KERNEL-BASED NONPARAMETRIC TESTING IN HIGH-DIMENSIONAL DATA WITH APPLICATIONS TO GENE SET ANALYSIS

By

Tao He

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Statistics – Doctor of Philosophy

2015

ABSTRACT

KERNEL-BASED NONPARAMETRIC TESTING IN HIGH-DIMENSIONAL DATA WITH APPLICATIONS TO GENE SET ANALYSIS

By

Tao He

The ultimate goal of genome-wide association studies (GWAS) is understanding the underlying relationship between genetic variants and phenotype. While the heretability is largely missing in univariate analysis of traditional GWAS, it is believed that the joint analysis of variants, that are interactively functioning in a biological pathway (gene set), is more beneficial in detecting association signals. With the fast developing pace of sequencing techniques, more detailed human genome variation will be observed and hence the dimension of variants in the pathway could be extremely high. To model the systematic mechanism and the potential nonlinear interactions among the variants, in this dissertation we propose to model the set effect though a flexible non-parametric function under the high-dimensional setup, which allows the dimension goes to infinity as the size goes to infinity.

Chapter 2 considers testing a nonparametric function of high-dimensional variates in a reproducing kernel Hilbert space (RKHS), which is a function space generated by a positive definite or semidefinite kernel function. We propose a test statistic to test the nonparametric function under the high-dimensional setting. The asymptotic distributions of the test statistic are derived under the null hypothesis and a series of local alternative hypotheses, the explicit power formula under which are also provided. We also develop a novel kernel selection procedure to maximize the power of the proposed test, as well as a kernel regularization procedure to further improve power. Extensive simulation studies and a real data analysis were conducted to evaluate the performance of the proposed method.

Chapter 3 is theoretical investigation on the statistical optimality of kernel-based test statistic under the high-dimensional setup, from the minimax point of view. In particularly, we consider a high-dimensional linear model as the initial study. Unlike the sparsity or independence assumptions existing in related literature, we discussed the minimax properties under a structure free setting. We characterize the boundary that separates the testable region from the non-testable region, and show the rate-optimality of the kernel-based test statistic, under certain conditions on the covariance matrix and the growing speed of dimension.

Our work in Chapter 4 fills the blank of kernel-based test using multiple candidate kernels under the high dimensional setting. Firstly, we extend the test statistic proposed in Chapter 2 to an inclusive form that allows the adjustment of covariants. The asymptotic distribution of the new test statistic under the null hypothesis is then provided. Two practical and efficient strategies are developed to incorporate multiple kernel candidates into the testing procedures. Through comprehensive simulation studies we show that both strategies can calibrates the type I error rate and improve the power over the the poor choice of kernel candidate in the set. Particularly, the maximum method, one of the two strategies, is shown having potential to boost the power close to one using the best candidate kernel. An application to Thai baby birth weight data further demonstrates the merits of our proposed methods.

To my beloved family, especially my little boy Evan.

ACKNOWLEDGMENTS

First and foremost, I would like to express my sincere gratitude to my advisors Dr. Yuehua Cui and Dr. Ping-shou Zhong, for their excellent guidance, great patience, continuous encouragement and support throughout my graduate study and research. Dr. Cui led me into the fantastic world of statistical genetics, and inspired me to work in the interdisciplinary area. Dr. Zhong's expertise in statistical methodologies helped me broaden my vision in frontier of statistic theory and see more beauty of statistics. With the combined training they gave me in the past four years, I have equipped myself with the knowledge in both biology and statistic, and am capable of doing research independently. I greatly appreciate the time and efforts that both of them dedicated to helping me complete the research study smoothly. Their enthusiasms in research, optimistic and energetic attitude will continue inspiring me in the future.

I would also like to thank Dr. Ning Jiang, my co-advisor in the Ph.D program of Quantitative Biology, for her further training in the interdisciplinary area, valuable suggestions, continuous faith in my research capabilities, as well as offering collaboration opportunities in her group. My sincere thanks also go to my thesis committee members: Dr. Vidyadhar Mandrekar and Dr. Juan P. Steibel, for their precious time, helpful discussions and being always approachable.

I am also grateful to all the faculty and staff members in the Department of statistics and Probability who have taught me or assisted me. My special thanks go to Dr. James Stapleton and Ms. Sue Watson. Thank Dr. Stapleton for inviting us to the Thanksgiving party every year and letting us feel warm, as well as for treating us like his own children. Thank Ms. Watson for giving me countless help and special care during my pregnancy.

I would like to thank my husband, Jikai Lei, for giving me unconditional love, endless support and encouragement during the challenges of graduate school and life, as well as for being a great father for our son. At the time of most helpless and depressed, he holds my hands, stands beside me, and gives me faith to face whatever life presents me. I am truly thankful for having him in my life. Thank my cute little boy, Evan Y. Lei, for making my life more complete, colorful and meaningful.

My thanks also go to my academic family members: Dr. Shaoyu Li, Dr. Cen Wu, Dr. Xu Liu, Honglang Wang, and Bin Gao, for their valuable help in my daily life and useful discussions all these years. I also thank my great friends, Tingqiao Chen, Yingjie Li, Liqian Cai, Xiaoqing Zhu, Yuzhen Zhou and all other students in the department for all their supports, encouragements and company throughout this journey.

Last but not least, I would express my profound gratitude to my beloved parents, Chunhai He and Xiuxiang Zhang, who always believe in me and give me courage to pursue my dreams. My deep gratitude also goes to my parents-in-law, Caixia Xi and Huaizheng Lei, who took very good care of our son when I was not there.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

With the advances in genotyping technologies and the cataloging of millions of single nucleotide polymorphisms (SNPs) in the past decades, genome-wide association studies (GWAS) become increasingly popular for investigating the relationship between genetic variants and the phenotypic traits of interest. The reason that SNPs are chosen as the genetic markers is because they are the most fundamental type of genetic variation and provide comprehensive coverage of the whole genome variation. In a typical genome-wide association study, hundreds of thousands of SNPs are firstly assayed among hundreds or thousands of individuals, then the single-SNP analysis, where individual SNPs are tested one at a time, is performed to seek possible association signals between SNPs and a trait.

In spite of tremendous knowledge gain and excited findings due to GWAS (Hirschhorn and Daly , 2005; Gardon and Bell, 2001; Edwards et al., 2005; Yasuda et al., 2008), it has been pointed out that the discovered genetic risk factors can only explain a small fraction of heritability for many complex traits, either individually or collectively (Manolio et al., 2009). The mystery of the missing heritability has been hotly discussed in the genetic research community and several credible opinions and search directions has been proposed (Eichler et al., 2010). Firstly, genome-wide genotype assays often fail to adequately cover the copy number variation or the rare variates (Manolio et al., 2009; Mefford and Eichler, 2009), which potentially play important roles in the genetic structure of the traits. Secondly, a failure of considering epigenetic factors such as imprinting, is a possible source of missing

heritability (Kong et al., 2009). Another possible explanation is named by the interaction between genetic variants and environmental factors. From statistical testing perspective, the correction for multiple testing problems normally leads to a stringent threshold such that SNPs with small marginal effects are not able to be claimed significant (Gibson, 2010). Finally, the missing heritability is a consequence of single-variant-based analysis, while the true architecture of generic effects is potentially a highly complex interacted network of variants for the manifestation of complex traits (Zuk, et al., 2012). Hence novel statistical methods with a systems biological perspective are needed to identify more genetic risk factors and therefore to explain more fraction of heritability. Moreover, testing the effects of multiple variants jointly immediately reduces the multiple testing burden.

Testing biological meaningful set of variants simultaneously has been proposed as the new strategy. The introduction of biological structure not only provides a potential boost of power, but also grants researchers meaningful interpretation and possible utilization in the future (Wang, et al. 2010). One of the attractive sets is gene, which is well annotated for human and mutations in which are known to directly impact the functionality of human organism. Another more frequently used unit is biological pathway, which represents a group of genes that form a complex network to regulate a particular outcome coordinately. Similar to genes, pathway information has been provided in a variety of online resources, such as such as Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa, 2000; Reactome (Joshi-Tope et al., 2005) and Gene Ontology (Ashburner et al., 2000).

A broad range of methods have been developed for the association analysis of sets of SNPs or other variants, targeted towards the detection of significant pathways (or gene sets). One important class of the approaches in pathway association is combined p-value methods. Namely, the effect of SNPs is individually tested and the signal of a pathway is

then evaluated by combing p-values corresponding to the marginal tests of all SNPs within the pathway, in a certain way. For example, in the gene-set enrichment analysis proposed by Wang, Li and Bucan (2007), the strongest signal (i.e., the smallest p-value) within a gene is assigned to the gene, then the pathway is accessed by a weighted Kolomogorov-Smirnov-like running sum statistic, which is used to combine the signals of genes that in a pathway. Recently Yu et al. (2009) introduced an adaptive rank truncated product statistic which extends the truncation rank from one (only the smallest p-value is utilized) to an adaptively selected number. However, the combined p-value method shares the disadvantage with the traditional single-variant-based GWAS: the effect of SNP is estimated independently of all others. Moreover, since only the p-values of individual tests enter the higher-level analysis, researchers are taking the risk of losing important information.

Being independent with the single-variant testing results and capable of simultaneously considering the multiple variants, kernel methods are introduced to address the pathway-association issue. The basis of kernel methods is a positive definite or semidefinite kernel function (Hofmann et al., 2008), which not only can be used to measure the genetic similarity for all pairs of subjects, but also can flexibly model certain (linear or non-linear) relationship between the variants and a trait. Kernel machine regression method is one of the well-known example from the category (Liu et al., 2007; Liu et al., 2008). Note that the kernel methods can be also generalized to many approaches using notions of pairwise similarity (Mukhopadhyay et al., 2010; Schaid, 2010; Tzeng et al., 2009; Wessel and Schork, 2006).

Known as a very complex network where enormous genetic factors are involved, pathway is of great dimension and this dimension might be getting extremely large with the development of sequencing technology in the coming years. Under such a high-dimensional setup, all the existing kernel methods do not consider effect of data dimension on the test

statistic, hence are of lack of unified theoretical investigation under the high-dimensional setting. In the dissertation, we mainly develop the kernel-based testing procedures and the corresponding methodologies under the high-dimensional setting, aiming at the detection of significant pathways (or gene sets) that are associated with a continuous trait of interest.

In Chapter 2, we consider testing a high-dimensional nonparametric function in a reproducing kernel Hilbert space (RKHS), which is generated by a positive definite or semi-definite kernel function. A test statistic is proposed to test the nonparametric function under the "large p, small n" setup. The asymptotic distributions of the test statistic are studied under the null hypothesis and a series of local alternative hypotheses. We also develop a novel kernel selection procedure to maximize the power of the proposed test via maximizing the signal-to-noise ratio. Moreover, a test with regularized kernel is constructed to further improve the power. It is shown that the proposed test could nearly achieve the power of an oracle test if the regularization parameter is properly chosen. Extensive simulation studies were conducted to evaluate the finite sample performance of the proposed method. We applied the proposed method to a Yolkshire gilt data set to identify pathways that are associated with triiodothyronine level.

In Chapter 3, we attempt to study the statistical optimality of the kernel-based test statistic in a linear model from the minimax point of view, under the high-dimensional setup. In particularly, we consider a high-dimensional linear model as the initial investigation. After introducing the basic notation and definition for minimax testing problem, we establish a lower bound of the detection boundary that separates the testable region and non-testable region, followed by the upper bound. We show that the introduced kernel-based test statistic is rate-optimal, under certain conditions on the covariance matrix of variants and increasing speed of dimension.

In Chapter 4, we extend the kernel-based test statistic such that covariants adjustment is allowed. We provide the asymptotic distribution of the new test statistic under the null hypothesis. Moreover, we proposed two practical and efficient strategies to utilize multiple kernel candidates in the test. We demonstrated in simulation studies and real data analysis that under high-dimensional setting both strategies not only calibrates the type I error but also leads to the improvement of the power over the poor choice of the kernel candidate in set. Especially, the maximum method we proposed is observed to enable the power to be close to the one using optimal kernel function out of the candidates, while the multiple kernel strategy (Wu et al., 2013) proposed under the kernel machine framework suffers from power loss under a high-dimensional setting.

Chapter 2

Testing high-dimensional nonparametric functions in RKHS using single kernel

2.1 Introduction

High-dimensional data arise nowadays in a wide range of areas, such as biology, imaging and climate. In genetic studies, millions of single nucleotide polymorphisms (SNPs) can be measured simultaneously using the high-throughput technologies. The identification of genes that are associated with certain traits, such as blood pressure and grain yield, is increasingly important in health and agriculture sciences. While the traditional methods focus on the single gene based analysis, the limitation of single gene based method has been realized by many researchers (Manolio et al., 2009). Gene-set based analysis (Subramanian et al., 2005; Newton et al., 2007) holds great promising because gene regulation is often very complex and genes tend to work together in a non-linear way (Liu et al., 2007; Li and Cui, 2012) to achieve certain biological functions. To model the association between certain trait Y and a gene set, we consider the following nonparametric regression

$$Y_i = \mu + h(\mathbf{X}_i) + \epsilon_i, \quad i = 1, \cdots, n \tag{2.1.1}$$

where $\mathbf{X}_1, \cdots, \mathbf{X}_n$ are IID p-dimensional variates generated from a probability measure $\boldsymbol{P}$ on R^p and ϵ_i are IID random errors with zero mean and variance σ^2. For model identification purpose, we assume $E\{h(\mathbf{X}_i)\} = 0$. Since the number of genes p in a gene-set could range from a few to a few thousand but the sample size n in a genetic study is often limited, we consider a "large p, small n" setup to allow p to be much greater than n. Our interest in this chapter is on testing

$$H_0 : h(\cdot) = 0 \text{ vs } H_1 : h(\cdot) \neq 0 \tag{2.1.2}$$

under the "large p, small n" setup where $p(n) \to \infty$ as $n \to \infty$.

The hypothesis testing for a non-parametric function (2.1.2) in a fixed dimensional case (p fixed) has been well studied in the literature (Härdle and Mammen , 1993; Chen et al., 2003; Gao and Gijbels, 2008). A vast majority of them applied either a kernel smoothing method or a local polynomial method to construct test statistics. However, all of these methods suffer the "curse of dimensionality" (Fan and Gilbels, 1996) and can not be easily generalized to test functions in a high-dimensional space without a specific structure. Recently, Liu et al. (2007) proposed a score test for functions in a reproducing kernel Hilbert space by using the relationship between (2.1.1) and a linear mixed effect model (see also Liu et al., 2008). These methods can be regarded as a generalization of the score test proposed by Goeman et al. (2006). See Goeman et al. (2011) for a similar test in a generalized linear model. Nevertheless, the existing methods do not consider testing functions in the "large p, small n" setup and the effect of data dimension on the test is largely unknown.

The focus of the current chapter is on testing non-parametric functions in a "large p, small n" setup. The hypothesis testing for a high-dimensional parameter vector has received increasing attention recently. Zhong and Chen (2011) and Lan et al. (2014) considered

testing high-dimensional regression coefficients in a linear regression model which is a special case of (2.1.1) by setting $h(\mathbf{X}_i) = \mathbf{X}_i^T\boldsymbol{\beta}$ (Wang and Cui, 2013; Feng et al., 2013). However, none of the above methods can be applied to test a high-dimensional non-parametric function.

Because of the curse of dimensionality, it is very challenging or even impossible to estimate a high-dimensional non-parametric function without imposing any specific structure. However, it is interesting to find that hypothesis testing for high-dimensional non-parametric functions is still feasible without specific structures. Our method can be applied to assess any nonparametric functions in the RKHS generated by a positive semi-definite kernel. By introducing the RKHS, we show that the high-dimensional non-parametric function evaluated at data points can be represented as a function in a linear manifold generated by the kernel. This successfully translates the problem of testing nonparametric functions into a test for a high-dimensional vector. We then propose a U-statistic based test statistic to test the high-dimensional vector. The asymptotic distributions of the test statistic are obtained under the null hypothesis and a series of local alternatives without a specific distribution assumption.

Kernel selection is an important issue in a kernel machine based testing procedure (Liu et al., 2007). However, less work is done in this direction. We propose a new procedure for selecting kernels in the hypothesis testing context. By obtaining an explicit power function of the proposed test, we choose the kernel that maximizes the power function. Unlike the BIC criterion proposed in Liu et al. (2007), our procedure is tailored to the hypothesis testing problem and is particularly designed for improving the power of the proposed test. Moreover, we are able to construct a regularized kernel to further improve the power of the test. A novel method for choosing the regularization parameter is introduced. We show that the proposed test with regularized kernel could achieve the power of an oracle test if the

regularization parameter is properly chosen.

The rest of the chapter is organized as follows. In Section 2.2, we introduce the expansion of the nonparametric function in the RKHS at data points and the equivalent hypothesis. Section 2.3 proposes a new test statistic and establishes the main asymptotic distribution of the proposed test statistic under the null hypothesis and local alternatives. The kernel selection and regularization are discussed in Section 2.4. The finite sample performance of the proposed test statistic is evaluated by extensive simulations in Section 2.5. In Section 2.6, we apply the proposed method to a Yolkshire gilt data set to identify gene sets that are associated with triiodothyronine level. A brief discussion is given in Section 2.7. All the technical details are relegated to the final section.

2.2 Functional space and equivalent hypothesis

In this chapter, we consider functions $h(\cdot)$ that belong to a functional space $\mathscr{H}_K$ generated by a kernel $K_{\theta_n}(\cdot,\cdot)$ where θ_n are tuning parameters that possibly depend on n. For notation convenience, we suppress n in θ_n in the rest of the chapter. The kernel $K_\theta(x_1, x_2) : R^p \times R^p \to R$ is any symmetric and positive semi-definite function defined on $R^p \times R^p$. A kernel $K_\theta(x_1, x_2)$ is said to be positive semi-definite if the associated kernel matrix $(K_\theta(x_i, x_j))_{i,j=1}^M$ is an $M \times M$ positive semi-definite matrix defined on any M distinct points $x_1, \cdots, x_M \in R^p$. Some commonly used kernel functions include linear kernel $K_\theta(z_1, z_2) = z_1^T z_2/\theta$ and Gaussian kernel $K_\theta(z_1, z_2) = \exp(-\|z_1 - z_2\|^2/\theta)$. More examples of kernel functions could be found in Liu et al. (2007).

The functional space $\mathscr{H}_K$ is determined by the kernel function K_θ. For the purpose of

defining the functional space $\mathscr{H}_K$, we define the following normalized kernel

$$\mathcal{K}_\theta(x_1, x_2) = \frac{K_\theta(x_1, x_2)}{\sqrt{E\{K_\theta(\mathbf{X}_1, \mathbf{X}_1)\}E\{K_\theta(\mathbf{X}_2, \mathbf{X}_2)\}}}$$

where $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent copies of $\mathbf{X}$ with probability measure $\boldsymbol{P}$. It is then obvious to see that $E\{\mathcal{K}_\theta(\mathbf{X}_i, \mathbf{X}_i)\} = 1$ and $\mathcal{K}_\theta(x_1, x_2)$ is still positive semi-definite and symmetric. The above normalization ensures $E\{\mathcal{K}_\theta(\mathbf{X}, \mathbf{X})\} < \infty$ so that the eigen-decomposition of K_θ can be properly defined according to Lemma 1 in the last section. The normalization is needed because $E\{K_\theta(\mathbf{X}, \mathbf{X})\}$ could diverge in the high-dimensional case. For instance, if $K_\theta(\mathbf{X}, \mathbf{X}) = \mathbf{X}^T\mathbf{X}$ and $\mathrm{Var}(\mathbf{X}) = \boldsymbol{\Sigma}$, then $E\{K_\theta(\mathbf{X}, \mathbf{X})\} \geq tr(\boldsymbol{\Sigma})$ which implies that $E\{K_\theta(\mathbf{X}, \mathbf{X})\}$ is at least at the order of p if all the eigenvalues of $\boldsymbol{\Sigma}$ are bounded away from 0.

By Lemma 1 in the last section, we can write $\mathcal{K}_\theta(x_1, x_2) = \sum_{m=1}^{\infty} \lambda_{\mathcal{K}_\theta, m}\psi_{\theta,m}(x_1)\psi_{\theta,m}(x_2)$ where $\{\psi_{\theta,m}(\cdot)\}$ form a complete orthogonal normal system on $L^2(\boldsymbol{P})$. Without causing much confusion, we will use $\lambda_{\mathcal{K},m}$ and $\psi_m(\cdot)$ to denote $\lambda_{\mathcal{K}_\theta,m}$ and $\psi_{\theta,m}(\cdot)$, respectively. Then the space $\mathscr{H}_K$ is defined to be (Poggio and Shelton, 2002)

$$\mathscr{H}_K = \{f(x) : f(x) = \sum_{m=1}^{\infty} \alpha_m\psi_m(x) \text{ for } \alpha_m \text{ with} \sum_m \alpha_m^2/\lambda_{\mathcal{K},m} < \infty\}.$$

Let $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_\infty)$ be the coefficients in the representation of $h(\cdot)$, i.e., $h(\cdot) = \sum_{m=1}^{\infty} \alpha_m\psi_m(\cdot)$. To distinguish H_0 from H_1, one need to find a measure to quantify the distance between $h(\cdot)$ and 0. Here we define a norm $\|\cdot\|_{\mathrm{K}}$ below as a measure,

$$\|h\|_{\mathrm{K}}^2 = \sum_{m=1}^{\infty} \lambda_m\alpha_m^2 \tag{2.2.1}$$

where $\lambda_m = E\{K_\theta(\mathbf{X}, \mathbf{X})\}\lambda_{\mathcal{K},m}$ that may be considered as the eigenvalues of the kernel function $K_\theta(x, y)$. Obviously, the null hypothesis in (2.1.2) is true if and only if $\|h\|^2_{\mathrm{K}} = 0$, and $\|h\|^2_{\mathrm{K}} > 0$ under the alternative hypothesis.

For ease of the presentation, in the rest of this chapter, we consider a centralized kernel K_θ that satisfies $\mu_K = E\{K_\theta(\mathbf{X}_1, \mathbf{X}_2)\} = 0$. The centralized kernel K_θ can be constructed from any positive definite kernel function K^*_θ by setting $K_\theta(\mathbf{x}_1, \mathbf{x}_2) = K^*_\theta(\mathbf{x}_1, \mathbf{x}_2) - K^*_{1,\theta}(\mathbf{x}_1) - K^*_{1,\theta}(\mathbf{x}_2) + \mu_{K^*}$ where $K^*_{1,\theta}(\mathbf{x}_1) = E\{K^*_\theta(\mathbf{x}_1, \mathbf{X}_2)\}$ is the first order projection of K^*_θ. By Lemma 1 in the last section, K_θ is still semi-positive definite with only one zero eigenvalue $\lambda_{m^*} = 0$ corresponding to eigenfunction $\psi_{m^*}(x) = 1$, if K^*_θ is positive definite. Some benefits of a centralized kernel are discussed in Lindsay et al. (2008). The practical construction of a centralized kernel will be discussed in the next section. We will use K_θ and bold font $\mathbf{K}$ to denote, respectively, the kernel function and an $n \times n$ kernel matrix defined by $\mathbf{K} = \big(K_\theta(\mathbf{X}_i, \mathbf{X}_j)\big)_{i,j=1}^n$.

2.3 Asymptotic distributions

By the orthonormal expansion of $\mathcal{K}_\theta(x, y)$ in Section 2.2, we observe that

$$E\{(Y_i - \mu)(Y_j - \mu)K_\theta(\mathbf{X}_i, \mathbf{X}_j)\} = \sum_{m=1}^{\infty} \lambda_m \alpha_m^2 = \|h\|^2_{\mathrm{K}},$$

for any (i, j) pair such that $i \neq j$. Motivated by this observation, we consider the following test statistic

$$T_n = \frac{1}{n(n-1)} \sum_{i \neq j} K_\theta(\mathbf{X}_i, \mathbf{X}_j)(Y_i - \bar{Y}_n)(Y_j - \bar{Y}_n)/\hat{\sigma}^2 \tag{2.3.1}$$

where $\bar{Y}_n = n^{-1}\sum_{i=1}^n Y_i$ is the sample mean and $\hat{\sigma}^2 = (n-1)^{-1}\sum_{i=1}^n (Y_i - \bar{Y}_n)^2$ is the sample variance estimator of σ^2 under the null hypothesis (2.1.2). It then can be checked that $E(T_n) = o(1)$ under the null hypothesis and $E(T_n) = \sum_{m=1}^{\infty} \lambda_m \alpha_m^2/\sigma^2\{1+o(1)\}$ under the alternative. Therefore, the test statistic T_n is able to distinguish the null and alternative hypotheses in (2.1.2).

Define $\tau_k = E(\epsilon^k)$ as the k-th moment of random error ϵ, $V_k = \sum_{m=1}^{\infty} \lambda_m^k$ for integers $k = 1, 2, ...$ The following theorem summarizes the asymptotic distribution of T_n under H_0.

Theorem 1. *Under the null hypothesis H_0 in (2.1.2) (i) Assume $\tau_4 < \infty$. Then $nT_n/V_1 \xrightarrow{d} \sum_{m=1}^{\infty} \lambda_{\mathcal{K},m}(\chi_m^2 - 1)$ where $\lambda_{\mathcal{K},m}$ are the eigenvalues of $\mathcal{K}_\theta$, the normalized kernel of K_θ and χ_m^2 are independent chi-square distributions with 1 degree of freedom; (ii) If we further assume that*

$$V_4/V_2^2 \to 0 \text{ as } p(n) \to \infty, \tag{2.3.2}$$

then $\sigma_{T_n}^{-1} nT_n \xrightarrow{d} N(0,1)$, where $\sigma_{T_n}^2 = 2V_2$.

Remark 1 If the centralized kernel K_θ is unknown and is constructed from a kernel function K_θ^*, it will contain unknown quantities μ_{K^*} and $K_{1,\theta}^*(\mathbf{x}_1)$. Thus, T_n is not directly applicable. In this case, we can replace $K_\theta(X_i, X_j)$ by $K_{n,\theta}(X_i, X_j)$, which is the (i,j) element of $\mathbf{K}_n = \mathbf{K}_\theta^* - (n-1)^{-1}\mathbf{J}(\mathbf{K}_\theta^*)^0 - (n-1)^{-1}(\mathbf{K}_\theta^*)^0\mathbf{J} + n^{-1}(n-1)^{-1}\mathbf{J}(\mathbf{K}_\theta^*)^0\mathbf{J}$. Here $\mathbf{J}$ is an $n \times n$ matrix with all elements as 1 and $\mathbf{A}^0 = (A_{ij}^0)$ is a zero-diagonal matrix with $A_{ij}^0 = A_{ij}$ for $i \neq j$ and $A_{ii}^0 = 0$. Let $\hat{T}_n$ be the test statistic with corresponding kernel $K_{n,\theta}$. It can be shown that $(nT_n - n\hat{T}_n)/\sqrt{V_2} = o_p(1)$ (see the proof of Remark 1 in the last section). Therefore, $n\hat{T}_n/V_1$ has the same limiting distribution as nT_n/V_1.

If condition (2.3.2) holds, then an α level test rejects the null hypothesis if

$$\hat{\sigma}_{T_n}^{-1} n T_n > z_{1-\alpha} \tag{2.3.3}$$

where $z_{1-\alpha}$ is the lower $1-\alpha$ quantile of the standard normal distribution, $\hat{\sigma}_{T_n}^2 = 2(n-1)^{-2}\mathrm{tr}(\mathbf{H}\mathbf{K}_n^0\mathbf{H}\mathbf{K}_n^0)$ is a ratio consistent estimator for $\sigma_{T_n}^2$.

The null distribution of T_n in the first part of Theorem 1 is applicable even if (2.3.2) does not hold. However, since the asymptotic distribution is a weighted sum of chi-square distributions, obtaining accurate estimators for all the eigenvalues $\lambda_{\mathcal{K},m}$ $(m = 1, 2, \cdots,)$ simultaneously is difficult. Nevertheless, one can apply a Satterthwaite approximation to the mixture of chi-squares by a scaled chi-square distribution $\hat{a}\chi_{\hat{g}}^2/\hat{V}_1 - 1$, where $\hat{g} = \hat{V}_1/\hat{a}$, $\hat{a} = \hat{\sigma}_{T_n}^2/(2\hat{V}_1)$ and $\hat{V}_1 = n^{-1}\mathrm{tr}(\mathbf{H}\mathbf{K}_n)$ is a consistent estimator of V_1. Then an asymptotic α level test rejects the null hypothesis if

$$(nT_n + \hat{V}_1)/\hat{a} > \chi_{\hat{g},1-\alpha}^2 \tag{2.3.4}$$

where $\chi_{g,1-\alpha}^2$ is the $1-\alpha$ quantile of a chi-square distribution with g degrees of freedom.

To achieve better accuracy in size approximation, we provide an adjustment to the variance estimator $\hat{\sigma}_{T_n}^2$ using the high order moments of ϵ in (2.1.1). The adjusted variance estimator $\hat{\sigma}_{T_n,adj}^2$ was used to replace the estimator $\hat{\sigma}_{T_n}^2$ in the simulation study in Section 2.5 and real data analysis in Section 2.6. Assume the density function of ϵ is symmetric around 0. The adjusted variance estimator $\hat{\sigma}_{T_n,adj}^2$ is

$$\hat{\sigma}_{T_n,adj}^2 = \frac{1}{n^2}\Big\{(2 - \frac{12}{n^2} + \frac{6\hat{\Delta}}{n})\mathrm{tr}(\mathbf{H}\mathbf{K}_n^0\mathbf{H}\mathbf{K}_n^0) - (\frac{2}{n} + \frac{\hat{\Delta}}{n})\mathrm{tr}^2(\mathbf{H}\mathbf{K}_n^0) + \hat{\Delta}\mathrm{tr}(\mathbf{A}\circ\mathbf{A})\Big\},$$

where $\circ$ denotes the Hadamard product, $\mathbf{A} = \mathbf{H}\mathbf{K}_n^0\mathbf{H}$, and $\hat{\Delta} = n^{-1}\sum_{i=1}^n[(Y_i - \bar{Y}_n)/\hat{\sigma}]^4 - 3$. The derivation of $\sigma^2_{T_n,adj}$ is provided in the last section.

The next theorem studies the asymptotic distribution of the test statistic T_n under a sequence of local alternative hypotheses

$$H_{1n} : h(\mathbf{x}) = d_n(\mathbf{x}) \tag{2.3.5}$$

where $d_n(\mathbf{x})$ is any unknown function that possibly depends on n. For the purpose of model identification, we assume $\mathrm{E}\big\{d_n(\mathbf{X})\big\} = 0$. As usual, we consider local alternatives that are close to the null hypothesis, which is more challenging to be detected than fixed alternatives. More specifically, assume that $d_n(\cdot)$ satisfies the following two conditions

$$n\delta_K = O(\sqrt{V_2}) \text{ and } n^2\mathrm{E}\{d_n^8(\mathbf{X})\} = o(V_2^2/V_1^4) \tag{2.3.6}$$

where $\delta_K = \mathrm{E}\left\{K_\theta(\mathbf{X}_1, \mathbf{X}_2)d_n(\mathbf{X}_1)d_n(\mathbf{X}_2)\right\}$.

Theorem 2. *Assume that $E\{\psi_m^4(\mathbf{X})\} < \infty$ for all integers m. Under the local alternative hypothesis H_{1n} in (2.3.5), we have*

$$V_1^{-1}\Big(nT_n - \sigma_{T_n}\Psi(d_n)\Big) \xrightarrow{d} \sum_{m=1}^{\infty} \lambda_{\mathcal{K},m}(\chi^2_m - 1),$$

where $\Psi(d_n) = n\delta_K/(\sigma^2\sigma_{T_n})$ is the signal-to-noise ratio (SNR). Moreover, if (2.3.2) holds, then $\sigma_{T_n}^{-1}nT_n - \Psi(d_n) \xrightarrow{d} N(0,1)$.

Applying Theorem 2, the power of an α-level test for the rejection region in (2.3.3) under the local alternative (2.3.5) is $\Omega(d_n) = 1 - \Phi\Big(z_{1-\alpha} - \Psi(d_n)\Big)$, where $\Phi(\cdot)$ is the CDF for

standard normal distribution. Therefore, the power of the proposed test is determined by the SNR $\Psi(d_n)$. If the α-level rejection region in (2.3.4) is used, the power of the test is $\Omega(d_n) = P\big(\chi^2_g > \chi^2_{g,1-\alpha} - \sigma_{T_n}\Psi(d_n)/a\big)$.

Let $d_n(\mathbf{x}) = b_n\Delta_n(\mathbf{x})$ such that $\mathrm{E}\left\{K_\theta(\mathbf{X}_1, \mathbf{X}_2)\Delta_n(\mathbf{X}_1)\Delta_n(\mathbf{X}_2)\right\}$ is a constant. Then the proposed test has a non-trivial power if $b_n = V_2^{1/4}/\sqrt{n}$. If V_2 is a constant, then the proposed test is able to detect alternatives of order $1/\sqrt{n}$. In the high-dimensional case, however, if $V_2 \to \infty$ at certain rate, the proposed test is only able to detect alternatives of order $V_2^{1/4}/\sqrt{n}$, which is larger than the order $1/\sqrt{n}$. This reveals an adverse effect of dimensionality on the test. We can also observe that as long as $V_2 = o(n^2)$, the proposed test is consistent so that the power of the test converges to 1. This implicitly imposes condition on p and n since $V_2 = E\{K^2_\theta(\mathbf{X}_1, \mathbf{X}_2)\}$ depends on p. For example, if K_θ is a linear kernel and $\mathbf{\Sigma}$ has all the eigenvalues bounded, then $V_2 \sim p$. In this case, the proposed test is consistent if $p = o(n^2)$ for functions $h(x) = d_n(x)$. To further improve the power, we consider the choice of kernel function and the construction of a regularized kernel in the next section.

2.4 Kernel selection and regularization

2.4.1 Kernel selection

In Sections 2.4.1 and 2.4.2, we assume that the kernel K which generates the functional space $\mathscr{H}_K$ is known in reality. However, the functional space $\mathscr{H}_K$ is typically unknown. Therefore, an important question in practice is on how to select kernels. Kernel selection problem has been studied for Fisher discriminant analysis (Kim et al., 2006) and semi-supervise learning (Dai and Yeung, 2007). However, no kernel selection method is tailored to the hypothesis testing problem (Liu et al., 2007).

We propose to select kernels by maximizing the SNR of the proposed test. To emphasize the role of the kernel K_θ, we write $\Psi(d_n)$ as $\Psi_{K_\theta}(d_n)$. Given a family of candidate kernels $\mathscr{F}_{\mathbb{K}}$, the kernel $\mathbb{K}_\theta$ may be selected by maximizing the SNR as follows

$$\hat{\mathbb{K}}_\theta = \arg \max_{\mathbb{K}_\theta \in \mathscr{F}_{\mathbb{K}}} \hat{\Psi}_{\mathbb{K}_\theta}(d_n) \tag{2.4.1}$$

where $\hat{\Psi}_{\mathbb{K}_\theta}(d_n)$ is an estimator of $\Psi_{\mathbb{K}_\theta}(d_n) = n\delta_{\mathbb{K}}/(\sigma^2\sigma_{T_n})$. For a candidate kernel $\mathbb{K}_\theta \in \mathscr{F}_{\mathbb{K}}$, the unknown parameters $\delta_{\mathbb{K}}, \sigma_{T_n}$ and σ^2 can be substituted using estimators, $\hat{\delta}_{\mathbb{K}}(d_n) = \frac{1}{n(n-1)} \sum_{i \neq j} \mathbb{K}_\theta(\mathbf{X}_i, \mathbf{X}_j)(Y_i - \bar{Y}_n)(Y_j - \bar{Y}_n)$, $\hat{\sigma}^2 = (n-1)^{-1} \sum_{i=1}^n (Y_i - \bar{Y}_n)^2$ and $\hat{\sigma}^2_{T_n}$, respectively. The following Proposition 1 shows that δ_K and σ^2 can be consistently estimated, hence the SNR can be consistently estimated.

Proposition 1. *Under condition (2.3.6),* $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$ *and* $\sigma_{T_n}^{-1}(\hat{\delta}_K - \delta_K) \xrightarrow{p} 0$.

The proof of proposition 1 is given in the the last section.

2.4.2 Kernel regularization

In this section, we show that the power of the proposed test could be further improved by using a regularized kernel. The power function is determined by the SNR $\Psi(d_n)$, which can be written as follows

$$\Psi(d_n) = n \sum_{m=1}^{\infty} \lambda_m a_m^2 / (\sigma^2 \sigma_{T_n})$$

where $a_m = \mathrm{E}\{d_n(\mathbf{X})\psi_m(\mathbf{X})\}$ is the projection of $d_n(\mathbf{X})$ onto the m-th eigenfunction $\psi_m(\mathbf{X})$ of K_θ. We observe that the numerator of $\Psi(d_n)$ (the signal part) is determined by the magnitude of eigenvalues λ_m and the projections a_m. For a given set of eigenfunctions $\{\psi_m(\mathbf{X})\}_{m=1}^{\infty}$ and a function $d_n(x)$, the projections a_m are fixed. To enlarge the numerator

of $\Psi(d_n)$, one could adjust the eigenvalues λ_m associated with projection a_m so that larger non-zero projections receive higher weights than small projections.

To adjust the eigenvalues of the kernel without changing the eigenfunctional space, we introduce a regularized kernel in the following. For any centralized kernel matrix $\mathbf{K}$, define the regularized kernel matrix $\mathbf{K}_{R,\gamma}$ as

$$\mathbf{K}_{R,\gamma} = \mathbf{K} - \mathbf{K}(n\gamma\mathbf{I} + \mathbf{K})^{-1}\mathbf{K}, \tag{2.4.2}$$

whose similar version in a two-sample problem was discussed in Harchaoui et al. (2006). Let $K_{R,\gamma}$ be the kernel function corresponding to the kernel matrix $\mathbf{K}_{R,\gamma}$. It can be proved (see Lemma 3 in the last section) that the eigenfunctions of kernel function $K_{R,\gamma}$ are still $\{\psi_m(\mathbf{X})\}_{m=1}^{\infty}$, which are the same as that of K_θ. However, the corresponding eigenvalues of $K_{R,\gamma}$ are $\{\gamma\lambda_m/(\lambda_m+\gamma)\}_{m=1}^{\infty}$.

We now show that how a regularized kernel $K_{R,\gamma}$ could improve the power of the proposed test. To see the point, we compare the SNRs $\Psi(d_n)$ and $\Psi_R(d_n,\gamma)$ corresponding to the kernels K_θ and $K_{R,\gamma}$ respectively. Let $C_n = n/(\sqrt{2}\sigma^2)$. Then we have

$$\Psi(d_n) = C_n\frac{\sum_{m=1}^{\infty}\lambda_m a_m^2}{\sqrt{\sum_{m=1}^{\infty}\lambda_m^2}} \quad \text{and} \quad \Psi_R(d_n,\gamma) = C_n\frac{\sum_{m=1}^{\infty}\lambda_m a_m^2/(\lambda_m+\gamma)}{\sqrt{\sum_{m=1}^{\infty}\lambda_m^2/(\lambda_m+\gamma)^2}}. \tag{2.4.3}$$

By comparing the above two expressions, we see that $\sup_\gamma \Psi_R(d_n,\gamma) \geq \Psi(d_n)$. Because we observe that

$$\Psi_R(d_n,\gamma) = C_n\frac{\sum_{m=1}^{\infty}\lambda_m a_m^2/(\lambda_m/\gamma+1)}{\sqrt{\sum_{m=1}^{\infty}\lambda_m^2/(\lambda_m/\gamma+1)^2}} \to \Psi(d_n) \quad \text{as} \quad \gamma\to\infty,$$

the regularized kernel $K_{R,\gamma}$ is the same as the non-regularized kernel K if $\gamma \to \infty$. Thus, the introduction of the regularization parameter γ allows us to strike a balance between the numerator and denominator so that $\Psi_R(d_n, \gamma)$ is larger than $\Psi(d_n)$ for some γ.

To select the best regularization parameter γ, it is natural to consider maximizing the SNR $\Psi_R(d_n, \gamma)$. That is $\hat{\gamma} = \arg\max_{\gamma \in \mathbb{S}} \hat{\Psi}_R(d_n, \gamma)$, where $\mathbb{S} = \{s_1, ..., s_B\}$ is a set of positive candidate regularization parameters ordered in an increasing order. It may be noted that the denominator of $\Psi_R(d_n, \gamma)$ in (2.4.3) goes to infinity and the numerator of SNR in (2.4.3) increases as $\gamma \to 0$. A reasonable estimate for the numerator of (2.4.3) should be non-decreasing as $\gamma \to 0$. However, the numerator may not be well-estimated if the sample size is small. We therefore propose a modification to the above approach. Let $s_l^* \in \mathbb{S}$ be the smallest regularization parameter in $\mathbb{S}$ such that $\hat{\delta}_{\mathbb{K},\gamma}(d_n)$, the numerator of $\Psi_R(d_n, \gamma)$, achieves its maximum value in $\mathbb{S}$. We then put our attention to the tuning parameters that are larger than s_l^* in the set of $\mathbb{S}$. Given the samples, we can find the optimal tuning parameter by maximizing the following criterion

$$\hat{\gamma} = \arg \max_{\gamma \in \{s_l^*, ..., s_B\}} \hat{\Psi}_R(d_n, \gamma). \tag{2.4.4}$$

For the stability selection consideration, we propose the following procedure to select the tuning parameter γ

1. Randomly divide the sample $\{Y_i, X_i\}_{i=1}^n$ into L parts with the same sample size.

2. We drop the lth ($l = 1, 2..., L$) part of the sample, select the tuning parameter $\hat{\gamma}_l$ using the remaining $L - 1$ parts of sample based on the criterion (2.4.4).

3. Repeat step 2 for $l = 1, ..., L$. The stabilized tuning parameter is defined as $\tilde{\gamma} =$

median$\{\hat{\gamma}_1, ..., \hat{\gamma}_L\}$.

Simulation studies in Section 2.5 demonstrate that the above tuning parameter selection method works well in practice. Based on our experience in simulation, one could choose L between 4 and 8.

The regularization is most effective in the "sparse" case where the non-zero projections reside only in the first N coordinates corresponding to the N largest eigenvalues. To appreciate that, we hereafter consider the setting where $\lambda_m = c_m \lambda_1$ and $\{c_m\}_{m=1}^{\infty}$ is a decreasing sequence satisfying $c_1 = 1$. Let $\{a_m^2 \sim B_p, m \in S_1\}$ be the set of non-zero projections whose squares are of the same order as B_p and S_1 is a subset of $\{1, \cdots, N\}$. Here $a \sim b$ means that a and b are of the same order.

To show the effectiveness of regularization, we compare the SNR $\Psi_R(d_n, \gamma)$ to an "oracle" SNR $\Psi_R^O(d_n, \gamma)$ using regularized kernel. The oracle SNR is an ideal SNR which eliminates all the coordinates with zero projections. The oracle SNR is used for comparison purpose but it cannot be realized by any test procedure in practice. The oracle SNR $\Psi_R^O(d_n, \gamma)$ is defined as

$$\Psi_R^O(d_n, \gamma) \;=\; C_n \cdot \frac{\sum_{m \in S_1} \lambda_m a_m^2/(\lambda_m + \gamma)}{\sqrt{\sum_{m \in S_1} \lambda_m^2/(\lambda_m + \gamma)^2}}.$$

The following theorem provides the maximum orders of $\Psi_R^O(d_n, \gamma)$ and $\Psi_R(d_n, \gamma)$.

Theorem 3. *Let $|S_1|$ be the cardinality of signal set S_1. Assume that the regularization parameter γ^* satisfies $\gamma^* = o(\lambda_N), \gamma^* = O(\lambda_{N_1}), \lambda_{N_2} = o(\gamma^*)$, and $R_2/N\gamma^{*2} = o(1)$ where $N_1 = [N \log\log N]$, $N_2 = [N \log N]$, and $R_2 = \sum_{m=N_2}^{\infty} \lambda_m^2$. Then (i) $\max_\gamma \Psi_R^O(d_n, \gamma) \sim \Psi_R^O(d_n, \gamma^*)$ and both at the order $\sqrt{|S_1|} C_n B_p$ for large p; (ii) there exist constants J_0, J_1*

and J_2 such that, for large p,

$$\frac{J_1|S_1|C_nB_p}{\sqrt{N\log N}} \leq \Psi_R(d_n,\gamma^*) \leq \frac{J_2|S_1|C_nB_p}{\sqrt{N\{1+J_0\log N(c_{N_2}/c_{N_1})^2\}}}.$$

From Theorem 3, if $|S_1| \sim N$, we have

$$\frac{J_1\sqrt{|S_1|}C_nB_p}{\sqrt{\log|S_1|}} \leq \Psi_R(d_n,\gamma^*) \leq \frac{J_2\sqrt{|S_1|}C_nB_p}{\sqrt{1+J_0\log|S_1|(c_{N_2}/c_{N_1})^2}}.$$

Therefore, the SNR $\Psi_R(d_n,\gamma^*)$ of the proposed test with regularized kernel can attain the SNR $\Psi_R^O(d_n,\gamma^*)$ of the oracle test within a factor of a slowly varying function $\log(N)$.

The above regularization could enhance the dimensionality that the proposed test could handle. Recall the local alternatives considered in Theorem 2. Let $d_n(\mathbf{x}) = b_{R,n}(\gamma^*)\Delta_n(\mathbf{x})$ where $\Delta_n(\mathbf{x})$ is a function such that $\mathrm{E}\{K_\theta(\mathbf{X}_1,\mathbf{X}_2)\Delta_n(\mathbf{X}_1)\Delta_n(\mathbf{X}_2)\}$ is a constant. Using the regularized kernel with regularization parameter γ^*, the proposed test has a non-trivial power if $b_{R,n}(\gamma^*)$ is at the order $b_{R,n}(\gamma^*) = \frac{V_2^{1/4}}{\sqrt{n}}\rho^{1/4}(\gamma^*)$ where

$$\rho(\gamma^*) = \Big(\frac{\sum_{m=1}^{\infty}\lambda_m^2/(\lambda_m/\gamma^*+1)^2}{\{\sum_{m\in S_1}\lambda_m a_m^2/(\lambda_m/\gamma^*+1)\}^2}\Big)\Big(\frac{V_2}{\{\sum_{m\in S_1}\lambda_m a_m^2\}^2}\Big)^{-1}.$$

Assume γ^* satisfies the conditions in Theorem 3. Then we have

$$\rho(\gamma^*) \sim \frac{N}{|S_1|^2}\cdot(\sum_{m\in S_1} c_m)^2.$$

If $|S_1| \sim N$ and $c_m = m^{-\alpha}$ for $\alpha > 1/2$, then $\rho(\gamma^*) = O(N^{-\min\{2\alpha-1,1\}}) = o(1)$. This means that the smallest detectable order using a regularized kernel is smaller than that of a unregularized kernel. The improvement is significant when N is large and $\alpha > 1$. Moreover,

the test is consistent if $V_2 = o\{n^2\rho^{-1}(\gamma^*)\}$. Comparing to the unregularized case which requires $V_2 = o\{n^2\}$, the regularized kernel is powerful for higher dimensional functions since $\rho(\gamma^*) \to 0$.

2.5 Simulation study

The simulation studies were designed to evaluate the finite sample performance of the proposed test, kernel selection and regularization methods. We simulated IID samples $\{\mathbf{X}_i, Y_i\}_{i=1}^n$ from the following model

$$Y_i = \mu + h(\mathbf{X}_i) + \epsilon_i \quad i = 1, \cdots, n \tag{2.5.1}$$

where the random error ϵ_is were chosen to be $N(0,1)$ or $Laplace(0, \sqrt{2}/2)$ distribution. To generate the covariates $\mathbf{X}$, we first generated a p-dimensional normally distributed random vector $\mathbf{Z}$ with mean 0 and covariance $\mathbf{\Sigma} = (0.6^{|i-j|})_{i,j=1}^p$. Then we obtained the covariates $\mathbf{X} = (X_1, \cdots, X_p)^T$ by setting the j-th component by $X_j = F_{nj}(Z_j)$ for $j = 1, \cdots, p$. Here F_{nj} is the empirical cumulative distribution of j-th component given by $F_{nj}(z) = n^{-1}\sum_{i=1}^n I(Z_{ij} \leq z)$. We considered two settings regarding the relationship between n and p: (i) $p < n$ and (ii) $p >> n$ with $n = 40, 60$ and 100. Specifically, $p = (3, 5, 10)$ in setting (i), $p = (150, 200, 250)$ in setting (ii).

We wish to test $H_0 : h(\cdot) = 0$. To assess the empirical size of the proposed test, we chose $h(\mathbf{x}) = 0$ under H_0. To evaluate the empirical power, we chose $h(\mathbf{x}) = h_L(\mathbf{x}) - E(h_L)$ in

setting (i) and $h(\mathbf{x}) = h_H(\mathbf{x}) - E(h_H)$ in setting (ii), where

$$h_L(\mathbf{x}) = c_1(x_1 + x_2 - x_3) + c_2\{\exp(-x_2^2)H_2(x_2) + \exp(-x_3^2)H_5(x_3)\} + c_3\{x_1x_3 + \cos(x_3^2)\},$$
$$h_H(\mathbf{x}) = c_1\sum_{k=1}^{100}(-1)^k x_k + c_2\sum_{k=1}^{100}\{\exp(-x_k^2/p)H_2(x_k/p)\} + c_3\{x_1x_3 + \cos(x_3^2)\}$$

where $H_k(\cdot)$ is the kth order Hermite polynomial and c_1, c_2 and c_3 are constants specified below. For each setting, we designed two scenarios $\mathcal{S}_k$ with different values of c_1, c_2 and c_3 for each setting. Specifically, in setting (i), we used $\mathcal{S}_1 = \{c_1 = 0.002, c_2 = 0.2, c_3 = 0.002\}$ and $\mathcal{S}_2 = \{c_1 = 1.2, c_2 = 0.012, c_3 = 0.012\}$ and, in setting (ii), we chose $\mathcal{S}_3 = \{c_1 = 0.01, c_2 = 10, c_3 = 0.01\}$ and $\mathcal{S}_4 = \{c_1 = 100u, c_2 = 0.1u, c_3 = 0.1u, u = 0.0015\}$. In scenarios $\mathcal{S}_1$ and $\mathcal{S}_3$, c_2 are chosen to be much larger than c_1 such that the non-linear parts dominate the functions while in $\mathcal{S}_2$ and $\mathcal{S}_4$, c_1 are much larger than c_2 so that the linear parts dominate.

All the results for evaluating empirical power are based on 1000 simulation replicates and that for empirical size are based on 5000 simulation replicates. Three types of commonly used kernels were compared in all the simulations: linear kernel $K_L(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T\mathbf{y}/\theta$, Gaussian kernel $K_G(\mathbf{x}, \mathbf{y}) = \exp\{-\|\mathbf{x}-\mathbf{y}\|^2/\theta\}$ and the exponential kernel $K_E(\mathbf{x}, \mathbf{y}) = \exp\{-(\|\mathbf{x}\|^2 + 3\|\mathbf{x} - \mathbf{y}\|^2 + \|\mathbf{y}\|^2)/\theta\}$, where tuning parameter θ was set to be p.

To illustrate the finite sample null distribution, the histograms of the standardized test statistic under the exponential kernel were plotted in Figures 2.1 and 2.2. The corresponding asymptotic chi-square approximations are plotted as density curves. It can be seen that the approximation performs reasonably well for various dimensions. Table 2.1 summarizes the empirical size of the proposed test with normally and Lapalce distributed errors at the nominal level 5%. We can see that the empirical size of the proposed test was reasonably controlled at the nominal level for all three types of kernels and different error distributions.

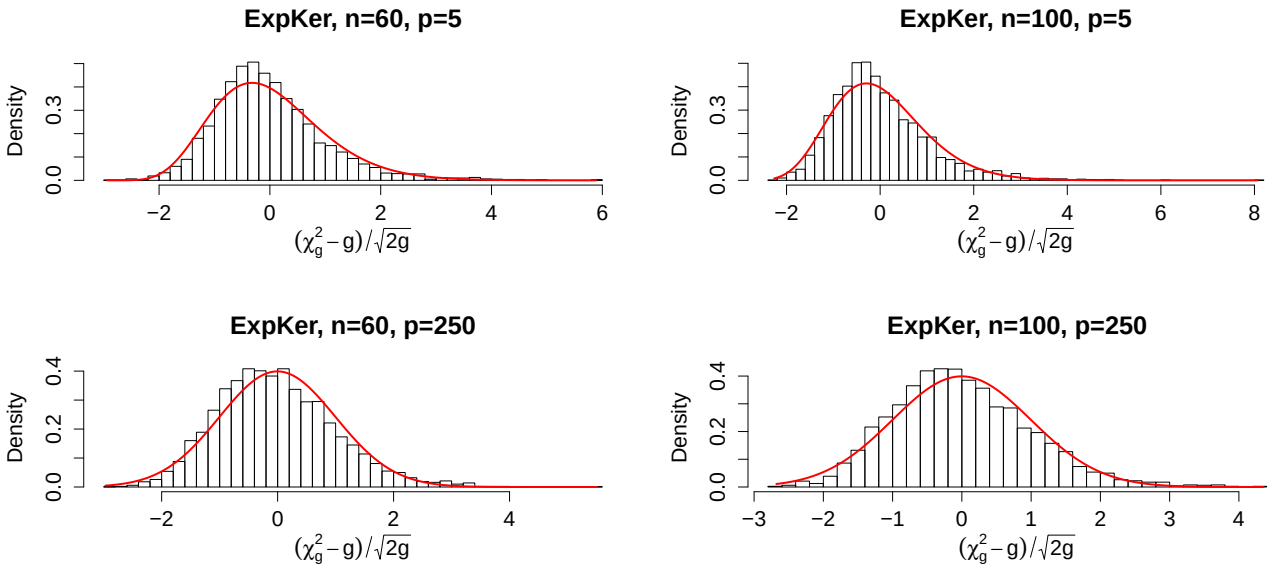

Figure 2.1 The null distribution of the standardized test statistic $(\chi^2_{\hat{g}} - \hat{g})/\sqrt{2\hat{g}}$ with Gaussian error ϵ_i and independent covariants, using centralized exponential kernel K_E, where $\chi^2_{\hat{g}} = (nT_n + \hat{V}_1)/\hat{a}$. The corresponding asymptotic chi-squared approximation are plotted as density curves.

Table 2.2 and 2.3, respectively, contain the empirical power of the proposed test for scenarios $\mathcal{S}_1$ and $\mathcal{S}_2$ under the setting (i). Several observations are given below. (1) There is a clear difference in power among the three types of kernels K_E, K_G and K_L, especially when p and n are relatively small. The power difference was especially striking in Table 2.2 for scenarios $\mathcal{S}_1$. The power based on the exponential and Gaussian kernels were both higher than that using the linear kernel. This is understandable since the non-linear parts dominate the function $h_L(\mathbf{x})$ in scenarios $\mathcal{S}_1$ and exponential kernel and Gaussian kernel contain richer non-linear eigenfunctions than that of the linear kernel, which can capture more information of non-linear functions; (2) The power increased as the sample size increased in all the cases; and (3) The proposed test was very robust to the change of error distributions. Similar patterns can be observed from Table 2.4 and 2.5 under setting (ii).

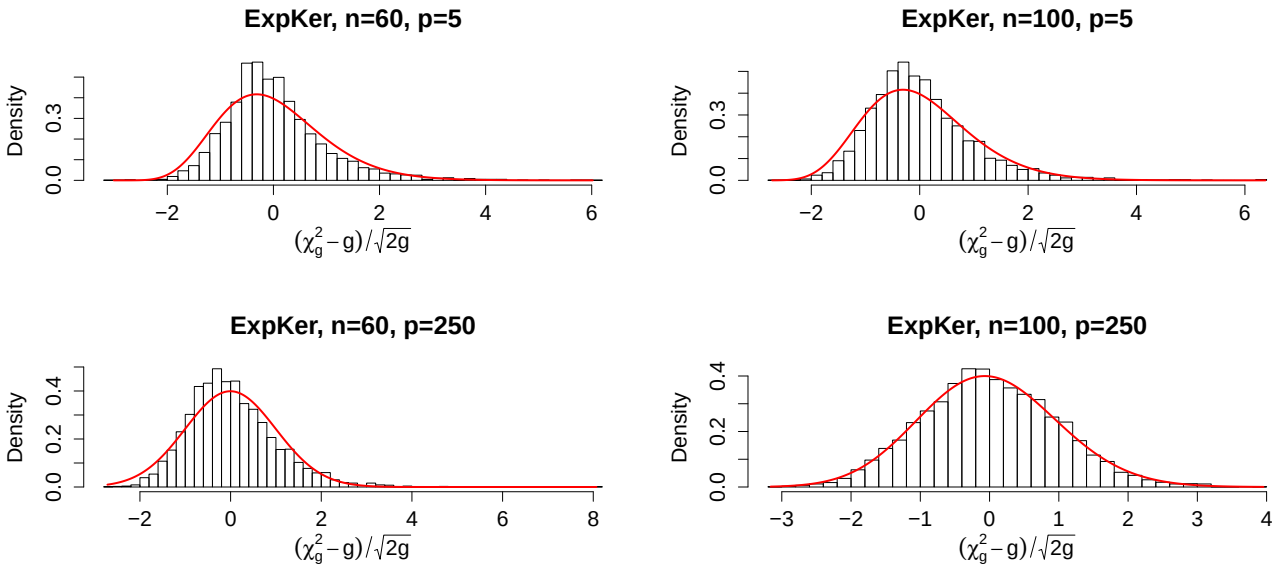

Figure 2.2 The null distribution of the standardized test statistic $(\chi^2_{\hat{g}} - \hat{g})/\sqrt{2\hat{g}}$ with Laplace error ϵ_i and dependent covariants, using centralized exponential kernel K_E, where $\chi^2_{\hat{g}} = (nT_n + \hat{V}_1)/\hat{a}$. The corresponding asymptotic chi-squared approximation are plotted as density curves.

2.5.1 Kernel selection

We observed from Table 2.2-2.5 that the empirical power of the test corresponding to different kernels could be very different. This naturally motivated us to select a kernel to improve the performance of the test. We applied the kernel selection method proposed in Section 4.1 to choose the optimal kernel among K_E, K_G and K_L for each simulation replicate.

We reported the percentage of each kernel being selected in 1000 simulation replicates among three candidate kernels K_E, K_G and K_L. We can see that almost in all cases in Table 2.2 and 2.4, the kernel selection method could choose the kernel with the highest power. This shows that the proposed kernel selection method worked very well in selecting

Table 2.1 Empirical size of the proposed test for Gaussian and Laplace errors with dependent covariates using different kernels

		Gaussian Error			Laplace Error		
n	p	K_E	K_L	K_G	K_E	K_L	K_G
40	3	0.058	0.055	0.056	0.051	0.046	0.047
	5	0.060	0.058	0.059	0.049	0.048	0.048
	10	0.060	0.062	0.062	0.047	0.044	0.046
	150	0.059	0.058	0.059	0.042	0.042	0.040
	200	0.064	0.064	0.064	0.047	0.047	0.047
	250	0.057	0.055	0.056	0.049	0.047	0.047
60	3	0.055	0.054	0.056	0.046	0.046	0.046
	5	0.058	0.055	0.056	0.053	0.054	0.055
	10	0.056	0.054	0.056	0.046	0.048	0.048
	150	0.054	0.053	0.053	0.049	0.047	0.048
	200	0.057	0.059	0.058	0.041	0.040	0.041
	250	0.060	0.059	0.059	0.049	0.047	0.048
100	3	0.053	0.054	0.056	0.048	0.046	0.047
	5	0.053	0.053	0.053	0.043	0.040	0.041
	10	0.055	0.053	0.054	0.049	0.045	0.046
	150	0.053	0.054	0.054	0.050	0.050	0.050
	200	0.055	0.054	0.054	0.054	0.052	0.053
	250	0.054	0.054	0.055	0.049	0.048	0.049

the optimal kernel. When the power among different kernels were similar, the percentages were evenly distributed among three kernels. To further confirm the validity of the proposed kernel selection method, for each simulation replicate, we estimated the theoretical power of the test using (2.4.1) for each kernel K_E, K_L and K_G. In Tables 2.2-2.5, we reported the mean of the estimated power for three kernels based on 1000 simulation replicates. We observed that, the estimated power was very close to the empirical power. In summary, the proposed kernel selection method is reliable for practical use.

2.5.2 Regularization

To show the impact of kernel regularization on power improvement, we generated data according to model (2.5.1) with random error ϵ following a Laplace distribution and the covariates $\mathbf{X}_i$ were IID random vectors with independently Uniform (0,1) components. The

Table 2.2 Empirical power of the proposed test for Gaussian and Laplace errors with dependent covariates using different kernels in the setting of $p < n$ under scenario $\mathcal{S}_1$. The estimated theoretical power is given in the parenthesis, and the percentage of a kernel being selected among the three candidate kernels using the proposed kernel selection method is displayed underneath it.

		Gaussian Error			Laplace Error		
n	p	K_E	K_L	K_G	K_E	K_L	K_G
40	3	0.726 (0.655)	0.105 (0.173)	0.296 (0.374)	0.760 (0.675)	0.117 (0.190)	0.320 (0.392)
		(96.8%)	(0.0%)	(3.2%)	(95.9%)	(0.0%)	(4.1%)
	5	0.270 (0.348)	0.099 (0.167)	0.152(0.241)	0.262 (0.340)	0.108 (0.175)	0.155 (0.241)
		(79.5%)	(0.0%)	(20.5%)	(84.2%)	(0.2%)	(15.6%)
	10	0.149 (0.218)	0.096 (0.160)	0.115 (0.186)	0.158 (0.231)	0.104 (0.168)	0.136 (0.200)
		(63.1%)	(1.7%)	(35.2%)	(64.1%)	(1.9%)	(34.0%)
60	3	0.992 (0.934)	0.126 (0.213)	0.704 (0.597)	0.990 (0.929)	0.139 (0.222)	0.708 (0.609)
		(99.5%)	(0.0%)	(0.5%)	(98.6%)	(0.0%)	(1.4%)
	5	0.515 (0.514)	0.117 (0.193)	0.241 (0.322)	0.523 (0.524)	0.124 (0.196)	0.249 (0.330)
		(90.9%)	(0.0%)	(9.1%)	(93.2%)	(0.0%)	(6.8%)
	10	0.185 (0.264)	0.098 (0.170)	0.124 (0.209)	0.196 (0.270)	0.109 (0.177)	0.142 (0.209)
		(72.0%)	(0.4%)	(27.6%)	(70.5%)	(0.2%)	(29.3%)
100	3	1.000 (1.000)	0.286 (0.359)	1.000 (0.988)	1.000 (1.000)	0.303 (0.367)	1.000 (1.000)
		(100.0%)	(0.0%)	(0.0%)	(100.0%)	(0.0%)	(0.0%)
	5	0.982 (0.888)	0.178 (0.266)	0.646 (0.577)	0.958 (0.871)	0.129 (0.272)	0.630 (0.578)
		(97.9%)	(0.0%)	(2.1%)	(97.8%)	(0.0%)	(2.2%)
	10	0.365 (0.420)	0.150 (0.218)	0.222 (0.300)	0.347 (0.404)	0.136 (0.207)	0.194 (0.285)
		(81.2%)	(0.1%)	(18.7%)	(80.1%)	(0.0%)	(19.9%)

function $h(\mathbf{x})$ was chosen to be 0 under H_0. Under the alternative, we chose $h(\mathbf{x}) = h_H(\mathbf{x})$ with constants c_1, c_2 and c_3 being set according to the scenario $\mathcal{S}_3$. In this simulation, the sample size was $n = 60$ and data dimension was $p = 200$. All the simulation results reported in this part are based on 1000 simulation replicates.

For each kernel K_E, K_L and K_G, we constructed the regularized kernels with regularization parameter γ using (2.4.2). We selected a sequence of regularization parameters of different orders ($\gamma = 10^{-a}/n$, $a \in (-5, 2)$) to check their effect on testing power. For each of the regularization parameter, we constructed the corresponding regularized test statistic and applied the test, respectively, to data generated under H_0 and H_1.

Figure 2.3 shows the empirical power and size of the proposed test using regularized kernel

Table 2.3 Empirical power of the proposed test for Gaussian and Laplace errors with dependent covariates using different kernels in the setting of $p < n$ under scenario $\mathcal{S}_2$. The estimated theoretical power is given in the parenthesis, and the percentage of a kernel being selected among the three candidate kernels using the proposed kernel selection method is displayed underneath it.

		Gaussian Error			Laplace Error		
n	p	K_E	K_L	K_G	K_E	K_L	K_G
40	3	0.605 (0.605)	0.621 (0.610)	0.634 (0.623)	0.602 (0.610)	0.610 (0.605)	0.621 (0.618)
		(37.5%)	(32.3%)	(30.2%)	(37.9%)	(35.5%)	(26.6%)
	5	0.496 (0.495)	0.494 (0.498)	0.507 (0.511)	0.434 (0.462)	0.450 (0.465)	0.463 (0.478)
		(38.6%)	(30.2%)	(31.2%)	(37.7%)	(29.6%)	(32.7%)
	10	0.338 (0.375)	0.329 (0.368)	0.346 (0.379)	0.298 (0.353)	0.315 (0.355)	0.329 (0.364)
		(39.7%)	(33.3%)	(27.0%)	(36.6%)	(32.7%)	(30.7%)
60	3	0.793 (0.782)	0.790 (0.776)	0.803 (0.788)	0.786 (0.774)	0.781 (0.771)	0.791 (0.783)
		(38.5%)	(29.4%)	(32.1%)	(35.6%)	(29.6%)	(34.8%)
	5	0.668 (0.656)	0.694 (0.663)	0.705 (0.678)	0.698 (0.683)	0.703 (0.685)	0.725 (0.700)
		(35.1%)	(24.3%)	(40.6%)	(33.9%)	(25.0%)	(41.1%)
	10	0.492 (0.501)	0.499 (0.507)	0.517 (0.520)	0.498 (0.505)	0.506 (0.506)	0.518 (0.519)
		(35.0%)	(29.9%)	(35.1%)	(35.0%)	(30.7%)	(34.3%)
100	3	0.971 (0.956)	0.969 (0.956)	0.974 (0.961)	0.968 (0.960)	0.965 (0.960)	0.968 (0.964)
		(39.4%)	(23.2%)	(37.4%)	(35.0%)	(26.2%)	(38.8%)
	5	0.921 (0.897)	0.928 (0.905)	0.932 (0.912)	0.896 (0.877)	0.897 (0.876)	0.907 (0.886)
		(33.2%)	(18.2%)	(48.6%)	(34.4%)	(18.2%)	(47.4%)
	10	0.786 (0.756)	0.800 (0.766)	0.810 (0.777)	0.781 (0.750)	0.780 (0.757)	0.794 (0.769)
		(31.4%)	(23.3%)	(45.3%)	(32.7%)	(20.1%)	(47.2%)

$K_{R,\gamma}$. The x-axis represents the $-\log_{10}(\gamma)$ and y-axis is the empirical power or size. The power with large regularization parameters γ was not displayed in the graph for a better view for small γ range. When γ is large ($-\log_{10}(\gamma) \in (-3.222, 1.778)$, not shown in Figure 2.3), the power of the test was initially the same as the one using non-regularized kernels (0.769, 0.674, and 0.672 for K_E, K_G and K_L), and then started to grow slowly. As for $-\log_{10}\gamma \in (1.778, 3.778)$, the power peak (0.810, 0.720 and 0.710, for K_E, K_G and K_L respectively) of the proposed test can be observed for all the three kernels. It can be seen from Figure 2.3 that the empirical size of the regularized test was all reasonably controlled.

To evaluate the method for selecting regularization parameters proposed in Section 4.2, we also marked the regularization parameter selection results in Figure 2.3. The three

Table 2.4 Empirical power of the proposed test for Gaussian and Laplace errors with dependent covariates using different kernels in the setting of $p >> n$ under scenario $\mathcal{S}_3$. The estimated theoretical power is given in the parenthesis, and the percentage of a kernel being selected among the three candidate kernels using the proposed kernel selection method is displayed underneath it.

		Gaussian Error			Laplace Error		
n	p	K_E	K_L	K_G	K_E	K_L	K_G
40	150	0.706 (0.716)	0.640 (0.648)	0.639 (0.653)	0.710 (0.726)	0.654 (0.726)	0.653 (0.665)
		(94.8%)	(2.9%)	(2.3%)	(95.3%)	(1.9%)	(2.8%)
	200	0.427(0.507)	0.396 (0.457)	0.396 (0.463)	0.390 (0.481)	0.344 (0.432)	0.348 (0.437)
		(89.4%)	(3.4%)	(7.2%)	(86.6%)	(4.8%)	(8.6%)
	250	0.261 (0.371)	0.238 (0.334)	0.265 (0.340)	0.270 (0.383)	0.246 (0.348)	0.246 (0.353)
		(81.8%)	(4.4%)	(13.8%)	(82.9%)	(5.0%)	(12.1%)
60	150	0.890 (0.869)	0.842 (0.815)	0.841 (0.818)	0.898 (0.876)	0.861 (0.826)	0.864 (0.828)
		(96.6%)	(2.4%)	(1.0%)	(96.9%)	(2.4%)	(0.7%)
	200	0.596 (0.615)	0.535 (0.560)	0.531 (0.564)	0.591 (0.613)	0.538 (0.562)	0.534 (0.566)
		(89.7%)	(6.2%)	(4.1%)	(89.6%)	(5.2%)	(5.2%)
	250	0.402 (0.457)	0.353 (0.419)	0.354 (0.423)	0.350 (0.430)	0.321(0.395)	0.321 (0.400)
		(81.9%)	(8.9%)	(9.2%)	(82.7%)	(8.1%)	(9.2%)
100	150	0.992 (0.984)	0.985 (0.969)	0.986 (0.970)	0.990 (0.982)	0.984 (0.969)	0.984 (0.969)
		(99.2%)	(0.8%)	(0.0%)	(99.1%)	(0.8%)	(0.1%)
	200	0.885 (0.857)	0.842 (0.818)	0.844 (0.820)	0.870 (0.840)	0.828 (0.798)	0.829 (0.799)
		(94.1%)	(3.9%)	(2.0%)	(94.9%)	(3.6%)	(1.5%)
	250	0.618 (0.631)	0.584 (0.590)	0.585 (0.593)	0.638 (0.640)	0.600 (0.602)	0.599 (0.604)
		(84.3%)	(10.7%)	(5.0%)	(87.4%)	(8.1%)	(4.5%)

Table 2.5 Empirical power of the proposed test for Gaussian and Laplace errors with dependent covariates using different kernels in the setting of $p >> n$ under scenario $\mathcal{S}_4$. The estimated theoretical power is given in the parenthesis, and the percentage of a kernel being selected among the three candidate kernels using the proposed kernel selection method is displayed underneath it.

		Gaussian Error			Laplace Error		
n	p	K_E	K_L	K_G	K_E	K_L	K_G
40	150	0.671 (0.691)	0.673 (0.682)	0.674 (0.687)	0.667 (0.687)	0.675 (0.680)	0.672 (0.684)
		(47.8%)	(20.8%)	(31.4%)	(44.8%)	(22.1%)	(33.1%)
	200	0.595 (0.616)	0.605 (0.626)	0.604 (0.631)	0.594 (0.634)	0.600 (0.632)	0.600 (0.637)
		(53.9%)	(14.2%)	(31.9%)	(53.8%)	(13.6%)	(32.6%)
	250	0.544 (0.597)	0.541 (0.580)	0.543 (0.587)	0.527 (0.589)	0.534 (0.600)	0.525 (0.584)
		(62.4%)	(8.8%)	(28.8%)	(63.1%)	(9.6%)	(27.3%)
60	150	0.903 (0.871)	0.903 (0.871)	0.904 (0.873)	0.860 (0.841)	0.865 (0.841)	0.866 (0.842)
		(36.1%)	(34.6%)	(29.3%)	(36.7%)	(33.1%)	(30.2%)
	200	0.817 (0.810)	0.821 (0.808)	0.823 (0.811)	0.824 (0.809)	0.832 (0.807)	0.833 (0.810)
		(46.6%)	(27.6%)	(27.8%)	(44.2%)	(26.6%)	(31.2%)
	250	0.733 (0.761)	0.733 (0.757)	0.744 (0.764)	0.768 (0.762)	0.773 (0.768)	0.772 0.765)
		(51.2%)	(19.7%)	(29.1%)	(50.7%)	(20.1%)	(29.2%)
100	150	0.993 (0.986)	0.991 (0.985)	0.992 (0.986)	0.989 (0.979)	0.990 (0.980)	0.990 (0.980)
		(24.9%)	(45.8%)	(29.3%)	(26.2%)	(46.4%)	(27.4%)
	200	0.983 (0.972)	0.986 (0.973)	0.986 (0.973)	0.985 (0.966)	0.985 (0.967)	0.985 (0.967)
		(31.6%)	(39.1%)	(29.3%)	(31.1%)	(41.9%)	(27.0%)
	250	0.980 (0.960)	0.981 (0.960)	0.981 (0.960)	0.972 (0.951)	0.972 (0.951)	0.972 (0.951)
		(33.3%)	(38.3%)	(28.4%)	(35.2%)	(34.8%)	(30.0%)

vertical lines correspond to the first quantile (Q_1), median and third quantile (Q_3) of the stabilized $\tilde{\gamma}$ obtained from the 1000 simulation replicates, where $L = 5$ were chosen in stability selection. It can be seen from Figure that the vertical lines were all very close to the place where the maximum power was achieved. This suggests that the proposed regularization selection method can locate the optimal regularization parameter to maximize the power of the proposed test.

2.6 An empirical study

We applied the proposed test to a Yolkshire gilt data set to find gene sets that are associated with triiodothyronine (T_3), which is an important thyroid hormone affecting growth and metabolism in the body. A total of 24,123 gene expressions were measured using liver tissues for 24 six-month-old Yolkshire gilts, whose T_3 levels in blood were also recorded. All the genes in the Yolkshire gilt data set were classified into 6176 Gene Ontology (GO) terms (gene sets), where each gene could be assigned to several GO terms according to its gene attributes in one of the three domains: cellular component, molecular function, and biological process. More details about the data set can be found in Lkhagvadori et al. (2009).

Let Y_i and $\mathbf{X}_i^{(k)} = (\mathbf{X}_{i1}^{(k)}, \mathbf{X}_{i2}^{(k)}, \cdots, \mathbf{X}_{ip_k}^{(k)})^T$ be, respectively, the measure of T_3 level for the i-th gilt and the standardized gene expression vector of the k-th GO term for the i-th gilt, where p_k is the total number of genes in the kth GO term. We consider the following nonparametric regression model

$$Y_i = \mu^{(k)} + h^{(k)}(\mathbf{X}_i^{(k)}) + \epsilon_i^{(k)}, \quad i = 1, ..., 24, \quad k = 1, ..., 6176. \tag{2.6.1}$$

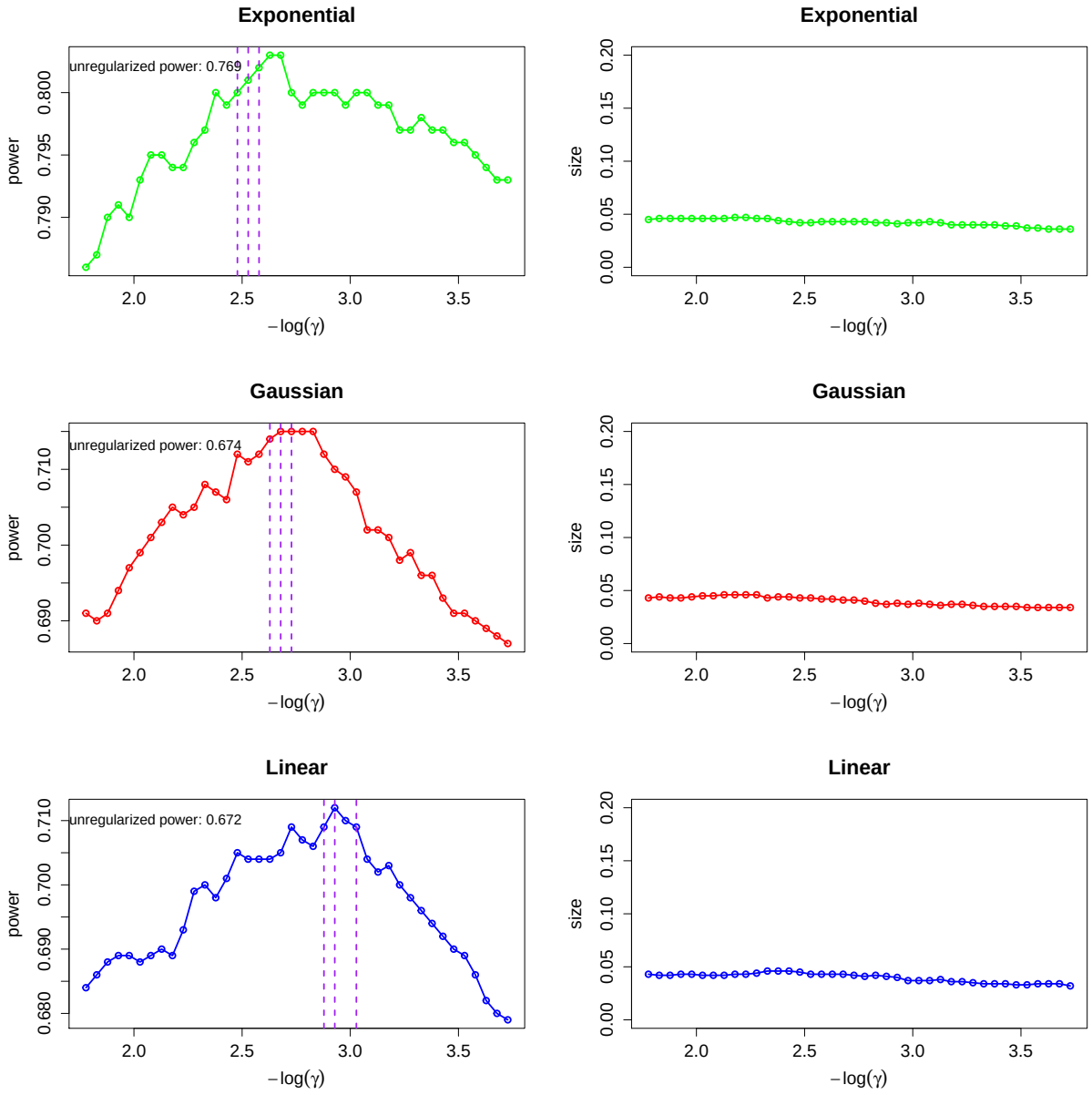

Figure 2.3 The empirical power (left panel) and size (right panel) for regularized kernels, where the vertical purple lines in the left panel denote the first, second and third quantile of the selected regularization parameters among 1000 simulation replicates. For each replicate, the regularization parameter was selected by the method introduced in Section 4.2.

For the k-th GO term, we are interested in testing

$$H_0 : h^{(k)}(\cdot) = 0 \text{ vs } H_1 : h^{(k)}(\cdot) \neq 0. \tag{2.6.2}$$

We conducted the proposed test to 6176 GO terms using four different centralized kernels: exponential kernel K_E, Gaussian kernel K_G, linear kernel K_L, and polynomial kernel K_P, where K_E, K_G and K_L were defined in Section 5 and $K_P(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^T \mathbf{x}_j/\theta)^2$. In all the kernels, the tuning parameter θ was set to be p_k. For each kernel, we obtained 6176 p-values and Benjamini-Hochberg procedure was applied for multiple test correction with false discovery rate controlled at the 1% level.

The proposed test with exponential kernel detected 43 significant GO terms, which contains all the GO terms detected by the other three kernels. Among the 43 significant GO terms, 28, 6, 6 were detected by Gaussian kernel, linear kernel and polynomial kernel, respectively, with 5 in common. The fact that more GO terms were detected by the exponential kernel might indicate a strong non-linear relationship between Y and $\mathbf{X}$, i.e., $h(\mathbf{X})$ is a non-linear function of $\mathbf{X}$.

2.7 Summary and discussions

In this study, we modeled the joint effect of high-dimensional variates in a set through a nonparametric function in an RKHS. We proposed a nonparmetric test for assessing the significance of the nonparametric function. Different from previous investigations, our method is applicable to the "large p, small n" setting. Our test is powerful in testing a non-parametric function even when the dimension p is much larger than the sample size n. We derived the

asymptotic distribution of the test statistic under the null hypothesis and a sequence of local alternative hypotheses under the "large p, small n" set up. Based on the obtained explicit power function, we proposed a kernel selection method which was designed to improve the power of the proposed test. Moreover, we introduced a test with regularized kernel which can further improve the power and enhance the dimensionality the test could handle. It was shown that the regularization kernel plays a similar role as a re-weighting method which adds higher weights to non-zero projections of the nonparametric function to the orthogonal bases of the RKHS. With a properly chosen regularization parameter, we demonstrate that the proposed test could achieve almost the same power as the oracle test. A practical method for selecting regularization parameter was also introduced in the chapter.

2.8 Lemmas and Proofs

Lemma 1. *Assume that K is a positive semi-definite and symmetric kernel defined on $\mathscr{X} \times \mathscr{X}$. Let μ be a finite measure on $\mathscr{X}$. If*

$$\int K(x,x)d\mu(x) < \infty, \tag{2.8.1}$$

then $K(x,y) = \sum_{j=1}^{\infty} \lambda_j \psi_j(x)\psi_j(y)$, $\{\psi_j(\cdot)\} \subset L^2(\mu)$ form a complete orthogonal normal system i.e., $E(\psi_j(X)\psi_k(X)) = \delta_{jk}$ where $\delta_{jk} = 1$ if $j = k$; $\delta_{jk} = 0$ if $j \neq k$, and $\sum_{j=1}^{\infty} \lambda_j < \infty$.

Proof: Given a positive definite kernel $K(x,y)$, we can construct a reproducing kernel

Hilbert space (RKHS) $\mathscr{H}_K$, and the reproducing property implies that

$$K(x,y) =< K(x,\cdot), K(y,\cdot) >_{\mathscr{H}_K} :=< K_x, K_y >_{\mathscr{H}_K} .$$

Since for any $K_x \in \mathscr{H}_K$, $\|K_x\| = \sqrt{< K_x, K_x >_{\mathscr{H}_K}} = \sqrt{K(x,x)}$, we have

$$|K(x,y)| =< K_x, K_y >_{\mathscr{H}_K} \leq \|K_x\|\|K_y\| = \sqrt{K(x,x)K(y,y)},$$

and

$$\int_{\mathscr{X}\times\mathscr{X}} K^2(x,y)d\mu(x)d\mu(y) \leq \int_{\mathscr{X}} K(x,x)d\mu(x) \cdot \int_{\mathscr{X}} K(y,y)d\mu(y) < \infty. \tag{2.8.2}$$

Therefore, $K(x,y)$ generates a compact operator on $L^2(\mu)$ through the integral operation $(Kf)(x) = \int_{\mathscr{X}} K(x,y)f(y)d\mu(y)$. Let $\{\lambda_j\}_{j=1}^{\infty}$ and $\{\psi_j(\cdot)\}_{j=1}^{\infty}$ be the eigenvalues and corresponding complete orthogonal normal system of kernel K under measure μ, i.e.,

$$\int K(x,y)\psi_i(y)d\mu(y) = \lambda_i\psi_i(x), \quad i = 1, 2, \cdots, \infty. \tag{2.8.3}$$

Since $K(x,y) \in L^2(\mu \bigotimes \mu)$, $K_x(\cdot) = K(x,\cdot) \in L^2(\mu)$, i.e., there exist $\{c_m(x)\}_{m=1}^{\infty}$ such that $K(x,y) = K_x(y) = \sum_m c_m(x)\psi_m(y)$, then we have $K_y(\cdot) = \sum_{m=1} c_m(\cdot)\psi_m(y)$. Because $K_y(\cdot) \in L^2(\mu)$ and $\{\psi_m(y)\}_{m=1}^{\infty}$ can be considered as constants once y is given, then $c_m(\cdot) \in L^2(\mu)$ and can be expanded using bases $\{\psi_m(\cdot)\}_{m=1}^{\infty}$. Therefore, we have $K(x,y) = \sum_{i,j=1}^{\infty} a_{ij}\psi_i(x)\psi_j(y)$, where $\sum_{i,j} a_{ij}^2 < \infty$ is due to (2.8.2).

It will be shown in the following that $a_{ij} = \lambda_i\delta_{ij}$, which implies $K(x,y) = \sum_{j=1}^{\infty} \lambda_j\psi_j(x)\psi_j(y)$.

Actually,

$$\int\int K(x,y)\psi_i(x)\psi_j(y)d\mu(x)d\mu(y) = \int\int \sum_{k,l} a_{kl}\psi_k(x)\psi_l(y)\psi_i(x)\psi_j(y)d\mu(x)d\mu(y)$$

where left hand side is $\int \lambda_i\psi_i(y)\psi_j(y)d\mu(y) = \lambda_i\delta_{ij}$ by using the eigen-decomposition property (2.8.3), and right hand side is $\sum_{k,l} a_{kl}\delta_{ki}\delta_{lj} = a_{ij}$. Moreover, we could conclude $\sum_{i=1}^{\infty}\lambda_i < \infty$, since $K(x,x) = \sum_i \lambda_i\psi_i^2(x)$, and $\sum_{i=1}^{\infty}\lambda_i = \int K(x,x)d\mu(x) < \infty$. This finishes the proof of Lemma 1. □

Proof of Theorem 1: (i) Under the null hypothesis, $Y_i = \mu + \epsilon_i$. Because the test statistic T_n is invariant to location shift, without loss of generality, we assume $\mu = 0$ in the following proof. Then $T_n^0 := T_n^{H_0}$, the reduced version under null hypothesis, can be written as

$$T_n^0 = \frac{1}{n(n-1)\sigma^2}\sum_{i\neq j} K(\mathbf{X}_i,\mathbf{X}_j)(\epsilon_i-\bar{\epsilon})(\epsilon_j-\bar{\epsilon})[1+(\frac{\sigma^2}{\hat{\sigma}^2}-1)] := T_{n1}^0[1+(\frac{\sigma^2}{\hat{\sigma}^2}-1)] \quad (2.8.4)$$

Since $\sigma^2/\hat{\sigma}^2 - 1 = o_p(1)$, under the null, we have $T_n^0 = T_{n1}^0\{1+o_p(1)\}$.

We now study the asymptotic distribution of T_{n1}^0/V_1 using the U-statistic theory (Lee , 1990). By plugging in the full expression of $\bar{\epsilon} = n^{-1}\sum_{i=1}^n \epsilon_i$, the leading order of T_{n1}^0/V_1 can be written as the sum of three U-statistics of different orders

$$\frac{T_{n1}^0}{V_1} = U_n^{(2)} + U_n^{(3)} + U_n^{(4)} + \Delta_n^0, \quad (2.8.5)$$

where $U_n^{(2)} = \frac{1}{P_n^2}\sum_{i\neq j}\Psi^{(2)}(\mathbf{Z}_i,\mathbf{Z}_j)$,

$$U_n^{(3)} = \frac{1}{P_n^3}\sum_{i\neq j\neq k}\Psi^{(3)}(\mathbf{Z}_i,\mathbf{Z}_j,\mathbf{Z}_k),\ \ U_n^{(4)} = \frac{1}{P_n^4}\sum_{i\neq j\neq k\neq l}\Psi^{(4)}(\mathbf{Z}_i,\mathbf{Z}_j,\mathbf{Z}_k,\mathbf{Z}_l),$$

$\Delta_n^0 = o_p(U_n^{(2)}+U_n^{(3)}+U_n^{(4)})$, and $\mathbf{Z} = (\mathbf{X},\epsilon)$. P_n^k is the number of k-permutations of n, $\Psi^{(k)}$ is the kernel function of k-th order U-statistic $U_n^{(k)}$ for $k = 2,3,4$ and of the following symmetric form $\Psi^{(2)}(\mathbf{Z}_i,\mathbf{Z}_j) = \mathcal{K}(\mathbf{X}_i,\mathbf{X}_j)[\epsilon_i\epsilon_j - n^{-1}(\epsilon_i+\epsilon_j)^2]/\sigma^2$, $\Psi^{(3)}(\mathbf{Z}_i,\mathbf{Z}_j,\mathbf{Z}_k) = \varphi^{(3)}(i,j,k) + \varphi^{(3)}(j,k,i) + \varphi^{(3)}(i,k,j)$ and

$$\begin{aligned}\Psi^{(4)}(\mathbf{Z}_i,\mathbf{Z}_j,\mathbf{Z}_k,\mathbf{Z}_l) =&\varphi^{(4)}(i,j,k,l) + \varphi^{(4)}(i,k,j,l) + \varphi^{(4)}(i,l,j,k) + \varphi^{(4)}(j,k,i,l)\\ &+ \varphi^{(4)}(j,l,i,k) + \varphi^{(4)}(k,l,i,j).\end{aligned}$$

where $\varphi^{(3)}(i,j,k) = -(3\sigma^2)^{-1}\mathcal{K}(\mathbf{X}_i,\mathbf{X}_j)[\epsilon_i\epsilon_k+\epsilon_j\epsilon_k-\epsilon_k^2/n]$, and $\varphi^{(4)}(i,j,k,l) = (6\sigma^2)^{-1}\mathcal{K}(\mathbf{X}_i,\mathbf{X}_j)\epsilon_k\epsilon_l$. To study the distribution of T_{n1}^0/V_1, we will look at the asymptotic properties of each U-statistic $U_n^{(k)}$ respectively. Specifically, we are going to show the following

$$(a):\quad nU_n^{(2)} \xrightarrow{d} \sum_{m=1}^{\infty}\lambda_{\mathcal{K},m}(\chi_m^2-1), \tag{2.8.6}$$

$$(b):\quad nU_n^{(3)} \xrightarrow{p} 0, \tag{2.8.7}$$

$$(c):\quad nU_n^{(4)} \xrightarrow{p} 0. \tag{2.8.8}$$

To see (a), we define the first-order and second-order projections of the kernel $\Psi^{(2)}(\cdot)$ as $\phi_1^{(2)}(\mathbf{z}_i) = E\{\Psi^{(2)}(\mathbf{z}_i,\mathbf{Z}_j)\} = 0$ and $\phi_2^{(2)}(\mathbf{z}_i,\mathbf{z}_j) = E\{\Psi^{(2)}(\mathbf{z}_i,\mathbf{z}_j)\} = \Psi_2^{(2)}(\mathbf{z}_i,\mathbf{z}_j)$, and their corresponding variances $\sigma_{2,1}^2 = \mathrm{Var}[\phi_1^{(2)}(\mathbf{Z}_i)], \sigma_{2,2}^2 = \mathrm{Var}[\phi_2^{(2)}(\mathbf{Z}_i,\mathbf{Z}_j)]$. It is not difficult to prove that $U_n^{(2)}$ is first-order degenerated, i.e., $\sigma_{2,1}^2 = 0$ and $\sigma_{2,2}^2 = 2V_{\mathcal{K},2}\{1+o(1)\} \neq 0$.

By the classical U-statistic theory, $nU_n^{(2)} \xrightarrow{d} \sum_{m=1}^{\infty} \lambda_{\mathcal{K}_z,m}(\chi_m^2 - 1)$, where $\{\lambda_{\mathcal{K}_z,m}\}_{m=1}^{\infty}$ are the eigenvalues of kernel function $\mathcal{K}_z(z_1, z_2) = \mathcal{K}(x_1, x_2)\epsilon_1\epsilon_2$ with respect to the distribution function F_z, i.e., solution of integral equations

$$\int \mathcal{K}_z(z_1, z_2)\psi_{\mathcal{K}_z,m}(z_1)dF_z(z_2) = \lambda_{\mathcal{K}_z,m}\psi_{\mathcal{K}_z,m}(z_2), m = 1, ..., \infty. \tag{2.8.9}$$

It remains to prove that $\lambda_{\mathcal{K}_z,m} = \lambda_{\mathcal{K},m}$. View kernel $\mathcal{K}_z(z_1, z_2)$ as the product of kernel $\mathcal{K}_z^1(z_1, z_2) := \mathcal{K}(x_1, x_2)$ and kernel $\mathcal{K}_z^2(z_1, z_2) := \epsilon_1\epsilon_2$, where $\mathcal{K}_z^2$ has only one non-zero eigenvalue 1 with eigenfunction $g(\epsilon) = \epsilon/\sigma$ under the null hypothesis. Through equations (2.8.9) above, it can be verified that eigenvalues and eigenfunctions of $\mathcal{K}_z(z_1, z_2)$ are $\{\lambda_{\mathcal{K},m}\}$ and $\{\psi_m(x) \cdot g(\epsilon)\}_{m=1}^{\infty}$ respectively. (b) and (c) can be achieved similarly by proving means and variances of the first- and second-order projections of $U_n^{(3)}$ and $U_n^{(4)}$ are all zero.

(ii) Let $Q_m = \lambda_{\mathcal{K},m}(\chi_m^2 - 1)$ for $m \in \mathcal{N}$, $S_n = \sum_{m=1}^{n} Q_m$, $\alpha_\infty := \{\mathrm{Var}(S_\infty)\}^{1/2} = \sqrt{2V_{\mathcal{K},2}}$ and $\{Q_m\}$ is the sum of independent random variables. Moreover, it is not difficult to see that $\mathrm{E}|Q_m|^4 \propto V_{\mathcal{K},4}$. By part (i), nT_n^0/V_1 has the same distribution as $\sum_{m=1}^{\infty} Q_m$. According to Lyapunov's Theorem, if $\alpha_\infty^{-4} \sum_{m=1}^{\infty} \mathrm{E}|Q_m|^4$ converges to 0 as $p \to \infty$ (i.e., condition (2.3.2)), then we can conclude asymptotic normality for nT_n, under the null hypothesis. In summary, $\sigma_{T_n}^{-1} nT_n$ weakly convergences to the standard normal distribution under condition (2.3.2).

Proof of Theorem 2: Under the alternative hypothesis H_{1n}, $Y_i = \mu + d_n(\mathbf{X}_i) + \epsilon_i$, where $\mathrm{E}d_n(\mathbf{X}_i) = 0$, $\mathrm{E}(\epsilon_i) = 0$, and $\mathrm{Var}(\epsilon_i) = \sigma^2$. Without loss of generality, we also assume $\mu = 0$

in the following proof. Similar to (2.8.4), considering the following expansion

$$T_n = \frac{1}{n(n-1)\sigma^2}\sum_{i\neq j} K(\mathbf{X}_i,\mathbf{X}_j)(Y_i-\bar{Y}_n)(Y_j-\bar{Y}_n)[1+(\frac{\sigma^2}{\hat{\sigma}^2}-1)] := T_{n1}[1+(\frac{\sigma^2}{\hat{\sigma}^2}-1)].$$

Under condition (2.3.6), $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$ (Proposition 1 in the last section), hence it is enough to study the behavior of T_{n1}. By plugging in expression of Y_i and Y_j under H_{1n}, T_{n1} could be decomposed into two parts: T_{n1}^0 and $\tilde{T}_{n1}^0$, where asymptotic distribution of T_{n1}^0 is the null distribution and has been studied in Theorem 1. The remainder term $\tilde{T}_{n1}^0$ can be expressed as the sum of the following three terms

$$\tilde{T}_{n1}^0 = \{\Theta_n^{(2)} + \Theta_n^{(3)} + \Theta_n^{(4)}\}\{1+o_p(1)\}, \tag{2.8.10}$$

where

$$\Theta_n^{(2)} = \frac{V_1}{n^2\sigma^2}\sum_{i\neq j}\mathcal{K}_\theta(\mathbf{X}_i,\mathbf{X}_j)\big[U_iU_j + U_i\epsilon_j + U_j\epsilon_i - \frac{1}{n}(U_i+U_j)^2 - \frac{2}{n}(U_i+U_j)(\epsilon_i+\epsilon_j)\big],$$
$$\Theta_n^{(3)} = \frac{V_1}{n^3\sigma^2}\sum_{i\neq j\neq k}\mathcal{K}_\theta(\mathbf{X}_i,\mathbf{X}_j)\big[-(U_i+U_j)(U_k+\epsilon_k) - (\epsilon_i+\epsilon_j)U_k + \frac{1}{n}(U_k^2+2U_k\epsilon_k)\big],$$
$$\Theta_n^{(4)} = \frac{V_1}{n^4\sigma^2}\sum_{i\neq j\neq k\neq l}\mathcal{K}_\theta(\mathbf{X}_i,\mathbf{X}_j)(U_kU_l + U_k\epsilon_l + U_l\epsilon_k),$$

and $U_i = d_n(\mathbf{X}_i)$. Denote the eigenvalues of normalized kernel $\mathcal{K}_\theta(x,y)$ as $\{\lambda_{\mathcal{K},m}\}_{m=1}^\infty$, where $\sum_m \lambda_{\mathcal{K},m} = 1$ and $\lambda_{\mathcal{K},m} \geq 0$ for each m. Notice that kernel K_θ and $\mathcal{K}_\theta$ have the same eigenfucntions $\{\psi_m(\mathbf{X})\}_{m=1}^\infty$. Besides, for centralized kernel, $\mu_m := \mathrm{E}(\psi_m(\mathbf{X})) = 0$ for $m \in \mathcal{N}\setminus\{m^*\}$, where $\mu_{m^*} = 1$ corresponds to zero eigenvalue ($\lambda_{m^*} = 0$) (See Lemma in the last section). Let $\mathbb{G} = \{m : \lambda_m > 0\}$, then $K_\theta(x_1,x_2) = \sum_{m\in\mathbb{G}}\lambda_m\psi_m(x_1)\psi_m(x_2)$

and $\mu_m = 0$ for all $m \in \mathbb{G}$. Define $a_m := \mathrm{E}\big(\psi_m(\mathbf{X})d_n(\mathbf{X})\big)$ representing the projection of function $d_n(\mathbf{X})$ onto the eigen-function $\psi_m(\mathbf{X})$.

In the following we will show that (a) $\mathrm{E}(\sigma_{T_n}^{-1} n\tilde{T}_{n1}^0) - \Psi(d_n) = o(1)$, and (b) $\mathrm{Var}\{n\tilde{T}_{n1}^0\} = o(\mathrm{Var}\{nT_{n1}^0\})$. To prove (a) and (b), let us study the asymptotic behavior of each term in $n\sigma_{T_n}^{-1}\tilde{T}_{n1}^0 = n\sigma_{T_n}^{-1}(\Theta_n^{(2)} + \Theta_n^{(3)} + \Theta_n^{(4)})$. Firstly split

$$\sigma_{T_n}^{-1} n\Theta_n^{(2)} = n\big(\tilde{S}_{11} + 2\tilde{S}_{12} + 2\tilde{S}_{13} + 2\tilde{S}_{14}\big)\{1 + o_p(1)\}, \tag{2.8.11}$$

where

$$\tilde{S}_{11} := \frac{V_1}{n^2\sigma^2\sigma_{T_n}} \sum_{i\neq j} \mathcal{K}_\theta(\mathbf{X}_i, \mathbf{X}_j)U_iU_j, \quad \tilde{S}_{12} = \frac{V_1}{n^2\sigma^2\sigma_{T_n}} \sum_{i\neq j} \mathcal{K}_\theta(\mathbf{X}_i, \mathbf{X}_j)U_i\epsilon_j,$$

$$\tilde{S}_{13} = -\frac{V_1}{n^3\sigma^2\sigma_{T_n}} \sum_{i\neq j} \mathcal{K}_\theta(\mathbf{X}_i, \mathbf{X}_j)U_i^2, \quad \tilde{S}_{14} = -\frac{V_1}{2n^3\sigma^2\sigma_{T_n}} \sum_{i\neq j} \mathcal{K}_\theta(\mathbf{X}_i, \mathbf{X}_j)U_i\epsilon_i.$$

We want to show $n\tilde{S}_{11} \xrightarrow{p} \Psi(d_n)$ and $n\tilde{S}_{1j} \xrightarrow{p} 0$ for $j = 2, 3, 4$. Actually, $\mathrm{E}(n\tilde{S}_{11}) = n\sigma_{T_n}^{-1}V_1 \sum_{m=1}^\infty a_m^2\lambda_{\mathcal{K},m} = \Psi(d_n)\{1 + o(1)\}$, and

$$n^2\tilde{S}_{11}^2 = \frac{V_1^2}{n^2\sigma^4\sigma_{T_n}^2} \sum_{\substack{i\neq j, k\neq l \\ m_1, m_2\in\mathbb{G}}} \lambda_{\mathcal{K},m_1}\lambda_{\mathcal{K},m_2}\psi_{m_1}(\mathbf{X}_i)\psi_{m_1}(\mathbf{X}_j)\psi_{m_2}(\mathbf{X}_k)\psi_{m_2}(\mathbf{X}_l)U_iU_jU_kU_l.$$

Define index subsets $I_c = \{(i,j,k,l)||\{i,j\}\cap\{k,l\}| = c, i,j,k,l \in \{1,\cdots,n\}, i\neq j, k\neq l\}$ for $c = 0, 1, 2$, where $|\cdot|$ denotes the set cardinality. For example, I_0 represents set $\{(i,j,k,l) \in |i,j,k,l \in \{1,\cdots,n\}, i\neq j\neq k\neq l\}$. Then $n^2\tilde{S}_{11}^2 = J_0 + J_1 + J_2$, where

$$J_c = \frac{V_1^2}{n^2\sigma^4\sigma_{T_n}^2} \sum_{\substack{i,j,k,l\in I_c \\ m_1, m_2\in\mathbb{G}}} \lambda_{\mathcal{K},m_1}\lambda_{\mathcal{K},m_2}\psi_{m_1}(\mathbf{X}_i)\psi_{m_1}(\mathbf{X}_j)\psi_{m_2}(\mathbf{X}_k)\psi_{m_2}(\mathbf{X}_l)U_iU_jU_kU_l.$$

By using the orthogonal and centralized properties of eigen-functions, it can be proved that

$$\begin{aligned}
\mathrm{E}(J_0) &= \mathrm{E}^2(n\tilde{S}_{11})\{1+o(1)\} = \Psi^2(d_n)\{1+o(1)\},\\
\mathrm{E}(J_1) &= \frac{4V_1^2}{n^2\sigma^4\sigma_{T_n}^2} \sum_{\substack{i\neq j\neq k\\ m_1,m_2\in\mathbb{G}}} \lambda_{\mathcal{K},m_1}\lambda_{\mathcal{K},m_2}\mathrm{E}\{\psi_{m_1}(\mathbf{X}_i)\psi_{m_2}(\mathbf{X}_i)U_i^2\}\cdot\mathrm{E}\{\psi_{m_1}(\mathbf{X}_j)U_j\}\cdot\mathrm{E}\{\psi_{m_2}(\mathbf{X}_k)U_k\}\\
&= 4nV_1^2\sigma^{-4}\sigma_{T_n}^{-2}\bigg(\sum_{m_1,m_2\in\mathbb{G}}\lambda_{\mathcal{K},m_1}\lambda_{\mathcal{K},m_2}a_{m_1}a_{m_2}b_{m_1,m_2}\bigg)\{1+o(1)\},\\
\mathrm{E}(J_2) &= 2nV_1^2\sigma^{-4}\sigma_{T_n}^{-2}\bigg(\sum_{m_1,m_2\in\mathbb{G}}\lambda_{\mathcal{K},m_1}\lambda_{\mathcal{K},m_2}b_{m_1,m_2}^2\bigg)\{1+o(1)\},
\end{aligned}$$

where $b_{m_1,m_2} = \mathrm{E}[\psi_{m_1}(\mathbf{X})\psi_{m_2}(\mathbf{X})d_n^2(\mathbf{X})]$. Under condition (2.3.6), we can prove that $|a_m| \le D_1\big[\mathrm{E}d_n^8(\mathbf{X})\big]^{1/8}$, and $|b_{m_1,m_2}| \le D_2\big[\mathrm{E}d_n^8(\mathbf{X})\big]^{1/4}$ for some finite constants D_1 and D_2, by using Cauchy-Schwartz inequality. Therefore, $\mathrm{E}(J_1) \le 4\sigma^{-4}D_1^2D_2\cdot n\big[\mathrm{E}d_n^8(\mathbf{X})\big]^{1/2}V_1^2/2V_2 = o(1)$, $\mathrm{E}(J_2) \le 2\sigma^{-4}D_2^2\cdot n\big[\mathrm{E}d_n^8(\mathbf{X})\big]^{1/2}V_1^2/2V_2 = o(1)$, and $\mathrm{Var}(n\tilde{S}_{11}) = o(1)$ under condition (2.3.6). Hence $n\tilde{S}_{11} \xrightarrow{d} \Psi(d_n) = O(1)$. It remains to prove $n\tilde{S}_{1j} \xrightarrow{p} 0$ for $j = 2,3,4$ in (2.8.11).

It is easy to see that $\mathrm{E}(n\tilde{S}_{12}) = \mathrm{E}(n\tilde{S}_{13}) = \mathrm{E}(n\tilde{S}_{14}) = 0$ by using the centralized kernel property. Moreover, it can be proved that $\mathrm{Var}(n\tilde{S}_{12}) = \mathrm{Var}(n\tilde{S}_{13}) = \mathrm{Var}(n\tilde{S}_{14}) = o(1)$. Actually,

$$\begin{aligned}
\mathrm{Var}(n\tilde{S}_{12}) &= \frac{V_1^2}{\sigma^2\sigma_{T_n}^2}\sum_{m\in\mathbb{G}}\lambda_{\mathcal{K},m}^2 b_{m,m}\{1+o(1)\} \le (\sqrt{2}\sigma)^{-2}D_2[\mathrm{E}d_n^8(\mathbf{X})]^{1/4},\\
\mathrm{Var}(n\tilde{S}_{13}) &= \frac{V_1^2}{n^2\sigma^4\sigma_{T_n}^2}\left\{\sum_{m\in\mathbb{G}}\lambda_{\mathcal{K},m}^2 e_m + \sum_{m_1,m_2}\lambda_{\mathcal{K},m_1}\lambda_{\mathcal{K},m_2}b_{m_1,m_2}^2\right\}\{1+o(1)\}\\
&\le (\sqrt{2}n\sigma^2)^{-2}[\mathrm{E}d_n^8(\mathbf{X})]^{1/2}(D_3 + D_2^2V_1^2/V_2)\\
\mathrm{Var}(n\tilde{S}_{14}) &= \frac{V_1^2}{n^2\sigma^2\sigma_{T_n}^2}\sum_{m\in\mathbb{G}}\lambda_{\mathcal{K},m}^2 b_{m,m}\{1+o(1)\} \le (\sqrt{2}n\sigma)^{-2}D_2[\mathrm{E}d_n^8(\mathbf{X})]^{1/4},
\end{aligned}$$

where $e_m = \mathrm{E}[\psi_m^2(\mathbf{X})d_n^4(\mathbf{X})] \leq D_3[\mathrm{E}d_n^8(\mathbf{X})]^{1/2}$ for some constant $D_3 > 0$. The variance above are all of order $o(1)$ under condition (2.3.6). For the triple sum terms $\Theta_n^{(3)}$ in (2.8.10),

$$\tilde{S}_{21} = \frac{V_1}{n^3\sigma^2\sigma_{Tn}} \sum_{i\neq j\neq k} \mathcal{K}_\theta(\mathbf{X}_i, \mathbf{X}_j)U_i\epsilon_k, \quad \tilde{S}_{22} = \frac{V_1}{n^3\sigma^2\sigma_{Tn}} \sum_{i\neq j\neq k} \mathcal{K}_\theta(\mathbf{X}_i, \mathbf{X}_j)U_iU_k,$$

$$\tilde{S}_{23} = \frac{V_1}{n^3\sigma^2\sigma_{Tn}} \sum_{i\neq j\neq k} \mathcal{K}_\theta(\mathbf{X}_i, \mathbf{X}_j)U_k\epsilon_i, \quad \tilde{S}_{24} = \frac{V_1}{n^4\sigma^2\sigma_{Tn}} \sum_{i\neq j\neq k} \mathcal{K}_\theta(\mathbf{X}_i, \mathbf{X}_j)U_k\epsilon_k,$$

$$\tilde{S}_{25} = \frac{V_1}{n^4\sigma^2\sigma_{Tn}} \sum_{i\neq j\neq k} \mathcal{K}_\theta(\mathbf{X}_i, \mathbf{X}_j)U_k^2.$$

Similarly, it is not difficult to see that $\mathrm{E}(n\tilde{S}_{2j}) = 0$ for $j = 1, ..., 5$. Furthermore, up to a factor of $\{1 + o(1)\}$, we have the following

$$\begin{aligned}
\mathrm{Var}(n\tilde{S}_{21}) =& \frac{V_1^2}{2\sigma^2V_2} \sum_{m\in\mathbb{G}} \lambda_{\mathcal{K},m}^2\big(a_m^2 + 2n^{-1}b_{m,m}\big),\\
\mathrm{Var}(n\tilde{S}_{22}) =& \frac{V_1^2}{2\sigma^4V_2} \sum_{m_1,m_2\in\mathbb{G}} \lambda_{\mathcal{K},m_1}\lambda_{\mathcal{K},m_2}\big(a_{m_1}^2a_{m_2}^2 + n^{-1}d_{m_1,m_2}^2\mathrm{E}[d_n^2(\mathbf{X})] + 2n^{-1}a_{m_1}c_{m_2}d_{m_1,m_2}\big)\\
&+ \frac{V_1^2}{2\sigma^4V_2} \sum_{m\in\mathbb{G}} \lambda_{\mathcal{K},m}^2\big(a_m^2\mathrm{E}[d_n^2(\mathbf{X})] + n^{-1}b_{m,m}\mathrm{E}[d_n^2(\mathbf{X})] + n^{-1}c_m^2\big),\\
\mathrm{Var}(n\tilde{S}_{23}) =& (2n\sigma^4)^{-1}\big(V_2^{-1}V_1^2 \sum_{m\in\mathbb{G}} \lambda_{\mathcal{K},m}^2a_m^2 + \mathrm{E}[d_n^2(\mathbf{X})]\big),
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}(n\tilde{S}_{24}) =& (n^3\sigma^4)^{-1}\mathrm{E}[d_n^2(\mathbf{X})],\\
\mathrm{Var}(n\tilde{S}_{25}) =& (2n^2\sigma^4)^{-1}\bigg(\mathrm{E}^2[d_n^2(\mathbf{X})] + 2n^{-1}\mathrm{E}[d_n^4(\mathbf{X})]\bigg) + \frac{2V_1^2}{n^3\sigma^4V_2} \sum_{m\in\mathbb{G}} \lambda_{\mathcal{K},m}^2a_m^2,
\end{aligned}$$

where $c_m = \mathrm{E}[\psi_m(\mathbf{X})d_n^2(\mathbf{X})]$, and $d_{m_1,m_2} = \mathrm{E}[\psi_{m_1}(\mathbf{X})\psi_{m_2}(\mathbf{X})d_n(\mathbf{X})]$. Under condition (2.3.6), all the triple sum terms are of small order. Finally consider the following quadruple

sum terms $\Theta_n^{(4)}$ in (2.8.10),

$$\tilde{S}_{31} = \frac{V_1}{n^4\sigma^2\sigma_{Tn}} \sum_{i\neq j\neq k\neq l} \mathcal{K}_\theta(\mathbf{X}_i, \mathbf{X}_j) U_k U_l, \quad \tilde{S}_{32} = \frac{V_1}{n^4\sigma^2\sigma_{Tn}} \sum_{i\neq j\neq k\neq l} \mathcal{K}_\theta(\mathbf{X}_i, \mathbf{X}_j) U_k \epsilon_l.$$

It can be shown that $\mathrm{E}(n\tilde{S}_{31}) = \mathrm{E}(n\tilde{S}_{31}) = 0$, and

$$\begin{aligned}
\mathrm{Var}(n\tilde{S}_{31}) =& \frac{V_1^2}{2n^2\sigma^2 V_2} \big(\mathrm{E}[d_n^2(\mathbf{X})] \sum_{m\in\mathbb{G}} \lambda_{\mathcal{K},m}^2 a_m^2 + \big(\sum_{m\in\mathbb{G}} \lambda_{\mathcal{K},m} a_m\big)^2 + \mathrm{E}^2[d_n^2(\mathbf{X})] V_{\mathcal{K},2}\big)\{1+o(1)\}, \\
\mathrm{Var}(n\tilde{S}_{32}) =& \frac{V_1^2}{2n^2\sigma^2 V_2} \big(\sum_{m\in\mathbb{G}} \lambda_{\mathcal{K},m}^2 a_m^2 + \mathrm{E}[d_n^2(\mathbf{X})] V_{\mathcal{K},2}\big)\{1+o(1)\},
\end{aligned}$$

are of order $o(1)$ under condition (2.3.6). Therefore, $n\sigma_{T_n}^{-1}\tilde{T}_{n1}^0 \xrightarrow{d} \Psi(d_n)$ under the local hypothesis H_{1n} (2.3.5). This finishes the proof of Theorem 2.

Proof to Theorem 3: First proving part (i). Consider the regularized oracle location shift $\Psi_R^O(d_n, \gamma)$, whose order is proportional to

$$f(\gamma) := \frac{\sum_{m\in S_1} g_m(\gamma)}{\sqrt{\sum_{m\in S_1} g_m^2(\gamma)}},$$

where $g_m(\gamma) = \lambda_m/(\lambda_m + \gamma)$. It can be shown that function $f(\gamma)$ is maximized when $g_m(\gamma)$ is a non-zero constant for $m \in S_1$. Denote $f_1(\gamma) = \sum_{m\in S_1} g_m(\theta)$, and $f_2(\gamma) = \sqrt{\sum_{m\in S_1} g_m^2(\gamma)}$. Since $f_1' = \sum g_m'(\gamma)$ and $f_2' = \sum g_m(\gamma) g_m'(\gamma)/f_2$, then $f'(\gamma) = 0$ (*i.e.*, $f_1' f_2 - f_1 f_2' = 0$) is equivalent to

$$\sum_{m_1\neq m_2\in S_1} g_{m_1}(\gamma) g_{m_2}'(\gamma) \Big(g_{m_1}(\gamma) - g_{m_2}(\gamma) \Big) = 0,$$

where $\hat{\gamma} = 0$ (i.e., $g_m(\hat{\gamma}) = 1$) is one of the solutions. Then we can show that $\mathrm{sgn}(f'')|_{\gamma=\hat{\gamma}} =$

$\mathrm{sgn}(f_1'' f_2 - f_1 f_2'')_{\gamma=\hat{\gamma}}$, where $(f_1'' f_2 - f_1 f_2'')_{\gamma=\hat{\gamma}} = -\sum_{m\in S_1} \lambda_m^{-2}|S_1|^2 + (\sum_{m\in S_1} \lambda_m^{-1})^2|S_1|$ is strictly less than zero when there exists at least one $m \in S_1$ such that $\lambda_m \neq 1$, by using Cauchy-Schwarz inequality. For the case where $\lambda_m = 1$ for all $m \in S_1$, $f(\gamma) = \sqrt{|S_1|}$ does not depend on $\hat{\gamma}$. On the other hand, using the Cauchy-Schwarz inequality, we have $|f(\gamma)| \leq \sqrt{|S_1|}$. Therefore, $\max_\gamma \Psi_R^O(d_n, \gamma) \sim \max_\gamma f(\gamma) C_n B_p = f(\hat{\gamma}) C_n B_p = \sqrt{|S_1|} C_n B_p$. Furthermore, if $\gamma^* = o(\lambda_N)$, then $g_m(\gamma^*)$ at the order of 1 for $m \leq N$ and $\mu_R^O(d_n, \gamma^*) \sim \sqrt{|S_1|} C_n B_p$. Hence $\max_\gamma \Psi_R^O(d_n, \gamma) \sim \Psi_R^O(d_n, \gamma^*) \sim \sqrt{|S_1|} C_n B_p$.

(ii): It is not difficult to see for a regularization parameter γ^* satisfying conditions in Theorem 3, $g_m(\gamma^*) \to 1$ for $m = 1, \cdots, N$, and $g_m(\gamma^*) \to 0$ for $m \geq N_2$. Since $\gamma^* = o(\lambda_N)$, there exists $\epsilon_1 > 0$ small enough s.t. $\gamma^* < \epsilon_1 \lambda_N$, hence $\dfrac{|S_1|}{1+\epsilon_1} \leq \sum_{m\in S_1} \dfrac{\lambda_m}{\lambda_m + \gamma^*} \leq |S_1|$. Similarly, there exists $\epsilon_2 > 0$ small enough s.t. $\lambda_{N_2} < \epsilon_2 \gamma^*$, and $\dfrac{R_2}{(1+\epsilon_2)^2 \gamma^{*2}} \leq \sum_{m=N_2}^{\infty} \left(\dfrac{\lambda_m}{\lambda_m + \gamma^*}\right)^2 \leq \dfrac{R_2}{\gamma^{*2}}$. Then, we have

$$\Psi_R(d_n, \gamma^*) \geq \frac{J_1 |S_1| C_n B_p}{\sqrt{N \log N + R_2/\gamma^{*2}}} \tag{2.8.12}$$

for some positive constant J_1. Assuming $\gamma^* = J_0 \lambda_{N_1}$, then

$$\Psi_R(d_n, \gamma^*) \leq \frac{J_2 |S_1| C_n B_p}{\sqrt{N(1+\epsilon_2)^2/(1+\epsilon_1)^2 + (N \log N - N) J_0^{-2} (c_{N_2}/c_{N_1})^2 + R_2/\gamma^{*2}}}. \tag{2.8.13}$$

Since ϵ_1 and ϵ_2 go to 0, and $R_2/\gamma^{*2} = o(N)$, we obtain the conclusion in part (ii) by combining (2.8.12) and (2.8.13).

Lemma 2. *If kernel $K_\theta^*(x_1, x_2)$ is a positive definite kernel, then the centralized kernel $K_\theta(x_1, x_2)$ is positive semi-definite.*

Proof: Assume kernel function $K_\theta^*(x, y)$ has eigen-deposition $\{\lambda_m^*, \psi_m^*(\cdot)\}_{m=1}^\infty$, and cen-

tralized kernel $K_\theta(x, y)$ has eigen-decomposition $\{\lambda_m, \psi_m(\cdot)\}_{m=1}^{\infty}$. Since the kernel can be normalized, we assume the sum of eigenvalues are bounded without loss of generality. Recall the definition of the centralized kernel function

$$K_\theta(\mathbf{x}_1, \mathbf{x}_2) = K^*_\theta(\mathbf{x}_1, \mathbf{x}_2) - K^*_{1,\theta}(\mathbf{x}_1) - K^*_{1,\theta}(\mathbf{x}_2) + \mu_{K^*}.$$

Then we have $\mathrm{E}[K_\theta(\mathbf{x}_1, \mathbf{X}_2)] = \mathrm{E}[K^*_\theta(\mathbf{x}_1, \mathbf{X}_2)] - K^*_{1,\theta}(\mathbf{x}_1) = 0$, or equivalently,

$$\int 1 \cdot K_\theta(\mathbf{x}_1, \mathbf{x}_2) d\mu(\mathbf{x}_2) = 0,$$

which implies that $\psi_{m^*}(\cdot) = 1$ is one of the eigenfunctions corresponding to zero eigenvalue. Due to the orthogonality of the system, $\mathrm{E}\{\psi_m(\mathbf{X})\} = 0$ for $m \neq m^*$. By the eigen-decomposition equality (2.8.3), we have

$$\lambda_m \psi_m(\mathbf{x}_1) = \mathrm{E}\{K_\theta(\mathbf{x}_1, \mathbf{X}_2)\psi_m(\mathbf{X}_2)\} = \mathrm{E}\{K^*_\theta(\mathbf{x}_1, \mathbf{X}_2)\psi_m(\mathbf{X}_2)\} - \mathrm{E}\{K^*_{1,\theta}(\mathbf{X}_2)\psi_m(\mathbf{X}_2)\}$$
$$+ \{\mu_{K^*} - K^*_{1,\theta}(\mathbf{x}_1)\}\mathrm{E}\{\psi_m(\mathbf{X}_2)\} = \mathrm{E}\{K^*_\theta(\mathbf{x}_1, \mathbf{X}_2)\psi_m(\mathbf{X}_2)\} - \mathrm{E}\{K^*_{1,\theta}(\mathbf{X}_2)\psi_m(\mathbf{X}_2)\},$$

for any $m \neq m^*$. By plugging in $K^*_\theta(\mathbf{x}_1, \mathbf{x}_2) = \sum_{m=1}^{\infty} \lambda^*_m \psi^*_m(\mathbf{x}_1)\psi^*_m(\mathbf{x}_2)$, and multiplying $\psi_m(\mathbf{x}_1)$ to both sides, we have

$$\lambda_m \psi^2_m(\mathbf{x}_1) = \sum_s \lambda^*_s \psi^*_s(\mathbf{x}_1)\psi_m(\mathbf{x}_1)\mathrm{E}\{\psi^*_s(\mathbf{X}_2)\psi_m(\mathbf{X}_2)\} - \mathrm{E}\{K^*_{1,\theta}(\mathbf{X}_2)\psi_m(\mathbf{X}_2)\}\psi_m(\mathbf{x}_1),$$

for $m \neq m^*$. Taking expectation with respect to $\mathbf{X}_1$ and using the orthogonal normal property,

$$\lambda_m = \sum_{m=1}^{\infty} \lambda^*_m \mathrm{E}^2[\psi^*_m(\mathbf{X})\psi_m(\mathbf{X})] \geq 0, \quad m \neq m^*.$$

In addition, $\lambda_{m^*} = 0$, then the positive semi-definiteness of centralized kernel function can be achieved. $\square$

Proof of Remark 1: Let $T_n = \frac{1}{n(n-1)} \sum_{i \neq j} K_\theta(\mathbf{X}_i, \mathbf{X}_j)(Y_i - \bar{Y})(Y_j - \bar{Y})/\hat{\sigma}^2$, and $T_{n1} = \frac{1}{n(n-1)} \sum_{i \neq j} K_\theta(\mathbf{X}_i, \mathbf{X}_j)(Y_i - \bar{Y})(Y_j - \bar{Y})/\sigma^2$ be the statistics using the true centralized kernel K_θ, $\hat{T}_n = \frac{1}{n(n-1)} \sum_{i \neq j} K_{n,\theta}(\mathbf{X}_i, \mathbf{X}_j)(Y_i - \bar{Y})(Y_j - \bar{Y})/\hat{\sigma}^2$ and $\hat{T}_{n1} = \frac{1}{n(n-1)} \sum_{i \neq j} K_{n,\theta}(\mathbf{X}_i, \mathbf{X}_j)(Y_i - \bar{Y})(Y_j - \bar{Y})/\sigma^2$ be the ones using empirically centralized kernel $K_{n,\theta}$. Using the similar arguments in proof of Theorem 2, $nT_n/\sqrt{V_2} = nT_{n1}/\sqrt{V_2}\{1 + o_p(1)\}$ and $n\hat{T}_n/\sqrt{V_2} = n\hat{T}_{n1}/\sqrt{V_2}\{1 + o_p(1)\}$. To show $nT_n/\sqrt{V_2} = n\hat{T}_n/\sqrt{V_2}\{1 + o_p(1)\}$, it remains to show $(nT_{n1} - n\hat{T}_{n1})/\sqrt{V_2} = o_p(1)$.

In fact, $\Delta_{n,D} := V_2^{-1/2}(nT_{n1} - n\hat{T}_{n1}) = \frac{1}{n-1} \sum_{i \neq j} \mathbf{D}_{ij}(Y_i - \bar{Y})(Y_j - \bar{Y})/\sigma^2$, where $\mathbf{D} = V_2^{-1/2}(\mathbf{K} - \mathbf{K}_n)$, $D_{ij} = V_2^{-1/2}\{K^*_{1,\theta}(\mathbf{X}_j) - (n-1)^{-1} \sum_{k \neq j} K^*_{kj} + K^*_{1,\theta}(\mathbf{X}_i) - (n-1)^{-1} \sum_{k \neq i} K^*_{ki} + n^{-1}(n-1)^{-1} \sum_{k \neq l} K^*_{kl} - \mu_{K^*}\}$ and $K^*_{ij} = K^*(\mathbf{X}_i, \mathbf{X}_j)$. Viewing $\Delta_{n,D}$ as a special case that was considered in proof of Theorem 2, it is not difficult to see that $\mathrm{E}\Delta_{n,D} = 0$, and the asymptotic variance of $\Delta_{n,D}$ is $2\mathrm{E}(D^2_{ij}) \leq CV_2^{-1}(2\sigma^2_{\Delta,1} + \sigma^2_{\Delta,2})$ for some constant C, where $\sigma^2_{\Delta,1} = \mathrm{E}[K^*_{1,\theta}(\mathbf{X}_j) - (n-1)^{-1} \sum_{k \neq j} K^*_{kj}]^2$ and $\sigma^2_{\Delta,2} = \mathrm{E}[n^{-1}(n-1)^{-1} \sum_{k \neq l} K^*_{kl} - \mu_{K^*}]^2$. In the following we will show that $V_2^{-1}(2\sigma^2_{\Delta,1} + \sigma^2_{\Delta,2}) = o(1)$.

Let $\{\lambda^*_m, \psi^*_m\}_{m=1}^\infty$ be the eigen-decomposition of kernel K^*. Denote $V_2^* = \sum_{m=1}^\infty \lambda^{*2}_m$, $\kappa_m = \mathrm{E}\{\psi^*_m(X)\}$, $\nu_1 = \sum_{m=1} \lambda^*_m \kappa^2_m$ and $\nu_2 = \sum_{m=1} \lambda^{*2}_m \kappa^2_m$. Since the $\Delta_{D,n}$ is invariant when the kernel is scaled, we can assume $\max_m \lambda^*_m = 1$ without loss of generality. Then it can be shown that $\sigma^2_{\Delta,1} = n^{-1}(V_2^* - \nu_2)$ and $\sigma^2_{\Delta,2} = 4n^{-1}\nu_2 + n^{-2}V_2^*$. Moreover, it has been studied in Lindsay *et al.* (2014) that $V_2 = V_2^* - 2\nu_2 + \nu_1^2$, where $\nu_2 \leq \nu_1 \leq 1 \leq \sqrt{V_2^*}$,

Therefore,

$$V_2^{-1}(2\sigma_{\Delta,1}^2+\sigma_{\Delta,1}^2)=\frac{2}{n}\frac{V_2^*+\nu_2}{V_2}\{1+o(1)\}\leq\frac{2}{n}\frac{V_2^*+\sqrt{V_2^*}}{V_2^*-2\nu_2+\nu_2^2},$$

which is $o(1)$ no matter V_2^* is infinite or finite. □

Proof of Proposition 1: Under alternative $H_{1n}: h(\mathbf{x})=d_n(\mathbf{x})$, we have $Y_i=\mu+d_n(\mathbf{X}_i)+\epsilon_i$ for $i=1,2...,n$, where $\mathrm{E}(\mathbf{Y}_i)=0$ and $\mathrm{E}(\mathbf{Y}_i^2)=\sigma^2+\mathrm{E}\{d_n^2(\mathbf{X}_i)\}=\sigma^2\{1+o(1)\}$, under condition (3.11). Therefore, by the law of large number

$$\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^n(Y_i-\bar{Y})^2=\frac{1}{n}\sum_{i=1}^n Y_i^2-(\bar{Y})^2\xrightarrow{p}\sigma^2.$$

Moreover,

$$\frac{\hat{\delta}_K-\delta_K}{\sigma_{T_n}}=\left(\frac{nT_n-n\delta_K/\hat{\sigma}^2}{\sigma_{T_n}}\right)\cdot\frac{\hat{\sigma}^2}{n}=O_p(1)\cdot\frac{\hat{\sigma}^2}{n}\xrightarrow{p}0,$$

under condition (3.11). □

Derivation of the adjusted variance $\sigma_{T_n,adj}^2$: Consider

$$n(n-1)T_{n1}=\frac{2}{\sigma^2}\mathbf{Y}^T\mathbf{H}\mathbf{K}^0\mathbf{H}\mathbf{Y}-\frac{1}{\sigma^4(n-1)}\mathbf{Y}^T\mathbf{H}\mathbf{K}^0\mathbf{H}\mathbf{Y}\mathbf{Y}^T\mathbf{H}\mathbf{Y}\triangleq G_1-G_2,$$

By using results from Zhong and chen (2011), we have $\mathrm{E}(G_1) = 2\mathrm{tr}(\mathbf{HK}^0)$,

$$\begin{aligned}
\mathrm{E}(G_2) &= \frac{1}{(n-1)}[\mathrm{tr}(\mathbf{HK}^0)\mathrm{tr}(\mathbf{H}) + 2\mathrm{tr}(\mathbf{HK}^0) + \Delta\mathrm{tr}(\mathbf{HK}^0\mathbf{H} \circ \mathbf{H})], \\
\mathrm{Var}(G_1) &= 8\mathrm{tr}(\mathbf{HK}^0\mathbf{HK}^0) + 4\Delta\mathrm{tr}(\mathbf{A} \circ \mathbf{A}), \\
\mathrm{Cov}(G_1, G_2) &= \frac{1}{(n-1)}[16\mathrm{tr}(\mathbf{HK}^0\mathbf{HK}^0) + 4\mathrm{tr}^2(\mathbf{HK}^0) + 4\mathrm{tr}(\mathbf{HK}^0\mathbf{HK}^0)\mathrm{tr}(\mathbf{H})] \\
&+ \frac{2\Delta}{(n-1)}[\mathrm{tr}(\mathbf{HK}^0)\mathrm{tr}(\mathbf{A} \circ \mathbf{H}) + \mathrm{tr}(\mathbf{H})\mathrm{tr}(\mathbf{A} \circ \mathbf{A}) + \mathrm{tr}(\mathbf{A}^2 \circ \mathbf{H}) + 2\mathrm{tr}(\mathbf{A} \circ \mathbf{A})] \\
&+ \frac{2(\tau_6 - 15 - 6\Delta)}{(n-1)}\mathrm{tr}(\mathbf{A} \circ \mathbf{A} \circ \mathbf{H})
\end{aligned}$$

where $\tau_k = E(\frac{Y-\mu}{\sigma})^k$ for any $k \in \mathcal{N}$. Applying the results from (Bao and Ullah, 2010),

$$\begin{aligned}
&\mathrm{E}(\frac{1}{\sigma^8}\mathbf{Y}^T\mathbf{AYY}^T\mathbf{AYY}^T\mathbf{HYY}^T\mathbf{HY}) \\
= \;& \mathrm{tr}^2(\mathbf{HK}^0)\mathrm{tr}^2(\mathbf{H}) + 10\mathrm{tr}^2(\mathbf{HK}^0)\mathrm{tr}(\mathbf{H}) + 2\mathrm{tr}^2(\mathbf{H})\mathrm{tr}(\mathbf{HK}^0\mathbf{HK}^0) + 20\mathrm{tr}(\mathbf{HK}^0\mathbf{HK}^0)\mathrm{tr}(\mathbf{H}) \\
&+24\mathrm{tr}^2(\mathbf{HK}^0) + 48\mathrm{tr}(\mathbf{HK}^0\mathbf{HK}^0) + R_n
\end{aligned}$$

where $R_n = \gamma_2 f_{\gamma_2} + \gamma_4 f_{\gamma_4} + \gamma_6 f_{\gamma_6} + \gamma_2^2 f_{\gamma_2^2}$, $\gamma_2 = \tau_4 - 3 = \Delta$, $\gamma_4 = \tau_6 - 15\Delta - 15$, $\gamma_6 = \tau_8 - 28\gamma_4 - 35\Delta^2 - 210\Delta - 105$, and

$$\begin{aligned}
f_{\gamma_2} &= \mathrm{tr}^2(\mathbf{HK}^0)\Big\{\frac{5(n-1)^2}{n} + 8 + 24\frac{(n-1)}{n}\Big\} + \mathrm{tr}(\mathbf{HK}^0\mathbf{HK}^0)\Big\{\frac{10(n-1)^2}{n} + \frac{64(n-1)}{n}\Big\} \\
&\quad + \mathrm{tr}(\mathbf{A} \circ \mathbf{A})\Big\{(n-1)^2 + \frac{2(n-1)^2}{n} + 16(n-1) + 48 + \frac{16(n-1)}{n}\Big\}, \\
f_{\gamma_4} &= 2\mathrm{tr}^2(\mathbf{HK}^0)(\frac{n-1}{n})^2 + 4\mathrm{tr}(\mathbf{HK}^0\mathbf{HK}^0)(\frac{n-1}{n})^2 + \mathrm{tr}(\mathbf{A} \circ \mathbf{A})(\frac{n-1}{n})(2n+18), \\
f_{\gamma_6} &= \mathrm{tr}(\mathbf{A} \circ \mathbf{A})(\frac{n-1}{n})^2, \\
f_{\gamma_2^2} &= \mathrm{tr}^2(\mathbf{HK}^0)\Big(2 - \frac{2}{n} + \frac{4}{n^2}\Big) + \mathrm{tr}(\mathbf{A} \circ \mathbf{A})\Big\{8 - \frac{16}{n} + \frac{(n-1)^2}{n}\Big\} \\
&\quad + \mathrm{tr}(\mathbf{HK}^0\mathbf{HK}^0)\Big(24 - \frac{32}{n} + \frac{12}{n^2}\Big).
\end{aligned}$$

Hence,

$$\begin{aligned}\mathrm{Var}(G_2) &= \frac{1}{(n-1)^2}\mathrm{E}(\frac{1}{\sigma^8}\mathbf{Y}^T\mathbf{A}\mathbf{Y}\mathbf{Y}^T\mathbf{A}\mathbf{Y}\mathbf{Y}^T\mathbf{H}\mathbf{Y}\mathbf{Y}^T\mathbf{H}\mathbf{Y}) - \frac{1}{(n-1)^2}\{\mathrm{E}(G_2)\}^2 \\ &= \mathrm{tr}^2(\mathbf{H}\mathbf{K}^0)\big\{\frac{6}{n-1} + \frac{20}{(n-1)^2}\big\} + \mathrm{tr}(\mathbf{H}\mathbf{K}^0\mathbf{H}\mathbf{K}^0)\big\{2 + \frac{20}{n-1} + \frac{48}{(n-1)^2}\big\} + \frac{R_n}{(n-1)^2} \\ &- \frac{1}{(n-1)^2}\big\{4\Delta\mathrm{tr}(\mathbf{H}\mathbf{K}^0)\mathrm{tr}(\mathbf{A}\circ\mathbf{H}) + \Delta^2\mathrm{tr}^2(\mathbf{A}\circ\mathbf{H}) + 2\Delta\mathrm{tr}(\mathbf{H}\mathbf{K}^0)\mathrm{tr}(\mathbf{H})\mathrm{tr}(\mathbf{A}\circ\mathbf{H})\big\}.\end{aligned}$$

Denote

$$S_1 = \mathrm{tr}^2(\mathbf{H}\mathbf{K}^0)\big(-\frac{2}{n-1}\big) + \mathrm{tr}(\mathbf{H}\mathbf{K}^0\mathbf{H}\mathbf{K}^0)\big(2 - \frac{12}{n-1}\big),$$

and

$$S_2 = \mathrm{tr}^2(\mathbf{H}\mathbf{K}^0)\big(-\frac{\Delta}{n}\big) + \mathrm{tr}(\mathbf{H}\mathbf{K}^0\mathbf{H}\mathbf{K}^0)\big(\frac{6\Delta}{n}\big) + \Delta\mathrm{tr}(\mathbf{A}\circ\mathbf{A}),$$

then we have

$$\begin{aligned}\mathrm{Var}(G_1 - G_2) &= S_1 + S_2 + o(S_1 + S_2) \\ &= \mathrm{tr}(\mathbf{H}\mathbf{K}^0\mathbf{H}\mathbf{K}^0)\{2 - \frac{12}{(n-1)^2} + \frac{6\Delta}{n}\} + \mathrm{tr}^2(\mathbf{H}\mathbf{K}^0)\{-\frac{2}{n-1} - \frac{\Delta}{n}\} \\ &+ \Delta\mathrm{tr}(\mathbf{A}\circ\mathbf{A}) + o(S_1 + S_2),\end{aligned}$$

where it can be proved $\mathrm{tr}(\mathbf{A}\circ\mathbf{A}) = \big\{\frac{2\mathrm{tr}(\mathbf{K}^2)}{n} - \frac{1}{n}\mathrm{tr}^2(\mathbf{H}\mathbf{K}^0) - \frac{2}{n}\mathrm{tr}(\mathbf{H}\mathbf{K}^0\mathbf{H}\mathbf{K}^0)\big\}\{1 + o(1)\}$.

Therefore,

$$\mathrm{Var}(nT_{n1}) = \frac{1}{(n-1)^2}\mathrm{Var}(G_1 - G_2),$$

which is an adjustment for the variance of test statistic nT_n, since $\mathrm{Var}(nT_n) = \mathrm{Var}(nT_{n1})\{1 + o(1)\}$. This finishes the proof. □

Let T_K be the integral operator defined using kernel K, i.e., $T_K f = \int K(x,\cdot)f(x)d\mu(x)$ for $f \in L^2(\mu)$. Then the eigenvalues corresponding to kernel function K are actually the ones correspond to integral operator, and we denote them by $\lambda(T_K)$ in the following.

Lemma 3. *For the given regularized kernel $K_{R,\gamma}$ in the chapter, we have*

$$\lambda_m(T_{K_{R,\gamma}}) = \frac{\gamma\lambda_m(T_K)}{\gamma + \lambda_m(T_K)}.$$

Proof: Applying the result from Dauxois *et al.* (1982), we have

$$n^{-1}\lambda_m(\mathbf{K}) - \lambda_m(T_K) \overset{a.s.}{\to} 0, \quad n \to \infty \tag{2.8.14}$$

and

$$n^{-1}\lambda_m(\mathbf{K}_{R,\gamma}) - \lambda_m(T_{K_{R,\gamma}}) \overset{a.s.}{\to} 0, \quad n \to \infty \tag{2.8.15}$$

for any integer m. Next we will show that $\lambda(\mathbf{K}_{R,\gamma}) = \gamma\lambda(\mathbf{K})/\{\lambda(\mathbf{K}) + \gamma\}$. If we assume kernel matrix $\mathbf{K}$ has eigen-decomposition $\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T$, where $\mathbf{\Lambda} = \text{diag}\{\Lambda_1, ..., \Lambda_n\}$, then

$$\begin{aligned}
\mathbf{K}_{R,\gamma} &= \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T - \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T(n\gamma\mathbf{Q}\mathbf{Q}^T + \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T)^{-1}\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T \\
&= \mathbf{Q}\big\{\mathbf{\Lambda} - \mathbf{\Lambda}(n\gamma\mathbf{I} + \mathbf{\Lambda})^{-1}\mathbf{\Lambda}\big\}\mathbf{Q}^T,
\end{aligned}$$

which implies $\lambda_m(\mathbf{K}_{R,\gamma}) = \Lambda_m - \dfrac{\Lambda_m^2}{n\gamma + \Lambda_m} = \dfrac{\gamma\Lambda_m}{\gamma + \Lambda_m/n}$. Hence we have

$$n^{-1}\lambda_m(\mathbf{K}_{R,\gamma}) - \frac{\gamma\lambda_m(T_K)}{\gamma + \lambda_m(T_K)} = \frac{\gamma\lambda_m(\mathbf{K})/n}{\gamma + \lambda_m(\mathbf{K})/n} - \frac{\gamma\lambda_m(T_K)}{\gamma + \lambda_m(T_K)} \overset{a.s.}{\to} 0. \tag{2.8.16}$$

Combing (2.8.15) and (2.8.16), we can see $\lambda_m(T_{K_{R,\gamma}}) = \dfrac{\gamma\lambda_m(T_K)}{\gamma + \lambda_m(T_K)}$. □

Chapter 3

A rate-optimal test for high-dimensional linear model

3.1 Introduction

In the previous chapter, we proposed a test statistic to test the nonparametric function of high-dimensional variates in a RKHS generated by a kernel function K. In particular, the functions are of form $h(x) = \sum_{m=1}^{S} \alpha_m \psi_m(x)$, where $\{\psi_m(\cdot)\}_{m=1}^{S}$ correspond to the eigen-decomposition of kernel function K and form a complete orthogonal normal system, S is the total number of positive eigen-values of kernel function K. Testing the nonparametric functions is essentially equivalent to the problem of detecting whether the unknown coefficient vector $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_S)$ is zero (i.e., Y_i are independent of covariants $\mathbf{X}_i$) or not. Specifically, given a predefined norm $\|\cdot\|$ on $\mathbb{R}^S$, let us consider testing the hypothesis

$$H_0 : \boldsymbol{\alpha} = \mathbf{0} \quad \text{vs} \quad H_1(\varepsilon_n^2) : \|\boldsymbol{\alpha}\| \geq \varepsilon_n \tag{3.1.1}$$

where $\varepsilon_n > 0$ is the separation radius. As we can observe, the smaller ε_n is, the more difficult to distinguish between H_0 and $H_1(\varepsilon_n^2)$. An interesting question is: what is the smallest rate ε_n such that it is still possible to successfully detect the alternative $H_1(\varepsilon_n^2)$ in (3.1.1)? Moreover, is the kernel-based test able to achieve the detection boundary?

As an initial investigation, in this chapter we mainly focus on a high-dimensional linear regression model as follows:

$$Y_i = \mu + \sum_{m=1}^{p} \beta_m X_{im} + \epsilon_i, i = 1, ..., n, \tag{3.1.2}$$

where $\beta_m \in \mathbb{R}$ are unknown coefficients, $\mathbf{X}_i = (X_{i1}, ..., X_{ip})^T$ is a p-dim Gaussian random variable with mean vector $\mathbf{0}_p$ and covariance matrix $\mathbf{\Sigma}$; the dimension p goes to infinity as n goes to infinity; ϵ_i are IID random Gaussian errors with mean 0 and known variance σ^2 (then assuming $\sigma^2 = 1$ without loss of generality); ϵ_i are independent of $\mathbf{X}_i (1 \leq i \leq n)$.

This optimal testing problem in (3.1.2) is closely related to several existing works on detection boundary for Gaussian models (Donoho and Jin, 2004; Donoho and Jin, 2008; Donoho and Jin, 2009; Hall and Jin, 2010; Ingster et al., 2010; Ingster et al., 2009). Specifically, Donoho and Jin (2004), Hall and Jin (2010) and Ingster et al. (2009) considered the model with $X_{ij} = 1_{\{i=j\}}$ and dimension $p = n$, and Donoho and Jin (2008, 2009) investigated the model (3.1.2) with $X_{ij} = Z_i 1_{\{i=j\}}$ ($Z_i = \pm 1$ is class label) and $p = n$ under a classification setting. As an extension to linear regression under high-dimensional setting $p >> n$, the detection boundary of model (3.1.2) with the covariance matrix $\mathbf{\Sigma} = \mathbf{I}$ was studied in (Ingster et al., 2010)). However, those works were all developed under sparsity assumption that the majority of the coefficients are zero (e.g., the proportion of the nonzero coefficients is $p^{-\gamma}$, $\gamma \in (0, 1)$). The major contribution of the present work lies on the investigation of detection boundary under a general structure-free (no sparsity assumption) high-dimensional linear model with correlated variables, which have been more commonly seen in practice.

The rest of the chapter is organized as follows. After introducing the basic notation

and definition for minimax testing problem in Section 3.2, we establish a lower bound of the detection boundary ε_n in Section 3.3. Section 3.4 introduces the kernel-based test and provides its asymptotic distributions under the null and a general alternative hypotheses, which leads to the establishment of its non-trivial power at the detection boundary ε_n, under certain conditions. Summary and further discussions are given in Section 3.5.

3.2 Minimax testing problem

Assume the covariance matrix has decomposition $\mathbf{\Sigma} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T$, $\mathbf{\Lambda} = diag(\lambda_1, ..., \lambda_p)$ is a diagonal matrix, and $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p > 0$. Noting linear transformations $\tilde{\mathbf{X}} = \mathbf{\Lambda}^{-1/2}\mathbf{Q}^T\mathbf{X}$ can generate p-dim Gaussian random variable $\tilde{\mathbf{X}}$ with mean $\mathbf{0}_p$ and identity covariance matrix, we can see that $(\tilde{X}_1, ..., \tilde{X}_p)$ forms orthogonal normal system (i.e., $\mathrm{E}(\tilde{X}_i\tilde{X}_j) = 1_{\{i=j\}}$, $i, j = 1, ..., p$). Moreover, linear kernel function $K(\mathbf{X}_1, \mathbf{X}_2) = \mathbf{X}_1^T\mathbf{X}_2$ has alternative expansion $K(\mathbf{X}_1, \mathbf{X}_2) = \tilde{\mathbf{X}}_1^T\mathbf{\Lambda}\tilde{\mathbf{X}}_2$. The above two observations together imply that the function space generated by the linear kernel includes functions of form

$$h(\mathbf{X}_i) = \boldsymbol{\alpha}^T\tilde{\mathbf{X}}_i = \boldsymbol{\alpha}^T\mathbf{\Lambda}^{-1/2}\mathbf{Q}^T\mathbf{X}_i, \quad i = 1, ..., n,$$

which is equivalent to the model (3.1.2) by letting $\boldsymbol{\beta} = \mathbf{Q}\mathbf{\Lambda}^{-1/2}\boldsymbol{\alpha}$. Therefore, testing the high dimensional linear function is equivalent to testing the high dimensional coefficient vector $\boldsymbol{\alpha} = \mathbf{0}$. In the following, we will formalize this problem from an asymptotic minimax point of view.

For any coefficient vector $\boldsymbol{\alpha} \in \mathbb{R}^p$, we define norm $\|\boldsymbol{\alpha}\|_{\mathbf{\Lambda}}^2 = \boldsymbol{\alpha}^T\mathbf{\Lambda}\boldsymbol{\alpha}$, and incorporate it

into the following testing problem

$$H_0 : \|\boldsymbol{\alpha}\|_{\boldsymbol{\Lambda}}^2 = 0 \quad \text{vs} \quad H_1(\varepsilon_n^2) : \|\boldsymbol{\alpha}\|_{\boldsymbol{\Lambda}}^2 \geq \varepsilon_n^2, \tag{3.2.1}$$

under the high-dimensional setting where $p \to \infty$ as $n \to \infty$.

We denote a test based on the n observations as $\phi_n \in \{0, 1\}$, where $\phi_n = 1$ represents a rejection of H_0, and $\phi_n = 0$ indicates a decision of retaining H_0. Therefore, the type I error of the test is

$$\alpha(\phi_n) = P_0(\phi_n = 1),$$

which is assumed to be asymptotically controlled at level α, i.e., $\alpha(\phi_n) \leq \alpha + o(1)$ and the minimax type II error of the test is defined as

$$\pi(\phi_n, \varepsilon_n^2) = \sup_{h \in H_1(\varepsilon_n^2)} P_h(\phi_n = 0),$$

where P_0 and P_h are the probability measures that correspond to observations under the null hypothesis $\boldsymbol{\alpha} = \mathbf{0}$ and the alternative hypothesis $h(\mathbf{X}_i) = \boldsymbol{\alpha}^T \tilde{\mathbf{X}}_i = \boldsymbol{\alpha}^T \boldsymbol{\Lambda}^{-1/2} \mathbf{Q}^T \mathbf{X}_i$. Then the minimax power, i.e., the worst power against $H_1(\varepsilon_n^2)$ is $1 - \pi(\phi_n, \varepsilon_n^2)$. Following Guerre and Levergne (2002), we call a test ϕ_n has an asymptotically trivial power against $H_1(\varepsilon_n^2)$ if the minimax power is no larger than the significance level $\alpha \in (0, 1)$, i.e., $1 - \pi(\phi_n, \varepsilon_n^2) \leq \alpha + o(1)$. Accordingly, a test ϕ_n has an asymptotically non-trivial power against $H_1(\varepsilon_n^2)$ if $1 - \pi(\phi_n, \varepsilon_n^2) > \alpha + o(1)$.

We want to find a separation rate $\tilde{\varepsilon}_n$ such that for any rate ε_n that goes to zero faster than $\tilde{\varepsilon}_n$ (i.e., $\varepsilon_n = o(\tilde{\varepsilon}_n)$), any test will have trivial power against $H_1(\varepsilon_n^2)$, while there existing a test ϕ_n^* that has non-trivial power against $H_1(\varepsilon_n^2)$ if $\varepsilon_n^2 = \kappa \tilde{\varepsilon}_n^2$ for some constant

$\kappa > 0$. This special separation rate $\tilde{\varepsilon}_n$ and the test ϕ_n^* are normally called optimal minimax rate (or detection boundary) and rate optimal test, respectively.

3.3 Lower bound of separation rate

In this section, we establish a lower bound for the optimal minimax rate $\tilde{\varepsilon}_n$, as stated in the following theorem.

Theorem 4. *Let* $\tilde{\varepsilon}_n^2 = \lambda_p/\sqrt{n}$ *and assume that* $\log p = O(n^{1/2})$. *For any test* ϕ_n *with* $\alpha(\phi_n) \leq \alpha + o(1)$, *the minimax type II error* $\pi(\phi_n, \varepsilon_n^2) \geq 1 - \alpha + o(1)$ *when* $\varepsilon_n = o(\tilde{\varepsilon}_n)$.

Proof. Note that it is enough to show $\pi(\phi_n, \varepsilon_n^2) + \alpha(\phi_n) \geq 1 + o(1)$. Following the lower bound investigations in Cai and Ma (2013) and Guerre and Levergne (2002), we consider a least favorable subset of the alternative hypothesis $H_1(\varepsilon_n^2)$ in (3.2.1)

$$H_1^*(\varepsilon_n^2) = \Big\{ \boldsymbol{\alpha}^T \tilde{\mathbf{x}}, \boldsymbol{\alpha} \in \Theta(\varepsilon_n^2) \Big\}, \tag{3.3.1}$$

where

$$\Theta(\varepsilon_n^2) = \Big\{ \boldsymbol{\alpha}_\nu \Big| \boldsymbol{\alpha}_\nu = \varepsilon_n \nu, \nu \in \{\lambda_1^{-1/2}\mathbf{e}_1, \lambda_2^{-1/2}\mathbf{e}_2, ..., \lambda_p^{-1/2}\mathbf{e}_p\} \Big\},$$

where $\mathbf{e}_k$ is the p-dim vector with kth element being 1 and others being 0. It is not difficult to verify that for any $\boldsymbol{\alpha}_\nu \in \Theta(\varepsilon_n^2)$, $\boldsymbol{\alpha}_\nu^T \boldsymbol{\Lambda} \boldsymbol{\alpha}_\nu = \varepsilon_n^2$.

Under a Bayesian setting, we introduce a prior probability over alternative $H_1(\varepsilon_n^2)$ as

$$\Pi_{1n}(h(\mathbf{x}) = \boldsymbol{\alpha}^T \tilde{\mathbf{x}} | \boldsymbol{\alpha} \in \Theta(\varepsilon_n^2)) = \frac{1}{p}.$$

By using the Bayesian prior measures, we can obtain a lower bound of the sum of errors

$$\begin{aligned}\alpha(\phi_n) + \pi(\phi_n, \varepsilon_n^2) &= P_0(\phi_n = 1) + \sup_{h \in H_1(\varepsilon_n^2)} P_h(\phi_n = 0) \\ &\geq P_0(\phi_n = 1) + \int P_{h \in H_1(\varepsilon_n^2)}(\phi_n = 0) d\Pi_{1n}(h), \qquad (3.3.2)\end{aligned}$$

where the lower bound represents the Bayes error of any test ϕ_n, which is larger than the error of the optimal Bayesian likelihood ratio test given below.

Let $\mathsf{X} = \{\mathbf{X}_i\}_{i=1}^n$, $\mathsf{Y} = \{Y_i\}_{i=1}^n$, and denote $p_h(\mathsf{X}, \mathsf{Y})$ as the density function of (X, Y), where the subscript h acknowledges the dependence of X and Y through the regression function h. Plugging in the prior measures on the function h, we can define the joint densities $\mathbf{p}_0(\mathsf{X}, \mathsf{Y})$ and $\mathbf{p}_{1n}(\mathsf{X}, \mathsf{Y})$, associated with H_0 and $H_1(\varepsilon_n^2)$ respectively, as

$$\mathbf{p}_0(\mathsf{X}, \mathsf{Y}) = p_0(\mathsf{X}, \mathsf{Y}) \text{ and } \mathbf{p}_{1n}(\mathsf{X}, \mathsf{Y}) = \int p_h(\mathsf{X}, \mathsf{Y}) d\Pi_{1n}(h).$$

The optimal Bayesian likelihood ratio test rejects the null hypothesis if

$$L_n = \frac{\mathbf{p}_{1n}(\mathsf{X}, \mathsf{Y})}{\mathbf{p}_0(\mathsf{X}, \mathsf{Y})} = \frac{\mathbf{p}_{1n}(\mathsf{Y}|\mathsf{X})}{\mathbf{p}_0(\mathsf{Y}|\mathsf{X})} \geq 1$$

and the corresponding Bayesian error is

$$\begin{aligned}& 1 - \frac{1}{2} \int \int |\mathbf{p}_{1n}(\mathsf{X}, \mathsf{Y}) - \mathbf{p}_0(\mathsf{X}, \mathsf{Y})| d\mathsf{X} d\mathsf{Y} \\ = \; & 1 - \frac{1}{2} \int \left(\int |L_n - 1| \cdot \mathbf{p}_0(\mathsf{Y}|\mathsf{X}) d\mathsf{Y} \right) \mathbf{p}(\mathsf{X}) d\mathsf{X} \\ = \; & 1 - \frac{1}{2} \mathrm{E}_{\mathsf{X}} \Big[\mathrm{E}_0 \big\{ |L_n - 1| \big| \mathsf{X} \big\} \Big] \qquad (3.3.3)\end{aligned}$$

where E_0 is the conditional expectation with respect to $\mathbf{p}_0(\mathsf{Y}|\mathsf{X})$, and E_{X} is the expectation with respect to the marginal density of $\mathbf{p}(\mathsf{X})$. Then through (3.3.2) and the optimal Bayesian error (3.3.3) we can get another lower bound

$$\pi(\phi_n, \varepsilon_n^2) + \alpha(\phi_n) \geq \liminf_{n\to\infty} \left\{ 1 - \frac{1}{2}\mathrm{E}_{\mathsf{X}}\Big[\mathrm{E}_0\big\{|L_n - 1|\big|\mathsf{X}\big\}\Big] \right\} + o(1),$$

which is based on the likelihood ratio L_n. By Fatou's lemma, it remains to show $\mathrm{E}_0\big\{|L_n - 1|\big|\mathsf{X}\big\} \to 0$ in probability, or equivalently, $\mathrm{E}_0\big\{(L_n-1)^2\big|\mathsf{X}\big\} \xrightarrow{p} 0$, which can be further reduced to $\mathrm{E}_0\big(L_n^2\big|\mathsf{X}\big) \xrightarrow{p} 1$ since $\mathrm{E}_0\big(L_n\big|\mathsf{X}\big) = 1$. Hence to complete the proof, we need to verify

$$\mathrm{E}_0\big(L_n^2\big|\mathsf{X}\big) \xrightarrow{p} 1. \tag{3.3.4}$$

In the following we will focus on the likelihood ratio $L_n = \dfrac{\mathbf{p}_{1n}(\mathsf{Y}|\mathsf{X})}{\mathbf{p}_0(\mathsf{Y}|\mathsf{X})}$, where the denominator

$$\mathbf{p}_0(\mathsf{Y}|\mathsf{X}) = (2\pi)^{-n/2} \exp\left\{ -\frac{1}{2}\sum_{i=1}^n Y_i^2 \right\},$$

and the numerator is

$$\begin{aligned}
\mathbf{p}_{1n}(\mathsf{Y}|\mathsf{X}) &= (2\pi)^{-n/2} \int \left[\exp\left\{ -\frac{1}{2}\sum_{i=1}^n (Y_i - h(\mathbf{X}_i))^2 \right\} \right] d\Pi_{1n}(h) \\
&= (2\pi)^{-n/2} \int \left[\exp\left\{ -\frac{1}{2}\sum_{i=1}^n Y_i^2 - \frac{1}{2}\sum_{i=1}^n h^2(\mathbf{X}_i) + \sum_{i=1}^n Y_i h(\mathbf{X}_i) \right\} \right] d\Pi_{1n}(h) \\
&= \mathbf{p}_0(\mathsf{Y}|\mathsf{X}) \int \left[\exp\left\{ -\frac{1}{2}\sum_{i=1}^n h^2(\mathbf{X}_i) + \sum_{i=1}^n Y_i h(\mathbf{X}_i) \right\} \right] d\Pi_{1n}(h). \qquad (3.3.5)
\end{aligned}$$

Specifically,

$$\sum_{i=1}^{n} h^2(\mathbf{X}_i) = \sum_{i=1}^{n} (\boldsymbol{\alpha}_\nu^T \tilde{\mathbf{X}}_i)^2 = \sum_{i=1}^{n} \tilde{X}_i^T \boldsymbol{\alpha}_\nu \boldsymbol{\alpha}_\nu^T \tilde{\mathbf{X}}_i$$
$$\sum_{i=1}^{n} Y_i h(\mathbf{X}_i) = \sum_{i=1}^{n} Y_i \boldsymbol{\alpha}_\nu^T \tilde{\mathbf{X}}_i.$$

Hence (3.3.5) implies that

$$L_n = \frac{1}{p} \sum_{\nu} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} \tilde{X}_i^T \boldsymbol{\alpha}_\nu \boldsymbol{\alpha}_\nu^T \tilde{\mathbf{X}}_i + \sum_{i=1}^{n} Y_i \boldsymbol{\alpha}_\nu^T \tilde{\mathbf{X}}_i \right) \tag{3.3.6}$$

and

$$L_n^2 = \frac{1}{p^2} \sum_{\nu,\nu'} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} \tilde{X}_i^T \boldsymbol{\alpha}_\nu \boldsymbol{\alpha}_\nu^T \tilde{\mathbf{X}}_i - \frac{1}{2} \sum_{i=1}^{n} \tilde{X}_i^T \boldsymbol{\alpha}_{\nu'} \boldsymbol{\alpha}_{\nu'}^T \tilde{\mathbf{X}}_i + \sum_{i=1}^{n} Y_i (\boldsymbol{\alpha}_\nu + \boldsymbol{\alpha}_{\nu'})^T \tilde{\mathbf{X}}_i \right) \tag{3.3.7}$$

Under the null hypothesis, $Y_i = \epsilon_i$ follows a standard normal distribution. Then $\sum_{i=1}^{n} \epsilon_i (\boldsymbol{\alpha}_\nu + \boldsymbol{\alpha}_{\nu'})^T \tilde{\mathbf{X}}_i$ is the sum of n independent centered normal random variables with variance $\sum_{i=1}^{n} \tilde{\mathbf{X}}_i^T (\boldsymbol{\alpha}_\nu + \boldsymbol{\alpha}_{\nu'})(\boldsymbol{\alpha}_\nu + \boldsymbol{\alpha}_{\nu'})^T \tilde{\mathbf{X}}_i$, conditioned on X. It is known that for any centered normal random variable Z_0 with variance σ_0^2, $\mathrm{E}\{\exp(Z_0)\} = \exp(\sigma_0^2/2)$. Therefore, (3.3.7) yields

$$\mathrm{E}(L_n^2 | \mathsf{X}) \;=\; \frac{1}{p^2} \sum_{\nu,\nu'} \exp\left(\frac{1}{2} \sum_{i=1}^{n} \tilde{\mathbf{X}}_i^T \left\{ \boldsymbol{\alpha}_\nu \boldsymbol{\alpha}_{\nu'}^T + \boldsymbol{\alpha}_{\nu'} \boldsymbol{\alpha}_\nu^T \right\} \tilde{\mathbf{X}}_i \right).$$

Before proceeding further, let us denote $a_n = n\lambda_p^{-1}\varepsilon_n^2$ and $b_n = n\lambda_p^{-2}\varepsilon_n^4$. On one hand,

$$\frac{1}{p^2} \sum_{\nu \neq \nu'} \exp\left(\frac{1}{2} \sum_{i=1}^{n} \tilde{\mathbf{X}}_i^T \left\{ \boldsymbol{\alpha}_\nu \boldsymbol{\alpha}_{\nu'}^T + \boldsymbol{\alpha}_{\nu'} \boldsymbol{\alpha}_\nu^T \right\} \tilde{\mathbf{X}}_i \right) = \frac{1}{p^2} \sum_{k \neq k'} \exp\left(\sum_{i=1}^{n} \frac{\varepsilon_n^2}{\sqrt{\lambda_k \lambda_{k'}}} \tilde{X}_{ik} \tilde{X}_{ik'} \right) := M_n,$$

where for any fixed $k \neq k'$,

$$\mathrm{E}\Big\{ \frac{\varepsilon_n^2}{\sqrt{\lambda_k \lambda_{k'}}} \sum_{i=1}^{n} \tilde{X}_{ik} \tilde{X}_{ik'} \Big\} = 0, \quad \mathrm{Var}\Big\{ \frac{\varepsilon_n^2}{\sqrt{\lambda_k \lambda_{k'}}} \sum_{i=1}^{n} \tilde{X}_{ik} \tilde{X}_{ik'} \Big\} \leq b_n.$$

Since $b_n = o(1)$ if $\varepsilon_n^2 = o(\tilde{\varepsilon}_n^2)$, $M_n \xrightarrow{p} 1 + p^{-1}$. On the other hand,

$$\frac{1}{p^2} \sum_{\nu = \nu'} \exp\left(\frac{1}{2} \sum_{i=1}^{n} \tilde{\mathbf{X}}_i^T \Big\{ \boldsymbol{\alpha}_\nu \boldsymbol{\alpha}_{\nu'}^T + \boldsymbol{\alpha}_{\nu'} \boldsymbol{\alpha}_\nu^T \Big\} \tilde{\mathbf{X}}_i \right) = \frac{1}{p^2} \sum_{k=1}^{p} \exp\left(\varepsilon_n^2 \sum_{i=1}^{n} \frac{\tilde{X}_{ik}^2}{\lambda_k} \right) := R_n.$$

Actually,

$$0 \leq R_n \leq \frac{1}{p} \sum_{k=1}^{p} \exp\left(\varepsilon_n^2 \sum_{i=1}^{n} \frac{\tilde{X}_{ik}^2}{\lambda_p} - \log p \right), \tag{3.3.8}$$

and

$$\exp\left(\varepsilon_n^2 \sum_{i=1}^{n} \frac{\tilde{X}_{ik}^2}{\lambda_p} - \log p \right) \xrightarrow{d} W_k, \quad k = 1, ..., p,$$

where $\{W_k\}_{k=1}^{p}$ are independent log-normal random variables with mean

$$\mu_{LN} = \exp(a_n + \frac{1}{2} b_n - \log p)$$

and variance

$$\sigma_{LN}^2 = \Big\{ \exp(b_n) - 1 \Big\} \mu_{LN}^2,$$

where $\mu_{LN} = o(1)$ and $\sigma_{LN}^2 = o(1)$ when $\varepsilon_n^2 = o(\tilde{\varepsilon}_n^2) = o(\frac{\lambda_p}{\sqrt{n}})$. Because the upper bound in (3.3.8) asymptotically converges to a distribution with mean μ_{LN} and variance $p^{-1}\sigma_{LN}^2$,

we can obtain $R_n = o_p(1)$ through (3.3.8), and conclude

$$\mathrm{E}(L_n^2|\mathsf{X}) = M_n + R_n \xrightarrow{p} 1,$$

under a high-dimensional setup. This finishes the proof.

□

Remark 3.1 (a). In the context of testing functions of specific forms under a fixed dimensional set-up, the minimax rate depends on the smoothness of the function class (Guerre and Levergne, 2002; Ingster, 1993). According to Guerre and Lavergne (2002), if the smooth index $s \geq p/4$, then $\tilde{\varepsilon}_n^2 = n^{-4s/(p+4s)}$, and $\tilde{\varepsilon}_n^2 = 1/\sqrt{n}$ if $s < p/4$, when the norm was defined as $\|h\|_{GL}^2 = \mathrm{E}(h^2(\mathbf{X})) = \boldsymbol{\alpha}^T\boldsymbol{\alpha}$. If assuming all the eigenvalues are bounded in the sense $0 < m \leq \lambda_p \leq \lambda_1 \leq M \leq \infty$, then the norm considered in our setting is of the same order as $\|h\|_{GL}^2$. Besides, the class of linear functions essentially has smooth index $s = 1$ and gives rate $\tilde{\varepsilon}_n^2 = 1/\sqrt{n}$ when $p > 4$. Hence the lower bound in Theorem 4 is well connected to existing results.

(b). Although $\mathbf{X}_i$ is assumed to follow a multivariate normal distribution in our model, the lower bound result in the above theorem still holds for any distribution with $\mathrm{E}(\tilde{X}_{ik}^4) < \infty$, $i = 1, ..., n$, $k = 1, ..., p$.

3.4 Upper bound of separation rate

In this section, we derive the upper bound of the separation rate by showing the existence of a test whose minimax power is nontrivial against $H_1(\kappa\tilde{\varepsilon}_n^2)$ for some constant $\kappa > 0$. The lower and upper bounds together characterize the detection boundary, which can be then

used as a minimax benchmark to evaluate the performance of a test asymptotically.

3.4.1 Test statistic

Given a random sample $\{\mathbf{X}_1, Y_1\}, \{\mathbf{X}_2, Y_2\}, ..., \{\mathbf{X}_n, Y_n\}$, we consider a test statistic

$$Q_n = \frac{1}{n(n-1)} \sum_{i \neq j} \mathbf{X}_i^T \mathbf{X}_j (Y_i - \bar{Y}_n)(Y_j - \bar{Y}_n). \tag{3.4.1}$$

Denote $V_k = \mathrm{tr}(\mathbf{\Sigma}^k)$ for $k \in \mathbb{Z}^+$. In parallel to the results in Chapter 2, we can similarly give the asymptotic distribution of Q_n under the null hypothesis, which is given in Theorem 5. The asymptotic distribution of Q_n under a single alternative hypothesis

$$H_1(\boldsymbol{\alpha}) : Y_i = \mu + h_{\boldsymbol{\alpha}}(\mathbf{X}_i) + \epsilon_i \tag{3.4.2}$$

is discussed in Theorem 6, where $h_{\boldsymbol{\alpha}}(\mathbf{X}_i) = \sum_{m=1}^{p} \alpha_m \tilde{X}_{im}$ $(1 \leq i \leq n)$. It should be pointed out that those results are derived based on some mild moment assumptions on $\mathbf{X}$ or ϵ, as described below.

A1. $\mathrm{E}(\tilde{X}_{ik}^4) < \infty$, $i = 1, ..., n$, $k = 1, ..., p$.

A2. $\mathrm{E}(\epsilon_i^4) < \infty$, $i = 1, ..., n$

Theorem 5. *Assume $V_4/V_2^2 \to 0$ as $p(n) \to \infty$. Then under assumptions A1 and A2, we have*

$$\frac{nQ_n}{\sqrt{2V_2}} \xrightarrow{d} N(0, 1) \tag{3.4.3}$$

under the null hypothesis.

Therefore, an α level test rejects the null hypothesis if $(2V_2)^{-1/2} n Q_n > z_{1-\alpha}$, where $z_{1-\alpha}$

is the lower $1-\alpha$ quantile of the standard normal distribution. Denote $\psi_m(\mathbf{X}_i) = \tilde{X}_{im}$, and $\boldsymbol{\psi}_i = \Big(\psi_1(\mathbf{X}_i), \psi_2(\mathbf{X}_i), ..., \psi_p(\mathbf{X}_i)\Big)$, $1 \le m \le p$, $1 \le i \le n$. Then we can see that $\boldsymbol{\psi}_i$ enjoys the nice properties of $\mathrm{E}\boldsymbol{\psi}_i = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{\psi}_i) = \mathbf{I}$.

Theorem 6. *Assume $V_4/V_2^2 \to 0$ as $p(n) \to \infty$. Then under assumptions A1 and A2, we have*

$$\frac{n^{1/2}(Q_n - \theta_{\boldsymbol{\alpha}})}{\sigma_n(\boldsymbol{\alpha})} \xrightarrow{d} N(0,1), \tag{3.4.4}$$

under the alternative hypothesis $H_1(\boldsymbol{\alpha})$, where

$$\begin{aligned}
\theta_{\boldsymbol{\alpha}} &= \sum_{m=1}^{p} \lambda_m \alpha_m^2, \\
\sigma_n^2(\boldsymbol{\alpha}) &= 4 \sum_{m_1,m_2=1}^{p} \lambda_{m_1}\lambda_{m_2}\alpha_{m_1}\alpha_{m_2} b_{m_1,m_2}(\boldsymbol{\alpha}) + 4\sum_{m=1}^{p} \lambda_m^2 \alpha_m^2 - 4\theta_{\boldsymbol{\alpha}}^2 \\
&\quad + 2n^{-1}\Bigg\{ \sum_{m_1,m_2=1}^{p} \lambda_{m_1}\lambda_{m_2} b_{m_1,m_2}^2(\boldsymbol{\alpha}) + 2\sum_{m=1}^{p} \lambda_m^2 b_{m,m}^2(\boldsymbol{\alpha}) + V_2 - \theta_{\boldsymbol{\alpha}}^2 \Bigg\},
\end{aligned}$$

and $b_{m_1,m_2}(\boldsymbol{\alpha}) = E\{\psi_{m_1}(\mathbf{X})\psi_{m_2}(\mathbf{X})h_{\boldsymbol{\alpha}}^2(\mathbf{X})\}$ for $m_1, m_2 = 1, ..., p$.

Proof. It is not difficult to see $Q_n = Q_n^0\{1 + o_p(1)\}$, where the leading order term is

$$Q_n^0 = \frac{1}{n(n-1)} \sum_{i \ne j} \mathbf{X}_i^T \mathbf{X}_j (h_{\boldsymbol{\alpha}}(\mathbf{X}_i) + \epsilon_i)(h_{\boldsymbol{\alpha}}(\mathbf{X}_j) + \epsilon_j). \tag{3.4.5}$$

To use the classical results of U-statistic (Lee, 1990), let us denote $\mathbf{Z} = (\mathbf{X}, \epsilon)$ and define the first-order and second-order projections of the symmetric function $\varphi(\mathbf{Z}_1, \mathbf{Z}_2) = \mathbf{X}_1^T \mathbf{X}_2 (h_{\boldsymbol{\alpha}}(\mathbf{X}_1) + \epsilon_1)(h_{\boldsymbol{\alpha}}(\mathbf{X}_2) + \epsilon_2)$ as

$$\varphi^{(1)}(\mathbf{z}_1) = \mathrm{E}\{\varphi(\mathbf{z}_1, \mathbf{Z}_2)\}, \quad \varphi^{(2)}(\mathbf{z}_1, \mathbf{z}_2) = \varphi(\mathbf{z}_1, \mathbf{z}_2)$$

and their variances

$$\sigma^2_{(1)} = \mathrm{Var}\{\varphi(\mathbf{z}_1, \mathbf{Z}_2)\}, \quad \sigma^2_{(2)} = \mathrm{Var}\{\varphi^{(2)}(\mathbf{z}_1, \mathbf{z}_2)\}.$$

Plugging the expansions

$$h_{\boldsymbol{\alpha}}(\mathbf{X}_i) = \sum_{m=1}^{p} \alpha_m \psi_m(\mathbf{X}_i)$$

and

$$\mathbf{X}_1^T \mathbf{X}_2 = \sum_{m=1}^{p} \lambda_m \psi_m(\mathbf{X}_1)\psi_m(\mathbf{X}_2)$$

into the function $\varphi(\mathbf{Z}_1, \mathbf{Z}_2)$, we can see that the first-order projection and its variance are

$$\begin{aligned} \varphi^{(1)}(\mathbf{z}_1) &= \sum_{m=1}^{p} \lambda_m \alpha_m \psi_m(\mathbf{x}_1)(h(\mathbf{x}_1) + \epsilon_1), \\ \sigma^2_{(1)} &= \sum_{m,m'=1}^{p} \lambda_m \lambda_{m'} \alpha_m \alpha_{m'} b_{m,m'}(\boldsymbol{\alpha}) - \theta^2_{\boldsymbol{\alpha}} + \sum_{m=1}^{p} \lambda_m^2 \alpha_m^2. \end{aligned}$$

The variance of second-order projection is

$$\begin{aligned} \sigma^2_{(2)} &= \mathrm{E}\Bigg\{ \sum_{m,m'=1}^{p} \lambda_m \lambda_{m'} \psi_m(\mathbf{X}_1)\psi_m(\mathbf{X}_2)\psi_{m'}(\mathbf{X}_1)\psi_{m'}(\mathbf{X}_2) \\ &\qquad \Big(h^2_{\boldsymbol{\alpha}}(\mathbf{X}_1)h^2_{\boldsymbol{\alpha}}(\mathbf{X}_2) + h^2_{\boldsymbol{\alpha}}(\mathbf{X}_1)\epsilon_2^2 + h^2_{\boldsymbol{\alpha}}(\mathbf{X}_2)\epsilon_1^2 + \epsilon_1^2\epsilon_2^2 \Big) \Bigg\} - \theta^2_{\boldsymbol{\alpha}} \\ &= \sum_{m,m'=1}^{p} \lambda_m \lambda_{m'} b^2_{m,m'}(\boldsymbol{\alpha}) + 2\sum_{m=1}^{p} \lambda_m^2 b_{m,m}(\boldsymbol{\alpha}) + V_2 - \theta^2_{\boldsymbol{\alpha}}. \end{aligned}$$

According to the classical results on U-statistic (Lee, 1990), $\mathrm{Var}(Q_n^0) = \{4n^{-1}\sigma^2_{(1)} + 2n^{-2}\sigma^2_{(2)}\}\{1 + o(1)\}$, and

$$\frac{n^{1/2}(Q_n^0 - \theta_{\boldsymbol{\alpha}})}{n\mathrm{Var}(Q_n^0)} \xrightarrow{d} N(0, 1).$$

Letting $\sigma_n^2(\boldsymbol{\alpha}) = 4\sigma_{(1)}^2 + 2n^{-1}\sigma_{(2)}^2$, we can obtain the claim. □

With the normal assumption on $\mathbf{X}_i$ and ϵ_i in model (3.1.2), assumptions A1 and A2 automatically hold. Moreover, we can obtain a further characterizations for the asymptotic variance $\sigma_n^2(\boldsymbol{\alpha})$.

Corollary 1. *Assume $\mathbf{X}_i$ follows a Gaussian distribution for $i = 1, 2, ..., n$, then*

$$\sigma_n^2(\boldsymbol{\alpha}) = (4 + 6n^{-1})(\boldsymbol{\alpha}^T\boldsymbol{\Lambda}\boldsymbol{\alpha})^2 + (4 + 8n^{-1})(\boldsymbol{\alpha}^T\boldsymbol{\alpha} + 1)(\boldsymbol{\alpha}^T\boldsymbol{\Lambda}^2\boldsymbol{\alpha}) + 2n^{-1}(\boldsymbol{\alpha}^T\boldsymbol{\alpha} + 1)^2 V_2.$$

Proof. Based on normality assumption, $\boldsymbol{\psi}_i \sim N(\mathbf{0}, \mathbf{I})$, $i = 1, ..., n$. Rewrite $2b_{m_1m_2}$ as a product of two quadratic forms

$$2b_{m_1,m_2}(\boldsymbol{\alpha}) = \mathrm{E}[\boldsymbol{\psi}_1^T(\mathbf{e}_{m_1}\mathbf{e}_{m_2}^T + \mathbf{e}_{m_2}\mathbf{e}_{m_1}^T)\boldsymbol{\psi}_1\boldsymbol{\psi}_1^T\boldsymbol{\alpha}\boldsymbol{\alpha}^T\boldsymbol{\psi}_1],$$

where $\mathbf{e}_k$ represents the p-dim vector with the k-th element 1 and others 0. By using classical results on expectation of quadratic forms in Kumar (1973), we have

$$2b_{m_1,m_2}(\boldsymbol{\alpha}) = (\mathbf{e}_{m_1}^T\mathbf{e}_{m_2} + \mathbf{e}_{m_2}^T\mathbf{e}_{m_1})\boldsymbol{\alpha}^T\boldsymbol{\alpha} + 2\boldsymbol{\alpha}^T(\mathbf{e}_{m_1}\mathbf{e}_{m_2}^T + \mathbf{e}_{m_2}\mathbf{e}_{m_1}^T)\boldsymbol{\alpha},$$

which yields $b_{m_1m_2}(\boldsymbol{\alpha}) = 2\alpha_{m_1}\alpha_{m_2}$ for $m_1 \neq m_2$ and $b_{mm}(\boldsymbol{\alpha}) = 2\alpha_m^2 + \boldsymbol{\alpha}^T\boldsymbol{\alpha}$. Thus, we

obtain

$$\begin{aligned}
\sigma_{(1)}^2 &= \sum_{m_1,m_2=1}^{p} \lambda_{m_1}\lambda_{m_2}\alpha_{m_1}\alpha_{m_2}b_{m_1,m_2}(\boldsymbol{\alpha}) + \sum_{m=1}^{p}\lambda_m^2\alpha_m^2 - \theta_{\boldsymbol{\alpha}}^2 \\
&= 2\sum_{m_1\neq m_2}\lambda_{m_1}\lambda_{m_2}\alpha_{m_1}^2\alpha_{m_2}^2 + \sum_m \lambda_m^2\alpha_m^2(\boldsymbol{\alpha}^T\boldsymbol{\alpha} + 2\alpha_m^2 + 1) - \big(\sum_m \lambda_m\alpha_m^2\big)^2 \\
&= \big(\sum_m \lambda_m\alpha_m^2\big)^2 + (\boldsymbol{\alpha}^T\boldsymbol{\alpha}+1)\sum_m \lambda_m^2\alpha_m^2 \\
&= (\boldsymbol{\alpha}^T\boldsymbol{\Lambda}\boldsymbol{\alpha})^2 + (\boldsymbol{\alpha}^T\boldsymbol{\alpha}+1)(\boldsymbol{\alpha}^T\boldsymbol{\Lambda}^2\boldsymbol{\alpha}),
\end{aligned}$$

and

$$\begin{aligned}
\sigma_{(2)}^2 &= \sum_{m_1,m_2=1}^{p}\lambda_{m_1}\lambda_{m_2}b_{m_1,m_2}^2(\boldsymbol{\alpha}) + 2\sum_{m=1}^{p}\lambda_m^2 b_{m,m}(\boldsymbol{\alpha}) + V_2 - \theta_{\boldsymbol{\alpha}}^2 \\
&= 4\sum_{m_1\neq m_2}\lambda_{m_1}\lambda_{m_2}\alpha_{m_1}^2\alpha_{m_2}^2 + \sum_m \lambda_m^2(2\alpha_m^2 + \boldsymbol{\alpha}^T\boldsymbol{\alpha})^2 + 2\sum_m\lambda_m^2(2\alpha_m^2+\boldsymbol{\alpha}^T\boldsymbol{\alpha}) \\
&\quad + V_2 - \big(\sum_m \lambda_m\alpha_m^2\big)^2 \\
&= 3\big(\sum_m \lambda_m\alpha_m^2\big)^2 + (\boldsymbol{\alpha}^T\boldsymbol{\alpha}+1)^2V_2 + (4\boldsymbol{\alpha}^T\boldsymbol{\alpha}+4)\sum_m\lambda_m^2\alpha_m^2 \\
&= 3(\boldsymbol{\alpha}^T\boldsymbol{\Lambda}\boldsymbol{\alpha})^2 + (\boldsymbol{\alpha}^T\boldsymbol{\alpha}+1)^2V_2 + 4(\boldsymbol{\alpha}^T\boldsymbol{\alpha}+1)(\boldsymbol{\alpha}^T\boldsymbol{\Lambda}^2\boldsymbol{\alpha}).
\end{aligned}$$

Noting that $\sigma_n^2(\boldsymbol{\alpha}) = 4\sigma_{(1)}^2 + 2n^{-1}\sigma_{(2)}^2$, we complete the proof of the claim. □

3.4.2 Rate-optimality of the test

Equipped with the asymptotic behaviors stated in Theorems 5 and 6, we now study the minimax power of the proposed test (denoted as ϕ_n^*) against the composite alternative hypothesis $H_1(\varepsilon_n^2)$. In particular, we confirmed the existence of the constant $\kappa > 0$ such that the test has nontrivial power against $H_1(\kappa\tilde{\varepsilon}_n^2)$, when the decreasing speed of eigenvalues is not too fast. Specifically, we put the following two assumptions on the decay speed of eigenvalues.

A3. $\lambda_1/\lambda_p = o(\sqrt{n})$.

A4. There exists a constant $b^* > 0$ such that $(n\omega_p)^{-1} \leq b^*$, where $\omega_p = \lambda_p^2/V_2$.

Theorem 7. *Under assumptions A3 and A4, ϕ_n^* is rate-optimal, i.e., for any given significance level $\alpha \in (0,1)$, there exists a constant $\kappa > 0$ such that*

$$1 - \pi(\phi_n^*, \kappa\tilde{\varepsilon}_n^2) = 1 - \sup_{h \in H_1(\kappa\tilde{\varepsilon}_n^2)} P_h(\phi_n^* = 0) > \alpha + o(1).$$

Proof. According to the definition (3.2.1), for any $h \in H_1(\kappa\tilde{\varepsilon}_n^2)$, there exists $\tau \geq \kappa$ such that $\|\alpha\|_P^2 = \boldsymbol{\alpha}^T\boldsymbol{\Lambda}\boldsymbol{\alpha} = \tau\tilde{\varepsilon}_n^2$. Suppose κ is large enough such that

$$\left(\frac{n\omega_p}{2}\right)^{1/2}\tau \geq \left(\frac{n\omega_p}{2}\right)^{1/2}\kappa \geq \frac{3}{2}z_{1-\alpha}. \tag{3.4.6}$$

Consider the type II error of test ϕ_n^* against single alternative $H_1(\boldsymbol{\alpha})$ where $\|\boldsymbol{\alpha}\|_{\boldsymbol{\Lambda}}^2 = \tau\tilde{\varepsilon}_n^2$ (i.e., $\boldsymbol{\alpha}^T\boldsymbol{\Lambda}\boldsymbol{\alpha} = \tau\tilde{\varepsilon}_n^2$)

$$\begin{aligned}
P_h(\phi_n^* = 0) &= P_h\left(\frac{nQ_n}{\sqrt{2V_2}} \leq z_{1-\alpha}\right) = P_h\left(n^{1/2}Q_n \leq n^{-1/2}\sqrt{2V_2}z_{1-\alpha}\right) \\
&= P_h\left(n^{1/2}(Q_n - \boldsymbol{\alpha}^T\boldsymbol{\Lambda}\boldsymbol{\alpha}) \leq n^{-1/2}\sqrt{2V_2}z_{1-\alpha} - n^{1/2}\boldsymbol{\alpha}^T\boldsymbol{\Lambda}\boldsymbol{\alpha}\right) \\
&= P_h\left(|n^{1/2}(Q_n - \boldsymbol{\alpha}^T\boldsymbol{\Lambda}\boldsymbol{\alpha})| \leq |\, n^{-1/2}\sqrt{2V_2}z_{1-\alpha} - n^{1/2}\boldsymbol{\alpha}^T\boldsymbol{\Lambda}\boldsymbol{\alpha}\,|\right).
\end{aligned}$$

By using Chebyshev's inequality,

$$
\begin{aligned}
&P_h\left(|n^{1/2}(Q_n-\boldsymbol{\alpha}^T\boldsymbol{\Lambda}\boldsymbol{\alpha})| \leq| \, n^{-1/2}\sqrt{2V_2}z_{1-\alpha}-n^{1/2}\boldsymbol{\alpha}^T\boldsymbol{\Lambda}\boldsymbol{\alpha} \,|\right)\\
&\leq \frac{\sigma_n^2(\boldsymbol{\alpha})}{\left[\sqrt{(2V_2/n)}z_{1-\alpha}-\sqrt{n}\boldsymbol{\alpha}^T\boldsymbol{\Lambda}\boldsymbol{\alpha}\right]^2}\{1+o(1)\}\\
&= \frac{n\sigma_n^2(\boldsymbol{\alpha})/(2V_2)}{\left[z_{1-\alpha}-n(2V_2)^{-1/2}\boldsymbol{\alpha}^T\boldsymbol{\Lambda}\boldsymbol{\alpha}\right]^2}(1+o(1)) \qquad (3.4.7)
\end{aligned}
$$

where the numerator

$$
\frac{n\sigma_n^2(\boldsymbol{\alpha})}{2V_2} = \left[\frac{4n(\boldsymbol{\alpha}^T\boldsymbol{\Lambda}\boldsymbol{\alpha})^2+4n(\boldsymbol{\alpha}^T\boldsymbol{\alpha}+1)\boldsymbol{\alpha}^T\boldsymbol{\Lambda}^2\boldsymbol{\alpha}}{2V_2}+(\boldsymbol{\alpha}^T\boldsymbol{\alpha}+1)^2\right]\{1+o(1)\}. \qquad (3.4.8)
$$

Since $\boldsymbol{\alpha}^T\boldsymbol{\Lambda}\boldsymbol{\alpha}=\tau\tilde{\varepsilon}_n^2=\frac{\tau\lambda_p}{\sqrt{n}}$, it is not difficult to see

$$
\boldsymbol{\alpha}^T\boldsymbol{\alpha}\leq \ \frac{\tau\tilde{\varepsilon}_n^2}{\lambda_p}=\frac{\tau}{\sqrt{n}},\ \boldsymbol{\alpha}^T\boldsymbol{\Lambda}^2\boldsymbol{\alpha}\leq \ \tau\lambda_1\tilde{\varepsilon}_n^2=\frac{\tau\lambda_1\lambda_p}{\sqrt{n}}. \qquad (3.4.9)
$$

Plugging (3.4.9) into (3.4.8), we can see

$$
\frac{n\sigma_n^2(\boldsymbol{\alpha})}{2V_2}\leq\left[\frac{4\tau^2\lambda_p^2+4\tau\lambda_1\lambda_p(\tau+\sqrt{n})}{2V_2}+(\frac{\tau}{\sqrt{n}}+1)^2\right]\{1+o(1)\}. \qquad (3.4.10)
$$

On the other hand, (3.4.6) implies

$$
z_{1-\alpha}\leq\frac{2}{3}\frac{n\boldsymbol{\alpha}^T\boldsymbol{\Lambda}\boldsymbol{\alpha}}{\sqrt{2V_2}},
$$

hence the denominator

$$\left[\frac{n\boldsymbol{\alpha}^T\boldsymbol{\Lambda}\boldsymbol{\alpha}}{\sqrt{2V_2}} - z_{1-\alpha}\right]^2 \geq \left[\frac{1}{3}\frac{n\boldsymbol{\alpha}^T\boldsymbol{\Lambda}\boldsymbol{\alpha}}{\sqrt{2V_2}}\right]^2 = \frac{n\tau^2\lambda_p^2}{18V_2}. \tag{3.4.11}$$

Combine (3.4.8) and (3.4.11), the upper bound in (3.4.7)

$$\begin{aligned}
&\frac{n\sigma_n^2(\boldsymbol{\alpha})/(2V_2)}{\left[z_{1-\alpha} - n(2V_2)^{-1/2}\boldsymbol{\alpha}^T\boldsymbol{\Lambda}\boldsymbol{\alpha}\right]^2} \\
\leq\ & 9\left[\frac{4\tau^2\lambda_p^2 + 4\tau\lambda_1\lambda_p(\tau+\sqrt{n})}{n\tau^2\lambda_p^2} + \frac{2V_2(\tau/\sqrt{n}+1)^2}{n\tau^2\lambda_p^2}\right] \\
=\ & 9\left[\frac{4}{n} + 4\left(\frac{1}{n} + \frac{1}{\tau\sqrt{n}}\right)\frac{\lambda_1}{\lambda_p} + \left(\frac{1}{\sqrt{n}} + \frac{1}{\tau}\right)^2\frac{2}{n\omega_p}\right] \\
=\ & 9\left[\frac{4}{n} + 4\left(\frac{1}{n} + \frac{1}{\kappa\sqrt{n}}\right)\frac{\lambda_1}{\lambda_p} + \left(\frac{1}{\sqrt{n}} + \frac{1}{\kappa}\right)^2(2b^*)\right] \\
=\ & 9\left[o(1) + \left(\frac{1}{\sqrt{n}} + \frac{1}{\kappa}\right)^2(2b^*)\right]
\end{aligned} \tag{3.4.12}$$

under assumptions A3 and A4. Noting that there exists a constant $\kappa = \kappa(\alpha, b^*) > 0$ such that the bound in (3.4.12) is uniformly controlled by $1-\alpha$, we can reach the conclusion that the test ϕ_n^* is rate-optimal in the sense that

$$1 - \pi(\phi_n^*, \kappa\tilde{\varepsilon}_n^2) = 1 - \sup_{h \in H_1(\kappa\tilde{\varepsilon}_n^2)} P_h(\phi_n^* = 0) > \alpha + o(1).$$

This finishes the proof. □

Remark 3 (a). Assumption A4 sometimes implicitly puts restriction on the dimension order. For example, in the special case where all the eigenvalues are bounded constant, then Assumptions A4 indicates that at most $p/n \to r \in (0, \infty)$.

(b). Assumptions A3 and A4 might not be necessary for the rate-optimality of test ϕ_n^*. The upper bounds given in (3.4.9) are probably too loose to get assumptions weaker than A3 and A4. However, it is not easy to bound the two quantities in (3.4.9) sharply without specific assumptions on the decay speed of eigen-values.

Recall in (Ingster et al. 2010), the authors consider the high-dimensional sparse linear model with independent and standardized variables (i.e., $\tilde{\mathbf{X}}_i = \mathbf{X}_i$ and $\lambda_1 = ... = \lambda_p = 1$). They derive the detection boundary $\varepsilon_n^* = \frac{p^{1/4}}{\sqrt{n}} \bigwedge \frac{1}{n^{1/4}}$ in moderately sparse case where the proportion of non-zero coefficient is $p^{-\beta}$ $(\frac{1}{2} < \beta < 1)$. Comparing our detection boundary $\tilde{\varepsilon}_n = \frac{1}{n^{1/4}}$ (if $\lambda_p = 1$) to ε_n^*, our rate does not contain the minimum with the $\frac{p^{1/4}}{\sqrt{n}}$ term. As is discussed in Cai and Ma (2013), the optimal test or detection boundary against a structured alternative is quite difficult from the one without structural assumption. Intuitively, with more information about the signals, the detection boundary is potentially tighter than the one against the structure-free alternative, which is also observed here. However, we can see that the difference disappears as the order of dimension p gets larger. This shows an adverse effect of dimensionality on the high-dimensional detection boundary.

3.5 Summary and discussion

In this chapter, we consider the detection boundary problem in testing a general linear model under the high-dimensional setting, where the p-dim variables are correlated and the dimension p can go to infinity as n goes to infinity. The problem is studied from a minimax point of view. We firstly establish the boundary that separates the detectable region and non-detectable region. Then a test is introduced and shown to be rate-optimal under certain conditions on the eigenvalue decay speed.

One of the most important directions of the future research is the study of detection boundary for a general model where $h(x) = \sum_{m=1}^{S} \alpha_m \psi_m(x)$, $\{\psi_m(\cdot)\}_{m=1}^{S}$ correspond to the eigen-decomposition of kernel function K and form a complete orthogonal normal system. We want to extend the optimal results under linear model (associated with linear kernel) to the general case. In addition, since the variance of noise σ^2 is normally unknown in many applications, finding a consistent estimator for σ^2 under both the null and alternative hypotheses is of great importance. Finally, it is still unclear if assumption A3 and A4 are violated, whether the current lower bound is still sharp enough and what is the corresponding optimal test. Answering these two questions is another interesting project for future investigation.

Chapter 4

Testing high-dimensional nonparametric functions in RKHS using multiple kernels

4.1 Introduction

Driven by advancements in microarray and next generation sequencing technologies, increasing number of genetic variants including small variations in single nucleotide polymorphisms (SNPs) and large variations, such as indel and copy number variation, are generated in a daily basis. The traditional genome-wide association studies (GWAS), aiming at detecting the SNPs that are associated with complex traits and accessing the effect of each SNP one at a time, has been proven to be a powerful tool to unveil the genetic architecture of a variety of complex traits. Although the traditional single-variant-based GWAS have successfully detected many genetic variants that are associated with the traits of interest, their power is still limited because of the weakness of individual signals and the lack of consideration of potential interactions among genetic variants.

The limitation of single variants analysis was overcome by the recent wave of set-based association studies. Such extension to set-based analysis is a natural choice because genetic

variants or genes in a set (i.e., a pathway) tend to work coordinately to fulfill their task. On one hand, the subtle effects in multiple variants can be combined so that the joint signal of the set could be potentially boosted. On the other hand, the set-based strategy improves the power to capture the complicated interactions among variants. There are a variety of public resources available to create the SNP-set or gene-set, such as the annotated gene models (for SNP-set), Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa, 2000), Reactome (Croft et al., 2010) and Gene Ontology (Ashburner et al., 2000).

Kernel-based testing (KBT) framework, which measures the similarity between genetic variants through a kernel function and comparing it to the phenotype similarity, is one of the most popular and powerful methods in set-based association studies (Liu et al. 2007; Liu et al. 2008). Moreover, KBT is a very general framework so that many other similarity based approaches (e.g., Reiss et al. 2010; Wessel and Schork 2006; Mukhopadhyay et al. 2010; Tzeng et al. 2009) are closely related to it. In Chapter 2, we have built a KBT framework with a single kernel function for quantitative traits, under the high dimensional setting where the total number of variants could be extremely large. As is observed in our simulations in Chapter 2 and other literature (Wessel and Schork 2006; Wu et al. 2010; Lin et al. 2011), the power of kernel-based test generally depends on the choice of kernel. Specifically, if the true function comes from the function space generated by the kernel, then utilizing the corresponding kernel will achieve high power. However, the underlying genetic architecture (the true function) is typically unknown. Given a few candidate kernels, one simple way is to choose the one with the smallest p-value. This, however, will inflate the type I error rate due to kernel selection. Based on the kernel machine testing proposed by Liu et al. (2007), Wu et al. (2010) proposed a perturbation method under multiple candidate kernels. However, this strategy is over-conservative in high-dimensional case and needs computational intensive

procedures to evaluate statistical significance.

Our interest in this chapter is to find an efficient multiple-kernel testing procedure that can maintain nominal type I error rate while achieving high power in a high-dimensional setting (i.e., $p > n$), under the KBT framework we developed in Chapter 2. Here we mainly focus on a high-dimensional setting and assume a set of candidate kernels are given. In the subsequent sections, we firstly extend our model in Chapter 2 by integrating covariates into the test statistic, and access the asymptotic distribution of the adjusted test statistic. We then propose two effective and efficient testing procedures when multiple kernel candidates are available. In the first procedure, we propose a test using the average of the standardized kernels in the candidate set, which is referred to as the simple average kernel method. In the second procedure, we introduce a new test statistic taking the maximum of the test statistics using the standardized kernels across the candidate set. We demonstrate the performance of the two strategies through a real data application and extensive simulation studies under both continuous and discrete variable settings. We show that under a high-dimensional setting, the proposed approaches not only calibrate the nominal type I error rate, but also enable the power to be close to the one using the best candidate function in the set, while the perturbation method proposed by Wu et al. (2010) suffers power loss. To make the work self-contained, some of the notations given in the earlier chapters are defined in this chapter again.

4.2 Statistical model

We assume that n unrelated subjects from a population were observed in a study design, where a quantitative (continuous) trait is of interest. For the ith subject, let Y_i be the

quantitative measurement, $\mathbf{W}_i = (W_{i1}, W_{i2}, ..., W_{iL})^T$ be the L-dim covariates, where L is finite and $i = 1, 2, ..., n$. In this chapter, we focus our attention to a p-dim SNP-set or gene-set, where p is assumed to be large and even allowed to be larger than sample size n. Let $\mathbf{X}_i = (X_{i1}, ..., X_{ip})^T \in \mathcal{X}$ be the vector of measurements of the set for subject i, which could be the genotype values for the SNPs in the SNP-set, or the gene expression profile for the genes in the gene-set (or pathway). For genotype values of SNPs, X_{ij} is typically coded as the number of minor alleles that subject i possesses at the jth specific position, hence a discrete variables takes three possible values 0, 1, or 2. The gene expression levels of gene-set are generally continuous measurements. Throughout the chapter, we do not assume specific assumption on the distribution of X_i.

To model the relationship between the quantitative trait and the SNP-set (or gene-set), we consider the following semi-parametric regression model,

$$Y_i = \mu + \boldsymbol{\alpha}^T \mathbf{W}_i + h(\mathbf{X}_i) + \epsilon_i, \quad i = 1, 2, ..., n, \tag{4.2.1}$$

where h is an unknown function, ϵ_i is a random subject-specific error term following a certain distribution (not necessarily normally distributed) with $\mathrm{E}(\epsilon_i) = 0$, $\mathrm{Var}(\epsilon_i) = \sigma^2$ and independent of $(\mathbf{X}_i, \mathbf{Z}_i)$. The identifiability of the h function is assured by side condition $\mathrm{E}[h(\mathbf{X}_i)] = 0$. We want to test the existence of association between the SNP-set (or gene-set) and the continuous trait of our interest, i.e.,

$$\mathrm{H}_0 : h(\cdot) = 0 \quad \text{vs} \quad \mathrm{H}_1 : h(\cdot) \neq 0. \tag{4.2.2}$$

4.2.1 Kernel function

Before proceeding to the test statistic, we want to introduce kernel function, which is widely used to measure the similarity between two subjects, as well as the functional space that is generated by the kernel function. A function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a kernel function if it is symmetric and positive semi-definite (i.e., $K(x_1, x_2) = K(x_2, x_1)$, and the kernel matrix $\mathbf{K}$ defined by $K_{ij} = K(x_i, x_j)$ is positive semi-definite, for any $x_1, x_2, ..., x_n \in \mathcal{X}$). In our context, $K(\mathbf{X}_i, \mathbf{X}_j)$ is a measure of similarity between the ith and the jth subject based on the genotypes of the SNPs in the SNP-set or the expression levels of the genes in the gene-set.

For any positive definite kernel K^* with corresponding matrix $\mathbf{K}^*$, we can defined its centralized kernel

$$K(x_1, x_2) = K^*(x_1, x_2) - K_1^*(x_1) - K_1^*(x_2) + \mu_{K^*} \tag{4.2.3}$$

satisfying $\mathrm{E}\{K(X_1, X_2)\} = 0$, where $K_1^*(x_1) = \mathrm{E}K^*\{(x_1, X_2)\}$, and $\mu_{K^*} = \mathrm{E}\{K^*(X_1, X_2)\}$. Empirically, centralized kernel matrix $\mathbf{K}_c$ can be replaced by its estimator

$$\mathbf{K}_{c,n} = \mathbf{K}^* - (n-1)^{-1}[\mathbf{J}(\mathbf{K}^*)^0 + (\mathbf{K}^*)^0\mathbf{J}] + n^{-1}(n-1)^{-1}\mathbf{J}(\mathbf{K}^*)^0\mathbf{J},$$

where $\mathbf{J}$ is an $n \times n$ matrix with all the elements as 1, and $\mathbf{D}^0 = \mathbf{D} - diag(\mathbf{D})$ is a zero-diagonal matrix sharing all non-diagonal elements with $\mathbf{D}$. For notation simplicity, hereafter we use K^*, K and $\mathbf{K}_n$ to represent the original kernel function, centralized kernel function and the empirical version of centralized kernel matrix.

Some commonly used kernel functions include linear kernel $K^*(x_1, x_2) = x_1^T x_2$, polyno-

mial kernel $K^*(x_1, x_2) = (x_1^T x_2 + c)^d$, Gaussian kernel $K^*(x_1, x_2) = \exp(-\|x_1 - x_2\|^2/\rho)$, and IBS kernel (for discrete genotype data only) $K^*(x_1, x_2) = (2p)^{-1} \sum_{j=1}^n IBS(x_{1j}, x_{2j}) = (2p)^{-1} \sum_{j=1}^n (2 - |x_{1j} - x_{2j}|)$, where $c, \rho > 0, d \in \mathbb{N}$ are tuning parameters. For a review of genomic similarity and more kernel functions, please refer to Schaid (2010a, 2010b).

In this chapter, we will utilize centralized kernel in the testing, as the asymptotic distribution shape of the test statistic using non-centralized kernel is fully determined by the centralized kernel. More benefits of using centralized kernel can be found in Lindsay et al. (2008, 2014). Furthermore, we can define the standardized kernel

$$\mathcal{K}(x_1, x_2) = K(x_1, x_2)/\mathrm{E}\{K(X, X)\}$$

from which it is easy to verify that $\mathrm{E}\{\mathcal{K}(\mathbf{X}, \mathbf{X})\} = 1$. Next let us briefly look at the eigen-decomposition of kernel function, which is an important way to characterize the kernel function. Assume $K(\cdot, \cdot)$ is a kernel function defined on $\mathcal{X} \times \mathcal{X}$, and μ is a probability measure on $\mathcal{X}$. Then the spectral decomposition theorems (Lemma 1 of Chapter 2, Steinwart and Scovel, 2012) implies that the standardized kernel $\mathcal{K}(\cdot, \cdot)$ enjoys the following representation

$$\mathcal{K}(x_1, x_2) = \sum_{m=1}^{S} \lambda_{\mathcal{K},m} \psi_m(x_1) \psi_m(x_2), \quad \forall x_1, x_2 \in \mathcal{X},$$

where the eigenfunctions $\{\psi_m(\cdot)\}$ form a complete orthonormal system (i.e., $\mathrm{E}\{\psi_m^2(X)\} = 1$ for any m, $\mathrm{E}\{\psi_m(X)\psi_{m'}(X)\} = 0$ for $m \neq m'$), and $\lambda_{\mathcal{K},1} \geq \lambda_{\mathcal{K},2} \geq ... \geq \lambda_{\mathcal{K},S} > 0$ are the non-zero eigenvalues satisfying $\sum_{m=1}^S \lambda_{\mathcal{K},m} = 1$. The standardization is required because $\mathrm{E}\{K(\mathbf{X}, \mathbf{X})\}$ could diverge in the high-dimensional case, and it ensures $\mathrm{E}\{\mathcal{K}(\mathbf{X}, \mathbf{X})\} < \infty$ so that the eigen-decomposition can be properly defined. By denoting $\lambda_m = \mathrm{E}(K(\mathbf{X}, \mathbf{X}))\lambda_{\mathcal{K},m}$,

we can get the pseudo eigen-decomposition of kernel function $K(\cdot, \cdot)$

$$K(x_1, x_2) = \sum_{m=1}^{S} \lambda_m \psi_m(x_1)\psi_m(x_2), \quad \forall x_1, x_2 \in \mathcal{X},$$

It should be noticed that the eigen-decomposition not only depends on the expression of the kernel, but also implicitly depends on the space $\mathcal{X}$ (e.g.,dimension p).

A functional space $\mathcal{H}_K$, named reproducing kernel Hilbert space (RKHS), can be generated for any kernel function $K(\cdot, \cdot)$, and the form of the functions that reside in $\mathcal{H}_K$ is characterized by the choice of the kernel K. Here we assume that the h function in model (4.2.1) is a member of the RKHS $\mathcal{H}_K$. Therefore, by specifying the kernel function, we are assuming some relationship between the trait and the SNP-set (or gene-set). For example, linear kernel indicates that the overall genetic effect is a linear combination of the individual effects in the set, i.e., $h(\mathbf{X}_i) = \beta^T \mathbf{X}_i$; polynomial kernel with $(c, d) = (1, 2)$ implies a quadratic model $h(\mathbf{X}_i) = \beta^T \mathbf{X}_i + \mathbf{X}_i^T \Lambda \mathbf{X}_i$, where simple product interactions and quadratic effects are modeled in addition to the linear effects. Now we can see the exciting and changeling aspect: many choices of kernels empower the model flexibility, while the truth in nature is largely unknown. It can be expected that the power is limited where a kernel is incorrectly assumed, i.e., the model is mis-specified. In the following sections we will start with the hypothesis testing using single kernel function, followed by the one using multiple kernel functions, through which the power can be greatly boosted over the choices of kernel functions in the candidate set.

4.2.2 Hypothesis test based on a single kernel

Consider the following kernel-based U-statistic (KU)

$$T_n = \frac{1}{n(n-1)} \sum_{i \neq j} K(\mathbf{X}_i, \mathbf{X}_j)(Y_i - \hat{Y}_i)(Y_j - \hat{Y}_j)/\hat{\sigma}^2, \tag{4.2.4}$$

where $\hat{Y}_i$ and $\hat{\sigma}^2$ are the sample estimates under the null model $Y_i = \mu + \boldsymbol{\alpha}^T W_i + \epsilon_i$. Specifically, let $\tilde{\mathbf{W}}_{n\times(L+1)} = [\mathbf{1}_n, \mathbf{W}_{n\times L}]$, $\mathbf{A} = \tilde{\mathbf{W}}(\tilde{\mathbf{W}}^T\tilde{\mathbf{W}})^{-1}\tilde{\mathbf{W}}^T$, then $\hat{\mathbf{Y}} = \mathbf{A}\mathbf{Y}$ and $\hat{\sigma}^2 = \mathbf{Y}^T(\mathbf{I} - \mathbf{A})\mathbf{Y}/(n - L - 1)$. Define $V_k = \sum_{m=1}^{\infty} \lambda_m^k$ for any positive integer k. Then the asymptotic normality of the test statistic T_n under the null hypothesis is stated in the following theorem.

Theorem 8. *Assume the density function of error ϵ is symmetric around 0 with $E(\epsilon_i^4) = \tau_4 < \infty$, then under the null hypothesis of no genetic effect,*

$$\sigma_{T_n}^{-1} n T_n \xrightarrow{d} N(0, 1), \tag{4.2.5}$$

if the following condition is satisfied

$$V_4/V_2^2 \to 0 \quad as \quad p(n) \to \infty, \tag{4.2.6}$$

where $\sigma_{T_n}^2$ is the variance of nT_n and can be estimated by the following estimator

$$\hat{\sigma}_{T_n}^2 = \frac{1}{n^2}\Big\{(2 - \frac{12}{n^2} + \frac{6\hat{\Delta}}{n})tr(\mathbf{B}^2) - (\frac{2}{n} + \frac{\hat{\Delta}}{n})tr^2(\mathbf{B}) + \hat{\Delta}tr(\mathbf{B} \circ \mathbf{B})\Big\},$$

where $\mathbf{B} = \mathbf{H}\mathbf{K}_n^0\mathbf{H}$, $\mathbf{H} = \mathbf{I} - \mathbf{A}$, $\circ$ denotes the Hadamard product (elementwise product) and

$\hat{\Delta} = n^{-1} \sum_{i=1}^{n} [(Y_i - \hat{Y}_i)/\hat{\sigma}]^4 - 3$.

Given the asymptotic normality, we can then obtain the p-value for testing $H_0 : h(.) = 0$, i.e.,

$$p\text{-value} = 1 - \Psi(\sigma_{T_n}^{-1} n T_n), \tag{4.2.7}$$

where $\Psi(\cdot)$ is the cumulative density function for a standard normal distribution. As we can see from the theorem, asymptotic normality relies on the key condition (4.2.6). It was mentioned earlier that this value depends on the kernel function, the dimension p of the space where the kernel is defined, and the probability measure μ. To highlight the effect of dimension, define $\pi_p = V_4/V_2^2$. In the following, let us take a further discussion on this condition.

Proposition 2. *Consider linear kernel $K^*(x_1, x_2) = x_1^T x_2$, and assume a multivariate random variable $\mathbf{X}_i = (X_{i1}, ..., X_{ip})$ follows some distribution with covariance matrix $\mathbf{\Sigma}$, $i = 1, ..., n$. Then $\pi_p = tr(\mathbf{\Sigma}^4)/tr^2(\mathbf{\Sigma}^2)$.*

Proposition 3. *Consider the quadratic kernel $K^*(x_1, x_2) = (x_1^T x_2 + 1)^2$, which is a special polynomial kernel. Denote $\tilde{\mathbf{X}}_i = (X_{i1}^2, ..., X_{ip}^2, \sqrt{2}X_{i1}X_{i2}, \cdots, \sqrt{2}X_{i(p-1)}X_{ip}, \sqrt{2}X_{i1}, ..., \sqrt{2}X_{ip})$ as a J-dim random vector with covariance matrix $\mathbf{\Sigma}$, where $J = (p^2 + 3p)/2$. Then $\pi_p = tr(\mathbf{\Sigma}^4)/tr^2(\mathbf{\Sigma}^2)$.*

Proposition 4. *Consider the IBS kernel*

$$K^*(x_1, x_2) = (2p)^{-1} \sum_{m=1}^{p} IBS(x_{1m}, x_{2m}) = (2p)^{-1} \sum_{m=1} (2 - |x_{1m} - x_{2m}|).$$

Denote $\tilde{\mathbf{X}}_i = (X_{i1}, ..., X_{ip}, 1_{\{X_{i1}=1\}}, ..., 1_{\{X_{ip}=1\}})$ as a $2p$-dim random vector with covariance matrix $\mathbf{\Sigma}$. Then $\pi_p = tr(\mathbf{\Sigma}^4)/tr^2(\mathbf{\Sigma}^2)$.

The proofs to Proposition 2-4 are relegated to the last section. From the above propositions we can see that under the three widely-used kernels, condition (4.2.6) is equivalent to a condition on the covariance matrix of a random vector whose length depends on p. Besides, it is a weak condition that brings little constraint to the growth rate of p relative to n. Moreover, if the covariance matrix is of constant order elementwisely, then it is not difficult to see that π_p is of orders p^{-1}, p^{-2} and p^{-1} for linear kernel, quadratic kernel and IBS kernel respectively, and $\pi_p \to 0$ as $p \to 0$. For more discussion on the condition $\mathrm{tr}(\mathbf{\Sigma}^4)/\mathrm{tr}^2(\mathbf{\Sigma}^2) \to 0$, please refer to Chen et al. (2010).

Unfortunately, for most of the complex kernel functions, the explicit condition in terms of a covariance matrix is still unknown or very difficult to derive. However, there do exist consistent estimators for V_2 and V_4 that can provide us the empirical version of π_p. Specifically, $\hat{V}_2 = (P_n^2)^{-1}\mathrm{tr}\{(\mathbf{K}_n^0)^2\}$, $\hat{V}_4 = (P_n^4)^{-1}\mathrm{tr}\{(\mathbf{K}_n^0)^4\}$, $\hat{\pi}_p = \hat{V}_4/\hat{V}_2^2$, and P_n^k is the number of k-permutations of n.

4.2.3 Hypothesis test under multiple candidate kernels

In the previous section, we proposed a test statistic based on a single kernel candidate, and we showed its asymptotic normality under a high-dimensional setting. Since the overall optimal kernel is always unknown, here we consider a set of M (finite) candidate kernel functions $K_1(\cdot,\cdot), K_2(\cdot,\cdot), ..., K_M(\cdot,\cdot)$ with kernel matrix $\mathbf{K}_{n,1}, \mathbf{K}_{n,2}, ..., \mathbf{K}_{n,M}$. Two testing methods are proposed under this setting. In the first one, a new kernel function is generated by taking the simple average of the normalized kernel candidates and then applied into the single-kernel testing procedure. The second method uses a maximum test statistic and the well-developed results on multivariate normal distribution. Both methods are computationally efficient and easy to implement in practice.

4.2.3.1 Test based on kernel average

Without any prior knowledge of the nonparametric function $h(\cdot)$ in (4.2.1), taking the simple average among a set of normalized kernels is a natural choice, where the normalization is necessary for equal-metric consideration. In particularly, denote the standardized kernels with their empirical matrix forms as

$$\mathcal{K}_m(\cdot,\cdot) = \frac{K_m(\cdot,\cdot)}{\mathrm{E}\{K_m(\mathbf{X},\mathbf{X})\}}, \quad \mathbb{K}_{n,m} = \frac{n\mathbf{K}_{n,m}}{\mathrm{tr}(\mathbf{K}_{n,m})}, \quad m = 1, 2, ..., M$$

and the simple average kernel with its matrix form as

$$\tilde{K}(\cdot,\cdot) = \frac{1}{M}\sum_{m=1}^{M} \mathcal{K}_m(\cdot,\cdot), \quad \tilde{\mathbf{K}}_n = \frac{1}{M}\sum_{m=1}^{M} \mathbb{K}_{n,m}.$$

Intuitively, the performance of the test using $\tilde{K}$ is most likely a compromise between the best and the worst ones. Its power will not be close to the optimal one among a candidate set, but it is a conservative option to improve the power over the weakest choice in the set.

4.2.3.2 Maximum test among a candidate set

An alternative idea to the average kernel testing is to perform test and obtain the p-value for individual kernels, then taking the minimum among all the p-values. Such a minimum p-value strategy has been proposed in literature. However, the minimum p-value method often requires computationally expensive techniques such as permutation or perturbation to evaluate the null distribution. Here we propose a maximum statistic among all the candidate kernels and take advantage of the derived asymptotic normality under the large dimensional assumption. Let $T_{n,m}$ and $\sigma^2_{T_{n,m}}$ be the test statistic and the corresponding variance using

the mth kernel function, and denote $Q_m = \sigma_{T_{n,m}}^{-1} n T_{n,m}$ for $m = 1, ..., M$. As we can see from (4.2.7), the p-value is fully determined by Q_m, hence minimizing p-values is essentially equivalent to maximizing Q_m. Therefore, we focus on the maximum statistic

$$Q_{max} = \max_{1 \le m \le M} Q_m.$$

Let $\rho_{kl,n} = \text{cov}(Q_k, Q_l)$ and $\rho_{kl,n} \to \rho_{kl}^0$ as $n \to \infty$, $k, l = 1, ..., M$. The following theorem states the the asymptotic distribution of the maximum statistic Q_{max}.

Theorem 9. *Assume condition (4.2.6) is satisfied for each candidate kernel K_m, then*

$$Q_{max} \xrightarrow{d} \max_{1 \le m \le M} Z_m,$$

where $\mathbf{Z} = (Z_1, Z_2, ..., Z_M)^T$ follows a multivariate normal distribution with mean $\mathbf{0}_M$ and covariance matrix $\mathbf{\Omega}^0 = (\rho_{kl}^0)$.

Based on Theorem 9, the p-value of maximum test is given by

$$\text{P}(Q_{max} > q_{max}) = 1 - \text{P}(Q_{max} \le q_{max}) = [1 - P(\mathbf{Z} \le q_{max} \mathbf{1}_M)] \{1 + o(1)\},$$

where the leading order term can be efficiently and accurately calculated in many popular platforms (e.g., *mvnorm* package in R). Although the true covariance matrix $\mathbf{\Omega}^0$ is unknown, it can be approximately substituted by its consistent estimator $\hat{\mathbf{\Omega}}_n = (\hat{\rho}_{kl,n})$, where

$$\hat{\rho}_{kl,n} = \frac{1}{n^2} \Big\{ (2 - \frac{12}{n^2} + \frac{6\hat{\Delta}}{n}) \text{tr}(\tilde{\mathbf{B}}_k \tilde{\mathbf{B}}_l) - (\frac{2}{n} + \frac{\hat{\Delta}}{n}) \text{tr}(\tilde{\mathbf{B}}_k) \text{tr}(\tilde{\mathbf{B}}_l) + \hat{\Delta} \text{tr}(\tilde{\mathbf{B}}_k \circ \tilde{\mathbf{B}}_l) \Big\},$$

where $\tilde{\mathbf{B}}_m = \mathbf{H} \mathbf{K}_{n,m}^0 \mathbf{H} / \hat{\sigma}_{T_{n,m}}$, $i = 1, ..., M$. This maximum statistic strategy enjoys several

merits. Firstly, the nature of maximum strategy tends to detect the most significant signal among all the kernels in the set. Moreover, the asymptotic normality results obtained under high-dimensional setting greatly reduce our computational burden, and protects the size from being inflated or over-conservative. It should be noted that the maximum method is designed for high-dimensional case only. Under the low-dimensional case, the distribution of Q_{max} can be approximately viewed as the maximum among M correlated chi-square random variables, whose asymptotic behaviors are still unclear to us, and we need to seek for Monte Carlo techniques like perturbation method to evaluate its significance.

4.3 Applications to real data

In this section, we illustrate our methods via the analysis of a Thai baby birth weight data to investigate the significant pathways that are associated with the birth weight. As part of Hyperglycemia and Adverse Pregnacy Outcome (HAPO) study, this data collect genotype and phenotype information for 1209 Thai infants and their mothers. For more details about the HAPO study, see `http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000096.v4.p1&phv=163690&phd=2831&pha=&pht=2446&phvf=&phdf=&phaf=&phtf=&dssp=1&consent=&temp=1`. In the data cleaning step, infants with large proportion of missing SNPs ($> 10\%$) were removed, and SNPs with minor allele frequency (MAF) less than 0.05 or showing deviation from Hardy-Weinberg equilibrium (p-value$<$ 0.001) were also excluded. The final data set contains 970,342 SNPs in 1189 infants (580 males, 509 females). The pathways were defined by Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000). SNPs that are within 5kb up- and downstream of a gene were firstly assigned to the corresponding gene based on Human Genome

Build v38, and then grouped into 186 SNP-sets using the KEGG pathway information retrieved from the Molecular Signature Database (MSigDB) (Subramanian et al., 2005). The length of the SNP-sets ranges from 167 to 9912, where $> 86\%$ of the gene sets are of dimension higher than 500.

We test each pathway (SNP-set) for the association with birth weight, adapting gender (1=male, 2=female) and baby's gestational age at delivery (in weeks) as two covariates. Since we have little knowledge about the underlying true model, we applied three different kernels in the test, including IBS kernel, linear kernel and polynomial kernel ($c = 1, d = 2$). In addition, we also applied simple average kernel test, perturbation method and maximum statistic method, with the three kernels as candidates. The false discovery rate was controlled using q-value (Storey and Tibshirani, 2003) significance levels (0.05, 0.1). Table 4.1 summarizes the significant KEGG pathway indexes using different methods. The corresponding p-values and information of the significant pathways are reported in Table 4.2.

Table 4.1 shows that the perturbation method (Wu et al. 2010) failed to detect any signal, which was probably due to the over-conservative behavior under the high-dimensional setting. Among the seven distinct pathways detected by the three kernels at q-level 0.1, the maximum test was able to capture five (71.4%) of them, while individual kernel and simple average kernel identified four (57.1%) and three (42.9%) of them, respectively. At q-level 0.05, the observations were quite similar. Pooling the significant pathways resulted a union of four, where the maximum test detected three (75%) of them, and the best kernel and simple average kernel identified three (75%) and one (25%) of them, respectively. From Table 4.2, we can obtain the impression that the p-value order of maximum test is generally close to the smallest p-value among the three kernels, which implies that the maximum test tends to improve the power over the weak choices of kernel.

Table 4.1 Significant KEGG pathway indexes using different methods.

q-level	IBS	Linear	Poly	SimAv	Perturb	Max
0.10	36,44,101,169	36,80,123,169	36,48,80,169	36,80,169	NA	36,44,80,101,169
0.05	101,169	36,80,169	36,169	169	NA	36,101,169

Table 4.2 List of significant KEGG pathways and the p-values using the corresponding kernel functions.

idx	# of SNPs	Name*	IBS	Linear	Poly	SimAv	Max
169	485	KTC	1.93E-05	1.42E-07	1.32E-08	1.95E-07	1.32E-08
101	914	KP	1.71E-04	2.27E-02	2.38E-02	5.50E-03	2.84E-04
36	785	KGBCS	3.13E-03	8.14E-04	5.25E-04	9.76E-04	8.54E-04
80	914	KPSP	1.16E-02	1.06E-03	1.53E-03	2.20E-03	1.77E-03
44	1052	KAAM	1.38E-03	8.06E-02	8.28E-02	2.68E-02	2.32E-03
48	419	KGBLANS	3.55E-02	5.18E-03	3.19E-03	7.56E-03	5.41E-03
123	555	KNLRSP	1.32E-02	3.78E-03	7.43E-03	6.02E-03	5.99E-03

*KTC: KEGG thyroid cancer; KP: KEGG peroxisome; KGBCS: KEGG glycosaminoglycan biosynthesis chondroitin sulfate; KPSP: KEGG ppar signaling pathway; KAAM: KEGG arachidonic acid metabolism; KGBLANS: KEGG glycosphingolipid biosynthesis lacto and neolacto series; KNLRSP: KEGG nod like receptor signaling pathway.

4.4 Simulation studies

Extensive simulation studies were conducted to evaluate the type I error rate and the empirical power of the proposed methods. The continuous trait are simulated from the following model

$$Y_i = 0.03W_{i1} + 0.5W_{i2} + h(\mathbf{X}_i) + \epsilon_i, \quad i = 1, ..., n,$$

where ϵ_i are independent and identically distributed random errors generated from $N(0, 1)$ distribution, $W_{i1} \sim N(2, 1)$ and $W_{i2} \sim Ber(0.6)$ are independent covariates, and $\mathbf{X}_i$ is p-dim continuous or discrete vectors representing the genotypes of the SNP-set or the expression profile of the gene-set, $i = 1, ..., n$. To evaluate the type I error, we generate data sets under the null hypothesis of no genetic association (i.e., $h(\cdot) = 0$), and record the proportion

that (incorrectly) reject the null hypothesis. To assess the power, we generate data sets by specifying the h function, and record the proportion that (correctly) reject the null hypothesis. For both power and type I error evaluations, we generate 1000 data sets, and set the significance level as 0.05. In the following two sections, we will assess the performance of the proposed methods under the continuous and discrete variant settings separately.

4.4.1 Continuous variants

Under the continuous variants setting, we simulate $\mathbf{X}_i = (X_{i1}, ..., X_{ip})$ from a multivariate normal distribution with mean $\mathbf{0}_p$ and covariance matrix $\Gamma = (0.6^{|j-k|})$, where $p = 100$ and $i = 1, ..., n$. We allow the sample size to vary as $n = 500, 1000, 2000$. The candidate set consists of three commonly used kernels, including linear kernel, polynomial kernel ($c = 1, d = 2$) and Gaussian kernel $K^*(x_1, x_2) = \exp(-\|x_1 - x_2\|^2/p)$. Besides, simple average kernel method, perturbation method and maximum method were also applied. Table 4.3 reports the type I error rates of tests with varying sample size. We can see that under this setting type I error was not well-protected using the perturbation method, and others are reasonably controlled (close to the nominal level 0.05). This finding implies that perturbation method is over-conservative under the high-dimensional setup, while our methods can control the size.

Table 4.3 Empirical type I error rates of testing with single kernel or multiple kernels under continuous variants setting

n	Gaussian	Linear	Poly	SimAv	Perturb	Max
500	0.055	0.051	0.063	0.051	0.012	0.058
1000	0.055	0.055	0.056	0.053	0.015	0.058
2000	0.052	0.051	0.046	0.051	0.021	0.047

For the power evaluation, we considered four different scenarios and under each scenario h function was set differently as follows:

$$
\begin{aligned}
A: &\quad h(x) = 0.4x_1x_3, \\
B: &\quad h(x) = 0.1x_1 + 0.1x_3 + 0.4x_1x_3 \\
C: &\quad h(x) = 0.1(x_1 - x_3) + 0.8\cos(x_3)\exp(-x_3^2/5), \\
D: &\quad h(x) = \sum_{k=1}^{S} \left\{(-0.01)^k x_k + 2\exp(-x_k^2/100)H_2(x_k/100)\right\} + 0.01\{x_1x_3 + \cos x_3^2\},
\end{aligned}
$$

where $H_k(\cdot)$ is the kth order Hermite polynomial and $S = 30$. For each scenario, 1000 data sets were simulated to estimate the empirical power. Figure 4.1 shows the empirical power under different scenarios. We can see that different kernels result in different powers, depending on the underlying trait architecture. Simple average kernel gives intermediate power among the candidate kernels, and the power of maximum test under each scenario was generally close to the optimal kernel. For example, under scenario A polynomial kernel was the best kernel in the sense of having highest power among the three candidate kernels, and the other two kernels experienced considerable power loss relative to the polynomial kernel. Particularly, the relative power loss for Gaussian and linear kernels were $(67\%, 71\%)$, $(80\%, 85\%)$, $(87\%, 93\%)$ at sample size of 500, 1000 and 2000 respectively, while correspondingly the relative power loss for the simple average kernel, maximum test and perturbation methods were $(66\%, 24\%, 90\%)$, $(80\%, 25\%, 95\%)$, $(87\%, 11\%, 60\%)$. Thus, the maximum test suffered the least power loss when compared to the one with the highest power (the polynomial kernel in this case). This trends were quite similar in other scenarios. Therefore, the maximum test was demonstrated as a good solution in practice to maintain proper power over the weak choices of kernels, under the high-dimensional setting.

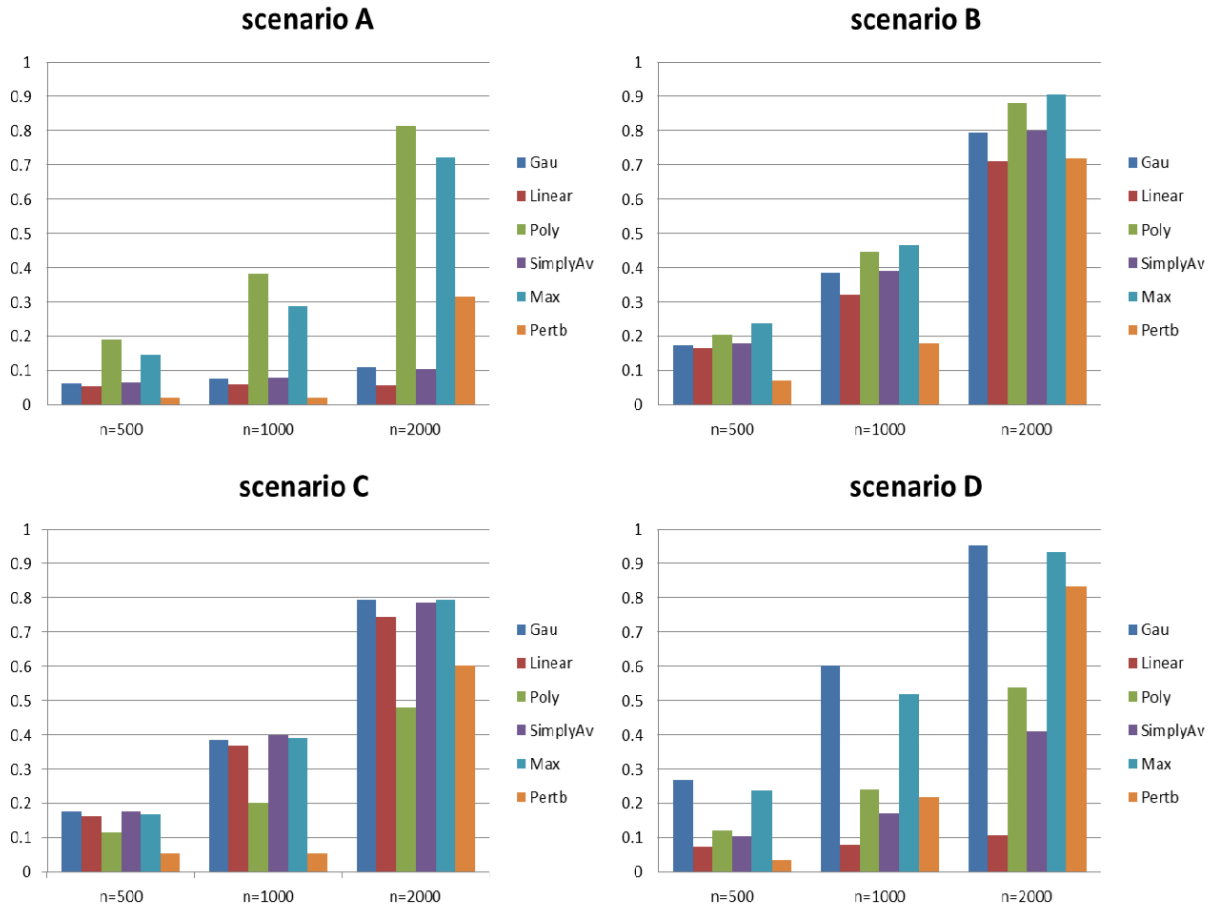

Figure 4.1 Empirical testing power with single kernel and multiple kernels with continuous variants.

4.4.2 Discrete variants

For the discrete variants setting, we generated genotypes based on 378 HAPMAP SNPs located within the KEGG thyroid cancer pathway, which is detected as a significant pathway associated with birth weight in our real data analysis part, using the HAPGEN software (Marchini et al. 2007). Then we simulate the quantitative traits for $n = 1000, 2000$, under scenarios E, F, and G. Under scenario E, we let the $h(\cdot)$ function take the form of

$$h(x) = 0.2(x_1 - x_4) + \cos(x_4)\exp(-x_4^2/5)$$

where the first and fourth SNPs in the set are causal with a nonlinear interaction effect, in addition to the main effects of different directions. To mimic the scenarios where large number of causal SNPs contributes to the trait variation, we consider the following model,

$$h(x) = a_M \sum_{k \in S_M} \beta_k x_k + a_I \sum_{(k,k') \in S_I} \alpha_{kk'} x_k x_{k'},$$

where S_M is a pre-defined set of 30 causal SNPs with main effects, S_I consists of 60 SNP-pairs representing 60 simple interactions. Both $\{\beta_k, k \in S_M\}$ and $\{\alpha_{kk'}, (k,k') \in S_I\}$ are independently generated from Uniform(0,0.02), and are fixed once generated for all simulation replicates. We set the coefficients $(a_M, a_I) = (0.01, 1.5)$ under scenario F, indicating the combination of (weak) main effects and (relatively strong) interaction effects. We let $(a_M, a_I) = (3.5, 0)$ under scenario G, which implies a pure main-effect model.

In addition to linear and polynomial kernels, we add the IBS kernel to the candidate set, since it is commonly used to measure the SNPs similarity between two subjects. Similar to the previous section, the simple average method, perturbation method and maximum method

Table 4.4 Empirical type I error rates of testing with single kernel and multiple kernels under the discrete variants setting

n	IBS	Linear	Poly	SimAv	Perturb	Max
1000	0.052	0.050	0.047	0.050	0.037	0.054
2000	0.053	0.045	0.046	0.043	0.038	0.050

were all applied. Table 4.4 displays the type I error rates of tests under different sample sizes. Similar as what we observed under the continuous variants setting, the perturbation method tends to be conservative under the discrete setting, while the average and the maximum methods maintain reasonable nominal level ($\alpha = 0.05$). The power simulation results are shown in Table 4.5, where the best and second best powers among all the tests are shown with the underline and bold font, respectively. Again, we observed the power difference of applying different kernels. Among the different methods, the perturbation method has the smallest power which might be due to the issue of high-dimensionality. The perturbation method cannot handle the high-dimensional case well. The maximum test has the second highest power. All the powers were improved as sample size increases from 1000 to 2000.

Table 4.5 Empirical power of testing with single kernel and multiple kernels under the discrete variants setting*

n	Scenario	IBS	Linear	Poly	SimAv	Perturb	Max
1000	E	<u>0.526</u>	0.457	0.429	0.480	0.388	**0.488**
	F	0.397	0.412	<u>0.475</u>	0.428	0.383	**0.452**
	G	0.390	0.423	<u>0.444</u>	0.422	0.356	**0.431**
2000	E	<u>0.967</u>	0.932	0.913	0.954	0.927	**0.961**
	F	0.738	0.753	<u>0.813</u>	0.781	0.745	**0.796**
	G	0.769	0.790	<u>0.808</u>	0.798	0.748	**0.799**

* The best power cross all the tests is underlined, and the second best is shown with bold font.

In summary, the simulation results indicate that it is generally safe to apply the maximum test strategy given a set of candidate kernels. The maximum test can control the type I

error well, while it also maintains relatively high power compared to the best one. Without knowing the underlying truth, one can apply the maximum test in practice.

4.5 Discussions

In this chapter, we developed testing procedures to test relationship between multiple variants in a gene set and a quantitative trait, while adjusting for other covariates' effects. We considered a general setting where the variants work coordinately in a non-linear way, and the dimension of the variants p is high in the sense that p can go to infinity as sample size n goes to infinity. We first proposed a test statistic based on a single kernel function, and derived its asymptotic distribution under the null hypothesis. Based on this, we proposed two practical and efficient testing strategies to when multiple candidate kernels are available. We demonstrated, via extensive simulation studies and real data analysis, that under a high-dimensional setting both the simple average method and the maximum method can reasonably control the false positive rate while they can also substantially improve the power over weaker choices of kernels. In particular, the maximum method performs as good as the optimal one given a set of candidate kernels. Compared to the perturbation method (Wu et al., 2013) based on the kernel machine framework, the maximum method outperformed it uniformly in various simulation settings.

Our methods enjoy several advantages as described below. The first advantage lies on the ability to accommodate high-dimensional variants and to maintain reasonable type I error rate, even if the utilized kernel functions do not reflect the underlying relationship between the variants and the trait. Another advantage is the flexibility, which is revealed in two aspects. On one hand, we consider a general model which can potentially capture

any complex interaction mechanism and is different from many models restricted to linear relation and/or linear interactions. On the other hand, when there are a range of kernels that can be selected to form the candidate set, the proposed maximum kernel testing strategy is shown to maintain improved power over the poor choices of kernels in the set, without the prior knowledge of the genetic system. Thirdly, our method is easy to implement and is free of computational burden, by applying the asymptotic result of the test statistic. This feature can greatly facility the applications in pathway (or gene-set) associations studies where the variants (SNPs or gene expression profiles) are typically in high dimensions. However, it should be noted that our method relies on the asymptotic results where the dimension of p is large. In low dimensional cases, the perturbation method by Wu et al. (2010) works well.

In our proposed methods, we only consider continuous responses. Therefore, it is one of our major interests to study the test procedures under a dichotomous response. Besides, our current methods were developed without prior knowledge. However, the kernel function actually allows for the inclusion of known information, such as the minor allele frequencies or association signals from an independent study. For example, weighted linear, quadratic, or IBS kernels can be constructed by assigning weights to variables individually. Thus, extension to weighted kernel is another direction that needs further investigation.

4.6 Proofs

Proof of Proposition 2: By the definition of centralized kernel in (4.2.3), we can obtain the centralized linear kernel as $K(x_1, x_2) = (x_1 - \boldsymbol{\mu})^T (x_2 - \boldsymbol{\mu})$, where $\boldsymbol{\mu} = (\mu_1, \mu_2, ..., \mu_p)^T$ is the mean of random vectors $\mathbf{X}_i$, $i = 1, ..., n$. Assume the covariance matrix has decomposition $\boldsymbol{\Sigma} = \boldsymbol{Q^T \Lambda Q}$ with $\boldsymbol{\Lambda}$ being the diagonal matrix. Let $\tilde{\mathbf{X}}_i = \boldsymbol{\Lambda^{-1/2} Q^T}(\mathbf{X}_i - \boldsymbol{\mu})$, where it

is obvious to see $\mathrm{E}(\tilde{\mathbf{X}}_i) = \mathbf{0}$ and $\mathrm{Var}(\tilde{\mathbf{X}}_i) = \mathbf{I}$, for $i = 1, ..., n$. Noting that the centralized kernel can be written as

$$K(x_1, x_2) = \tilde{\mathbf{X}}_1^T \mathbf{\Lambda} \tilde{\mathbf{X}}_2 = \sum_{m=1}^{p} \Lambda_{mm} \tilde{\mathbf{X}}_{1m} \tilde{\mathbf{X}}_{2m}.$$

we can obtain our claim by letting $\lambda_m = \Lambda_{mm}$ and $\phi_m(x_1) = \tilde{\mathbf{X}}_{1m}$, $m = 1, ..., p$.

Proof of Proposition 3: Let us firstly derive the closed form of the centralized kernel function for the quadratic kernel $K^*(x_1, x_2) = (x_1^T x_2 + 1)^2$. Decompose the kernel K^* into the sum of three parts

$$K^*(x_1, x_2) = (x_1^T x_2)^2 + 2x_1^T x_2 + 1. \tag{4.6.1}$$

In the following we will study each part separately, because the centralized function of the K^* essentially is the sum of the individual centralized functions. For the constant 1 part, the corresponding centralized version is 0. Since we have studied the centralized version of inner product $x_1^T x_2$ in Proposition 2, it remains to investigate the first term $(x_1^T x_2)^2$. Obviously, we have

$$\mathrm{E}(x_1^T \mathbf{X}_2)^2 = x_1^T \mathbf{R} x_1,$$
$$\mathrm{E}(\mathbf{X}_1^T \mathbf{X}_2)^2 = \mathrm{E}\{\mathbf{X}_1^T \mathbf{R} \mathbf{X}_1\} = \mathrm{tr}(\mathbf{R}\mathbf{\Sigma}_0) + \boldsymbol{\mu}^T \boldsymbol{R} \boldsymbol{\mu},$$

where $\mathbf{R} = (R_{ij}) = \mathbf{\Sigma}_0 + \boldsymbol{\mu}\boldsymbol{\mu}^T$ is a constant matrix, and $\boldsymbol{\mu}$, $\mathbf{\Sigma}_0$ are the mean and covariance

matrix of $\mathbf{X}_i$ respectively, $i = 1, ..., n$. Thus the centralized version of $(x_1^T x_2)^2$ is

$$\begin{aligned} & (x_1^T x_2)^2 - x_1^T \mathbf{R} x_1 - x_2^T \mathbf{R} x_2 + \mathrm{tr}(\mathbf{R}\boldsymbol{\Sigma}_0) + \boldsymbol{\mu}^T \boldsymbol{R} \boldsymbol{\mu} \\ = & \sum_{i,j=1}^{p} (x_{1i}x_{1j} - R_{ij})(x_{2i}x_{2j} - R_{ij}) \\ = & \sum_{i=1}^{p} (x_{1i}^2 - R_{ii})(x_{2i}^2 - R_{ii}) + \sum_{i<j} (\sqrt{2}x_{1i}x_{1j} - \sqrt{2}R_{ij})(\sqrt{2}x_{2i}x_{2j} - \sqrt{2}R_{ij}). \end{aligned}$$

Combing the centralized expansions for the three terms in (4.6.1), we can rewrite

$$\begin{aligned} K(x_1, x_2) = & \sum_{i=1}^{p} (x_{1i}^2 - R_{ii})(x_{2i}^2 - R_{ii}) + \sum_{i<j} (\sqrt{2}x_{1i}x_{1j} - \sqrt{2}R_{ij})(\sqrt{2}x_{2i}x_{2j} - \sqrt{2}R_{ij}) \\ & + \sum_{i=1}^{p} (\sqrt{2}x_{1i} - \sqrt{2}\mu_i)(\sqrt{2}x_{2i} - \sqrt{2}\mu_i). \end{aligned}$$

Assume S-dim random vectors

$$\tilde{\mathbf{X}}_i = (X_{i1}^2, ..., X_{ip}^2, \sqrt{2}X_{i1}X_{i2}, ..., \sqrt{2}X_{i(p-1)}X_{ip}, \sqrt{2}X_{i1}, ..., \sqrt{2}X_{ip}),\ i = 1, ..., n$$

follow some distribution with covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{Q^T \Lambda Q}$, then we can achieve our conclusion, i.e., $\pi_p = \mathrm{tr}(\boldsymbol{\Sigma}^4)/\mathrm{tr}^2(\boldsymbol{\Sigma}^2)$, by performing the similar orthogonal transformations we proposed in the proof of Proposition 2.

Proof of Proposition 4: IBS kernel, taking the form of

$$K^*(x_1, x_2) = \frac{1}{2p} \sum_{m=1} (2 - |x_{1m} - x_{2m}|)$$

is defied based on the total number of alleles shared identical by state (IBS) by two subjects at the SNPs within a SNP-set. Noticing $X_{im} \in \{0, 1, 2\} (1 \le i \le n, 1 \le m \le p)$, it is not

difficult to verify that K^* has an alternative form of

$$K^*(x_1, x_2) = \frac{1}{2p}\sum_{m=1}^{p}\frac{1}{2}(x_{1m}-2)(x_{2m}-2) + \frac{1}{2}x_{1m}x_{2m} + 1_{\{x_{1m}=1\}}1_{\{x_{2m}=1\}},$$

hence the centralized kernel has the following expansion

$$K(x_1, x_2) = \frac{1}{2p}\sum_{m=1}^{p}(x_{1m}-2q_m)(x_{2m}-2q_m) + \left[1_{\{x_{1m}=1\}}-\theta_m\right]\left[1_{\{x_{2m}=1\}}-\theta_m\right],$$

where q_m is the minor allele frequency of the mth SNP, and $\theta_m = \mathrm{P}(x_{im}=1) = 2q_m(1-q_m)$. Using the similar arguments as the proofs of Proposition 2, we can obtain the result.

Proof of Theorem 9: Assume condition (4.2.6) is satisfied for each candidate kernel K_m, then

$$Q_m \xrightarrow{d} Z_m,\ m = 1, ..., M.$$

By using Cramèr-Wold device, $(Q_1, ..., Q_M)^T \xrightarrow{d} \mathbf{Z}$. Then the conclusion can be immediately obtained through the continuous mapping theorem.

Chapter 5

Conclusions and future directions

"High-dimension" is one of the new characteristics of high-throughout data. A common feature of high-dimensional data is that the number of features could be much larger than the sample size, the so-called "large p, small n" problem. A specific example in genomic studies is encountered when detecting the significant genes/gene sets that are associated with certain trait, where the number of genetic variants within a gene or gene set could range from a few to a few thousand or even larger, but the sample size is often limited. Such a setup fails most existing methods which are developed for a fixed dimensional case, or do not consider effect of data dimension on the test statistic. To model the systematic mechanism and potential complex interaction among the genetic variants, we proposed to model the gene set effect via a flexible non-parametric regression function under a "large p, small n" setup.

In Chapter 2, we proposed a nonparametric U-statistic for testing the high-dimensional non-parametric function in a reproducing kernel Hilbert space generated by a positive definite or semi-definite kernel. We derived the asymptotic distributions of the test statistic under the null hypothesis and a sequence of local alternatives under a "large p, small n" setting without assuming specific error distribution. We derived the explicit power function of the test based on which we can empirically select optimal kernel function that provides a solution to a long-standing question in literature about optimal kernel selection. To further improve the testing power while maintaining appropriate testing size, a kernel regularization

technique was proposed. Unlike the BIC criterion proposed for kernel selection in kernel machine testing procedure, our approach is tailored to a hypothesis-testing problem and particularly designed for improving the power of the proposed test. Both simulation and real data analysis demonstrate the power of our method. In addition to strong practical motivation, our method contributes to the theory and methodology of kernel-based testing of nonparametric functions, especially under the "large p, small n" set up.

Chapter 3 considers the optimality of test procedure we proposed in the previous chapter, from the minimax point of view. Especially, we discuss the optimal test under the high-dimensional linear model (corresponds to the linear kernel), where the p-dim variables are correlated and the dimension p can go to infinity as n goes to infinity. Without the sparsity assumption that only a small proportion (goes to zero as sample size goes to infinity) of the variants contribute to the phenotype or the independent variants assumption presented in existing literature, we consider a structure-free scenario. We firstly establish the boundary that separates the detectable region and non-detectable region. Then the test statistic using linear kernel is introduced and shown to be rate-optimal under certain conditions on the increasing speed of dimensional p and the decay speed of eigenvalues of the covariance matrix.

We start Chapter 4 by upgrading the kernel-based test proposed in Chapter 2 to a general version that allows the adjustment of covariants, under the high-dimensional setting. Then we provide the asymptotic distribution of the general test statistic under the null hypothesis. Motivated by the testing problem using multiple kernel candidates, we develop two practical and efficient testing procedures: simple average method and maximum method. Unlike other computational-intensive approaches using Monte Carlo p-value to evaluate significance, both strategies are purely based on the asymptotic results and easy to implement. In the application to Thai baby birth weight data, we demonstrated that both strategies lead

to the detection of more signals than testing using the poor choice of kernel, as well as more findings than the perturbation method (Wu et al., 2013) that is proposed under the kernel machine framework. We further confirm the merits of our strategies through the comprehensive simulation studies under continuous or discrete variants settings, where the maximum method further displays its competitive performance in the sense that only a small difference in power versus using the best candidate kernel.

In this dissertation, we focus on continuous traits. However, many traits of interest in practice are qualitative. For example, the traits in case-control studies might be the disease status for individuals. Therefore, extension to dichotomous traits is an important direction of our future research.

As observed empirically in our simulation studies in Chapter 2, the high-dimensional non-parametric test intuitively suffers from power loss when lots of "noise" variables are included under a sparse alternative. Therefore, it is natural to consider removing those noise variables to enhance the power, where the challenge remaining is how to perform the test and to eliminate the noises at the same time. The exploration of simultaneously testing and removing the influence of noise variants can be an area of potential future research.

Another direction of my interest is to introduce the kernel-based testing into gene-environment interaction context. It has been increasingly recognized that many complex disease are not triggered by genetic factor only, but rather through interaction between genetic and environmental factor. However, most of current methods on gene-environment interaction are not applicable under high-dimensional setting.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Ashburner, M., et al. (2000). Gene Ontology: tool for the unification of biology. *Nature genetics*, **25**, 25-29.

[2] Bao, Y. and Ullah, A. (2010). Expectation of quadratic forms in normal and nonnormal variables with econometric applications. *Journal of Statistical Planning and Inference*, **140**, 1193-1205.

[3] Cai, T., and Ma, Z. (2013). Optimal hypothesis testing for high dimensional covariance matrices.*Bernoulli*, **19**, 2359-2388.

[4] Cantor, R. and Bell, J. (2001). Association study designs for complex diseases. *Nature Reviews Genetics*, **2**, 91-99.

[5] Chen, S., Härdle, W., and Li, M. (2003). An empirical likelihood goodness-of-fit test for time series. *Journal of the Royal Statistical Society, Series B*, **65**, 663-678.

[6] Chen, S., Zhang, L. and Zhong, P. (2010). Tests for high-dimensional covariance matrices. *Journal of the American Statistical Association* **105**, 810-819.

[7] Croft, D., O'Kelly, G., Wu, G., et al. (2010). Reactome: a database of reactions, pathways and biological processes. *Nucleic acids research*, gkq1018.

[8] Dai, G., Yeung, D., and Qian, Y. (2007). Face recognition using a kernel fractional-step discriminant analysis algorithm. *Pattern Recognition*, **40**, 229-243.

[9] Dauxois, J., Pousse, A., and Romain, Y. (1982). Asymptotic theory for the principal component analysis of a vector random function: Some applications to statistical inference. *Journal of Multivariate Analysis*, **12**, 136-154.

[10] Donoho, D. and Jin, J. (2004) High criticism for detecting sparse heterogeneous mixtures. *Annals of Statistics*, **32**, 962-994.

[11] Donoho, D. and Jin, J. (2008) High criticism thresholding: Optimal feature selection when useful features are rare and weak. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 14790-14795.

[12] Donoho, D. and Jin, J. (2009). Feature selection by higher criticism thresholding achieves optimal phase diagram. *Royal Society Philosophical Transaction A*, **367**, 4449-4470.

[13] Eichler, et al. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics*, **11**, 446-460.

[14] Fan, J. and Gilbels, I. (1996). *Local polynomial modeling and its applications*, Chapman and Hall, Suffolk.

[15] Feng, L., Zou, C., Wang, Z. and Chen, B.(2013). Rank-based score tests for high-dimensional regression coefficients. *Electronic Journal of Statistics*, **7**, 2131-2149.

[16] Gao, J. and Gijbels, I. (2008). Bandwidth selection in nonparametric kernel testing. *Journal of the American Statistical Association*, **103**, 1584-1594.

[17] Gibson, G. (2010). Hints of hidden heritability in GWAS. *Nature Genetics*, **42**, 558-560.

[18] Goeman, J., Geer, S. and Houwelingen, H. (2006). Testing against a high dimensional alternative. *Journal of the Royal Statistical Society, Series B*, **68**, 477-493.

[19] Goeman, J. and Houwelingen, H. (2011). Testing against a high-dimensional alternative in the generalized linear model: asymptotic type I error control. *Biometrika*, **98**, 381-390.

[20] Guerre, E. and Lavergne, P. (2002) Optimal minimax rates for nonparametric specification testing in regression models. *Econometric Theory*, **18**, 1139-1171.

[21] Hall, P. and Jin, J. (2010) Innovated higher criticism for detecting sparse signals in correlated noise. *Annals of Statistics*, **38**, 1686-1732.

[22] Harchaoui, Z., Bach, F. and Moulines, E. (2008). Testing for homogeneity with kernel fisher discriminant analysis. *Advances in Neural Information Processing Systems (NIPS)*, 609-616.

[23] Härdle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *The Annals of Statistics*, **21**, 1926-1947.

[24] Hirschhorn, J. and Daly, M. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, **6**, 95-108.

[25] Hofmann, T., Schölkopf, B. and Smola, A. (2008). Kernel methods in machine learning. *The Annals of Statistics*, **36**, 1171-1220.

[26] Ingster, Y., Tsybakov, A. and Verzelen, N. (2010) Detection boundary in sparse regression. *Electronic Journal of Statistics*, **4**, 1476-1526.

[27] Ingster, Y., Pouet, C. and Tsybakov. (2009) Classification of sparse high-dimensional vectors. *Royal Society Philosophical Transactions A*, **367**, 4427-4448.

[28] Ingster, Y. (1993) Asymptotically minimax hypothesis testing for nonparametric alternatives, parts I, II, and III. *Mathematical Methods of Statistics*, **2**, 85-114,171-189, and 249-268.

[29] Kanehisa, M. and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, **28**, 27-30.

[30] Kim, S., Magnani, A., and Boyd, S. (2006). Optimal kernel selection in kernel fisher discriminant analysis. *Proceeding ICML '06 Proceedings of the 23rd international conference on Machine learning*, 465-472.

[31] Kong, A. et al. (2009). Parental origin of sequence variants associated with complex diseases. *Nature*, **462**, 868-874.

[32] Kumar, A. (1973). Expectation of product of quadratic forms. *Sankhy: The Indian Journal of Statistics, Series B*, **35**, 359-362.

[33] Lan, W., Wang, H. and Tsai, C. (2014). Testing covariates in high dimensional regression. *Annals of Institute Statistical Mathematics*, **66**, 279-301.

[34] Lee, A. (1990). *U-statistics: Theory and Practice*. New York, Basel: Marcel Dekker, Inc.

[35] Li, S. and Cui, Y. (2012). Gene-centric gene-gene interaction: a model-based kernel machine method. *Annals of Applied Statistics*, **6**, 1134-1161.

[36] Lin, X., Cai, T., Wu, M., Zhou, Q., Liu, G., Christiani, D. and Lin, X. (2011). Kernel machine SNP-set analysis for censored survival outcomes in genome-wide association studies. *Genetic Epidemiology*, **35** , 82-93

[37] Lindsay, B., Markatou, M., Ray, S., Yang, K. and Chen, S.(2008). Quadratic distances on probabilities: a unified foundation. *The Annals of Statistics*, **36**, 983-1006.

[38] Lindsay, B., Markatou, M., and Ray, S. (2014). Kernels, Degrees of Freedom, and Power Properties of Quadratic Distance Goodness-of-Fit Tests. *Journal of the American Statistical Association*, **109**, 395-410.

[39] Liu, D., Ghosh, D. and Lin, X. (2008). Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. *BMC Bioinformatics*,**9**, 292.

[40] Liu, D., Lin, X. and Ghosh, D. (2007). Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. *Biometrics*, **63**, 1079-1088.

[41] Lkhagvadorj, S., Qu, L., Cai, W., Couture, O., Barb, C., Hausman, G., Nettleton, D., Anderson, L., Dekkers, J., and Tuggle, C. (2009). Microarray gene expression profiles of fasting induced changes in liver and adipose tissues of pigs expressing the melanocortin-4 receptor D298N variant. *Physiological Genomeics*, **38**, 98-111.

[42] Marchini, J., Howie, B., Meyers, S., McVean, G. and Donnelly, P. (2007). A new multipoint method for genome-wide asoociation studies by imputation of genotypes. *Nature Genetics*, **39**, 906-913.

[43] Manolio, T., et al. (2009). Finding The missing heritability of complex diseases. *Nature*, **461**, 747-753.

[44] Mefford, H. and Eichler, E. (2009). Duplication hotspots, rare genomic disorders and common disease. *Current Opinion in Genetics & Development*, **19**, 196-204.

[45] Mukhopadhyay, I., Feingold, E., Weeks, D. and Thalamuthu, A. (2010). Association tests using kernel-based measures of multi-locus genotype similarity between individuals. *Genetic Epidemiology*, **34**, 213-221.

[46] Newton, M., Quintana, F., Boon, J., Sengupta, S. and Ahlquist, P. (2007). Random-set methods identify distinct aspects of the enrichment signals in gene-set analysis. *The Annals of Applied Statistics*, **1**, 85-106.

[47] Poggio, T., and C. R. Shelton. (2002). "On the mathematical foundations of learning," *American Mathematical Society*, **39**, 1-49.

[48] Reiss, P., Stevens, M., Shehzad, Z., Petkova, E and Milham, M. (2010). On distance-based permutation tests for between-group comparisons. *Biometrics*, **66**, 636-643.

[49] Schaid, D. (2010) Genomic Similarity and Kernel Methods I: Advancements by Building on Mathematical and Statistical Foundations.*Human heredity*, **70**, 109-131.

[50] Schaid, D. (2010) Genomic Similarity and Kernel Methods II: Methods for Genomic Information. *Human heredity*, **70**, 132-140.

[51] Steinwart, I. and Scovel, C. (2012) Mercer's theorem on general domains: On the interaction between measures, kernels, and RKHSs. *Constructive Approximation*, **35**, 363-417.

[52] Storey, J. and Tibshirani, R. (2003) Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, **100**, 9440-9445.

[53] Subramanian, A. et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 15545-15550.

[54] Tzeng, J., Zhang, D., Chang, S., Thomas, D. and Davidian, M. (2009). Gene-trait similarity regression for multimarker-based association analysis. *Biometrics*, **65**, 822-832.

[55] Wang, K., Li, M. and Bucan, M. (2007). Pathway-based approaches for analysis of genomewide association studies. *The American Journal of Human Genetics*, **81**, 1278-1283.

[56] Wang, K., Li, M. and Hakonarson, H. (2010). Analyzing biological pathway in genome-wide association studies. *Nature Reviews Genetics*, **11**, 843-854.

[57] Wang, S. and Cui, H. (2013). Generalized F test for high dimensional linear regression coefficients. *Journal of Multivariate Analysis*, **117**, 134-149.

[58] Wessel, J. and Schork, N. (2006). Generalized genomic distance-based regression methodology for multilocus association analysis. *The American Journal of Human Genetics*, **79**, 792-806.

[59] Wu, M. et al. (2010) Powerful SNP-set analysis for case-control genome-wide association studies. *The American Journal of Human Genetics*, **86**, 929-942.

[60] Yu, K. et al. (2009). Pathway analysis by adaptive combination of p-values. *Genetic Epidemiology*, **33**, 700-709.

[61] Zhong, P. and Chen, S. (2011). Tests for high dimensional regression coefficients with factorial designs. *Journal of the American Statistical Association*, **106**, 260-274.

[62] Zuk, O., Hechter, E., Sunyaev, S. and Lander, E. (2012). They mystery of missing heritability: Genetic interactions creat phantom heritability. *Proceedings of the National Academy of Sciences*, **109**, 1193-1198.