (2) $\epsilon_i \sim t(2)$ independently for $i = 1, ..., p$. P_i^s and P_a are computed for model size $d = [n/log(n)]$. We examine the efficacy of our marginal sure screening procedure ECH-SIS when the response is categorical, in compare with DC-SIS, HR-SIS, PSIS (Pan, Wang and Li 2016) and MV-SIS. PSIS and PSIS* are implemented as in Cui, Li and Zhong (2015). The results are presented in **Table 3.3** and **Table 3.4** based on 500 simulations.

Both **Table 3.3** and **Table 3.4** indicate that the proposed ECH-SIS is robust and has decent performance, especially when the error term is heavy-tailed and the number of categories increases.

Example 3.2. This is example assembles the numerical studies in Zhu et al. (2011), Example 1 in Li, Zhong and Zhu (2012) and the Example 4.1 in Cui, Li and Zhong (2015), where we simulate a continuous response with a mix of continuous and cat-

Table 3.3. Example 3.1: MMS and accuracy

		Case (1): $\epsilon_{ij} \sim N(0,1)$				Case (1): $\epsilon_{ij} \sim t(2)$			
Pr	Method	MMS	P_1^s	P_2^s	P_a	MMS	P_1^s	P_2^s	P_a
Balanced	DC-SIS	2.0(0.0)	1.00	1.00	1.00	2.0(0.0)	0.99	0.98	0.97
	HR-SIS	2.0(0.0)	1.00	1.00	1.00	2.0(0.0)	1.00	1.00	1.00
	PSIS	2.0(0.0)	1.00	1.00	1.00	2.5(9.1)	0.79	0.88	0.71
	MV-SIS	2.0(0.0)	1.00	1.00	1.00	2.0(0.0)	1.00	0.99	0.99
	ECH-SIS	2.0(0.0)	1.00	1.00	1.00	2.0(0.0)	1.00	1.00	1.00
Unbalanced	DS-SIS	2.0(0.0)	1.00	1.00	1.00	2.0(1.1)	0.95	0.96	0.92
	HR-SIS	2.0(0.0)	1.00	1.00	1.00	2.0(0.0)	1.00	0.99	0.99
	PSIS	2.0(0.0)	1.00	1.00	1.00	5.5(48.8)	0.75	0.75	0.55
	MV-SIS	2.0(0.0)	1.00	1.00	1.00	2.0(0.7)	0.96	0.99	0.95
	ECH-SIS	2.0(0.0)	1.00	1.00	1.00	2.0(0.0)	1.00	0.99	0.99

Table 3.4. Example 3.1: MMS and accuracy

Method	MMS	P_1^s	P_2^s	P_3^s	P_4^s	P_5^s	P_6^s	P_7^s	P_8^s	P_9^s	P_{10}^s	P_a
Balanced, $\epsilon_{ij} \sim N(0,1)$												
DC-SIS	10.0(0.0)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99
HR-SIS	11.0(2.2)	0.99	1.00	0.97	1.00	0.99	0.99	0.99	1.00	0.99	1.00	0.92
PSIS*	10.0(0.0)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
MV-SIS	10.0(0.0)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
ECH-SIS	10.0(0.0)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Balanced, $\epsilon_{ij} \sim t(2)$												
DC-SIS	15.0(21.8)	0.86	0.99	0.99	0.99	0.97	0.98	0.99	0.99	0.99	0.98	0.74
HR-SIS	15.5(20.1)	0.97	0.95	0.97	0.97	0.95	1.00	0.94	0.96	0.99	0.91	0.74
PSIS*	365.2(563.6)	0.73	0.75	0.76	0.73	0.75	0.75	0.75	0.73	0.76	0.79	0.05
MV-SIS	11.0(3.7)	1.00	1.00	1.00	0.99	0.99	1.00	1.00	0.99	0.99	0.99	0.95
ECH-SIS	10.0(0.7)	1.00	0.99	0.99	1.00	0.99	1.00	0.99	1.00	0.98	1.00	0.95
Unbalanced, $\epsilon_{ij} \sim N(0,1)$												
DC-SIS	13.0(14.9)	0.82	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.82
HR-SIS	19.5(37.5)	0.84	0.87	0.97	0.95	0.98	0.99	1.00	1.00	1.00	1.00	0.65
PSIS*	10.0(0.0)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
MV-SIS	10.0(0.0)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
ECH-SIS	10.0(0.0)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Unbalanced, $\epsilon_{ij} \sim t(2)$												
DC-SIS	126.5(248.3)	0.35	0.90	0.93	0.93	0.96	1.00	0.99	1.00	1.00	1.00	0.22
HR-SIS	55.5(119.0)	0.65	0.83	0.88	0.91	0.98	0.98	0.99	1.00	1.00	1.00	0.41
PSIS*	343.5(444.9)	0.68	0.66	0.56	0.58	0.64	0.63	0.60	0.73	0.61	0.67	0.05
MV-SIS	13.0(9.8)	0.93	0.98	0.98	0.98	0.98	1.00	1.00	1.00	1.00	1.00	0.85
ECH-SIS	11.0(7.8)	0.93	0.95	0.97	0.99	0.99	1.00	1.00	1.00	1.00	1.00	0.83

egorical predictors. We generate continuous predictors $\mathbf{X}^{(1)} = (X_1, ..., X_{\frac{p}{2}})$ from a multivariate normal distribution $N(\mathbf{0}, \Sigma_1)$ and $\Sigma_1 = (\sigma_{ij})_{\frac{p}{2}\times\frac{p}{2}}$ with $\sigma_{ii} = 1$, $\sigma_{ij} = 0.4$ if both $i, j \in \mathcal{D}$ or $i, j \in \bar{\mathcal{D}}$, and $\sigma_{ij} = 0.1$ otherwise. We also mimic the behavior of SNPs in Genome-Wide Association Studies (GWAS) to simulate categorical predictors $\mathbf{X}^{(2)}$ in the following way:

$$X_j^{(2)} = \begin{cases} 1, & \text{if } Z_j < Q_1 \\ 0, & \text{if } Q_1 \le Z_j < Q_3, \\ -1, & \text{if } Z_j \ge Q_3 \end{cases}$$

where $X_j^{(2)}$ indicates the dominant effect of the jth SNP (typically denoted by $\{AA, Aa, aa\}$), $j = 1, ..., \frac{p}{2}$, $\mathbf{Z} = (Z_1, ..., Z_{\frac{p}{2}}) \sim N(0, \Sigma_2)$, $\Sigma_2 = (\rho_{ij})_{\frac{p}{2}\times\frac{p}{2}}$ with $\rho_{ij} = 0.8^{|i-j|}$, and Q_1 and Q_3 are the first and third quartiles of a standard normal distribution, respectively. Let $\mathbf{X} = (\mathbf{X}^{(1)}, \mathbf{X}^{(2)})$ and consider the following model

$$Y = 3\beta_1 X_1 X_2 + 2\beta_2 X_{\frac{p}{2}+1} - 2\beta_3 |X_{\frac{p}{2}+20}| + \epsilon,$$

where $\beta_j = (-1)^{\zeta}(a + |\kappa|)$ for $j = 1, 2, 3$, $a = 2\log n/\sqrt{n}$, $\zeta \sim$Bernoulli(0.4), $\kappa \sim N(0, 1)$ and $\epsilon \sim N(0, 1)$. The model contains a interaction term and has a non-linear relation between $X_{\frac{p}{2}+20}$ and Y. We set $d = [n/log(n)]$. The results are reported in **Table 3.5** based on 500 replicates. ECH-SIS has higher accuracy and hence, smaller and more stable MMS to include all active predictors than other procedures.

Table 3.5. Example 3.2: MMS and accuracy

Method	MMS	P_1^s	P_2^s	$P_{\frac{p}{2}+1}^s$	$P_{\frac{p}{2}+20}^s$	P_a
$(n, p) = (100, 500)$						
DC-SIS	18.0(20.2)	1.00	1.00	0.68	1.00	0.68
HR-SIS	15.0(5.2)	0.95	1.00	1.00	1.00	0.95
ECH-SIS	6.0(2.2)	1.00	1.00	1.00	1.00	1.00
$(n, p) = (200, 2000)$						
DC-SIS	5.0(2.2)	0.95	1.00	1.00	1.00	0.95
HR-SIS	5.0(0.9)	1.00	0.95	1.00	1.00	0.95
ECH-SIS	4.0(0.0)	1.00	1.00	1.00	1.00	1.00

Example 3.3. This example is similar to Example 4.2.2 in Fan and lv (2008). We examine the efficacy of our SVS method for a categorical response and mixed-type predictors. We generate correlated continuous predictors $\mathbf{X}^{(1)} = (X_1, ..., X_{p_1})$ from a multivariate normal distribution $N(\mathbf{0}, \Sigma)$ and $\Sigma = (\sigma_{ij})_{p_1\times p_1}$ with $\sigma_{ii} = 1$, $\sigma_{ij} = 0.5$ if both $i, j \neq 5$, and $\sigma_{ij} = \sqrt{0.5}$ otherwise. Discrete predictors are generated as follows: $X_1^{(2)} = Z_1$, where $Z_1 \sim Poisson(2)$, $X_2^{(2)} = Z_1 + Z_2$, where $Z_2 \sim Poisson(1)$, $X_i^{(2)}$ independently follows a Beta-binomial distribution with size 5, success probability $pr = 0.9 - 0.8\frac{i-3}{p_2-3}$ and overdispersion parameter $s = -2 + 8\frac{i-3}{p_2-3}$, for $i = 3, ..., p_2$. Let

$\mathbf{X} = (\mathbf{X}^{(1)}, \mathbf{X}^{(2)})$ and consider the following model

$$Y = I\{5X_3 + 5X_4 - 10\sqrt{0.5}X_5 > 0\} + 2I\{X_{p_1+2} - X_{p_1+1} > 1\},$$

where $I\{\cdot\}$ is an indicator function. The model is designed such that $X_5 \perp\!\!\!\perp Y$ and $X_{p_1+1} \perp\!\!\!\perp Y$. Therefore, marginal sure screening methods cannot pick up the true model except by chance. To calculate P_i^s and P_a, d_1 is set to $[0.95n]$ and d_2 is set to $n - 1 - d_1$. The results are presented in **Table 3.6** based on 500 simulations. As we can expect, all the SIS procedures fail to detect X_5 and X_{p_1+1}. Our SVS procedure significantly improves P_5^s and $P_{p_1+1}^s$ and hence, P_a, while both DC-SVS and HR-SVS have difficulty selecting the discrete predictor X_{p_1+1}.

Table 3.6. Example 3.3: accuracy

Method	P_3^s	P_4^s	P_5^s	$P_{p_1+1}^s$	$P_{p_1+2}^s$	P_a
$(n, p_1, p_2) = (100, 480, 20)$						
DC-SIS	0.97	0.96	0.10	0.33	0.94	0.01
DC-SVS	0.98	0.98	1.00	0.29	0.94	0.28
HR-SIS	0.94	0.93	0.14	0.27	0.95	0.03
HR-SVS	0.95	0.94	1.00	0.25	0.94	0.22
ECH-SIS	0.92	0.94	0.12	0.18	0.89	0.03
ECH-SVS	0.93	0.92	1.00	0.97	1.00	0.84
$(n, p_1, p_2) = (200, 1950, 50)$						
DC-SIS	1.00	1.00	0.01	0.17	1.00	0.00
DC-SVS	1.00	1.00	1.00	0.15	0.99	0.15
HR-SIS	1.00	1.00	0.05	0.14	1.00	0.01
HR-SVS	1.00	1.00	1.00	0.17	1.00	0.13
ECH-SIS	0.99	1.00	0.05	0.08	0.99	0.01
ECH-SVS	1.00	1.00	1.00	0.43	1.00	0.43

Example 3.4. This example is similar to Example 4.2.2 in Fan and lv (2008). We evaluate the performance of our SVS method for a continuous response and mixed-type predictors. We generate continuous predictors $\mathbf{X}^{(1)} = (X_1, ..., X_{p_1})$ from a multivariate normal distribution $N(\mathbf{0}, \Sigma)$ and $\Sigma = (\sigma_{ij})_{p_1 \times p_1}$ with $\sigma_{ii} = 1$, $\sigma_{ij} = 0.3$ if both $i, j \neq 5$, and $\sigma_{ij} = 0.5$ otherwise. Discrete predictors are generated as follows: $X_1^{(2)} = Z_1$, $X_2^{(2)} = Z_1 + Z_2$, $X_i^{(2)} = Z_1 + Z_i$ for $i = 3, ..., p_2$, where $Z_1 \sim Poisson(1)$, $Z_2 \sim Poisson(3)$ and $Z_i \sim Poisson(2)$ for $i = 3, ..., p_2$. Let $\mathbf{X} = (\mathbf{X}^{(1)}, \mathbf{X}^{(2)})$ and

consider the following model

$$Y = 5X_3 + 5X_4 - 5X_5 + e^{X_{p_1+1} - X_{p_1+2}}\epsilon,$$

where $\epsilon \sim N(0,1)$. The model is designed such that $X_5 \perp\!\!\!\perp Y$ and $X_{p_1+1} \perp\!\!\!\perp Y$. The heteroscedastic error term increases the difficulty of variable selection. To calculate P_i^s and P_a, d_1 is set to $[0.95n]$ and d_2 is set to $n-1-d_1$. The results are presented in **Table 3.7** based on 500 simulations. ECH-SVS is much more powerful than DC-SVS and HR-SVS as the latter two barely can detect the important discrete predictors.

Table 3.7. Example 3.4: accuracy

Method	P_3^s	P_4^s	P_5^s	$P_{p_1+1}^s$	$P_{p_1+2}^s$	P_a
$(n, p_1, p_2) = (100, 450, 50)$						
DC-SIS	1.00	1.00	0.15	0.15	0.13	0.00
DC-SVS	1.00	1.00	1.00	0.28	0.13	0.05
HR-SIS	1.00	1.00	0.18	0.18	0.17	0.01
HR-SVS	1.00	1.00	1.00	0.44	0.14	0.07
ECH-SIS	1.00	1.00	0.06	0.54	0.99	0.03
ECH-SVS	1.00	1.00	1.00	0.97	1.00	0.97
$(n, p_1, p_2) = (200, 1800, 200)$						
DC-SIS	1.00	1.00	0.09	0.03	0.09	0.01
DC-SVS	1.00	1.00	1.00	0.38	0.07	0.04
HR-SIS	1.00	1.00	0.06	0.05	0.06	0.00
HR-SVS	1.00	1.00	1.00	0.74	0.04	0.03
ECH-SIS	1.00	1.00	0.01	0.09	0.96	0.00
ECH-SVS	1.00	1.00	1.00	0.98	0.98	0.97

3.6 Discussion

We propose a novel two-stage sufficient variable selection procedure based on a newly developed independence measure. The procedure is model-free and capable of handling different types of data with a simple sequential implementation. In this chapter, the selected model size has to be specified beforehand. To avoid an ad hoc choice of the model size, one may follow an algorithm proposed by Kong, Wang and Wahba (2015) or use independent test, such as permutation or bootstrap test to determine an appropriate number. Another possible improvement over our current procedure

is to incorporate the idea of Balasubramanian, Sriperumbudur and Lebanon (2013), where they consider taking the supremum of HSIC over a family of kernels.

Chapter 4 Sufficient Dimension Reduction via Expected Conditional Hilbert-Schmidt Independence Criterion

4.1 Introduction

Sufficient dimension reduction (SDR) has been a rapidly developed research area which has wide applications in regression, machine learning, genomics, et al., where data of high dimension are common. SDR aims to capture regression information based on the notion of sufficiency, meaning that a set of functions of the predictors contains all the information about the response and no other predictors can provide any additional information. Note that SDR is distinct from sufficient variable selection in the sense that variable selection reduces the number of predictors while dimension reduction downsizes the data to a few linear combinations or nonlinear functions of the predictors.

In general, most approaches for linear SDR can be briefly classified into three categories: inverse regression methods, forward regression methods and joint relation methods. The term "inverse regression" refers to the conditional distribution of $\mathbf{X}|Y$, where $\mathbf{X} \in \mathbb{R}^p$ and $Y \in \mathbb{R}$, which reverses the typical conditional distribution of $Y|\mathbf{X}$ of interest for regression models. Sliced inverse regression (SIR) introduced by Li (1991) is the most well-known SDR method. SIR fails when the regression function is symmetric about 0 as it is based on the first order moment, which motivates the development of second-order methods like sliced average variance estimation (SAVE, Cook and Weisberg 1991). These inverse methods and related others require linearity or constant covariance conditions of the predictors that are difficult to verify in practice. Forward methods including minimum average variance estimator (MAVE, Xia et al. 2002), outer product of gradients (OPG, Xia et al. 2002) and their extensions, relax those conditions. However, they involve high-dimensional smoothing

kernel. Joint relation methods mainly focus on the conditional mean $E(Y|\mathbf{X})$. Representative work include principal hessian direction (PHD; Li 1992), an informational method (Yin and Cook 2005) and a Fourier method (Zhu and Zeng 2006). PHD cannot detect linear trend, while the latter two either use smoothing technique or impose strong conditions on predictors. In recent years, more correlation-based SDR methods have emerged such as approaches relying on likelihood (Cook and Forzani 2009) and distance covariance (Sheng and Yin 2013, 2016).

In this chapter, we propose a new correlation-based SDR method using expected conditional Hilbert-Schmidt independence criterion (ECHSIC) that we introduce in chapter 2. Our method is model-free and inherits the advantages of ECHSIC when dealing with either continuous or discrete responses. In contrast to traditional SDR methods, ours does not impose strong assumptions on the predictors and can exhaustively recover the central subspace.

The rest of the chapter is organized as follows. Section 2 reviews the independence measure ECHSIC and basic concepts of SDR. Section 3 introduces a new SDR method via ECHSIC with an algorithm. Section 4 includes two numerical studies of a single-index model with a continuous response and a multi-index model with a categorical response. Section 5 concludes the chapter with a short discussion. All proofs are delayed in the appendix.

4.2 Preliminaries

4.2.1 ECHSIC

For a given translation-invariant positive kernel K, the ECHSIC for testing $\mathbf{X} \perp\!\!\!\perp Y$ proposed in Chapter 2 is defined as

$$\mathcal{H}_K^2(\mathbf{X}|Y) \equiv \int |\varphi_{\mathbf{X}|Y}(\mathbf{u}) - \varphi_{\mathbf{X}}(\mathbf{u})|^2 \omega(\mathbf{u}) d\mathbf{u},$$

where $\varphi_{\mathbf{X}|Y}$ and $\varphi_{\mathbf{X}}$ are the conditional characteristic function of $\mathbf{X}|Y$ and the marginal characteristic function of $\mathbf{X}$, respectively, and $\omega(\mathbf{u})d\mathbf{u}$ is a finite nonnegative Borel measure corresponding to K such that the condition in Bochner Theorem (Wendland 2004, Theorem 6.6) is satisfied. Equivalently, we have

$$\mathcal{H}_K^2(\mathbf{X}|Y) = E_Y E_{\mathbf{X}_Y,\mathbf{X}'_Y} K(\mathbf{X},\mathbf{X}') - E_{\mathbf{X},\mathbf{X}'} K(\mathbf{X},\mathbf{X}'),$$

where $\mathbf{X}'$ is an i.i.d. copy of $\mathbf{X}$.

If Y is finite discrete, given a sample $(\mathbf{X}_t, Y_t)$, $t = 1, ..., n$, a natural estimator of $\mathcal{H}_K^2(\mathbf{X}|Y)$ is

$$\mathcal{H}_{K,n}^2(\mathbf{X}|Y) \equiv \frac{1}{n}\sum_{l=1}^{L}\frac{1}{n_l}\sum_{i,j=1}^{n_l} K(\mathbf{X}_i^{(l)} - \mathbf{X}_j^{(l)}) - \frac{1}{n^2}\sum_{i,j=1}^{n} K(\mathbf{X}_i, \mathbf{X}_j), \tag{4.1}$$

where Y has L levels $\{y^{(1)}, y^{(2)}, ..., y^{(L)}\}$ and within each level, we have n_l observations $(\mathbf{X}_t^{(l)}, y^{(l)})$, $t = 1, ..., n_l$ for $l = 1, ..., L$. This estimator is more appropriate and powerful than DCOV/HSIC when $\mathbf{Y}$ is discrete and leads to a kernel ANOVA test.

If Y is continuous, one can either slice on Y to apply the above estimator or use an alternative kernel regression approach. The kernel estimator with a selected smoothing kernel $G : \mathbb{R} \to \mathbb{R}$ and a bandwidth $h \equiv h(n)$ is given by

$$\mathcal{H}_{K,G,n}^2(\mathbf{X}|Y) \equiv \frac{1}{n^3}\sum_{t_1t_2t_3t_4t_5} \frac{G_{t_1t_2}G_{t_1t_3}d_{t_2t_3t_4t_5}}{\sum_{s_1s_2} G_{t_1s_1}G_{t_1s_2}}, \tag{4.2}$$

where $G_{ts} \equiv G_h(Y_t - Y_s)$, $G_h(y) \equiv h^{-p_2}G(y/h)$, $d_{t_2t_3t_4t_5} \equiv K_{t_2t_3} - K_{t_2t_4} - K_{t_3t_5} + K_{t_4t_4}$ and $K_{ts} \equiv K(\mathbf{X}_t - \mathbf{X}_s)$. As pointed out in Chapter 2, different approaches can be employed to address the issue of the random denominator in (4.2). In this Chapter, we simply assume that the density of Y is bounded below by some positive number and hence, no trimming function or weight function needs to be applied.

4.2.2 A Review of SDR

Let $\boldsymbol{\beta}$ be a $p \times q$ matrix $(1 \le q \le p)$. If

$$Y \perp\!\!\!\perp \mathbf{X}|\boldsymbol{\beta}^T\mathbf{X}, \tag{4.3}$$

that is, Y depends on $\mathbf{X}$ only through $\boldsymbol{\beta}^T\mathbf{X}$, then we can reduce the dimension of the data from p to q without loss of information. The columns of $\boldsymbol{\beta}$ spans a dimension reduction subspace denoted as $\mathcal{S}(\boldsymbol{\beta})$. If the intersection of all dimension reduction subspace is itself a dimension reduction subspace, then it is called the central subspace (CS, Li 1991, Cook 1994, Cook 1996), denoted by $\mathcal{S}_{Y|\mathbf{X}}$, with the structural dimension d defined as the dimension of $\mathcal{S}_{Y|\mathbf{X}}$. The existence and uniqueness of the central subspace have been established by Cook (1996) and Yin, Li and Cook (2008) under mild conditions. Throughout this chapter, we assume that CS exists and is unique. Our goal is to recover the CS, which includes determining d and finding a basis, so that the dimension is reduced to the most extent while preserving complete regression information.

4.3 A New SDR Method via ECHSIC

In the classical linear models, we find the ordinary least square (OLS) estimator of $\boldsymbol{\beta}$ that minimizes the Euclidean distance between Y and $\boldsymbol{\beta}^T\mathbf{X}$, which is the simplest dimension reduction method. In a more general sense, we can achieve SDR by finding $\boldsymbol{\beta}$ such that $\boldsymbol{\beta}^T\mathbf{X}$ is most related to Y if a model is not pre-specified. Then intuitively, a powerful independence measure or correlation index like ECHSIC can be useful.

Essentially, we identify $\mathcal{S}_{Y|\mathbf{X}}$ by solving

$$\max_{\boldsymbol{\beta}^T\Sigma_{\mathbf{X}}\boldsymbol{\beta}=\mathbf{I}_d} \mathcal{H}_K^2(\boldsymbol{\beta}^T\mathbf{X}|Y), \tag{4.4}$$

with respect of $\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is a $p \times d$ matrix. We claim that the solution is a basis of $S_{\mathbf{Y}|\mathbf{X}}$.

We consider two cases, $d = 1$ (single-index model) and $d > 1$ (multi-index model) assuming d is known.

4.3.1 Single-Index

Let $\Sigma_{\mathbf{X}}$ be the covariance matrix of $\mathbf{X}$, $P_{\boldsymbol{\beta}(\Sigma_{\mathbf{X}})}$ denote the projection onto $\boldsymbol{\beta}$ with respect to the inner product $< \mathbf{a}, \mathbf{b} >= \mathbf{a}^T\Sigma_{\mathbf{X}}\mathbf{b}$, that is, $P_{\boldsymbol{\beta}(\Sigma_{\mathbf{X}})} = \boldsymbol{\beta}(\boldsymbol{\beta}^T\Sigma_{\mathbf{X}}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^T\Sigma_{\mathbf{X}}$ and $Q_{\boldsymbol{\beta}(\Sigma_{\mathbf{X}})} = \mathbf{I} - P_{\boldsymbol{\beta}(\Sigma_{\mathbf{X}})}$. The following proposition ensures that the solution to (4.4) indeed spans the CS.

Proposition 4.1. *Let $\boldsymbol{\eta}$ be a basis of the central subspace $\mathcal{S}_{Y|\mathbf{X}}$ with $\boldsymbol{\eta}^T\Sigma_{\mathbf{X}}\boldsymbol{\eta} = 1$. If $P^T_{\boldsymbol{\eta}(\Sigma_{\mathbf{X}})}\mathbf{X} \perp\!\!\!\perp Q^T_{\boldsymbol{\eta}(\Sigma_{\mathbf{X}})}\mathbf{X}$, then $\mathcal{H}^2_K(\boldsymbol{\beta}^T\mathbf{X}|Y) \leq \mathcal{H}^2_K(\boldsymbol{\eta}^T\mathbf{X}|Y)$ for any $\boldsymbol{\beta} \in \mathbb{R}^p$ with $\boldsymbol{\beta}^T\Sigma_{\mathbf{X}}\boldsymbol{\beta} = 1$. The equality holds if and only if $\mathcal{S}(\boldsymbol{\beta}) = \mathcal{S}(\boldsymbol{\eta})$.*

4.3.2 Multi-Index

We adopt the previous notations and extend the results to multi-index models.

Proposition 4.2. *Let $\boldsymbol{\eta}$ be a basis of the central subspace $\mathcal{S}_{Y|\mathbf{X}}$ and $\boldsymbol{\eta}^T\Sigma_{\mathbf{X}}\boldsymbol{\eta} = \mathbf{I}_d$. Suppose $\boldsymbol{\beta}$ is a $p \times d_1$ matrix with $d_1 \leq d$, $dim(\mathcal{S}(\boldsymbol{\beta})) = d_1$, and $\boldsymbol{\beta}^T\Sigma_{\mathbf{X}}\boldsymbol{\beta} = \mathbf{I}_{d_1}$. Assuming $\mathcal{S}(\boldsymbol{\beta}) \subseteq \mathcal{S}(\boldsymbol{\eta})$, then $\mathcal{H}^2_K(\boldsymbol{\beta}^T\mathbf{X}|Y) \leq \mathcal{H}^2_K(\boldsymbol{\eta}^T\mathbf{X}|Y)$. The equality holds if and only if $\mathcal{S}(\boldsymbol{\beta}) = \mathcal{S}(\boldsymbol{\eta})$.*

Proposition 4.3. *Let $\boldsymbol{\eta}$ be a basis of the central subspace $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ and $\boldsymbol{\eta}^T\Sigma_{\mathbf{X}}\boldsymbol{\eta} = \mathbf{I}_d$. Suppose $\boldsymbol{\beta}$ is a $p \times d_2$ matrix. If $P^T_{\boldsymbol{\eta}(\Sigma_{\mathbf{X}})}\mathbf{X} \perp\!\!\!\perp Q^T_{\boldsymbol{\eta}(\Sigma_{\mathbf{X}})}\mathbf{X}$ and $\mathcal{S}(\boldsymbol{\beta}) \not\subseteq \mathcal{S}(\boldsymbol{\eta})$, then $\mathcal{H}^2_K(\boldsymbol{\beta}^T\mathbf{X}|\mathbf{Y}) < \mathcal{H}^2_K(\boldsymbol{\eta}^T\mathbf{X}|\mathbf{Y})$.*

4.3.3 An algorithm

Naturally, an estimator of $\boldsymbol{\eta}$, denoted by $\boldsymbol{\eta}_n$, is

$$\boldsymbol{\eta}_n \equiv \arg \max_{\boldsymbol{\beta}^T\hat{\Sigma}_{\mathbf{X}}\boldsymbol{\beta}=\mathbf{I}_d} \mathcal{H}^2_n(\boldsymbol{\beta}^T\mathbf{X}|\mathbf{Y}),$$

where $\mathcal{H}^2_n(\boldsymbol{\beta}^T\mathbf{X}|\mathbf{Y})$ can be either (4.1) or (4.2) depending on Y.

We present our algorithm for $d = 1$ first then modify it for $d > 1$ later. Sequential quadratic programming (SQP) adopted in our algorithm is an iterative method

commonly used for constrained nonlinear optimization. We set Lagrangian

$$\mathcal{L}(\boldsymbol{\beta}, \lambda) \equiv \mathcal{H}_n^2(\boldsymbol{\beta}^T\mathbf{X}|\mathbf{Y}) + \lambda(\boldsymbol{\beta}^T\hat{\Sigma}_{\mathbf{X}}\boldsymbol{\beta} - 1)$$

as our objective function, where λ is a Lagrange multiplier, and propose the following algorithm to find $\boldsymbol{\eta}_n$:

1. Randomly generate 1000 standard normal vectors and choose the one that gives the highest ECHSIC to be our initial of SQP.

2. Construct and solve QP sub-problem to obtain a search direction $d_{\boldsymbol{\beta}}$. Given a current iterate $(\boldsymbol{\beta}^{(j)}, \lambda^{(j)})$, the QP sub-problem is defined as

$$\min_{d_{\boldsymbol{\beta}}} - \left(\nabla\mathcal{L}(\boldsymbol{\beta}^{(j)}, \lambda^{(j)})^T d_{\boldsymbol{\beta}} + \frac{1}{2} d_{\boldsymbol{\beta}}^T H^{(k)} d_{\boldsymbol{\beta}}\right)$$

   subject to $\nabla h(\boldsymbol{\beta}^{(j)})^T d_{\boldsymbol{\beta}} + h(\boldsymbol{\beta}^{(j)}) = 0$, where $d_{\boldsymbol{\beta}} = \boldsymbol{\beta} - \boldsymbol{\beta}^{(j)}$, ∇ represents gradient, $H^{(k)}$ is the quasi-Newton approximation of the Hessian of $\mathcal{L}$ at $\boldsymbol{\beta}^{(k)}$, and $h(\boldsymbol{\beta}) = \boldsymbol{\beta}^T\Sigma_{\mathbf{X}}\boldsymbol{\beta} - 1$. Then the solution $\hat{d}_{\boldsymbol{\beta}}$ serves as the search direction to construct a new iterate.

3. Choose the step length α. Then $\boldsymbol{\beta}^{(j+1)} = \boldsymbol{\beta}^{(j)} + \alpha\hat{d}_{\boldsymbol{\beta}}$. Similarly, we update $\lambda^{(j)}$ by computing $\lambda^{(j+1)} = \lambda^{(j)} + \alpha\hat{d}_{\lambda}$.

4. Repeat step 2 and step 3 until $\left|\boldsymbol{\beta}^{(j+1)} - \boldsymbol{\beta}^{(j)}\right|$ is sufficiently small, say less than 10^{-6}.

Gill, Murray and Wright (1981) gives more details of SQP regarding how to solve the QP sub-problem and choose the step length.

The algorithm for $d > 1$ is very similar except for choosing the initial. We can use SIR, SAVE and other SDR methods to find an initial that gives the highest ECHSIC.

4.4 Numerical Studies

In this section, we examine the performance of our method, in compare with other existing approaches including SIR (Li 1991), SAVE (Cook and Weisberg 1991), PHD (Li 1992) and DCOV (Sheng and Yin 2013, 2016). We use $\Delta(\mathcal{S}_{\mathbf{X}|Y}, \widehat{\mathcal{S}}_{\mathbf{X}|Y}) = ||P_{\mathcal{S}_{Y|\mathbf{X}}} - P_{\widehat{\mathcal{S}}_{Y|\mathbf{X}}}||$ (Li, Zha and Chiaromonte 2005) to evaluate the accuracy of CS estimates, where $||\cdot||$ gives the maximum singular values of a matrix, $P_{\mathcal{S}_{Y|\mathbf{X}}}$ and $P_{\widehat{\mathcal{S}}_{Y|\mathbf{X}}}$ are orthogonal projections onto $\mathcal{S}_{Y|\mathbf{X}}$ and its estimate $\widehat{\mathcal{S}}_{\mathbf{X}|Y}$, respectively. A small Δ indicates an accurate estimate of the CS. We report the mean and the standard error of Δ based on 100 replicates.

Example 4.1. This example is the same as model (C) part (1) in Section 4.1 of Sheng and Yin (2013). Let $\boldsymbol{\beta} = (1, 1, 1, 1, 1, 0, 0, 0, 0, 0)^T$ and $\mathbf{X} \sim N(\mathbf{0}, \mathbf{I}_{10})$. Consider model

$$Y = 0.2(\boldsymbol{\beta}^T\mathbf{X})^2\epsilon,$$

where $\epsilon \sim N(0, 1)$. The model has a nonlinear structure in the regression variance function. Simulation results are listed in **Table 4.1**. Since the predictors are normal, SAVE and PHD perform better than SIR. Our method recovers the CS more accurately and stably than all others.

Table 4.1. Example 4.1: estimation accuracy

Method	$n = 100$		$n = 200$	
	Δ	SE	Δ	SE
SIR	0.9351	0.0970	0.9478	0.0734
SAVE	0.4706	0.1716	0.2897	0.0734
PHD	0.5555	0.1475	0.4240	0.1132
DCOV	0.4867	0.3451	0.2177	0.2473
ECHSIC	0.2275	0.0695	0.1659	0.0236

Example 4.2. In this example, we consider a multi-index model with categorical response. Let $\boldsymbol{\beta}_1 = (2, 1, 0, 0, 0, 0, 0, 0, 0, 0)/\sqrt{5}$ and $\boldsymbol{\beta}_2 = (1, -1, 0, 0, 0, 0, 0, 0, 0, 0)$.

Consider a multinomial model

$$Pr(Y=1|\mathbf{X}) = \frac{\exp\{g(\boldsymbol{\beta}_1^T\mathbf{X})\}}{1+\exp\{g(\boldsymbol{\beta}_1^T\mathbf{X})\}+\exp\{\boldsymbol{\beta}_2^T\mathbf{X}\}}$$
$$Pr(Y=2|\mathbf{X}) = \frac{\exp\{\boldsymbol{\beta}_2^T\mathbf{X}\}}{1+\exp\{g(\boldsymbol{\beta}_1^T\mathbf{X})\}+\exp\{\boldsymbol{\beta}_2^T\mathbf{X}\}}$$
$$Pr(Y=3|\mathbf{X}) = \frac{1}{1+\exp\{g(\boldsymbol{\beta}_1^T\mathbf{X})\}+\exp\{\boldsymbol{\beta}_2^T\mathbf{X}\}}$$

where $g(z) = \exp(5z-2)/\{1+\exp(5z-3)\}-1.5$ and $X_j \overset{iid}{\sim} U(-2,2)$ for $j=1,...,10$. Simulation results are presented in **Table 4.2**. We can observe that our method has decent performance comparable to others.

Table 4.2. Example 4.2: estimation accuracy

Method	$n=100$		$n=200$	
	$\overline{\Delta}$	SE	$\overline{\Delta}$	SE
SIR	0.6131	0.1248	0.4806	0.1127
SAVE	0.9565	0.0653	0.9299	0.0900
PHD	0.9618	0.0438	0.9230	0.0954
DCOV	0.6129	0.1252	0.4806	0.1127
ECHSIC	0.6070	0.1291	0.4663	0.1184

4.5 Discussion

In this chapter, we propose a novel linear SDR method via ECHSIC assuming d is known. However, in practice, d is unknown and must be inferred from data. Different approaches can be used to estimating d such as Chi-squared test (Li 1991), permutation test (Cook and Yin 2001) and bootstrap test (Ye and Weiss 2003), which will be investigated in the near future. Furthermore, theoretical properties of the estimator will also be studied later.

Since ECHSIC can be derived using reproducing kernel Hilbert space, which is also the underlying theory for nonlinear SDR, our method can be generalized to deal with non-linearity in the future. In addition, the optimization problem (4.4) can be combined with penalized methods like LASSO (Tibshirani 1996) to achieve variable

selection simultaneously. To handle ultrahigh dimensional data with large p and small n, we can plug in our method to the framework proposed by Yin and Hilafu (2015). These topics are under our future research plan.

Appendices

A. Appendix of Chapter 2

1. **Proof of Theorem 2.2.**

$$\begin{aligned}
\psi_\omega^2(\mathbf{y}) &= \int |\varphi_{\mathbf{X}|\mathbf{Y}=\mathbf{y}}(\mathbf{u}) - \varphi_{\mathbf{X}}(\mathbf{u})|^2 d\omega(\mathbf{u}) \\
&= \int \Big\{ \varphi_{\mathbf{X}|\mathbf{Y}=\mathbf{y}}(\mathbf{u}) \bar{\varphi}_{\mathbf{X}|\mathbf{Y}=\mathbf{y}}(\mathbf{u}) - \varphi_{\mathbf{X}|\mathbf{Y}=\mathbf{y}}(\mathbf{u}) \bar{\varphi}_{\mathbf{X}}(\mathbf{u}) \\
&\quad - \bar{\varphi}_{\mathbf{X}|\mathbf{Y}=\mathbf{y}}(\mathbf{u}) \varphi_{\mathbf{X}}(\mathbf{u}) + \varphi_{\mathbf{X}}(\mathbf{u}) \bar{\varphi}_{\mathbf{X}}(\mathbf{u}) \Big\} d\omega(\mathbf{u}) \\
&= \int \Big\{ E\left[e^{i(\mathbf{X}-\mathbf{X}')^T\mathbf{u}} | \mathbf{Y}=\mathbf{y}, \mathbf{Y}'=\mathbf{y} \right] - E\left[e^{i(\mathbf{X}-\mathbf{X}')^T\mathbf{u}} | \mathbf{Y}=\mathbf{y} \right] \\
&\quad - E\left[e^{i(\mathbf{X}-\mathbf{X}')^T\mathbf{u}} | \mathbf{Y}'=\mathbf{y} \right] + E\left[e^{i(\mathbf{X}-\mathbf{X}')^T\mathbf{u}} \right] \Big\} d\omega(\mathbf{u}) \\
&= E_{\mathbf{X}_\mathbf{y}, \mathbf{X}'_\mathbf{y}} K(\mathbf{X}-\mathbf{X}') - 2E_{\mathbf{X}_\mathbf{y}, \mathbf{X}'} K(\mathbf{X}-\mathbf{X}') + E_{\mathbf{X}, \mathbf{X}'} K(\mathbf{X}-\mathbf{X}'). \qquad \square
\end{aligned}$$

2. **Proof of Theorem 2.3.** Since K generates ρ, we can write $\rho(\mathbf{x}, \mathbf{x}') = K(\mathbf{x}, \mathbf{x}) + K(\mathbf{x}', \mathbf{x}') - 2K(\mathbf{x}, \mathbf{x}')$. Denote $\nu = \mathbf{P} - \mathbf{Q}$, then

$$\begin{aligned}
D_\rho(\mathbf{y}) &= -\int\int [K(\mathbf{x}, \mathbf{x}) + K(\mathbf{x}', \mathbf{x}') - 2K(\mathbf{x}, \mathbf{x}')] d\nu(\mathbf{x}) d\nu(\mathbf{x}') \\
&= -2\int K(\mathbf{x}, \mathbf{x}) d\nu(\mathbf{x}) \int d\nu(\mathbf{x}') + 2\int\int K(\mathbf{x}, \mathbf{x}') d\nu(\mathbf{x}) d\nu(\mathbf{x}') \\
&= 2\int\int K(\mathbf{x}, \mathbf{x}') d\nu(\mathbf{x}) d\nu(\mathbf{x}') \\
&= 2\gamma_K^2(\mathbf{y}). \qquad \square
\end{aligned}$$

3. **Proof of Theorem 2.5.**

(1) By definition, $\mathcal{H}_K^2(\mathbf{X}|\mathbf{Y}) \geq 0$.

For arbitrary $\mathbf{x}$ and $\mathbf{x}'$ in $\mathbb{R}^p$, $\begin{pmatrix} K(\mathbf{x}, \mathbf{x}) & K(\mathbf{x}, \mathbf{x}') \\ K(\mathbf{x}, \mathbf{x}') & K(\mathbf{x}', \mathbf{x}') \end{pmatrix}$ is positive semi-definite since K is a symmetric and positive definite kernel. Hence, $K(\mathbf{x}, \mathbf{x}') \leq$

$\sqrt{K(\mathbf{x},\mathbf{x})K(\mathbf{x}',\mathbf{x}')}$, $\forall \mathbf{x},\mathbf{x}' \in \mathbb{R}^p$. Note that

$$\begin{aligned} E_{\mathbf{Y}} E_{\mathbf{X_Y},\mathbf{X'_Y}} K(\mathbf{X},\mathbf{X}') &\le E_{\mathbf{Y}} E_{\mathbf{X_Y},\mathbf{X'_Y}} \sqrt{K(\mathbf{X},\mathbf{X})K(\mathbf{X}',\mathbf{X}')} \\ &= E_{\mathbf{Y}} \left[E_{\mathbf{X_Y}} \sqrt{K(\mathbf{X},\mathbf{X})} \right]^2 \\ &\le E_{\mathbf{Y}} E_{\mathbf{X_Y}} K(\mathbf{X},\mathbf{X}) \\ &= E_{\mathbf{X}} K(\mathbf{X},\mathbf{X}). \end{aligned}$$

Therefore, $\mathcal{H}^2_K(\mathbf{X}|\mathbf{Y}) \le \mathcal{H}^2_K(\mathbf{X}|\mathbf{X})$.

(2) $\rho_K(\mathbf{X}|\mathbf{Y}) = 1$ iff $E_{\mathbf{Y}} E_{\mathbf{X_Y},\mathbf{X'_Y}} K(\mathbf{X},\mathbf{X}') = E_{\mathbf{X}} K(\mathbf{X},\mathbf{X})$. All the inequalities in (1) become equalities iff $\mathbf{X}$ is a function of $\mathbf{Y}$.

4. **Proof of Theorem 2.6.**

Let $K_{ij} \equiv K(\mathbf{X}_i,\mathbf{X}_j)$ and $I_i^{(l)} \equiv I\{\mathbf{Y}_i = \mathbf{y}^{(l)}\}$ for $l = 1,...,L$.

$$\begin{aligned} \mathcal{H}^2_{K,n}(\mathbf{X}|\mathbf{Y}) &= \sum_{l=1}^{L} \frac{n}{n_l} \frac{1}{n^2} \sum_{i,j=1}^{n_l} K(\mathbf{X}_i^{(l)},\mathbf{X}_j^{(l)}) - \frac{1}{n^2} \sum_{i,j=1}^{n} K(\mathbf{X}_i,\mathbf{X}_j) \\ &= \sum_{l=1}^{L} \frac{n}{n_l} \frac{1}{n^2} \sum_{i,j=1}^{n} K_{ij} I_i^{(l)} I_j^{(l)} - \frac{1}{n^2} \sum_{i,j=1}^{n} K_{ij} \\ &\equiv \sum_{l=1}^{L} \frac{n}{n_l} V_n^{(l)} - V_n^{(0)}, \end{aligned}$$

where $V_n^{(1)},...,V_n^{(L)}$ and $V_n^{(0)}$ are V-statistics. We denote the corresponding U-statistics by $U_n^{(1)},...,U_n^{(L)}$ and $U_n^{(0)}$, respectively. Applying the Strong Law of Large Numbers for U-statistic (Hoeffding, 1961), we have

$$\begin{aligned} & U_n^{(l)} \overset{a.s.}{\to} E\left[K_{12} I_1^{(l)} I_2^{(l)}\right],\ l = 1,...,L, \\ & U_n^{(0)} \overset{a.s.}{\to} E K_{12}, \\ & \frac{n_l}{n} = \frac{1}{n} \sum_{i=1}^{n} I_i^{(l)} \overset{a.s.}{\to} p_l. \end{aligned}$$

Note that $E_{\mathbf{X}_{\mathbf{y}^{(l)}},\mathbf{X}'_{\mathbf{y}^{(l)}}} K(\mathbf{X},\mathbf{X}') = \frac{1}{p_l^2} E\left[K_{12} I_1^{(l)} I_2^{(l)}\right]$. Therefore,

$$\sum_{l=1}^{L} \frac{n}{n_l} U_n^{(l)} - U_n^{(0)} \overset{a.s.}{\to} \mathcal{H}^2_K(\mathbf{X}|\mathbf{Y}).$$

Since $EK_{12} \leq E_{\mathbf{X}}K(\mathbf{X},\mathbf{X}) < \infty$, we have $E|U_n^{(l)} - V_n^{(l)}| = O(n^{-1})$ by Lemma 5.7.3 in Serfling (1980) and hence, $\mathrm{P}(|U_n^{(l)} - V_n^{(l)}| > \epsilon) \leq \frac{E|U_n^{(l)} - V_n^{(l)}|}{\epsilon} \to 0$ as $n \to \infty$ for any $\epsilon > 0$, i.e., $U_n^{(l)} - V_n^{(l)} \xrightarrow{P} 0$ by Markov's inequality for $l = 0, 1, .., L$. Consequently,

$$\mathcal{H}^2_{K,n}(\mathbf{X}|\mathbf{Y}) \xrightarrow{P} \mathcal{H}^2_K(\mathbf{X}|\mathbf{Y}). \qquad \square$$

5. **Proof of Theorem 2.7.**

Let $K_{ij} \equiv K(\mathbf{X}_i, \mathbf{X}_j)$, $I_i^{(l)} \equiv I\{\mathbf{Y}_i = \mathbf{y}^{(l)}\}$ for $l = 1, ..., L$ and $\widetilde{K}_{ij} \equiv \widetilde{K}(\mathbf{X}_i, \mathbf{X}_j) \equiv K(\mathbf{X}_i, \mathbf{X}_j) - E_{\mathbf{X}}K(\mathbf{X}_i, \mathbf{X}) - E_{\mathbf{X}}K(\mathbf{X}, \mathbf{X}_j) + E_{\mathbf{X},\mathbf{X}'}K(\mathbf{X}, \mathbf{X}')$.

$$\begin{aligned}
\mathcal{H}^2_{K,n}(\mathbf{X}|\mathbf{Y}) &= \sum_{l=1}^{L} \frac{n}{n_l}\frac{1}{n^2}\sum_{i,j=1}^{n} K_{ij}I_i^{(l)}I_j^{(l)} - \frac{1}{n^2}\sum_{i,j=1}^{n} K_{ij} \\
&= \sum_{l=1}^{L} \frac{n}{n_l}\frac{1}{n^2}\sum_{i,j=1}^{n} \widetilde{K}_{ij}I_i^{(l)}I_j^{(l)} - \frac{1}{n^2}\sum_{i,j=1}^{n} \widetilde{K}_{ij} \\
&\equiv \sum_{l=1}^{L} \frac{n}{n_l}V_n^{(l)} - V_n^{(0)}.
\end{aligned}$$

where $V_n^{(l)}$ $(l = 1, ..., L)$ and $V_n^{(0)}$ are V-statistics. We denote the corresponding U-statistics by $U_n^{(l)} \equiv \frac{1}{n(n-1)}\sum_{i\neq j} h_{ij}^{(l)}$ with kernel $h_{12}^{(l)} \equiv \widetilde{K}_{12}I_1^{(l)}I_2^{(l)}$ for $l = 1, ..., L$, and $U_n^{(0)} \equiv \frac{1}{n(n-1)}\sum_{i\neq j} h_{ij}^{(0)}$ with kernel $h_{12}^{(0)} \equiv \widetilde{K}_{12}$, respectively. Let $\widetilde{\mathcal{H}}^2_{K,n}(\mathbf{X}|\mathbf{Y}) \equiv \sum_{l=1}^{L} \frac{n}{n_l}U_n^{(l)} - U_n^{(0)}$, then

$$\begin{aligned}
& n\mathcal{H}^2_{K,n}(\mathbf{X}|\mathbf{Y}) - n\widetilde{\mathcal{H}}^2_{K,n}(\mathbf{X}|\mathbf{Y}) \\
=& \sum_{l=1}^{L} \frac{n}{n_l} n(V_n^{(l)} - U_n^{(l)}) - n(V_n^{(0)} - U_n^{(0)}) \\
=& \sum_{l=1}^{L} \frac{n}{n_l}\left[\frac{1}{n}\sum_{i=1}^{n} \widetilde{K}_{ii}I_i^{(l)} - U_n^{(l)}\right] - \left[\frac{1}{n}\sum_{i=1}^{n} \widetilde{K}_{ii} - U_n^{(0)}\right] \\
\xrightarrow{P}& (L-1)\mathcal{H}^2_K(\mathbf{X}|\mathbf{X}).
\end{aligned}$$

The above limit holds due to the null hypothesis. Thus our objective is to show

$$n\widetilde{\mathcal{H}}^2_{K,n}(\mathbf{X}|\mathbf{Y}) \xrightarrow{d} (L-1)\mathcal{H}^2_K(\mathbf{X}|\mathbf{X})(Q-1), \qquad (*)$$

where $Q = \sum_{i=1}^{\infty} \lambda_i Z_i^2$, $Z_i \overset{i.i.d}{\sim} N(0,1)$ and λ_i are positive constants with $\sum_{i=1}^{\infty} \lambda_i = 1$.

A representation for $h_{12}^{(l)}$, the same as in Serfling (1980, p.196), will be used. Let $\{\phi_m^{(l)}(\cdot)\}$ denote orthonormal eigenfunctions corresponding to the eigenvalues $\{\zeta_m^{(l)}\}$ defined in connection with $h_{12}^{(l)}$, i.e., $\{\phi_m^{(l)}(\cdot)\}$ satisfies the followings for $l = 0, ..., L$:

(i) $E_{(\mathbf{X}_2,\mathbf{Y}_2)}\left[h_{12}^{(l)}\phi_m^{(l)}(\mathbf{X}_2,\mathbf{Y}_2)\right] = \zeta_m^{(l)}\phi_m^{(l)}(\mathbf{X}_1,\mathbf{Y}_1)$

(ii) $E\left[\phi_{m_1}^{(l)}\phi_{m_2}^{(l)}\right] = \begin{cases} 1, & m_1 = m_2 \\ 0, & m_1 \neq m_2 \end{cases}$

(iii) $\lim_{M\to\infty} E\left[h_{12}^{(l)} - \sum_{m=1}^{M} \zeta_m^{(l)}\phi_m^{(l)}(\mathbf{X}_1,\mathbf{Y}_1)\phi_m^{(l)}(\mathbf{X}_2\mathbf{Y}_2)\right]^2 = 0.$

Then we write $h_{12}^{(l)} = \sum_{m=1}^{\infty} \zeta_m^{(l)}\phi_m^{(l)}(\mathbf{X}_1,\mathbf{Y}_1)\phi_m^{(l)}(\mathbf{X}_2\mathbf{Y}_2)$. In the same sense, we have $h_1^{(l)}(\mathbf{X}_1,\mathbf{Y}_1) \equiv E_{(\mathbf{X}_2,\mathbf{Y}_2)}h_{12}^{(l)} = \sum_{m=1}^{\infty} \zeta_m^{(l)}\phi_m^{(l)}(\mathbf{X}_1,\mathbf{Y}_1)E\phi_m^{(l)}(\mathbf{X}_2\mathbf{Y}_2)$. Note that $E_{\mathbf{X}_2}\widetilde{K}_{12} = 0$ so $h_1^{(l)}(\mathbf{X}_1,\mathbf{Y}_1) = 0$. Therefore, $E\phi_m^{(l)} = 0$ as $\mathrm{Var}h_1^{(l)} = 0$, for all m.

Let $\{\phi_m(\cdot)\}$ denote orthonormal eigenfunctions corresponsing to the eigenvalues $\{\zeta_m\}$ defined in connection with $\widetilde{K}_{12}$. Similarly, $E\phi_m = 0$. We can deduce from (i) and (ii) that:

(a) $\phi_m^{(l)}(\mathbf{X}_1,\mathbf{Y}_1) = \frac{1}{\sqrt{p_l}}\phi_m(\mathbf{X}_1)I_1^{(l)}$ for $l = 1, ..., L$

(b) $\phi_m^{(0)}(\mathbf{X}_1,\mathbf{Y}_1) = \phi_m(\mathbf{X}_1)$,

(c) $\frac{\zeta_m^{(l)}}{p_l} = \zeta_m^{(0)} = \zeta_m$, $l = 1, ..., L$.

We explain (a) and (c) only and the rest follows from the same logic. From (i),

$$\begin{aligned}\zeta_m^{(l)}\phi_m^{(l)}(\mathbf{X}_1,\mathbf{Y}_1) &= E_{(\mathbf{X}_2,\mathbf{Y}_2)}\left[h_{12}^{(l)}\phi_m^{(l)}(\mathbf{X}_2,\mathbf{Y}_2)\right] \\ &= \begin{cases} 0, & \text{if } \mathbf{Y}_1 \neq \mathbf{y}^{(l)}, \\ p_l E_{\mathbf{X}_2}\left[\widetilde{K}_{12}\phi_m^{(l)}(\mathbf{X}_2,\mathbf{y}^{(l)})\right], & \text{if } \mathbf{Y}_1 = \mathbf{y}^{(l)}, \end{cases}\end{aligned}$$

for $l = 1, ..., L$. Hence, $\phi_m^{(l)}(\mathbf{X}_1,\mathbf{Y}_1) = c\phi_m(\mathbf{X}_1)I_1^{(l)}$ for some constant c and $\frac{\zeta_m^{(l)}}{p_l} = \zeta_m$, for l=1,...,L. Required by (ii), $c = \frac{1}{\sqrt{p_l}}$.

Let $\widetilde{H} \equiv \sum_{m=1}^{\infty} \zeta_m [\sum_{j=1}^{L-1} Z_{m,j}^2 - (L-1)]$ and $\widetilde{H}_M \equiv \sum_{m=1}^{M} \zeta_m [\sum_{j=1}^{L-1} Z_{m,j}^2 - (L-1)]$, where $Z_{m,j} \overset{i.i.d}{\sim} N(0,1)$. Putting $T_n^{(l)} \equiv \frac{1}{n} \sum_{i \neq j} h_{ij}^{(l)}$, we have $nU_n^{(l)} = \frac{n}{n-1} T_n^{(l)}$. In terms of the above representation for $h^{(l)}$, $T_n^{(l)} = \frac{1}{n} \sum_{i \neq j} \sum_{m=1}^{\infty} \zeta_m^{(l)} \phi_m^{(l)}(\mathbf{X}_i, \mathbf{Y}_i) \phi_m^{(l)}(\mathbf{X}_j, \mathbf{Y}_j)$ and let $T_{n,M}^{(l)} \equiv \frac{1}{n} \sum_{i \neq j} \sum_{m=1}^{M} \zeta_m^{(l)} \phi_m^{(l)}(\mathbf{X}_i, \mathbf{Y}_i) \phi_m^{(l)}(\mathbf{X}_j, \mathbf{Y}_j)$ for $l = 0, ..., L$. Eventually, we will show that

$$n\widetilde{\mathcal{H}}_{K,n}^2(\mathbf{X}|\mathbf{Y}) = \frac{n}{n-1} \left[\sum_{l=1}^{L} \frac{n}{n_l} T_n^{(l)} - T_n^{(0)} \right] \overset{d}{\to} \widetilde{H} \qquad (**)$$

by using characteristic functions. The proof is decomposed into 3 parts.

(1) Given $\epsilon > 0$ and s, $\left| Ee^{is\left(\sum_{l=1}^{L} \frac{n}{n_l} T_n^{(l)} - T_n^{(0)}\right)} - Ee^{is\left(\sum_{l=1}^{L} \frac{n}{n_l} T_{n,M}^{(l)} - T_{n,M}^{(0)}\right)} \right| < \epsilon$ for M and n sufficiently large.

Using the inequality $|e^{iz} - 1| \leq |z|$, we have

$$\begin{aligned}
&\left| Ee^{is\left(\sum_{l=1}^{L} \frac{n}{n_l} T_n^{(l)} - T_n^{(0)}\right)} - Ee^{is\left(\sum_{l=1}^{L} \frac{n}{n_l} T_{n,M}^{(l)} - T_{n,M}^{(0)}\right)} \right| \\
\leq & |s| E \left| \sum_{l=1}^{L} \frac{n}{n_l} \left(T_n^{(l)} - T_{n,M}^{(l)} \right) - \left(T_n^{(0)} - T_{n,M}^{(0)} \right) \right| \\
\leq & |s| \left\{ \sum_{l=1}^{L} \frac{n}{n_l} E \left| T_n^{(l)} - T_{n,M}^{(l)} \right| + E \left| T_n^{(0)} - T_{n,M}^{(0)} \right| \right\} \\
\leq & |s| \left\{ \sum_{l=1}^{L} \frac{n}{n_l} \left[E \left(T_n^{(l)} - T_{n,M}^{(l)} \right)^2 \right]^{1/2} + \left[E \left(T_n^{(0)} - T_{n,M}^{(0)} \right)^2 \right]^{1/2} \right\}.
\end{aligned}$$

Similar to Serfling (1980, p.197 - p.198), we can show that $\sum_{m=1}^{\infty} \zeta_m^2 = E\widetilde{K}_{12}^2 < \infty$ since $E_{\mathbf{X}} K(\mathbf{X}, \mathbf{X}) < \infty$, and $E \left(T_n^{(l)} - T_{n,M}^{(l)} \right)^2 \leq 2^2 \sum_{m=M+1}^{\infty} \left[\zeta_m^{(l)} \right]^2$ for $l = 0, ..., L$. Combining with the fact that $\frac{n}{n_l} \overset{a.s.}{\to} \frac{1}{p_l}$, the conclusion follows.

(2) $\sum_{l=1}^{L} \frac{n}{n_l} T_{n,M}^{(l)} - T_{n,M}^{(0)} \overset{d}{\to} \widetilde{H}_M$.

We may write

$$T_{n,M}^{(l)} = \sum_{m=1}^{M} \zeta_m^{(l)} \left[\left(W_{n,m}^{(l)} \right)^2 - R_{n,m}^{(l)} \right],$$

where $W_{n,m}^{(l)} \equiv n^{-1/2} \sum_{t=1}^{n} \phi_K^{(l)}(\mathbf{X}_t, \mathbf{Y}_t)$ and $R_{n,m}^{(l)} = n^{-1} \sum_{t=1}^{n} \left[\phi_K^{(l)}(\mathbf{X}_t, \mathbf{Y}_t) \right]^2$.

From the foregoing considerations, it can be seen that

$$\mathbf{W}_{n,m} \equiv \begin{pmatrix} W_{n,m}^{(1)} & \cdots & W_{n,m}^{(L)} & W_{n,m}^{(0)} \end{pmatrix}^T \overset{d}{\to} \mathbf{W}_m,$$

where $\mathbf{W}_m \sim N(\mathbf{0}, \Sigma)$ with

$$\Sigma = \begin{pmatrix} 1 & & & \sqrt{p_1} \\ & \ddots & & \vdots \\ & & 1 & \sqrt{p_L} \\ \sqrt{p_1} & \cdots & \sqrt{p_l} & 1 \end{pmatrix},$$

and $\mathrm{Cov}(\mathbf{W}_{n,m_1}, \mathbf{W}_{n,m_2}) = \mathbf{0}$ for $m_1 \neq m_2$.

Also, $R^{(l)}_{n,m} \xrightarrow{P} 1$ for $l = 0, ..., L$. Let $i^2 = -1$.

Let $\mathbf{A}_n \equiv \mathrm{diag}\left(\sqrt{\frac{n}{n_1}p_1}, ..., \sqrt{\frac{n}{n_L}p_L}, i\right)$, then $\mathbf{A}_n \xrightarrow{P} \mathbf{A} \equiv \mathrm{diag}(1 \ \cdots \ 1 \ i)$ and $\mathbf{A}\Sigma\mathbf{A}^T$ has and only has non-zero eigenvalue 1 with multiplicity $L-1$. Therefore, $(\mathbf{A}_n\mathbf{W}_{n,m})^T\mathbf{A}_n\mathbf{W}_{n,m} \xrightarrow{d} \sum_{i=1}^{L-1} Z_i^2 \sim \chi^2_{L-1}$ as $Z_i \overset{iid}{\sim} N(0,1)$ and

$$\sum_{l=1}^{L} \frac{n}{n_l} T^{(l)}_{n,M} - T^{(0)}_{n,M} = \sum_{m=1}^{M} \zeta_m \left[(\mathbf{A}_n\mathbf{W}_{n,m})^T\mathbf{A}_n\mathbf{W}_{n,m} - \left(\sum_{l=1}^{L} p_l \frac{n}{n_l} R^{(l)}_{n,m} - R^{(0)}_{n,m} \right) \right]$$
$$\xrightarrow{d} \widetilde{H}_M.$$

(3) Given $\epsilon > 0$, $\left| Ee^{is\widetilde{H}} - Ee^{is\widetilde{H}_M} \right| < \epsilon$ for M sufficiently large.

This can be seen by Serfling (1980, p.199).

Combining (1) to (3), we establish (**). To finish the proof of (*), note that

$$\begin{aligned} \widetilde{H} &= \sum_{m=1}^{\infty} \zeta_m [\sum_{j=1}^{L-1} Z^2_{m,j} - (L-1)] \\ &= \sum_{i=1}^{\infty} \widetilde{\zeta}_i (Z_i^2 - 1), \text{ where } Z_i^2 \overset{i.i.d}{\sim} N(0,1) \text{ and } \sum_{i=1}^{\infty} \widetilde{\zeta}_i = (L-1) \sum_{i=1}^{\infty} \zeta_i, \\ &= [(L-1) \sum_{i=1}^{\infty} \zeta_i](Q-1), \text{ where } Q = \sum_{i=1}^{\infty} \lambda_i Z_i^2 \text{ and } \sum_{i=1}^{\infty} \lambda_i = 1. \end{aligned}$$

Thus, we need to show that $\sum_{m=1}^{\infty} \zeta_m = \mathcal{H}^2_K(\mathbf{X}|\mathbf{X})$. Indeed,

$$\mathcal{H}^2_K(\mathbf{X}|\mathbf{X}) = E_{\mathbf{X}} \widetilde{K}(\mathbf{X}, \mathbf{X}) = \sum_{m=1}^{\infty} \zeta_m E\phi^2_m(\mathbf{X}) = \sum_{m=1}^{\infty} \zeta_m. \qquad \square$$

6. **Proof of Corollary 2.1.**

$\mathcal{H}^2_{K,n}(\mathbf{X}|\mathbf{X}) \xrightarrow{P} \mathcal{H}^2_K(\mathbf{X}|\mathbf{X})$ by Theorem 6. If $\mathbf{X}$ and $\mathbf{Y}$ are independent, the conclusion follows from Theorem 6. If $\mathbf{X}$ and $\mathbf{Y}$ are dependent, $\mathcal{H}^2_{K,n}(\mathbf{X}|\mathbf{Y}) \xrightarrow{P} \mathcal{H}^2_K(\mathbf{X}|\mathbf{Y}) > 0$, and therefore, $n\mathcal{H}^2_{K,n}(\mathbf{X}|\mathbf{Y}) \xrightarrow{P} \infty$. $\square$

7. **Proof of Theorem 2.8.**

Let $\mathbf{W} \equiv (\mathbf{X}, \mathbf{Y})$ and $P_n(\mathbf{W}_1, .., \mathbf{W}_5) \equiv G_{12}G_{13}d_{2345}$. Note that $P_n(\mathbf{W}_1, ..., \mathbf{W}_5)$ is not symmetric in its arguments, we hence re-write Γ_n^U as a U-statistic with a symmetric kernel, i.e.

$$\Gamma_n^U(\mathbf{X}|\mathbf{Y}) = \binom{n}{5} \sum_{t_1<...<t_5} \mathcal{P}_n(\mathbf{W}_{t_1}, ..., \mathbf{W}_{t_5}),$$

where $\mathcal{P}_n(\mathbf{W}_{t_1}, ..., \mathbf{W}_{t_5}) \equiv \frac{1}{5!}\sum_\pi P_n(\mathbf{W}_{i_1}, ..., \mathbf{W}_{i_5})$ and $\sum_\pi$ denotes summation over the 5! permutations $(i_1, ..., i_5)$ of $(t_1, ..., t_5)$. Let $\theta_n = E\mathcal{P}_n(\mathbf{W}_1, ..., \mathbf{W}_5)$ and $\theta = E_{\mathbf{Y}}\left[\gamma_K^2(\mathbf{Y})f^2(\mathbf{Y})\right]$. Our goal is to show that $\Gamma_n^U(\mathbf{X}|\mathbf{Y}) \xrightarrow{P} \theta$. The proof involves two steps.

(1) $\theta_n = \theta + o_p(1)$.

First note that

$$\begin{aligned}
\theta_n &= E(G_{12}G_{13}d_{2345}) \\
&= E(G_{12}G_{13}K_{23}) - 2E(G_{12}G_{13}K_{24}) + E(G_{12}G_{13}K_{45}) \\
&= \theta_{n1} + \theta_{n2} + \theta_{n3}
\end{aligned}$$

Consider the first term,

$$\begin{aligned}
\theta_{n1} = &\int K(\mathbf{x}_2, \mathbf{x}_3)h^{-2q}G(\frac{\mathbf{y}_1 - \mathbf{y}_2}{h})G(\frac{\mathbf{y}_1 - \mathbf{y}_3}{h}) \\
&\quad f(\mathbf{x}_1, \mathbf{y}_1)f(\mathbf{x}_2, \mathbf{y}_2)f(\mathbf{x}_3, \mathbf{y}_3)d\mathbf{x}_1 d\mathbf{x}_2 d\mathbf{x}_3 d\mathbf{y}_1 d\mathbf{y}_2 d\mathbf{y}_3 \\
= &\int K(\mathbf{x}_2, \mathbf{x}_3)f(\mathbf{x}_2|\mathbf{y}_1 + h\mathbf{u})f(\mathbf{x}_3|\mathbf{y}_1 + h\mathbf{u})d\mathbf{x}_2 d\mathbf{x}_3 \\
&\quad G(\mathbf{u})G(\mathbf{v})f(\mathbf{y}_1 + h\mathbf{u})f(\mathbf{y}_1 + h\mathbf{u})d\mathbf{u}d\mathbf{v}f(\mathbf{y}_1)d\mathbf{y}_1 \\
= &\int K(\mathbf{x}_2, \mathbf{x}_3)f(\mathbf{x}_2|\mathbf{y}_1)f(\mathbf{x}_3|\mathbf{y}_1)d\mathbf{x}_2 d\mathbf{x}_3 f^3(\mathbf{y}_1)d\mathbf{y}_1 + O_p(h^\nu)
\end{aligned}$$

by Taylor expansion and condition (C1), (C3) and (C4). Similarly,

$$\theta_{n2} = \int K(\mathbf{x}_2, \mathbf{x}_4)f(\mathbf{x}_2|\mathbf{y}_1)f(\mathbf{x}_4)d\mathbf{x}_2 d\mathbf{x}_4 f^3(\mathbf{y}_1)d\mathbf{y}_1 + O_p(h^\nu)$$

$$\theta_{n2} = \int K(\mathbf{x}_4, \mathbf{x}_5) f(\mathbf{x}_4) f(\mathbf{x}_5) d\mathbf{x}_4 d\mathbf{x}_5 f^3(\mathbf{y}_1) d\mathbf{y}_1 + O_p(h^\nu)$$

Combining the three terms, we have $\theta_n = \theta + o_p(1)$.

(2) $\Gamma_n^U(\mathbf{X}|\mathbf{Y}) = \theta_n + o_p(1)$.

We adopt the H-decomposition in Lee (1990) and denote

$$\mathcal{P}_{nc}(\mathbf{W}_1, ..., \mathbf{W}_c) \equiv E[\mathcal{P}_n(\mathbf{W}_1, ..., \mathbf{W}_5)|\mathbf{W}_1, ..., \mathbf{W}_c],$$

$$P_{nc}(\mathbf{W}_{i_1}, ..., \mathbf{W}_{i_c}) \equiv E[P_n(\mathbf{W}_1, ..., \mathbf{W}_5)|\mathbf{W}_{i_1}, ..., \mathbf{W}_{i_c}],$$

where $1 \leq i_1 < ... < i_c \leq 5$. Let $\phi_n^{(1)} \equiv \mathcal{P}_{n1}(\mathbf{W}_1) - \theta_n$ and

$$\phi_n^{(c)} \equiv \mathcal{P}_{nc}(\mathbf{W}_1, ..., \mathbf{W}_c) - \sum_{j=1}^{c-1} \sum_{(c,j)} \phi_n^{(j)}(\mathbf{W}_{i_1}, ..., \mathbf{W}_{i_j}) - \theta_n,$$

where the $\sum_{(c,j)}$ is taken over all subsets $1 \leq i_1 < ... < i_j \leq c$ of $\{1, ..., c\}$. Then

$$\Gamma_n^U(\mathbf{X}|\mathbf{Y}) = E\mathcal{P}_n(\mathbf{W}_1, ..., \mathbf{W}_5) + \sum_{c=1}^{5} \binom{5}{c} \Phi_n^{(c)},$$

where $\Phi_n^{(c)} = \binom{n}{c}^{-1} \sum_{(n,c)} \phi_n^{(c)}(\mathbf{W}_{i_1}, ..., \mathbf{W}_{i_c})$ satisfies the following properties:

(i) $\Phi_n^{(c)}$ are uncorrelated with $E[\Phi_n^{(c)}] = 0$, $c = 1, ..., 5$.

(ii) $Var[\Phi_n^{(c)}] = \binom{n}{c}^{-1} Var[\phi_n^{(c)}(\mathbf{W}_1, ..., \mathbf{W}_c)]$.

(iii) $Var[\phi_n^{(c)}(\mathbf{W}_1, ..., \mathbf{W}_c)] = \sum_{j=1}^{c} (-1)^{c-j} \binom{c}{j} Var[\mathcal{P}_{nj}(\mathbf{W}_1, ..., \mathbf{W}_j)]$.

We first show that $Var[\Phi_n^{(1)}] = \frac{1}{n} Var[\mathcal{P}_{n1}(\mathbf{W}_1)] = O_p(\frac{1}{n})$. Note that $E[\mathcal{P}_{n1}^2(\mathbf{W}_1)]$ can be expanded into several terms as

$$E[\mathcal{P}_{n1}^2(\mathbf{W}_1)] = \frac{1}{5} \sum_{i=1}^{5} E[P_{n1}^2(\mathbf{W}_i)],$$

and each of these terms can be shown to be $O_p(1)$. For example,

$$\begin{aligned}
&E[P_{n1}^2(\mathbf{W}_1)] \\
=&E[E^2(G_{12}G_{13}K_{23}|\mathbf{W}_1)] + 4E[E^2(G_{12}G_{13}K_{24}|\mathbf{W}_1)] + E[E^2(G_{12}G_{13}K_{45}|\mathbf{W}_1)] \\
&- 4E[E(G_{12}G_{13}K_{23}|\mathbf{W}_1)E(G_{12}G_{13}K_{24}|\mathbf{W}_1)]
\end{aligned}$$

$$
\begin{aligned}
&- 4E[E(G_{12}G_{13}K_{45}|\mathbf{W}_1)E(G_{12}G_{13}K_{24}|\mathbf{W}_1)] \\
&+ 2E[E(G_{12}G_{13}K_{23}|\mathbf{W}_1)E(G_{12}G_{13}K_{45}|\mathbf{W}_1)] \\
=&E_{11} + 4E_{12} + E_{13} - 4E_{14} - 4E_{15} + E_{16},
\end{aligned}
$$

where

$$
\begin{aligned}
E_{11} =& \int \Bigg[\int K(\mathbf{x}_2, \mathbf{x}_3) h^{-2q} G(\frac{\mathbf{y}_1 - \mathbf{y}_2}{h}) G(\frac{\mathbf{y}_1 - \mathbf{y}_3}{h}) f(\mathbf{x}_2|\mathbf{y}_2) f(\mathbf{x}_3|\mathbf{y}_3) d\mathbf{x}_2 d\mathbf{x}_3 \\
& f(\mathbf{y}_2) f(\mathbf{y}_3) d\mathbf{y}_2 d\mathbf{y}_3 \Bigg]^2 f(\mathbf{y}_1) d\mathbf{y}_1 \\
=& \int \Bigg[\int K(\mathbf{x}_2, \mathbf{x}_3) G(\mathbf{u}) G(\mathbf{v}) f(\mathbf{x}_2|\mathbf{y}_1 + h\mathbf{u}) f(\mathbf{x}_3|\mathbf{y}_1 + h\mathbf{v}) d\mathbf{x}_2 d\mathbf{x}_3 \\
& f(\mathbf{y}_1 + h\mathbf{u}) f(\mathbf{y}_1 + h\mathbf{v}) d\mathbf{u} d\mathbf{v} \Bigg]^2 f(\mathbf{y}_1) d\mathbf{y}_1 \\
\leq& \int \Bigg\{ \int \Bigg[\int \sqrt{K(\mathbf{x}_2, \mathbf{x}_2)} f(\mathbf{x}_2|\mathbf{y}_1 + h\mathbf{u}) d\mathbf{x}_2 \Bigg] \Bigg[\int \sqrt{K(\mathbf{x}_3, \mathbf{x}_3)} f(\mathbf{x}_3|\mathbf{y}_1 + h\mathbf{v}) d\mathbf{x}_3 \Bigg] \\
& G(\mathbf{u}) G(\mathbf{v}) f(\mathbf{y}_1 + h\mathbf{u}) f(\mathbf{y}_1 + h\mathbf{v}) d\mathbf{u} d\mathbf{v} \Bigg\}^2 f(\mathbf{y}_1) d\mathbf{y}_1 \\
=& O_p(1),
\end{aligned}
$$

and $E_{1i} = O_p(1)$ for $i = 2, ..., 6$, which can be shown analogously to above. Therefore, $Var[\Phi_n^{(1)}] = O_p(\frac{1}{n})$. Futhermore, we can obtain that $Var[\Phi_n^{(2)}] = O_p(\frac{1}{n^2 h^q})$ and $Var[\Phi_n^{(c)}] = O_p(\frac{1}{n^c h^{2q}})$ for $c \geq 3$ by similar logic. Then by Chebyshev's inequality, $\Phi_n^{(c)} = o_p(1)$ for $c = 1, ..., 5$ and hence, $\Gamma_n^U(\mathbf{X}|\mathbf{Y}) = \theta_n + o_p(1)$.

8. **Proof of Theorem 2.9.**

We continue to use the notations in the proof of Theorem 8. This proof is built upon Lemma B.4 in Fan and Li (1996). We first examine three prerequisites for $nh^{q/2}\Gamma_n^U(\mathbf{X}|\mathbf{Y})$ to be asymptotically normally distributed.

(1) Under H_0 and assumption $E_{\mathbf{X}}K^2(\mathbf{X},\mathbf{X}) < \infty$, it is easy to show that

$$E[\mathcal{P}_{n1}(\mathbf{W}_1)] = 0$$

and

$$E[\mathcal{P}_n^2(\mathbf{W}_1, ..., \mathbf{W}_5)] < \infty.$$

(2) When $n \to \infty$,

$$\frac{E[\mathcal{G}_n^2(\mathbf{W}_1,\mathbf{W}_2)] + n^{-1}E[\mathcal{P}_{n2}^4(\mathbf{W}_1,\mathbf{W}_2)]}{E^2[\mathcal{P}_{n2}^2(\mathbf{W}_1,\mathbf{W}_2)]} \to 0,$$

where

$$\mathcal{G}_n(\mathbf{W}_1,\mathbf{W}_2) = E[\mathcal{P}_{n2}(\mathbf{W}_1,\mathbf{W}_3)\mathcal{P}_{n2}(\mathbf{W}_2,\mathbf{W}_3)|\mathbf{W}_1,\mathbf{W}_2].$$

Indeed, we can verify that $E[\mathcal{G}_n^2(\mathbf{W}_1,\mathbf{W}_2)] = O_p(h^{-q})$, $E[\mathcal{P}_{n2}^4(\mathbf{W}_1,\mathbf{W}_2)] = O_p(h^{-3q})$ and $E^2[\mathcal{P}_{n2}^2(\mathbf{W}_1,\mathbf{W}_2)] = O_p(h^{-2q})$. The conclusion follows from the assumptions that $h^q \to 0$ and $nh^q \to \infty$ as $n \to \infty$.

(3) $\dfrac{E[\mathcal{P}_{nc}^2]}{E[\mathcal{P}_{n2}^2]} = \dfrac{O_p(h^{-2q})}{O_p(h^{-q})} = O_p(n^{(c-2)})$ for $c = 3, 4, 5$.

According to Lemma B.4 in Fan and Li (1996), with (1)-(3) verified, it follows that $n\Gamma_n^U(\mathbf{X}|\mathbf{Y})$ is asymptotically distributed as $N(0, \frac{5^2(5-1)^2}{2}E[\mathcal{P}_{n2}^2(\mathbf{W}_1,\mathbf{W}_2)])$. Note that $E[\mathcal{P}_{n2}^2(\mathbf{W}_1,\mathbf{W}_2)] = \frac{1}{100}E[P_{n2}^2(\mathbf{W}_2,\mathbf{W}_3)] + O_p(1)$ and $E[P_{n2}^2(\mathbf{W}_2,\mathbf{W}_3)] = h^{-q}(\sigma^2 + o_p(1))$, where

$$\sigma^2 = C^q\left[EK_{12}^2 - 2EK_{12}K_{13} + E^2K_{12}\right]Ef^3(\mathbf{Y})$$

and $C = \int_{\mathbb{R}}[\int_{\mathbb{R}} g(\mu+\nu)g(\mu)d\mu]^2 dv$. Therefore,

$$nh^{q/2}\Gamma_n^U(\mathbf{X}|\mathbf{Y}) \xrightarrow{d} N(0, 2\sigma^2).$$

B. Appendix of Chapter 3

This appendix contains two sections. Section B.1 presents two lemmas that are repeatedly used in Section B.2. Section B.2 includes proofs of theorems in Chapter 3.

B.1 Lemmas

Lemma A.1 (Deviation bound for U-statistics, Hoeffding 1963)**.** *Let* $g(\mathbf{U}_1, .., \mathbf{U}_r)$ *be a kernel of a U-statistics* U_n*, i.e.,* $U_n \equiv \frac{1}{(n)_r} \sum_{i_r^n} g(\mathbf{U}_{i_1}, .., \mathbf{U}_{i_r})$*, where* $n > r$*,* $(n)_r \equiv \frac{n!}{(n-r)!}$ *and* $\sum_{i_r^n}$ *is taken over all r-tuples* $\{i_1, ..., i_r\}$ *drawn without replacement from* $\{1, ..., n\}$*. If* $b_1 \leq g(\mathbf{U}_1, .., \mathbf{U}_r) \leq b_2$*, then for any* $\epsilon > 0$*, the following bound holds:*

$$P\left\{|U_n - EU_n| \geq \epsilon\right\} \leq 2\exp\{-2w\epsilon^2/(b_2 - b_1)^2\},$$

where $w \equiv [n/r]$*, the largest integer contained in* n/r*.*

Lemma A.2. *Under condition (C1) and (C2), for any* $\epsilon \in (0, 1)$*, we have*

$$P\{|\widehat{w}_j^M - w_j^M| \geq \epsilon\} \leq 2L\exp\left\{-\frac{an^{1-4\tau}}{L^2}\epsilon^2\right\},$$

where $j = 1, ..., p$ *and* $a > 0$ *is a constant depending on* c_1*.*

Proof. We aim to show the uniform consistency of the denominator and the numerator of $\widehat{w}_j^M$ under regularity conditions respectively. Because the denominator of $\widehat{w}_j^M$ has a similar form as the numerator, we deal with its numerator only below. Let

$$\begin{aligned}
\widehat{\mathcal{H}} \equiv & \mathcal{H}_{K,n}^2(X_j|\mathbf{Y}) \\
= & \sum_{l=1}^{L} \frac{n}{n_l}\frac{1}{n^2}\sum_{i_1,i_2=1}^{n} K(X_{i_1,j}, X_{i_2,j}) I\{\mathbf{Y}_{i_1} = \mathbf{y}^{(l)}\} I\{\mathbf{Y}_{i_2} = \mathbf{y}^{(l)}\} \\
& - \frac{1}{n^2}\sum_{i_1,i_2=1}^{n} K(X_{i_1,j}, X_{i_2,j}) \\
\equiv & \sum_{l=1}^{L} \frac{1}{\hat{p}_l} V_n^{(l)} - V_n^{(0)},
\end{aligned}$$

where $V_n^{(l)}$ $(l = 0, ..., L)$ are V-statistics. Let $U_n^{(l)}$ $(l = 0, ..., L)$ be corresponding U-statistics with $E_l \equiv EU_n^l$ $(l = 0, ..., L)$. Under condition (C1), without loss of generality, we assume that the kernel K is bounded by 1. Hence, $|E_l| \leq 1$ for $l = 0, ..., L$. Denote $\mathcal{H} \equiv \mathcal{H}_K^2(X_j|\mathbf{Y}) = \sum_{l=1}^L \frac{1}{p_l} E_l - E_0$. For arbitrary $\epsilon < 1$,

$$\begin{aligned}
P\left\{|\widehat{\mathcal{H}} - \mathcal{H}| \geq \epsilon\right\} =& P\left\{\left|\sum_{l=1}^L \frac{1}{\hat{p}_l}\left(V_n^{(l)} - E_l\right) + \sum_{l=1}^L \left(\frac{1}{\hat{p}_l} - \frac{1}{p_l}\right) E_l - \left(V_n^{(0)} - E_0\right)\right| \geq \epsilon\right\} \\
\leq& P\left\{\sum_{l=1}^L \frac{1}{\hat{p}_l}\left|V_n^{(l)} - E_l\right| \geq \frac{\epsilon}{3}\right\} + P\left\{\sum_{l=1}^L \left|\frac{1}{\hat{p}_l} - \frac{1}{p_l}\right| E_l \geq \frac{\epsilon}{3}\right\} \\
&+ P\left\{\left|V_n^{(0)} - E_0\right| \geq \frac{\epsilon}{3}\right\} \\
\equiv& \Gamma_1 + \Gamma_2 + \Gamma_3.
\end{aligned}$$

We deal with Γ_1 first.

$$\begin{aligned}
\Gamma_1 &\leq P\left\{L \max_l \frac{1}{\hat{p}_l}\left|V_n^{(l)} - E_l\right| \geq \frac{\epsilon}{3}\right\} \\
&\leq P\left\{\max_l \frac{1}{\hat{p}_l}\left|V_n^{(l)} - E_l\right| \geq \frac{\epsilon}{3L},\ \min_l \hat{p}_l \geq c_1 n^{-\tau}\right\} + P\left\{\min_l \hat{p}_l < c_1 n^{-\tau}\right\} \\
&\leq P\left\{\max_l \left|V_n^{(l)} - E_l\right| \geq \frac{c_1 n^{-\tau}\epsilon}{3L}\right\} + P\left\{\max_l |\hat{p}_l - p_l| \geq c_1 n^{-\tau}\right\} \\
&\leq \sum_{l=1}^L P\left\{\left|V_n^{(l)} - E_l\right| \geq \frac{c_1 n^{-\tau}\epsilon}{3L}\right\} + \sum_{l=1}^L P\left\{|\hat{p}_l - p_l| \geq c_1 n^{-\tau}\right\} \\
&\equiv \sum_{l=1}^L \Gamma_{11}^{(l)} + \sum_{l=1}^L \Gamma_{12}^{(l)},
\end{aligned}$$

where the third inequality holds because $\max_l |\hat{p}_l - p_l| \geq p_l - \hat{p}_l \geq 2c_1 n^{-\tau} - c_1 n^{-\tau} = c_1 n^{-\tau}$ by condition (C2).

$$\begin{aligned}
\Gamma_{11}^{(l)} &= P\left\{\left|\frac{n-1}{n} U_n^{(l)} + \frac{1}{n^2}\sum_{i=1}^n K(X_{i,j}, X_{i,j}) I\{\mathbf{Y}_i = \mathbf{y}^{(l)}\} - E_l\right| \geq \frac{c_1 n^{-\tau}\epsilon}{3L}\right\} \\
&\leq P\left\{\frac{n-1}{n}\left|U_n^{(l)} - E_l\right| + \left|\frac{1}{n^2}\sum_{i=1}^n K(X_{i,j}, X_{i,j}) I\{\mathbf{Y}_i = \mathbf{y}^{(l)}\} - \frac{1}{n} E_l\right| \geq \frac{c_1 n^{-\tau}\epsilon}{3L}\right\} \\
&\leq P\left\{\left|U_n^{(l)} - E_l\right| \geq \frac{c_1 n^{-\tau}\epsilon}{3L} - \frac{1}{n}\right\} \\
&\leq P\left\{\left|U_n^{(l)} - E_l\right| \geq \frac{c_1 n^{-\tau}\epsilon}{6L}\right\}, \text{ for } n \text{ large enough} \\
&\leq 2\exp\left\{-\frac{c_1^2 n^{1-2\tau}\epsilon^2}{36L^2}\right\},
\end{aligned}$$

where the third inequality holds as $|E_l| \leq 1$ and the last inequality follows from Lemma A.1. Also,

$$\Gamma_{12}^{(l)} \leq 2\exp\left\{-2c_1^2 n^{1-2\tau}\epsilon^2\right\}$$

due to Lemma A.1.

Similarly, we have

$$\begin{aligned}
\Gamma_2 &\leq P\left\{\max_l \frac{|\hat{p}_l - p_l|}{\hat{p}_l p_l} \geq \frac{\epsilon}{3L}\right\} \\
&\leq P\left\{\max_l \frac{|\hat{p}_l - p_l|}{\hat{p}_l} \geq \frac{2c_1 n^{-\tau}\epsilon}{3L}\right\} \\
&\leq P\left\{\max_l \frac{|\hat{p}_l - p_l|}{\hat{p}_l} \geq \frac{2c_1 n^{-\tau}\epsilon}{3L},\ \min_l \hat{p}_l \geq c_1 n^{-\tau}\right\} + P\left\{\min_l \hat{p}_l < c_1 n^{-\tau}\right\} \\
&\leq P\left\{\max_l |\hat{p}_l - p_l| \geq \frac{2c_1^2 n^{-2\tau}\epsilon}{3L}\right\} + 2L\exp\left\{-2c_1^2 n^{1-2\tau}\epsilon^2\right\} \\
&\leq 2L\exp\left\{-\frac{4c_1^2 n^{1-4\tau}\epsilon^2}{9L^2}\right\} + 2L\exp\left\{-2c_1^4 n^{1-2\tau}\epsilon^2\right\},
\end{aligned}$$

and

$$\begin{aligned}
\Gamma_3 &= P\left\{\left|V_n^{(0)} - E_0\right| \geq \frac{\epsilon}{3}\right\} \\
&= P\left\{\left|\frac{n-1}{n}U_n^{(0)} + \frac{1}{n^2}\sum_{i=1}^n K(X_{i,j}, X_{i,j}) - E_0\right| \geq \frac{\epsilon}{3}\right\} \\
&\leq P\left\{\frac{n-1}{n}\left|U_n^{(0)} - E_0\right| + \left|\frac{1}{n^2}\sum_{i=1}^n K(X_{i,j}, X_{i,j}) - \frac{1}{n}E_0\right| \geq \frac{\epsilon}{3}\right\} \\
&\leq P\left\{\left|U_n^{(0)} - E_0\right| \geq \frac{\epsilon}{3} - \frac{1}{n}\right\} \\
&\leq P\left\{\left|U_n^{(0)} - E_0\right| \geq \frac{\epsilon}{6}\right\},\ \text{for } n \text{ large enough} \\
&\leq 2\exp\left\{-\frac{n\epsilon^2}{36}\right\}.
\end{aligned}$$

Combining Γ_1, Γ_2 and Γ_3, we have

$$P\left\{|\widehat{\mathcal{H}} - \mathcal{H}| \geq \epsilon\right\} \leq 2L\exp\left\{-\frac{an^{1-4\tau}}{L^2}\epsilon^2\right\},$$

where a is a positive constant depending on c_1.

B.2 Proofs of Theorems

1. **Proof of Theorem 3.1.**

Following from Lemma A.2 and condition (C2),

$$P\{\max_{1\le j\le p}|\widehat{w}_j^M - w_j^M| \ge cn^{-\gamma}\} \le 2pL\exp\left\{-\frac{\alpha n^{1-4\tau}}{L^2}c^2n^{-2\gamma}\right\}$$
$$\le O\left(p\exp\left\{-bn^{1-2\gamma-2\kappa-4\tau}+\kappa\log n\right\}\right),$$

where $b>0$ is a constant depending on c_1 and c.

Note that $\min_{j\in\mathcal{A}} w_j^M - \max_{j\in\bar{\mathcal{A}}} w_j^M \ge 2cn^{-\gamma}$ by condition (C3). Consequently,

$$\begin{aligned}
&P\left\{\max_{j\in\bar{\mathcal{A}}}\widehat{w}_j^M \ge \min_{j\in\mathcal{A}}\widehat{w}_j^M\right\}\\
\le& P\left\{\max_{j\in\bar{\mathcal{A}}}\widehat{w}_j^M - \max_{j\in\bar{\mathcal{A}}} w_j^M \ge \min_{j\in\mathcal{A}}\widehat{w}_j^M - \min_{j\in\mathcal{A}w_j^M} + 2cn^{-\gamma}\right\}\\
\le& P\left\{\max_{j\in\bar{\mathcal{A}}}\left|\widehat{w}_j^M - w_j^M\right| \ge cn^{-\gamma}\right\} + P\left\{\max_{j\in\mathcal{A}}\left|\widehat{w}_j^M - w_j^M\right| \ge cn^{-\gamma}\right\}\\
\le& O\left(p\exp\left\{-bn^{1-2\gamma-2\kappa-4\tau}+\kappa\log n\right\}\right)
\end{aligned}$$

by Lemma A.2, that is,

$$P\left\{\max_{j\in\bar{\mathcal{A}}}\widehat{w}_j^M < \min_{j\in\mathcal{A}}\widehat{w}_j^M\right\} \ge 1 - O\left(p\exp\left\{-bn^{1-2\gamma-2\kappa-4\tau}+\kappa\log n\right\}\right).$$

If $\mathcal{A}\not\subseteq\widehat{\mathcal{A}}$, there must exist some $j\in\mathcal{A}$ such that $w_j^M \ge 2cn^{-\gamma}$ but $\widehat{w}_j^M < cn^{-\gamma}$, then $\left|\widehat{w}_j^M - w_j^M\right| > cn^{-\gamma}$, which implies that $\min_{j\in\mathcal{A}}\left|\widehat{w}_j^M - w_j^M\right| > cn^{-\gamma}$. Therefore,

$$\begin{aligned}
P\left\{\mathcal{A}\not\subseteq\widehat{\mathcal{A}}\right\} &\le P\left\{\min_{j\in\mathcal{A}}\left|\widehat{w}_j^M - w_j^M\right| > cn^{-\gamma}\right\}\\
&\le \sum_{j\in\mathcal{A}} P\left\{\left|\widehat{w}_j^M - w_j^M\right| > cn^{-\gamma}\right\}\\
&\le O\left(s\exp\left\{-bn^{1-2\gamma-2\kappa-4\tau}+\kappa\log n\right\}\right).
\end{aligned}$$

Equivalently,

$$P\left\{\mathcal{A}\subseteq\widehat{\mathcal{A}}\right\} \ge 1 - O\left(s\exp\left\{-bn^{1-2\gamma-2\kappa-4\tau}+\kappa\log n\right\}\right).$$

2. **Proof of Theorem 3.2.**

The proof is identical to Theorem 3.1 since the same marginal measure is used.

3. **Proof of Theorem 3.3.**

Note that $\widehat{\mathcal{H}} \equiv \mathcal{H}^2_{K,n}(\mathbf{X}_{-j}|X_j;\mathbf{Y}) = \sum_{l=1}^{L} \frac{n_l}{n} \mathcal{H}^2_{K,n}(\mathbf{X}^{(l)}_{-j}|X^{(l)}_j) \equiv \sum_{l=1}^{L} \hat{p}_l \widehat{\mathcal{H}}^{(l)}$. Denote $\mathcal{H} \equiv E\widehat{\mathcal{H}}$ and $\mathcal{H}^{(l)} \equiv E\widehat{\mathcal{H}}^{(l)}$, then $\mathcal{H} = \sum_{l=1}^{L} p_l \mathcal{H}^{(l)}$. Let $\tilde{L} \equiv \max_j L_j$.

$$
\begin{aligned}
P\left\{|\widehat{\mathcal{H}} - \mathcal{H}| \geq \epsilon\right\} &= P\left\{\left|\sum_{l=1}^{L} \hat{p}_l \widehat{\mathcal{H}}^{(l)} - \sum_{l=1}^{L} p_l \mathcal{H}^{(l)}\right| \geq \epsilon\right\} \\
&\leq P\left\{\sum_{l=1}^{L} \hat{p}_l \left|\widehat{\mathcal{H}}^{(l)} - \mathcal{H}^{(l)}\right| \geq \frac{\epsilon}{2}\right\} + P\left\{\sum_{l=1}^{L} |\hat{p}_l - p_l|\, \mathcal{H}^{(l)} \geq \frac{\epsilon}{2}\right\} \\
&\leq P\left\{\max_l \left|\widehat{\mathcal{H}}^{(l)} - \mathcal{H}^{(l)}\right| \geq \frac{\epsilon}{2}\right\} + P\left\{\max_l |\hat{p}_l - p_l| \geq \frac{\epsilon}{4L}\right\} \\
&\leq \sum_{l=1}^{L} P\left\{\left|\widehat{\mathcal{H}}^{(l)} - \mathcal{H}^{(l)}\right| \geq \frac{\epsilon}{2}\right\} + \sum_{l=1}^{L} P\left\{|\hat{p}_l - p_l| \geq \frac{\epsilon}{4L}\right\} \\
&\leq 2L\tilde{L} \exp\left\{-\frac{a_1 n^{1-4\tau}}{\tilde{L}^2}\epsilon^2\right\} + 2L \exp\left\{-\frac{n}{8L^2}\epsilon^2\right\} \\
&\leq 2L\tilde{L} \exp\left\{-\frac{a_2 n^{1-4\tau}}{\max\{L^2, \tilde{L}^2\}}\epsilon^2\right\},
\end{aligned}
$$

by condition (C7), Lemma A.1 and Lemma A.2, where a_1 and a_2 are some positive constants. Then since $L = O(n^\kappa)$ and $\tilde{L} = O(n^\kappa)$,

$$
\begin{aligned}
P\{\max_{1\leq j\leq p} |\widehat{w}_j^{S,2} - w_j^{S,2}| \geq cn^{-\gamma}\} &\leq 2pL\tilde{L} \exp\left\{-\frac{a_2 n^{1-4\tau}}{\max\{L^2, \tilde{L}^2\}}\epsilon^2\right\} \\
&\leq O\left(p \exp\left\{-bn^{1-2\gamma-2\kappa-4\tau} + 2\kappa \log n\right\}\right),
\end{aligned}
$$

where $b > 0$ is a constant depending on c_1 and c.

The other two inequalities in Theorem 3.3 can be easily showed following the proof of Theorem 3.1.

C. Appendix of Chapter 4

1. **Proof of Proposition 4.1.**

Let $\boldsymbol{\eta}_0$ be the projection of $\boldsymbol{\beta}$ onto $\boldsymbol{\eta}$, that is, $\boldsymbol{\eta}_0 = P_{\boldsymbol{\eta}(\Sigma_{\mathbf{X}})}\boldsymbol{\beta} = c\boldsymbol{\eta}$, where c is a scalar. Let $\boldsymbol{\eta}_0^{\perp} = \boldsymbol{\beta} - \boldsymbol{\eta}_0$, then $1 = \boldsymbol{\beta}^T\Sigma_{\mathbf{X}}\boldsymbol{\beta} = c^2 + \boldsymbol{\eta}_0^{\perp,T}\Sigma_{\mathbf{X}}\boldsymbol{\eta}_0^{\perp} \geq c^2$. Then we have

$$
\begin{aligned}
&\mathcal{H}^2(\boldsymbol{\beta}^T\mathbf{X}|Y) \\
=&\int \left| E[e^{i<t,\boldsymbol{\beta}^T\mathbf{X}>}|Y] - Ee^{i<t,\boldsymbol{\beta}^T\mathbf{X}>} \right|^2 \omega(\mathbf{u})d\mathbf{u} \\
=&\int \left| E\{E[e^{i<t,(\boldsymbol{\eta}_0^T+\boldsymbol{\eta}_0^{\perp,T})\mathbf{X}>}|Y,\boldsymbol{\eta}^T\mathbf{X}]|Y\} - Ee^{i<t,(\boldsymbol{\eta}_0^T+\boldsymbol{\eta}_0^{\perp,T})\mathbf{X}>} \right|^2 \omega(\mathbf{u})d\mathbf{u} \\
=&\int \left| E\{E[e^{i<t,(\boldsymbol{\eta}_0^T+\boldsymbol{\eta}_0^{\perp,T})\mathbf{X}>}|\boldsymbol{\eta}^T\mathbf{X}]|Y\} - Ee^{i<t,(\boldsymbol{\eta}_0^T+\boldsymbol{\eta}_0^{\perp,T})\mathbf{X}>} \right|^2 \omega(\mathbf{u})d\mathbf{u} \\
=&\int \left| E\{e^{i<t,\boldsymbol{\eta}_0^T\mathbf{X}>}E[e^{i<t,\boldsymbol{\eta}_0^{\perp,T}\mathbf{X}>}|Y,\boldsymbol{\eta}^T\mathbf{X}]|Y\} - Ee^{i<t,\boldsymbol{\eta}_0^T\mathbf{X}>}Ee^{i<t,\boldsymbol{\eta}_0^{\perp,T}\mathbf{X}>} \right|^2 \omega(\mathbf{u})d\mathbf{u} \\
=&\int \left| E[e^{i<t,\boldsymbol{\eta}_0^T\mathbf{X}>}|Y]Ee^{i<t,\boldsymbol{\eta}_0^{\perp,T}\mathbf{X}>} - Ee^{i<t,\boldsymbol{\eta}_0^T\mathbf{X}>}Ee^{i<t,\boldsymbol{\eta}_0^{\perp,T}\mathbf{X}>} \right|^2 \omega(\mathbf{u})d\mathbf{u} \\
=&\int \left| Ee^{i<t,\boldsymbol{\eta}_0^{\perp,T}\mathbf{X}>} \right|^2 \left| E[e^{i<t,\boldsymbol{\eta}_0^T\mathbf{X}>}|Y] - Ee^{i<t,\boldsymbol{\eta}_0^T\mathbf{X}>} \right|^2 \omega(\mathbf{u})d\mathbf{u} \\
\leq&\int \left| E[e^{i<t,\boldsymbol{\eta}_0^T\mathbf{X}>}|Y] - Ee^{i<t,\boldsymbol{\eta}_0^T\mathbf{X}>} \right|^2 \omega(\mathbf{u})d\mathbf{u} \\
=&\mathcal{H}^2(\boldsymbol{\eta}_0^T\mathbf{X}|Y) \\
=&\mathcal{H}^2(\boldsymbol{\eta}^T\mathbf{X}|Y).
\end{aligned}
$$

The third equality follows from the assumption that $Y \perp\!\!\!\perp \mathbf{X}|\boldsymbol{\eta}^T\mathbf{X}$ and $\boldsymbol{\eta}_0 = c\boldsymbol{\eta}$. The fourth equality follows from the assumption $P_{\boldsymbol{\eta}(\Sigma_{\mathbf{X}})}^T\mathbf{X} \perp\!\!\!\perp Q_{\boldsymbol{\eta}(\Sigma_{\mathbf{X}})}^T\mathbf{X}$. The last equality assumes that K is scale-free.

Bibliography

Bach, F. R. and Jordan, M. I. (2002a). Kernel independent component analysis. *Journal of machine learning research*, 3(Jul):1–48.

Bach, F. R. and Jordan, M. I. (2002b). Tree-dependent component analysis. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, pages 36–44. Morgan Kaufmann Publishers Inc.

Balasubramanian, K., Sriperumbudur, B., and Lebanon, G. (2013). Ultrahigh dimensional feature screening via rkhs embeddings. In *Artificial Intelligence and Statistics*, pages 126–134.

Bowman, A. W. and Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis: the Kernel Approach with S-Plus Illustrations*, volume 18. OUP Oxford, Oxford.

Candes, E., Tao, T., et al. (2007). The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351.

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

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

Cook, R. D. and Forzani, L. (2009). Likelihood-based sufficient dimension reduction. *Journal of the American Statistical Association*, 104(485):197–208.

Cook, R. D. and Weisberg, S. (1991). Sliced inverse regression for dimension reduction: Comment. *Journal of the American Statistical Association*, 86(414):328–332.

Cook, R. D. and Yin, X. (2001). Dimension reduction and visualization in discriminant analysis (with discussion). *Australian & New Zealand Journal of Statistics*, 43(2):147–199.

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

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.

Dauxois, J. and Nkiet, G. M. (1998). Nonlinear canonical analysis and independence tests. *The Annals of Statistics*, 26(4):1254–1278.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*, volume 1. Cambridge university press, Cambridge.

Dueck, J., Edelmann, D., Gneiting, T., and Richards, D. (2014). The affinely invariant distance correlation. *Bernoulli*, 20(4):2305–2330.

Efron, B. and Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. CRC press, Boca Raton.

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, Y. and Li, Q. (1996). Consistent model specification tests: omitted variables and semiparametric functional forms. *Econometrica: Journal of the Econometric Society*, pages 865–890.

Fan, Y. and Li, Q. (1999). Root-n-consistent estimation of partially linear time series models. *Journal of Nonparametric Statistics*, 11(1-3):251–269.

Fukumizu, K., Bach, F. R., and Gretton, A. (2007). Statistical consistency of kernel canonical correlation analysis. *Journal of Machine Learning Research*, 8(Feb):361–383.

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

Fukumizu, K., Gretton, A., Lanckriet, G. R., Schölkopf, B., and Sriperumbudur, B. K. (2009). Kernel choice and classifiability for rkhs embeddings of probability distributions. In Bengio, Y., Schuurmans, D., Lafferty, J. D., Williams, C. K. I., and Culotta, A., editors, *Advances in Neural Information Processing Systems*, pages 1750–1758, Red Hook, NY.

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. (2008). Kernel measures of conditional dependence. In Platt, J. C., Koller, D., Singer, Y., and Roweis, S. T., editors, *Advances in Neural Information Processing Systems*, pages 489–496, Red Hook, NY. Curran Associates, Inc.

Gill, P. E., Murray, W., and Wright, M. H. (1981). *Practical Optimization*. Academic press, New York.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012a). A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005b). Measuring statistical dependence with hilbert-schmidt norms. In Jain, S., Simon, H. U., and Tomita, E., editors, *Algorithmic Learning Theory*, pages 63–77, Berlin, Heidelberg. Springer.

Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., and Smola, A. J. (2008). A kernel statistical test of independence. In Platt, J. C., Koller, D., Singer, Y., and Roweis, S. T., editors, *Advances in Neural Information Processing Systems*, pages 585–592, Red Hook, NY. Curran Associates, Inc.

Gretton, A., Herbrich, R., Smola, A., Bousquet, O., and Schölkopf, B. (2005a). Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6(Dec):2075–2129.

Gretton, A., Sejdinovic, D., Strathmann, H., Balakrishnan, S., Pontil, M., Fukumizu, K., and Sriperumbudur, B. K. (2012b). Optimal kernel choice for large-scale two-sample tests. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in neural information processing systems*, pages 1205–1213, Red Hook, NY. Curran Associates, Inc.

Gretton, A., Smola, A. J., Bousquet, O., Herbrich, R., Belitski, A., Augath, M., Murayama, Y., Pauls, J., Schölkopf, B., and Logothetis, N. K. (2005). Kernel constrained covariance for dependence measurement. In *AISTATS*, volume 10, pages 112–119.

Hoeffding, W. (1961). The strong law of large numbers for u-statistics. Technical report, North Carolina State University. Dept. of Statistics.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30.

Kong, J., Wang, S., and Wahba, G. (2015). Using distance covariance for improved variable selection with application to learning genetic risk models. *Statistics in Medicine*, 34(10):1708–1720.

Lee, A. J. (1990). *U-Statistics: Theory and Practice*. Marcel Dekker, New York.

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

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

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

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

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

Mai, Q. and Zou, H. (2013). The kolmogorov filter for variable screening in high-dimensional binary classification. *Biometrika*, 100(1):229–234.

Mai, Q., Zou, H., et al. (2015). The fused kolmogorov filter: a nonparametric model-free screening method. *The Annals of Statistics*, 43(4):1471–1497.

Pan, R., Wang, H., and Li, R. (2016). Ultrahigh-dimensional multiclass linear discriminant analysis by pairwise sure independence screening. *Journal of the American Statistical Association*, 111(513):169–179.

Rizzo, M. L. and Székely, G. J. (2010). Disco analysis: A nonparametric extension of analysis of variance. *The Annals of Applied Statistics*, 4(2):1034–1055.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, 56(4):931–954.

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. J. (1980). *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, Hoboken.

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.

Silva, P. F. B. (2013). Development of a system for automatic plant species recognition. *Master Thesis, University of Porto*.

Silva, P. F. B., Marcal, A. R. S., and da Silva, R. M. A. (2013). Evaluation of features for leaf discrimination. In Kamel, M. and Campilho, A., editors, *International Conference Image Analysis and Recognition*, pages 197–204, Berlin, Heidelberg. Springer.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, volume 26. CRC press, Boca Raton.

Smola, A., Gretton, A., Song, L., and Schölkopf, B. (2007). A hilbert space embedding for distributions. In Hutter, M., Servedio, R. A., and Takimoto, E., editors, *Algorithmic Learning Theory*, pages 13–31, Berlin, Heidelberg. Springer.

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Lanckriet, G., and Schölkopf, B. (2008). Injective hilbert space embeddings of probability measures. pages 111–122, Madison, WI. Omnipress.

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. R. (2010). Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11(Apr):1517–1561.

Su, L. and White, H. (2003). Testing conditional independence via empirical likelihood. *Department of Economics, UCSD*.

Su, L. and White, H. (2007). A consistent characteristic function-based test for conditional independence. *Journal of Econometrics*, 141(2):807–834.

Su, L. and White, H. (2008). A nonparametric hellinger metric test for conditional independence. *Econometric Theory*, 24(04):829–864.

Sun, X., Janzing, D., Schölkopf, B., and Fukumizu, K. (2007). A kernel-based causal learning algorithm. In Ghahramani, Z., editor, *Proceedings of the 24th international conference on Machine learning*, pages 855–862, New York, NY. ACM.

Székely, G. J. and Bakirov, N. K. (2003). Extremal probabilities for gaussian quadratic forms. *Probability Theory and Related Fields*, 126(2):184–202.

Székely, G. J. and Rizzo, M. L. (2009). Brownian distance covariance. *The Annals of Applied Statistics*, 3(4):1236–1265.

Székely, G. J. and Rizzo, M. L. (2013). The distance correlation t-test of independence in high dimension. *Journal of Multivariate Analysis*, 117:193–213.

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

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Wang, X., Jiang, B., and Liu, J. S. (2017). Generalized r-squared for detecting dependence. *Biometrika*, 104(1):129–139.

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.

Wendland, H. (2004). *Scattered Data Approximation*, volume 17. Cambridge university press, Cambridge.

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

Yamanishi, Y., Vert, J.-P., and Kanehisa, M. (2004). Heterogeneous data comparison and gene selection with kernel canonical correlation analysis. In Schlkopf, B., Tsuda, K., and Vert, J. P., editors, *Kernel Methods in Computational Biology*, pages 209–229, Cambridge, MA. MIT Press.

Yang, B., Yin, X., and Zhang, N. (2019). Sufficient variable selection using independence measures for continuous responses. *Journal of Multivariate Analysis*, Accepted.

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

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

Yin, X. and Yuan, Q. (2019). A new class of measures for testing independence. *Statistica Sinica [online]*, DOI: 10.5705/ss.202017.0538.

Yuan, Q. and Yin, X. (2017). Informational index and its applications in high dimensional data. *Ph.D. Dissertation, University of Kentucky.*

Zhu, L.-P., Li, L., Li, R., and Zhu, L.-X. (2011). Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, 106(496):1464–1475.

Zhu, L.-P., Zhu, L.-X., and Feng, Z.-H. (2010). Dimension reduction in regressions through cumulative slicing estimation. *Journal of the American Statistical Association*, 105(492):1455–1466.

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.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.

Vita

Chenlu Ke

Education

- **M.S. in Statistics** University of Kentucky, 2014-2016
- **B.S. in Mathematics** East China Normal University, 2010-2014

Professional Experience

- **Biostatistician** AMAG Pharmaceuticals, summer 2018
- **Research Assistant** University of Kentucky, 2016-2019
- **Teaching Assistant** University of Kentucky, 2014-2016

Scholastic Honors

- **R.L Anderson Outstanding Research Award**
  University of Kentucky, 2018
- **Industrial Math/Stat Modeling Workshop Travel Award**
  The Statistical and Applied Mathematical Sciences Institute (SAMSI), 2017
- **Boyd Harshbarger Travel Award**
  Southern Regional Council on Statistics (SRCOS), 2017
- **Student Presentation Award**
  KY-ASA Chapter Meeting, 2017
- **David Allen Scholarship**
  University of Kentucky, 2015

Publications

- Ke, C. and Yin, X. (2019) "Expected Conditional Characteristic Function-based Measures for Testing Independence". *Journal of the American Statistical Association*. DOI: 10.1080/01621459.2019.1604364

- Villamar, M., Cook, A., Ke, C., Xu, Y., Clay, J., Dolbec, K., Ward-Mitchell, R., Goldstein, L. and BensalemOwen, M. (2018) "Alert Protocol Reduces Time To Administration of Second-line Anti-seizure Medications for Status Epilepticus". *Neurology: Clinical Practice*, 8(6), 486-491.

- Young, D., Ke, C. and Zeng, X. (2018) "The Mixturegram: A Visualization Tool for Assessing the Number of Components in Finite Mixture Models". *Journal of Computational and Graphical Statistics*. 27(3):564-575.