Journal of Machine Learning Research 21 (2020) 1-47 Submitted 9/19; Revised 4/20; Published 5/20

Minimax Nonparametric Parallelism Test

Xin Xing XINXING@VT.EDU

Meimei Liu MEIMEILIU@VT.EDU

Department of Statistics
Virginia Tech, Blacksburg, VA, 24061, USA

Ping Ma PINGMA@UGA.EDU

Wenxuan Zhong WENXUAN@UGA.EDU

Department of Statistics
University of Georgia, Athens, GA 30601, USA

Editor: Xiaotong Shen

Abstract

Testing the hypothesis of parallelism is a fundamental statistical problem arising from many applied sciences. In this paper, we develop a nonparametric parallelism test for inferring whether the trends are parallel in treatment and control groups. In particular, the proposed nonparametric parallelism test is a Wald type test based on a smoothing spline ANOVA (SSANOVA) model which can characterize the complex patterns of the data. We derive that the asymptotic null distribution of the test statistic is a Chi-square distribution, unveiling a new version of Wilks phenomenon. Notably, we establish the minimax sharp lower bound of the distinguishable rate for the nonparametric parallelism test by using the information theory, and further prove that the proposed test is minimax optimal. Simulation studies are conducted to investigate the empirical performance of the proposed test. DNA methylation and neuroimaging studies are presented to illustrate potential applications of the test. The software is available at `https://github.com/BioAlgs/Parallelism`.

Keywords: asymptotic distribution, minimax optimality, nonparametric inference, parallelism test, penalized least squares, smoothing spline ANOVA, Wald test

1. Introduction

The assessment of parallelism is a fundamental problem in statistical inference and arises from many applications. For example, in genomic studies, one of primary interest is to detect genes with nonparallel expression patterns in time course studies (Storey et al., 2005; Ma et al., 2009). Another motivating example is in epigenomics, researchers are interested in testing whether the patterns of DNA methylation intensities along genome in the treatment and control groups are parallel or not (Hansen et al., 2012). The abnormal DNA methylation patterns are associated with changes in many important biological processes such as imprinting, X-chromosome inactivation, and aging (Schübeler, 2015). In functional neuroimaging, a common problem is to detect nonparallel signals (Nichols and Holmes, 2002; Orrison et al., 2017) among different brain regions.

There is an immense literature focused on the analysis of the parallelism of trends using linear model-based approaches, ranging from simple ANOVA (Sthle and Wold, 1989) to linear mixed models (Vossoughi et al., 2016). However, the linear model-based approaches have a limited ability to parsimoniously represent non-linear structures in complex data. Nonparametric parallelism comparison methods have drawn huge attention due to the modeling flexibility. Munk and Dette (1998) developed a test statistic through a weighted L_2 distances for the regression functions based on similar equal-spaced fixed design. Degras et al. (2011) tested the parallelism of multiple time series based on the L_2 distances between the local linear estimator of each individual curve and the global one for time series data when the time points are evenly spaced. Wang (1998) proposed a wavelet-based method to measure the changes of curves. Liu and Wang (2004) compared different nonparametric testing methods and showed that the performances of these tests depend on the shape of the true function. Ma et al. (2009) proposed an approximate F-test to detect nonparallel patterns in time course gene expression data with a more flexible random design.

However, rigorous testing methods with optimal power guarantees are still lacking in the existing nonparametric parallelism literature. The key cause of such research gap is that distinguishing from the simple/linear/polynomial null hypothesis, the parameter space of the null hypothesis for the nonparametric parallelism testing is a nonparametric function class with infinite dimension. How to conduct a rigorous test for such composite functional null hypothesis is still an open question. A major motivation of this article is on developing a nonparametric parallelism testing approach that detects the significance of the nonparallel effect, while guarantees statistical optimality in the sense of minimax testing rate, facilitating the power performance analysis.

In this article, we develop a nonparametric parallelism test based on the decomposition of tensor product reproducing Hilbert space (RKHS) (Wahba, 1990; Gu, 2013; Wang, 2011) under both fixed and random design. Tensor product RKHS provides a flexible space for modeling complex functions; see Wahba et al. (1995), Wood (2003) and reference therein. For the simplicity of description, we consider the case that there are two predictors only. Suppose the response variable Y_{ij} is the observed value of the jth subject at the ith time or spatial location for $i = 1, \cdots, n$ and $j = 1, \cdots, s$. Y_{ij} depends on two predictors $x_i^{\langle 1\rangle}$ and $x_j^{\langle 2\rangle}$ through an unknown bivariate function $f(\cdot,\cdot) \in \mathcal{H}$, the tensor product RKHS, where $x_i^{\langle 1\rangle} \in \mathcal{X}_1 = [0, 1]$ is a continuous variable representing the ith time or the i-th spatial location, and $x_j^{\langle 2\rangle} \in \mathcal{X}_2 = \{0, 1\}$ is a discrete variable representing the jth subject in different groups, $x_j^{\langle 2\rangle} = 1$ represents the jth subject in treatment group, otherwise in control group. That is,

$$Y_{ij} = f(x_i^{\langle 1\rangle}, x_j^{\langle 2\rangle}) + \epsilon_{ij}, i = 1, \cdots n, j = 1, \cdots, s, \tag{1}$$

where ϵ_{ij}s are i.i.d. random noise following a normal distribution with mean zero, variance σ^2, and s is the number of subjects. Each subject can be represented by a curve. When $s = 2$, there are two curves in total and each group has only one curve. When $s > 2$, we have multiple curves in each group. We assume the i.i.d. random noise since, in many scientific experiments, the random errors are attributed to environmental factors independent of the time points or spatial location. For example, in the fMRI data analysis in Section 6, the error is mostly attributed to the random movement of the head and imaging noise which are independent with the time.

Analogous to the classical ANOVA decomposition, $f \in \mathcal{H}$ has the smoothing spline ANOVA (SSANOVA) decomposition (Wahba, 1990):

$$f(x_i^{\langle 1\rangle}, x_j^{\langle 2\rangle}) = f_{00} + f_{10}(x_i^{\langle 1\rangle}) + f_{01}(x_j^{\langle 2\rangle}) + f_{11}(x_i^{\langle 1\rangle}, x_j^{\langle 2\rangle}), \tag{2}$$

where f_{00} is the grand mean, f_{10} and f_{01} are the main effects, and f_{11} is the nonparallel effect. When $f_{11} = 0$ (see the left panel in Figure 1), the curves in two groups are parallel. Then $f_{11} = 0$ is equivalent to that $f(x^{\langle 1\rangle}, 0)$ and $f(x^{\langle 1\rangle}, 1)$ are parallel. When $f_{11} \neq 0$ (see the right panel in

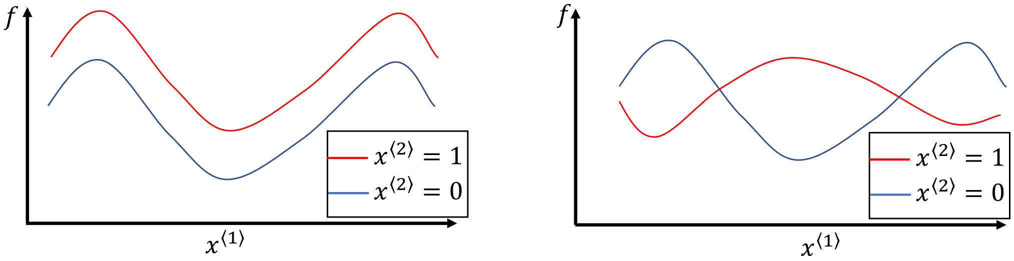


Figure 1: An illustration of two scenarios of a bivariate function $f(x^{\langle 1\rangle}, x^{\langle 2\rangle})$, where $x^{\langle 1\rangle}$ is continuous, $x^{\langle 2\rangle}$ only takes two values, 0 and 1. Left panel: the scenario with $f_{11} = 0$, i.e., $f(x^{\langle 1\rangle}, 0)$ and $f(x^{\langle 1\rangle}, 1)$ are parallel. Right panel: the scenario with $f_{11} \neq 0$, i.e., $f(x^{\langle 1\rangle}, 0)$ and $f(x^{\langle 1\rangle}, 1)$ are nonparallel.

Figure 1), the magnitude of $||f_{11}||_2^2$ characterizes the significance of the non-parallelism between the treatment and control groups, where $\|f_{11}\|_2^2 = \sum_{x^{\langle 2\rangle}=0}^{1} \int_0^1 f^2(x^{\langle 1\rangle}, x^{\langle 2\rangle}) d\omega_1$, with ω_1 as the marginal density of $x^{\langle 1\rangle}$. Statistically, the hypothesis testing for parallelism can be formulated as

$$H_0 : f_{11} = 0 \quad \text{vs} \quad H_1 : f_{11} \neq 0. \tag{3}$$

We introduce two concrete examples which motivate our study.

Example 1. *DNA methylation in case-control study.* DNA methylation is an essential epigenetic mechanism that regulates gene expression. Aberrant DNA methylation contributes to a number of human diseases including cancer (Stach et al., 2003). In a typical case-control study of DNA methylation (Filarsky et al., 2016), the DNA methylation level, denoted as Y_{ij} at the ith location $x_i^{\langle 1\rangle}$ on the genome for the jth individual in group $x_j^{\langle 2\rangle}$, can be modeled using Equation (1), where f is an unknown function with the SSANOVA decomposition in Equation (2). A primary focus is to infer whether the DNA methylation levels have different profiles along the genome between the case and control groups, i.e., testing the presence/absence of nonparallel effect f_{11} as in Equation (3).

Example 2. *Neuroimaging using functional magnetic resonance imaging (fMRI).* fMRI is a powerful neuroimaging technology for the diagnosis of many brain-related diseases. It measures brain activity by detecting changes associated with blood flow. The primary form of fMRI uses the blood-oxygen-level dependent (BOLD) as signal (Huettel et al., 2004). In many case-control studies, the BOLD signal, Y_{ij}, at the ith time $x_i^{\langle 1\rangle}$ for the jth subject in group $x_j^{\langle 2\rangle}$ is measured for a particular region of interest (ROI), and can be modeled using Equation (1), where f is an unknown function with the SSANOVA decomposition in Equation (2). The goal is to test whether the BOLD signals in two groups have same patterns along the time, i.e., test the significance of nonparallel effect f_{11} in Equation (3).

We first establish the minimax lower bound for nonparametric parallelism test in Equation (3) for general testing rules with the aid of tensor product decomposition of RKHS and the information theory. The tensor product decomposition in Equation (2) enables us to quantify the magnitude of nonparalelism by $||f_{11}||_2$, where $|| \cdot ||_2$ is the L_2 norm. Intuitively, the smaller $||f_{11}||_2$ is, the harder it is to distinguish the alternative hypothesis from the null. In analyzing the power performance, we consider a slightly different alternative hypothesis,

$$H_1^* : ||f_{11}||_2 \geq d_n, \tag{4}$$

where we remove the neighborhood within the d_n distance of $f_{11} = 0$ from the original alternative H_1. Here the sequence d_n is called the distinguishable rate (or separation rate) (Ingster and Suslina, 2012; Giné and Nickl, 2015). We first introduce a geometric interpretation of the testing problem in Equation (3), and then establish a general minimax lower bound for the distinguishable rate for the nonparametric parallelism test using the Bernstein k-width in information theory (Pinkus, 2012). Bernstein k-width provides a geometric measure of the distinguishable rate and is easy to evaluate in the tensor product RKHS. Recently, similar technique was also used in analyzing the testing problems over cones and studied in Gaussian sequence models (Wei and Wainwright, 2020).

In addition, we propose a Wald-type test statistic as the squared empirical norm of the penalized least square estimator of f_{11}. We derive its asymptotic null distribution, which satisfies the Wilks phenomenon. The asymptotic distribution of our test statistic is Gaussian, and the testing rule does not depend on any unknown quantities, thus is easy to compute. We can further reduce the computational cost by applying many popular fast computation methods such as fast random kernel methods Alaoui and Mahoney (2015) and subsampling methods such as Ma et al. (2015); Kim and Gu (2004). We note that our proposed Wald-type test distinguishes from the existing nonparametric testing methods as follows. The existing testing procedures mostly consider simple null hypothesis, such as the generalized likelihood ratio test in Fan et al. (2001), the penalized likelihood ratio test in Shang and Cheng (2013), the wavelet based method in Shen et al. (2002), and kernelized Stein method in Liu et al. (2016), whereas we consider a composite null hypothesis. More importantly, there is a nontrivial technical complication in addition to the above model setting difference. The composite null hypothesis $H_0 : f_{11} = 0$ here defines a nonparametric function in an infinite-dimensional functional space rather than a parametric function in a finite-dimensional parameter space as required in Shang and Cheng (2013), because testing $H_0 : f_{11} = 0$ is equivalent to testing $H_0 : f \in \{f_{00} + f_{10} + f_{01}\}$. Developing the limiting distribution of the test statistic in an infinite-dimensional null hypothesis space and quantifying the testing difficulty are very challenging since the distribution relies on the more delicate tensor product decomposition of the RKHS.

We further prove that the upper bound of the distinguishable rate for the proposed Wald type test matches the established minimax lower bound. Thus the proposed Wald-type test is minimax optimal. To the best of our knowledge, our work is the first one in establishing the minimax nonparametric parallelism test. Based on the Wald-type test statistic, we propose a data-adaptive choice of the regularization parameter with testing optimality guarantee.

The rest of the paper is organized as follows. We introduce the background of tensor product RKHS in Section 2. In Section 3, we introduce a minimax principle and a geometric interpretation of the parallelism testing problem. In Section 4, we derive the minimax lower bound of the distinguishable rate for general parallelism test using the information theory. Section 5 presents various simulation studies demonstrating substantial performance of our testing method, and Section 6 applies the methods to genome-wide anomaly of DNA methylation in *chronic lymphocytic leukemia* patients and brain function change in patients with *Alzheimer disease*. We conclude with a few remarks in Section 7. All technical proofs are relegated to the Appendix and Supplementary Material.

2. Background

In this section, we introduce some background of the tensor product RKHS, its tensor product decomposition, together with the penalized least square estimation.

2.1. Reproducing Kernel Hilbert Space

Given an RKHS $\mathcal{H}$ with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, there exists a symmetric and square integrable function $\mathcal{K}(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that

$$\langle f, \mathcal{K}(x, \cdot) \rangle_{\mathcal{H}} = f(x), \text{ for all } f \in \mathcal{H} \text{ and } x \in \mathcal{X}.$$

We call $\mathcal{K}$ as the reproducing kernel of $\mathcal{H}$. By Mercer's theorem, any continuous kernel has the following decomposition

$$\mathcal{K}(x, y) = \sum_{\nu=0}^{\infty} \lambda_\nu \varphi_\nu(x) \varphi_\nu(y), \tag{5}$$

where λ_νs are non-negative descending eigenvalues and φ_νs are eigen-functions.

We consider the bivariate function f in Equation (1) on the product domain $\mathcal{X}_1 \times \mathcal{X}_2$. We assume that f is a function in a tensor product RKHS (Lin, 2000)

$$\mathcal{H} = \mathcal{H}^{\langle 1 \rangle} \otimes \mathcal{H}^{\langle 2 \rangle}. \tag{6}$$

Given the Hilbert space $\mathcal{H}^{\langle 1 \rangle}$ and $\mathcal{H}^{\langle 2 \rangle}$, $\mathcal{H}^{\langle 1 \rangle} \otimes \mathcal{H}^{\langle 2 \rangle}$ is defined as the completion of the class of functions with the form $\sum_{i=1}^{M} \eta_{1i}(x)\eta_{2i}(y)$, for $\eta_{1i} \in \mathcal{H}^{\langle 1 \rangle}$, $\eta_{2i} \in \mathcal{H}^{\langle 2 \rangle}$, and M is any positive integer. We consider $\mathcal{H}^{\langle 1 \rangle}$ as an mth order homogeneous Sobolev space, i.e.,

$$\mathcal{H}^{\langle 1 \rangle} = \{\eta_1 \in L_2[0,1] \mid \eta_1^{(k)} \text{ is absolutely continuous and } \eta_1^{(k)}(0) = \eta_1^{(k)}(1) \\ \text{for } k = 0, 1, \ldots, m-1, \eta_1^{(m)} \in L_2[0,1]\},$$

and $\mathcal{H}^{\langle 2 \rangle}$ is a two-dimensional Euclidean space with standard Euclidean norm.

Assume that $\mathcal{H}^{\langle 1 \rangle}$ has the eigenvalue and eigenvector pairs $\{\mu_i, \phi_i\}_{i=0}^{\infty}$ and $\mathcal{H}^{\langle 2 \rangle}$ has the eigenvalue and eigenvector pairs $\{\nu_j, \psi_j\}_{j=1}^{2}$. Then we have the eigenvalue and eigenvector pairs for the kernel function $\mathcal{K}$ in $\mathcal{H}$ as

$$\{\mu_i \nu_j, \phi_i \psi_j\} \qquad \text{for } i = 0, \ldots, \infty, j = 1, 2, \tag{7}$$

in the decomposition in Equation (5). We refer Equation (7) as the eigensystem for $\mathcal{H}$. We further denote $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ as the product norm induced by the norm on the marginal space $\mathcal{H}_1$ and $\mathcal{H}_2$ (Lin, 2000).

Using the Riesz representation theorem (Schölkopf et al., 2001), we can easily represent any function $f \in \mathcal{H}$ as in the following Lemma.

Lemma 1 *Given the sampling points $\mathbf{x}_{ij} = (x_i^{\langle 1 \rangle}, x_j^{\langle 2 \rangle}), i = 1, \cdots, n$ and $j = 1, \cdots s$, for any f in a reproducing kernel Hilbert space $\mathcal{H}$, there exists a set of reproducing kernels $\mathcal{K}_{\mathbf{x}_{ij}}(\cdot, \cdot)$ such that*

$$f(x^{\langle 1 \rangle}, x^{\langle 2 \rangle}) = \sum_{i=1}^{n} \sum_{j=1}^{s} \alpha_{ij} \mathcal{K}_{\mathbf{x}_{ij}}(x^{\langle 1 \rangle}, x^{\langle 2 \rangle}) + \rho(x^{\langle 1 \rangle}, x^{\langle 2 \rangle}). \tag{8}$$

Lemma 1 implies that f can be expressed as a sum of a linear expansion of $\mathcal{K}_{\mathbf{x}_{ij}}$ and a nonlinear function ρ. Notice that when $(x^{\langle 1 \rangle}, x^{\langle 2 \rangle}) \in \{\mathbf{x}_{ij}\}_{i=1,\cdots,n}^{j=1,\cdots,s}$, we have $\rho(x^{\langle 1 \rangle}, x^{\langle 2 \rangle}) = 0$. Thus, $\rho(\cdot, \cdot)$ can be considered as a residual that quantifies the unknown information of function f. To get an estimate of f, we only need to specify $\mathcal{K}_{\mathbf{x}_{ij}}(\cdot, \cdot)$ and estimate α_{ij}. Next, we provide a way to construct the reproducing kernels $\mathcal{K}_{\mathbf{x}_{ij}}(\cdot, \cdot)$. In order to do that, we need the following two lemmas.

Lemma 2 *Suppose $\mathcal{K}^{\langle 1\rangle}$ is the reproducing kernel of $\mathcal{H}^{\langle 1\rangle}$ on $\mathcal{X}_1$, and $\mathcal{K}^{\langle 2\rangle}$ is the reproducing kernel of $\mathcal{H}^{\langle 2\rangle}$ on $\mathcal{X}_2$. Then the reproducing kernels of $\mathcal{H}^{\langle 1\rangle}\otimes\mathcal{H}^{\langle 2\rangle}$ on $\mathcal{X}=\mathcal{X}_1\times\mathcal{X}_2$ is $\mathcal{K}(\mathbf{x},\mathbf{z})=\mathcal{K}^{\langle 1\rangle}(x^{\langle 1\rangle},z^{\langle 1\rangle})\mathcal{K}^{\langle 2\rangle}(x^{\langle 2\rangle},z^{\langle 2\rangle})$ with $\mathbf{x}=(x^{\langle 1\rangle},x^{\langle 2\rangle})$ and $\mathbf{z}=(z^{\langle 1\rangle},z^{\langle 2\rangle})$.*

Lemma 3 *For every Sobolev space $\mathcal{H}$ of functions on $\mathcal{X}$, there corresponds a unique reproducing kernel $\mathcal{K}$, which is non-negative definite. If $\mathcal{K}_0$ and $\mathcal{K}_1$ are both non-negative definite reproducing kernels for $\mathcal{H}_0$ and $\mathcal{H}_1$, and $\mathcal{H}_0\bigcap\mathcal{H}_1=\{0\}$, then $\mathcal{H}_0\oplus\mathcal{H}_1$ has a reproducing kernel $\mathcal{K}=\mathcal{K}_0+\mathcal{K}_1$.*

Lemmas 2 and 3 can be easily proved based on Theorems 2.3 to 2.6 in Gu (2013). Lemma 2 states that the reproducing kernel of the tensor product space is the product of the reproducing kernels. Lemma 3 states that the reproducing kernel of a tensor sum space is the sum of the reproducing kernels. Therefore, to construct $\mathcal{K}_{\mathbf{x}_{ij}}(\cdot,\cdot)$, we introduce the decomposition of tensor product space in the following part.

2.2. Decomposition of Tensor Product Space

For any $\eta_1\in\mathcal{H}^{\langle 1\rangle}$ and $\eta_2\in\mathcal{H}^{\langle 2\rangle}$, define the averaging operators $\mathcal{A}_1:\eta_1\to\int_0^1\eta_1(x)dx$ and $\mathcal{A}_2:\eta_2\to\frac{1}{2}\sum_{k=1}^2\eta_2(k)$ where $\eta_2(k)=e_k^T\eta_2$, e_k is the unit vector with the kth element one and all other elements zeros. We have $\mathcal{H}^{\langle 1\rangle}$ and $\mathcal{H}^{\langle 2\rangle}$ with the following tensor sum decomposition $\mathcal{H}_0^{\langle 1\rangle}\oplus\mathcal{H}_1^{\langle 1\rangle}$ and $\mathcal{H}_0^{\langle 2\rangle}\oplus\mathcal{H}_1^{\langle 2\rangle}$ respectively, where $\mathcal{H}_0^{\langle 1\rangle}=\{\mathcal{A}_1\eta_1\mid\eta_1\in\mathcal{H}^{\langle 1\rangle}\}$, $\mathcal{H}_0^{\langle 2\rangle}=\{\mathcal{A}_2\eta_2\mid\eta_2\in\mathcal{H}^{\langle 2\rangle}\}$, $\mathcal{H}_1^{\langle 1\rangle}=\{(\mathcal{I}-\mathcal{A}_1)\eta_1\mid\eta_1\in\mathcal{H}^{\langle 1\rangle}\}$, $\mathcal{H}_1^{\langle 2\rangle}=\{(\mathcal{I}-\mathcal{A}_2)\eta_2\mid\eta_2\in\mathbb{R}^2\}$, and $\mathcal{I}$ is the identity operator. Thus $\mathcal{H}$ has the following tensor sum decomposition

$$\mathcal{H}=(\mathcal{H}_0^{\langle 1\rangle}\otimes\mathcal{H}_0^{\langle 2\rangle})\oplus(\mathcal{H}_1^{\langle 1\rangle}\otimes\mathcal{H}_0^{\langle 2\rangle})\oplus(\mathcal{H}_0^{\langle 1\rangle}\otimes\mathcal{H}_1^{\langle 2\rangle})\oplus(\mathcal{H}_1^{\langle 1\rangle}\otimes\mathcal{H}_1^{\langle 2\rangle}),\tag{9}$$

and for any $f\in\mathcal{H}^{\langle 1\rangle}\otimes\mathcal{H}^{\langle 2\rangle}$, we have

$$f=f_{00}+f_{10}+f_{01}+f_{11},\tag{10}$$

where $f_{00}=\mathcal{A}_1\mathcal{A}_2f\in\mathcal{H}_0^{\langle 1\rangle}\otimes\mathcal{H}_0^{\langle 2\rangle}$, $f_{10}=(\mathcal{I}-\mathcal{A}_1)\mathcal{A}_2f\in\mathcal{H}_1^{\langle 1\rangle}\otimes\mathcal{H}_0^{\langle 2\rangle}$, $f_{01}=\mathcal{A}_1(\mathcal{I}-\mathcal{A}_2)f\in\mathcal{H}_0^{\langle 1\rangle}\otimes\mathcal{H}_1^{\langle 2\rangle}$ and $f_{11}=(\mathcal{I}-\mathcal{A}_1)(\mathcal{I}-\mathcal{A}_2)f\in\mathcal{H}_1^{\langle 1\rangle}\otimes\mathcal{H}_1^{\langle 2\rangle}$. Thus, any function $f\in\mathcal{H}$ can be decomposed uniquely as : f_{00} the interception, f_{10} and f_{01} the marginal effects and f_{11} the two-way interaction term.

Denote the reproducing kernels of $\mathcal{H}_0^{\langle 1\rangle},\mathcal{H}_0^{\langle 2\rangle},\mathcal{H}_1^{\langle 1\rangle},\mathcal{H}_1^{\langle 2\rangle}$ as $\mathcal{K}_0^{\langle 1\rangle}$, $\mathcal{K}_1^{\langle 1\rangle}$, $\mathcal{K}_0^{\langle 2\rangle}$, $\mathcal{K}_1^{\langle 2\rangle}$, respectively. Specifically, $\mathcal{K}_0^{\langle 1\rangle}(x^{\langle 1\rangle},z^{\langle 1\rangle})=1$ and $\mathcal{K}_1^{\langle 1\rangle}(x^{\langle 1\rangle},z^{\langle 1\rangle})$ is defined as $(-1)^{m-1}k_{2m}(z^{\langle 1\rangle}-x^{\langle 1\rangle})$ for the mth order homogeneous subspace where $k_r(\cdot)$ is the rth order scaled Bernoulli polynomials (Abramowitz and Stegun, 1964; Gu, 2013) and $\mathbf{1}_{(\cdot)}$ is the indicator function. $\mathcal{K}_0^{\langle 2\rangle}(x^{\langle 2\rangle},z^{\langle 2\rangle})=1/2$ and $\mathcal{K}_1^{\langle 2\rangle}(x^{\langle 2\rangle},z^{\langle 2\rangle})=\mathbf{1}_{(z^{\langle 2\rangle}=x^{\langle 2\rangle})}-1/2$ on $\mathcal{X}_2$. Let $\mathcal{H}_{\ell\ell'}=\mathcal{H}_\ell^{\langle 1\rangle}\otimes\mathcal{H}_{\ell'}^{\langle 2\rangle}$ with reproducing kernel $\mathcal{K}^{\ell\ell'}$, where

$$\mathcal{K}^{\ell\ell'}(\mathbf{x}_{ij},\mathbf{x}_{i'j'})=\mathcal{K}_\ell^{\langle 1\rangle}(x_i^{\langle 1\rangle},x_{i'}^{\langle 1\rangle})\mathcal{K}_{\ell'}^{\langle 2\rangle}(x_{j'}^{\langle 2\rangle},x_{j'}^{\langle 2\rangle}),$$

for $\ell,\ell'\in\{0,1\}$. The induced inner product of $\mathcal{H}_{\ell\ell'}$ is denoted as $\langle f_{\ell\ell'},g_{\ell\ell'}\rangle_{\ell\ell'}$, where $f_{\ell\ell'}$ and $g_{\ell\ell'}$ are projections of f and g on $\mathcal{H}_{\ell\ell'}$ respectively, $\ell,\ell'\in\{0,1\}$. Notice that the metrics induced by inner products $\langle f_{\ell\ell'},g_{\ell\ell'}\rangle_{\ell\ell'}$ are not necessarily of the same scale for different $\ell\ell'$. The inner product for $\mathcal{H}$ can be defined as

$$\langle f,g\rangle_{\mathcal{H}}=\sum_{\ell\ell'}\theta_{\ell\ell'}^{-1}\langle f_{\ell\ell'},g_{\ell\ell'}\rangle_{\ell\ell'},\tag{11}$$

where $\theta_{\ell\ell'}$s re-scale the metrics on different $\mathcal{H}_{\ell\ell'}$, $\langle\cdot,\cdot\rangle_{\ell\ell'}$ is the restricted norm of $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ on $\mathcal{H}_{\ell\ell'}$.

Based on Lemmas 2 and 3, we can easily show that the reproducing kernels associated with Equation (11) is $\mathcal{K}(\mathbf{x}_{ij},\mathbf{x}_{i'j'}) = \sum_{\ell,\ell'}\theta_{\ell\ell'}\mathcal{K}^{\ell\ell'}(\mathbf{x}_{ij},\mathbf{x}_{i'j'})$ with $\ell,\ell' = 0,1$. Thus, given the sampling points $\mathbf{x}_{ij} = (x_i^{\langle 1\rangle}, x_j^{\langle 2\rangle})$ for $i = 1,\cdots,n$ and $j = 1,\cdots,s$, the kernel function in $\mathcal{H}$ is a bivariate function depending on $\mathbf{x}_{ij}$, i.e.,

$$\begin{aligned}\mathcal{K}_{\mathbf{x}_{ij}}(x^{\langle 1\rangle},x^{\langle 2\rangle}) = \frac{\theta_{00}}{2} + \theta_{01}(\mathbf{1}_{(x^{\langle 2\rangle}=x_j^{\langle 2\rangle})} - \frac{1}{2}) + \theta_{10}\frac{1}{2}\mathcal{K}_1^{\langle 1\rangle}(x^{\langle 1\rangle},z^{\langle 1\rangle})\\ + \theta_{11}\frac{1}{2}(\mathbf{1}_{(x^{\langle 2\rangle}=x_j^{\langle 2\rangle})} - \frac{1}{2})\mathcal{K}_1^{\langle 1\rangle}(x^{\langle 1\rangle},z^{\langle 1\rangle}), \quad (12)\end{aligned}$$

and accordingly $f(x^{\langle 1\rangle},x^{\langle 2\rangle}) = \sum_{ij}\alpha_{ij}\mathcal{K}_{\mathbf{x}_{ij}}(x^{\langle 1\rangle},x^{\langle 2\rangle}) + \rho(x^{\langle 1\rangle},x^{\langle 2\rangle})$ by Lemma 1.

In the function decomposition in Equation (10), it is easy to verify that $f_{00} \in \mathcal{H}_{00} = \big\{g : g = \{(\theta_{00}-\theta_{01})/2\}\sum_{ij}\alpha_{ij}\big\}$. As f_{00} is a constant for any $x^{\langle 1\rangle}$ and $x^{\langle 2\rangle}$, it is analogous to the ground mean in classical ANOVA models. Similarly, we have $f_{01} \in \mathcal{H}_{01} = \{g : g = \theta_{01}\sum_{ij}\alpha_{ij}\mathbf{1}_{(x^{\langle 2\rangle}=x_j^{\langle 2\rangle})}\}$. Recall that $x_j^{\langle 2\rangle}$ can only be either 0 or 1, we can rewrite f_{01} as $\mathbf{1}_{(x^{\langle 2\rangle}=0)}\beta_0 + \mathbf{1}_{(x^{\langle 2\rangle}=1)}\beta_1$, where $\beta_0 = \sum_{j=1}^{s}(\sum_{i=1}^{n}\alpha_{ij})\mathbf{1}_{(x_j^{\langle 2\rangle}=0)}$ and $\beta_1 = \sum_{j=1}^{s}(\sum_{i=1}^{n}\alpha_{ij})\mathbf{1}_{(x_j^{\langle 2\rangle}=1)}$.

We remark that f_{00} and f_{01} are all in a finite-dimensional space. The space $\mathcal{H}_{10}$ (where f_{10} lies in) spanned by the third term in the right hand side of Equation (12) is, however, an infinite-dimensional space, because we have uncountable $x \in \mathcal{X}_1$. The function can be expressed as a linear combination of the observed reproducing kernels plus a residual that quantifies the unobserved reproducing kernels, i.e., $\mathcal{H}_{10} = \{g : g = \frac{1}{2}\sum_{i=1}^{n}(\theta_{10}\sum_{j=1}^{s}\alpha_{ij})\mathcal{K}_1^{\langle 1\rangle}(x^{\langle 1\rangle},z^{\langle 1\rangle}) + \rho_2\}$. Notice that function in this space only changes as we change $x^{\langle 1\rangle}$. Thus, the third term in right hand side of (12) can be used to quantify the effect of the continuous variable such as the temporal effect. The forth term in the right hand side of Equation (12) varies for both continuous variable and the case-control indicator, thus it is the term that can catch different functional patterns between the case and control. Similarly, the space spanned by the last addend is also an infinite-dimensional space because we still have an infinite number of unobserved kernel functions in addition to the $n \times s$ observed kernel functions. Thus, we have $f_{11} \in \mathcal{H}_{11} = \{g : g = \frac{1}{2}\theta_{11}\sum_{ij}\alpha_{ij}(\mathbf{1}_{(x_j^{\langle 2\rangle}=x^{\langle 2\rangle})} - \frac{1}{2})\mathcal{K}_1^{\langle 1\rangle}(x^{\langle 1\rangle},z^{\langle 1\rangle}) + \rho_{12}\}$. Clearly, to test if two functions are parallel to each other, we only need to test if $f_{11} = 0$.

2.3. Penalized Least Squares

Here we introduce the penalized least square estimate of $f \in \mathcal{H}$, and the interaction term f_{11} in Equation (10). Given the sampling points $\mathbf{x}_{ij} = (x_i^{\langle 1\rangle}, x_j^{\langle 2\rangle})$ for $i = 1,\ldots,n$ and $j = 1,\ldots,s$, consider the model space

$$\mathcal{H}_{model} = \{g : g = \sum_{i=1}^{n}\sum_{j=1}^{s}\alpha_{ij}\mathcal{K}_{\mathbf{x}_{ij}}(x^{\langle 1\rangle},x^{\langle 2\rangle})\},$$

a closed linear subspace of $\mathcal{H}$. α_{ij}s are the regression coefficients, and the bivariate residual function $\rho(\cdot,\cdot)$ in Lemma 1 is in $\mathcal{H}_{residual} = \mathcal{H} \ominus \mathcal{H}_{model}$. Notice that $\rho(x_i^{\langle 1\rangle}, x_j^{\langle 2\rangle}) = \langle\mathcal{K}_{\mathbf{x}_{ij}}(x^{\langle 1\rangle},x^{\langle 2\rangle}),\rho\rangle = 0$ because of the orthogonality constraint between $\mathcal{H}_{model}$ and $\mathcal{H}_{residual}$. Then, f can be estimated by minimizing the penalized least squares functional as follows:

$$\frac{1}{ns}\sum_{i=1}^{n}\sum_{j=1}^{s}(Y_{ij}-\sum_{i'j'}\alpha_{i'j'}\mathcal{K}_{\mathbf{x}_{i'j'}}(x_i^{\langle 1\rangle},x_j^{\langle 2\rangle}))^2+\lambda J(f_{10}+f_{11}), \tag{13}$$

where the quadratic functional $J(f)=J(f_{10}+f_{11})=\|f_{10}+f_{11}\|_{\mathcal{H}}^2$ quantifies the roughness of f_{10} and f_{11}, the smoothing parameter λ controls the trade-off between the goodness-of-fit and the roughness of f_{10} and f_{11}. Recall ρ and $\mathcal{K}_{\mathbf{x}_{i'j'}}(\cdot,\cdot)$ are orthogonal to each other. Plugging Equation (8) into $J(f)$, we have

$$J(f)=\langle\sum_{i'j'}\alpha_{i'j'}(\theta_{10}\mathcal{K}^{10}_{\mathbf{x}_{i'j'}}+\theta_{11}\mathcal{K}^{11}_{\mathbf{x}_{i'j'}}),\sum_{i'j'}\alpha_{i'j'}(\theta_{10}\mathcal{K}^{10}_{\mathbf{x}_{i'j'}}+\theta_{11}\mathcal{K}^{11}_{\mathbf{x}_{i'j'}})\rangle_{\mathcal{H}}+\langle\rho,\rho\rangle_{\mathcal{H}}.$$

Further notice that $\langle\mathcal{K}^{\ell\ell'}_{\mathbf{x}_{ij}},\mathcal{K}^{\ell\ell'}_{\mathbf{x}_{i'j'}}\rangle=\mathcal{K}^{\ell\ell'}_{\mathbf{x}_{ij}}(x_{i'}^{\langle 1\rangle},x_{j'}^{\langle 2\rangle})$ by the reproducible property of reproducing kernels (Gu, 2013). Thus, substituting $\mathcal{K}$ and $\mathcal{K}^{\ell\ell'}$ by (12) and f in $J(f)$ by Equation (8), Equation (13) can be rewritten as

$$||\mathbf{y}-nsK\boldsymbol{\alpha}||_2^2+ns\lambda\boldsymbol{\alpha}^TQ\boldsymbol{\alpha}+ns\lambda\langle\rho,\rho\rangle_{\mathcal{H}}, \tag{14}$$

where $\mathbf{y}=(Y_{11},Y_{21},\ldots,Y_{ns})^T$, K is the $ns\times ns$ matrix with $(i+n(j-1),i'+n(j'-1))$th entry $\frac{1}{ns}\mathcal{K}_{\mathbf{x}_{ij}}(x_{i'}^{\langle 1\rangle},x_{j'}^{\langle 2\rangle})$, Q is the $ns\times ns$ matrix with $(i+n(j-1),i'+n(j'-1))$th entry $\frac{1}{ns}(\theta^{10}\mathcal{K}^{10}_{\mathbf{x}_{ij}}(x_{i'}^{\langle 1\rangle},x_{j'}^{\langle 2\rangle})+\theta^{11}\mathcal{K}^{11}_{\mathbf{x}_{ij}}(x_{i'}^{\langle 1\rangle},x_{j'}^{\langle 2\rangle}))$ and $\boldsymbol{\alpha}=(\alpha_{11},\alpha_{21},\ldots,\alpha_{ns})^T$. Similar to Chapter A3 in Gu (2013), we set the rescale parameter θ_{10} and θ_{11} to make $\theta_{10}\mathcal{K}^{10}$ and $\theta_{11}\mathcal{K}^{11}$ contribute equally in penalty term of Equation (14) (see Appendix A.1 for details) and set θ_{00} and θ_{01} as one since $\mathcal{H}_{00}$ and $\mathcal{H}_{01}$ are simply one-dimensional Euclidean space. Since ρ does not rely on α, the optimizer of $\boldsymbol{\alpha}$ in minimizing Equation (14) is equivalent to minimizing

$$\widehat{\boldsymbol{\alpha}}=\underset{\boldsymbol{\alpha}\in\mathbb{R}^{ns}}{\arg\min}\,||\mathbf{y}-nsK\boldsymbol{\alpha}||_2^2+ns\lambda\boldsymbol{\alpha}^TQ\boldsymbol{\alpha}. \tag{15}$$

The penalized least square estimate of f is then $\widehat{f}(x_i^{\langle 1\rangle},x_j^{\langle 2\rangle})=\sum_{i,j}^{n,s}\widehat{\alpha}_{i,j}\mathcal{K}_{\mathbf{x}_{i,j}}(x_i^{\langle 1\rangle},x_j^{\langle 2\rangle})$.

As n goes to infinity, we have countable number of kernels and $f(x^{\langle 1\rangle},x^{\langle 2\rangle})$ that the minimizer of Equation (13) resides in an infinite dimensional space spanned by a countable number of kernels, i.e.,

$$\mathcal{H}^{\infty}_{model}=\{g:g(x^{\langle 1\rangle},x^{\langle 2\rangle})=\sum_{ij}^{\infty}\alpha_{ij}\mathcal{K}_{\mathbf{x}_{ij}}(x^{\langle 1\rangle},x^{\langle 2\rangle})\}.$$

The nonparallel effect f_{11} also resides in a subspace that is spanned by a countable number of kernels. We denote the subspace by

$$\mathcal{H}^{\infty}_{11}=\{f_{11}:f_{11}(x^{\langle 1\rangle},x^{\langle 2\rangle})=\sum_{ij}^{\infty}\alpha_{ij}\frac{(-1)^{m-1}}{2}(\mathbf{1}_{(x_j^{\langle 2\rangle}=x^{\langle 2\rangle})}-\frac{1}{2})k_{2m}(x_i^{\langle 1\rangle}-x^{\langle 1\rangle})\}.$$

Here, we did not normalize f_{11} by the constant scale parameter θ_{11} for the simplicity of description. The penalized least square estimate of $f_{11}\in\mathcal{H}^{\infty}_{11}$ is

$$\hat{f}_{11}(x^{\langle 1\rangle},x^{\langle 2\rangle})=\sum_{i,j}^{n,s}\widehat{\alpha}_{ij}\frac{(-1)^{m-1}}{2}(\mathbf{1}_{(x_j^{\langle 2\rangle}=x^{\langle 2\rangle})}-\frac{1}{2})k_{2m}(x_i^{\langle 1\rangle}-x^{\langle 1\rangle}). \tag{16}$$

With a little abuse of notation, we use $\hat{\boldsymbol{f}}_{11}$ to denote the vector version evaluation of $\hat{f}_{11}$ on ns data points from now on. Plugging in $\widehat{\boldsymbol{\alpha}}$ to (16), we have an explicit expression of $\hat{\boldsymbol{f}}_{11}$ as

$$\hat{\boldsymbol{f}}_{11} = K_{11}M^{-1}(I_{ns} - S(S^TM^{-1}S)^{-1}S^TM^{-1})\mathbf{y}, \tag{17}$$

where I_{ns} is the ns dimensional identity matrix, S, M and K_{11} are reparametrization of the kernel matrices with explicit forms provided in Appendix A.1–"Notation Clarification".

In Section 4, we will construct a Wald type test statistics based on $\hat{f}_{11}$ for the parallelism test $H_0 : f_{11} = 0$, and derive its null asymptotic distribution. Before that, we first establish the minimax principle of the parallelism test for general testing rules in the following Section 3.

3. Minimax Principle of the Nonparametric Parallelism Test

Consider the test problem as follows

$$H_0 : f_{11} = 0 \quad vs \quad H_1 : ||f_{11}||_2 > 0. \tag{18}$$

Given a decision rule ϕ_n for the testing problem (18), $\phi_n = 0$ if H_0 is preferred and 1 otherwise. Then the zero-one loss function is

$$\text{Loss}(\phi_n) = \begin{cases} \phi_n & \text{if } H_0 \text{ is true,} \\ 1 - \phi_n & \text{if } H_1 \text{ is true.} \end{cases} \tag{19}$$

The minimax principle requires ϕ_n to minimize the maximum possible risk, i.e.,

$$\min_{\phi_n} \max_{\mathcal{H}} \mathbb{E}[\text{Loss}(\phi_n)] = \min_{\phi_n}[\max_{H_0} \mathbb{E}(\phi_n | H_0 \text{ is true}) + \max_{H_1} \mathbb{E}(1 - \phi_n | H_1 \text{ is true})]. \tag{20}$$

Notice $\mathbb{E}(\phi_n|H_0 \text{ is true})$ is the probability of making a type I error and $\mathbb{E}(1 - \phi_n|H_1 \text{ is true})$ is the probability of making a type II error. Intuitively, we choose ϕ_n to minimize the maximum possible type I error and type II error. Notice that if H_0 and H_1 are contiguous, we cannot ensure that Equation (20) can be controlled, because there may lie some f_{11} on the boundary of H_0 and H_1 for which strikes the balance between acceptance and rejection of the null hypothesis, and an appropriate decision cannot be made. Thus, instead of H_1, we consider a slightly different alternative hypothesis (4) and partition the parameter space into three sets: $H_0 + H_1^* + I$, of which I designates the indifference zone $0 < ||f_{11}||_2 < d_n$. Because d_n clearly separates H_0 from H_1^*, it is referred to as the distinguishable rate (a.k.a the separation rate) (Ingster and Suslina, 2012; Giné and Nickl, 2015). Let

$$\text{pseudo.risk}(\phi_n, d_n) = \sup_{H_0} \mathbb{E}(\phi_n|H_0 \text{ is true}) + \sup_{H_1^*} \mathbb{E}(1 - \phi_n|H_1^* \text{ is true}). \tag{21}$$

Then pseudo.risk(ϕ_n, d_n) converges to the risk function $\mathbb{E}[\text{Loss}(\phi_n)]$ as d_n goes to zero.

Compared to the risk function, the pseudo.risk is not only a function of a decision rule ϕ_n but also a function of the distinguishable rate d_n. When ϕ_n is given, we have $\sup_{H_1^*} \mathbb{E}(1 - \phi_n|H_1^* \text{ is true}) \leq \sup_{H_1} \mathbb{E}(1 - \phi_n|H_1 \text{ is true})$ because H_1^* is a subset of H_1. Thus, finding the largest pseudo.risk on H_1^* for a given ϕ_n is equivalent to finding the smallest d_n with a tolerable pseudo.risk. In another word, finding the maximum possible pseudo.risk over the parameter space can be considered as finding the smallest boundary of H_1^* such that an appropriate decision ϕ_n can be made and the risk can be controlled. Meanwhile, for an adequately large given d_n, we can always find a decision rule such that the pseudo.risk can reach its minimum value. Let $\phi_n^\dagger(d_n) = \arg\min_{\phi_n}$ pseudo.risk(ϕ_n, d_n). Then, if d_n can reach the smallest value $d_n^\dagger$, the corresponding $\phi_n^\dagger(d_n^\dagger)$ is the minimax decision. Thus, the essential step to find the minimax decision of pseudo.risk(ϕ_n, d_n) is to find $d_n^\dagger$ such that

$$d_n^\dagger = \arg\min_{d_n} \phi_n^\dagger(d_n). \tag{22}$$

Because $d_n^\dagger$ is an estimate of the distinguishable rate to obtain the minimax test, it is referred to as the minimax distinguishable rate. Clearly, the corresponding decision rule $\phi_n^\dagger$ is the minimax decision rule.

We first introduce a geometric interpretation of the testing problem (18). Geometrically, we can treat $\mathcal{E} = \{f \in \mathcal{H} : ||f||_{\mathcal{H}} < 1/2\}$ as an ellipse with eigenvalues in Equation (7) as axis lengths as shown in Figure 2. For any $f \in \mathcal{E}$, the projection of f on $\{f : f \in \mathcal{E}_{11} := \mathcal{H}_{11} \cap \mathcal{E}\}$ is f_{11}. The magnitude of nonparallelism can be qualified by $||f_{11}||_2$. The distinguishable rate d_n is the radius of the sphere centered at $f_{11} = 0$ in $\mathcal{H}_{11}$.

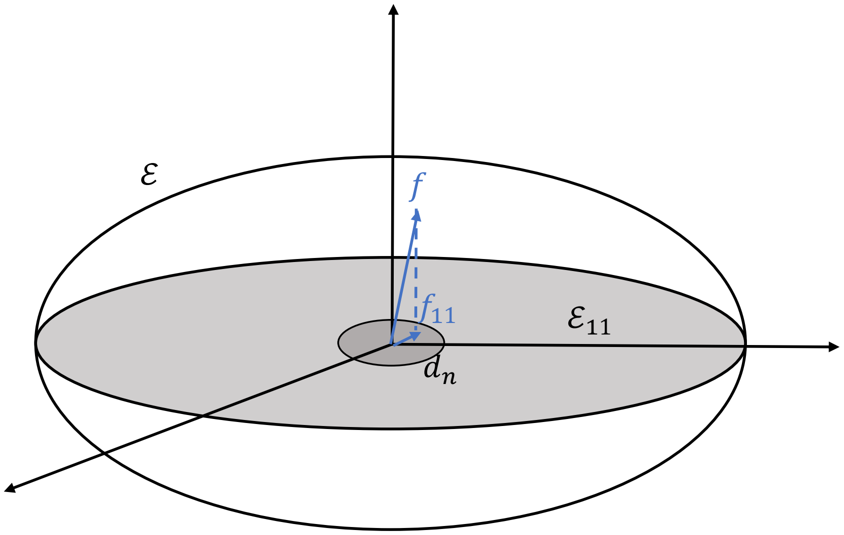


Figure 2: Geometric interpretation of the distinguishable rate of the parallelism test.

Intuitively, the testing will be harder when the projection of f on $\mathcal{E}_{11}$ is closer to the origin $f_{11} = 0$. We use the Bernstern width in Pinkus (2012) to characterize the testing difficulty. Let S_{k+1} be the set of all $(k+1)$-dimensional subspaces for any $k \geq 1$. For a compact set C, the Bernstein k-width is defined as

$$b_{k,2}(C) := \arg\max_{r \geq 0}\{\mathbb{B}_2^{k+1}(r) \subset C \cap S \text{ for some subspace } S \in S_{k+1}\}, \tag{23}$$

where $\mathbb{B}_2^{k+1}(r)$ is a $(k+1)$-dimensional l_2-ball with radius r centered at $f_{11} = 0$ in $\mathcal{E}_{11}$. The Bernstein width characterizes the largest ball that can be inserted into a $(k+1)$-dimensional subspace in $\mathcal{E}_{11}$. Based on the Bernstein width, we give an upper bound of the testing radius, i.e., for any f

projected in the ball with radius less than this upper bound, the minimum pseudo.risk is larger than $1/2$.

Lemma 4 *For any $f \in \mathcal{H}$, we have*

$$\inf_{\phi_n} \textit{pseudo.risk}(\phi_n, d_n) \geq 1/2$$

for all

$$d_n \leq r_B := \sup\{\delta \,|\, \delta \leq \frac{1}{2\sqrt{n}} \sigma(k_B(\delta))^{1/4}\}$$

where $k_B(\delta) := \arg\max_k\{b^2_{k-1,2}(\mathcal{H}_{11}) \geq \delta^2\}$ is the Bernstein lower critical dimension, and r_B is called the Bernstein lower critical radius.

Lemma 4 shows that when d_n is less than r_B, there has no test can distinguish the alternative hypothesis from the null. In order to achieve a non-trivial power, we need d_n to be larger than the Bernstein lower critical radius r_B, which is determined by the Bernstein lower critical dimension $k_B(\delta)$. In the next lemma, we provide the lower bound for $k_B(\delta)$.

Lemma 5 *Let $\{\rho_i\}_{i=1}^{\infty}$ be eigenvalues of $\mathcal{H}_{11}$. We have*

$$k_B(\delta) > \arg\max_k\{\sqrt{\rho_k} \geq \delta\} \tag{24}$$

Plugging in the lower bound of $k_B(\delta)$ derived in Lemma 5 to Lemma 4, we calculate a lower bound for r_B based on the decay rate of eigenvalues. r_B is served as a minimax lower bound for the distinguishable rate. The following theorem summarizes the minimax distinguishable rate for the testing problem (18).

Theorem 6 ***(Minimax lower bound for distinguishable rate)*** *In the nonparametric model (1) with SSANOVA (2). Suppose $f \in \mathcal{H}$, where $\mathcal{H} = \mathcal{H}^{\langle 1 \rangle} \otimes \mathcal{H}^{\langle 2 \rangle}$ with $\mathcal{H}^{\langle 1 \rangle}$ as the mth order Sobolev space*[1]*, and $\mathcal{H}^{\langle 2 \rangle}$ as a two-dimensional Euclidean space. The minimax distinguishable rate for testing hypotheses (18) is achieved at $d_n^{\dagger} \gtrsim n^{-2m/(4m+1)}$.*

Theorem 6 provides a general guidance for justifying a local minimax test, i.e., there is no test can distinguish the alternative from null if $d_n \lesssim n^{-2m/(4m+1)}$. The proof of Theorem 6 is presented in the Appendix. Essentially, for any test ϕ_n that is defined by a family of type I error $\alpha = \mathbb{E}(\phi_n)$ and by the supremum of the type II error $\delta = \sup_{H_1^*} \mathbb{E}(1 - \phi_n | H_1^* \text{ is true})$, we need ϕ_n converges to zero faster than d_n to ensure the distinguishability of the null distribution. We further remark that the minimax rate for nonparametric estimation is $n^{-m/(2m+1)}$ (Yang et al. (2017)) which is higher than the minimax distinguishable rate $n^{-2m/(4m+1)}$. In the next section, we will introduce a Wald type test for the hypothesis testing (18) with the separation rate d_n achieves the lower bound $n^{-2m/(4m+1)}$ indicating our proposed test is minimax optimal.

[1] The mth order Sobolev space is defined as $\mathcal{H}^{\langle 1 \rangle} = \{\eta_1 \in L_2[0,1] \mid \eta_1^{(k)}$ is absolutely continuous for $k = 0, 1, \ldots, m-1\}$.

4. Wald Type Parallism Test

In this section, we propose a Wald type test statistics based on the penalized least squares estimate of f_{11}, and derive the asymptotic distribution of the test statistics. We further prove an upper bound of the distinguishable rate of the Wald type test which matches the minimax lower bound established in Theorem 6.

4.1. Wald Type Test and Asymptotic Distribution

The nonparallel effect of the curves between the case group and the control group is measured by the magnitude of $||f_{11}||_2^2$. The nonparallel test in Equation (3) is equivalent to

$$H_0 : f \in \mathcal{H}^{\infty}_{model} \ominus \mathcal{H}^{\infty}_{11} \qquad vs \qquad H_1 : f \in \mathcal{H}^{\infty}_{model}$$

or equivalently, $H_1 : f_{11} \in \mathcal{H}^{\infty}_{11}$. First, notice that the null hypothesis in Equation (18) is a composite hypothesis as the null hypothesis defines a class of functions in $\mathcal{H}^{\infty}_{model} \ominus \mathcal{H}^{\infty}_{11}$. Second, H_0 defines an infinite dimensional parameter spaces as $n \to \infty$, the assumptions of Neyman-Pearson Lemma cannot be satisfied. Thus the uniformly most powerful test may not exist in general. To overcome the difficulty, we propose a Wald-type test

$$T_{n,\lambda} = \frac{1}{ns}||\hat{\boldsymbol{f}}_{11}||_2^2 \tag{25}$$

and show its minimax optimality.

Since Y_{ij} follows Equation (1) with f satisfying the SSANOVA decomposition in Equation (2), we can replace each element in vector $\mathbf{y}$ by $f_{00}(x_i^{\langle 1\rangle}, x_j^{\langle 2\rangle}) + f_{10}(x_i^{\langle 1\rangle}, x_j^{\langle 2\rangle}) + f_{01}(x_i^{\langle 1\rangle}, x_j^{\langle 2\rangle}) + f_{11}(x_i^{\langle 1\rangle}, x_j^{\langle 2\rangle}) + \epsilon_{ij}$. Then plug in the expression of $\hat{\boldsymbol{f}}_{11}$ in Equation (17) to $T_{n,\lambda}$, we have

$$T_{n,\lambda} = \frac{1}{ns}||K_{11}M^{-1}(I_n - S(S^TM^{-1}S)^{-1}S^TM^{-1})(\boldsymbol{f}_{00} + \boldsymbol{f}_{10} + \boldsymbol{f}_{01} + \boldsymbol{f}_{11} + \boldsymbol{\epsilon})||_2^2,$$

where $\boldsymbol{f}_{00}$, $\boldsymbol{f}_{10}$, $\boldsymbol{f}_{01}$ and $\boldsymbol{f}_{11}$ are ns dimensional vectors with the ijth entry $f_{00}(x_i^{\langle 1\rangle}, x_j^{\langle 2\rangle})$, $f_{10}(x_i^{\langle 1\rangle}, x_j^{\langle 2\rangle})$, $f_{01}(x_i^{\langle 1\rangle}, x_j^{\langle 2\rangle})$ and $f_{11}(x_i^{\langle 1\rangle}, x_j^{\langle 2\rangle})$ respectively, and $\boldsymbol{\epsilon}$ is the ns dimensional stochastic error that follows a normal distribution with mean 0 and variance $\sigma^2 I_{ns}$. Because f_{00}, f_{10} and f_{01} are in the space that is orthogonal to the space spanned by K_{11}, and $f_{11} = 0$ under the null hypothesis, $T_{n,\lambda}$ can be further simplified as

$$T_{n,\lambda} = \frac{1}{ns}||K_{11}M^{-1}(I_{ns} - S(S^TM^{-1}S)^{-1}S^TM^{-1})\boldsymbol{\epsilon}||_2^2. \tag{26}$$

A detailed discussion of this simplification will be provided in Lemma 12 in Appendix.

Next, we develop the null limiting distribution of $T_{n,\lambda}$ as n goes to infinity. In the derivation, we only require the number of subjects s to be finite. This requirement is desired in real applications since the number of subjects in an experiment is usually limited. For example, due to the high sequencing cost, there are usually only tens of sample sequenced in the DNA methylation studies.

We consider the following two designs.

Quasi-Uniform Design : $x_1^{\langle 1\rangle}, x_2^{\langle 1\rangle}, \ldots, x_n^{\langle 1\rangle} \overset{iid}{\sim} \omega(x^{\langle 1\rangle})$ where ω is the marginal density of $x^{\langle 1\rangle}$. For any $x^{\langle 1\rangle} \in [0, 1]$, there exist two constants $c_1, c_2 > 0$ such that $c_1 \leq \omega(x^{\langle 1\rangle}) \leq c_2$ (Eggermont and LaRiccia, 2001).

Uniform Design: $x_1^{\langle 1\rangle}, x_2^{\langle 1\rangle}, \ldots, x_n^{\langle 1\rangle}$ are evenly spaced on $[0, 1]$.

The above two designs are commonly used in scientific investigations. For example, in fMRI experiments, the sampling points on the time domain are usually measured with equal-time intervals. Thus, they are assumed to follow uniform design. On the other hand, the DNA methylation sites are randomly scattered on DNA sequence. Therefore, they are assumed to follow a quasi-uniform design.

Theorem 7 *For both the uniform design and the quasi-uniform design, if the smoothing parameter $\lambda = \mathcal{O}(n^{c-1})$ for any fixed $c \in (0, 1)$, we have*

$$\frac{T_{n,\lambda} - \mu_{n,\lambda}}{\sigma_{n,\lambda}} \xrightarrow{d} N(0, 1) \qquad \text{as } n \to \infty,$$

where $\mu_{n,\lambda} = \sigma^2 \operatorname{Tr}(\Delta)/(ns)$ and $\sigma^2_{n,\lambda} = 2\sigma^4 \operatorname{Tr}(\Delta^2)/(ns)^2$ with $\Delta = M^{-1}K_{11}^2 M^{-1}$.

In practice, we estimate the variance σ^2 via $\widehat{\sigma}^2$ defined as

$$\widehat{\sigma}^2 = \frac{\mathbf{y}^\top (I - A(\lambda))^2 \mathbf{y}}{\operatorname{Tr}(I - A(\lambda))},$$

where $A(\lambda) = K(nsK^2 + \lambda Q)^{-1}\mathbf{y}$, and $(I - A(\lambda))\mathbf{y}$ is the residual $\mathbf{y} - \widehat{\boldsymbol{f}}$ based on the objective function in equation (15). The consistency of the variance estimate $\widehat{\sigma}^2$ is established in Theorem 3.4 in Gu (2013).

The proof of Theorem 7 is provided in Appendix and sketched below. Notice that $T_{n,\lambda} = T_1 + T_2 - 2T_3$, where

$$\begin{aligned} T_1 =& \frac{1}{ns}\boldsymbol{\epsilon}^T M^{-1} K_{11}^2 M^{-1}\boldsymbol{\epsilon}, \\ T_2 =& \frac{1}{ns}||K_{11}M^{-1}S(S^T M^{-1}S)^{-1}S^T M^{-1}\boldsymbol{\epsilon}||_2^2, \\ T_3 =& \frac{1}{ns}\boldsymbol{\epsilon}^T M^{-1}S(S^T M^{-1}S)^{-1}S^T M^{-1}K_{11}^2 M^{-1}\boldsymbol{\epsilon}. \end{aligned} \tag{27}$$

We show that T_2 and T_3 are higher order small perturbation terms compared to T_1. Thus, the null distribution of $T_{n,\lambda}$ and the distribution of T_1 are asymptotically equivalent. We only need to focus on the distribution of the quadratic form $T_1 = \frac{1}{ns}\boldsymbol{\epsilon}^T \Delta \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon}$ having a mean zero normal distribution. To prove the normality of T_1, we show that the log-characteristic function of the standardized T_1 is asymptotically $-\sigma^2 t^2/2$, provided that $\operatorname{Tr}(\Delta^2)$ diverges as $\lambda \to 0$. Lemma 15 shows that $\operatorname{Tr}(\Delta^2) \succeq \hat{\tau}_\lambda$, where $\hat{\tau}_\lambda = \max\{i \mid \hat{\mu}_i \geq \lambda\}$ as the effective dimension (Bartlett et al., 2005; Liu et al., 2019) with $\hat{\mu}_1 \geq \cdots \geq \hat{\mu}_n$ the empirical eigenvalues of kernel matrix $K_1^{\langle 1\rangle}$ which is the kernel matrix of $\mathcal{H}_1^{\langle 1\rangle}$ with (i, i')th entry as $\frac{1}{n}\mathcal{K}_1^{\langle 1\rangle}(x_i^{\langle 1\rangle}, x_{i'}^{\langle 1\rangle})$. We further show in Lemma 13 and 14 that $\hat{\tau}_\lambda$ is of the same order as its population counterpart τ_λ defined as $\tau_\lambda = \max\{i \mid \mu_i \geq \lambda\}$, under both the quasi-uniform design and the uniform design, where $\mu_1 \geq \cdots \geq 0$ are a sequence of ordered eigenvalues satisfying $\mathcal{K}_1^{\langle 1\rangle}(x, x') = \sum_{i=1}^\infty \mu_i \phi_i(x)\phi_i(x')$. Since μ_i has a polynomial decay rate i^{-2m} (Gu, 2013), we have $\operatorname{Tr}(\Delta^2) \succeq \hat{\tau}_\lambda \asymp \tau_\lambda \asymp \lambda^{-1/(2m)}$ diverges as $\lambda \to 0$. Consequently, the testing consistency in Theorem 7 holds.

Theorem 7 characterizes the distribution of the test statistic $T_{n,\lambda}$ for $f \in \mathcal{H}^{\infty}_{model} \ominus \mathcal{H}^{\infty}_{11}$. The distribution turns out to be fairly simple and easy to calculate as the test statistic does not depend on any unknown nuisance functions such as f_{00}, f_{10} and f_{01}. Critical value can be easily found based on the known null distribution $N(\mu_{n,\lambda}, \sigma^2_{n,\lambda})$. Consequently, one can make a statistical decision by comparing $T_{n,\lambda}$ with the critical value. This nuisance-parameter free property is referred to as the "Wilks phenomenon" in statistics literature (Fan et al., 2001; Fan and Zhang, 2004).

4.2. Upper Bound of the Distingushiable Rate

Given type I error α, we show that our Wald-type testing rule $\phi_{n,\lambda} = \mathbf{1}_{(|T_{n,\lambda}-\mu_{n,\lambda}| \geq z_{\alpha/2}\sigma_{n,\lambda})}$ achieves the local minimax distinguishable rate. Without loss of generality, we assume $\|f\|_{\mathcal{H}} \leq 1$.

Theorem 8 *Let the minimum distinguishable rate of the test $\phi_{n,\lambda}$ be $d_n(\phi_{n,\lambda})$. Suppose $\lambda = \mathcal{O}(n^{c-1})$ for any fixed $c \in (0,1)$. Then for any $\delta > 0$, there exist positive constants C_δ and N_δ such that, when $n \geq N_\delta$, the tolerable pseudo.risk$(\phi_{n,\lambda}, d_n) = \alpha + \delta$, with $d_n(\phi_{n,\lambda}) := C_\delta\sqrt{\lambda + \sigma_{n,\lambda}}$.*

Theorem 8 shows that for a controlled type I error, $T_{n,\lambda}$ can achieve arbitrary small type II error provided that the local alternative is separated from the null by at least an amount of $d_n(\phi_{n,\lambda})$. The proof of Theorem 8 is collected in Appendix.

Note that $d_n^2(\phi_{n,\lambda})$ consists of two components: $\sigma_{n,\lambda}$ representing the standard variation of the test statistic $T_{n,\lambda}$, and λ representing the squared bias of $\hat{f}_{1,2}$ (see the proof of Lemma S.1 in the Supplementary). Through approximating $\sigma_{n,\lambda}$ by the Rademacher complexity (Bartlett et al., 2005; Liu et al., 2019), we show that $\sigma_{n,\lambda} \asymp \sqrt{\tau_\lambda}/n$, which is a decreasing function of λ. Hence, the minimum distinguishable rate for $\phi_{n,\lambda}$ is achieved by the trade-off between the bias of $\hat{f}_{1,2}$ and the standard derivation of $T_{n,\lambda}$, i.e., choosing appropriate λ such that $\lambda \asymp \sigma_{n,\lambda}$. Next, we prove that our proposed Wald-type test is minimax under two special design conditions: the quasi-uniform design and the uniform design in the next two corollaries.

Corollary 9 ***[Quasi-Uniform Design]*** *Let $\lambda \asymp n^{-4m/(4m+1)}$ and suppose $x^{\langle 1 \rangle}$ follows a quasi-uniform design. We have*

$$P(d_n(\phi_{n,\lambda}) \asymp n^{-2m/(4m+1)}) \geq 1 - 4\exp(-n^{1/(2m+1)}).$$

Corollary 10 ***[Uniform Design]*** *Let $\lambda \asymp n^{-4m/(4m+1)}$, and suppose $x^{\langle 1 \rangle}$ follows a uniform design, we have*

$$d_n(\phi_{n,\lambda}) \asymp n^{-2m/(4m+1)} \quad a.s.$$

Corollaries 9 and 10 suggest that if $\lambda \asymp n^{-4m/(4m+1)}$, our Wald-type test $\phi_{n,\lambda}$ can achieve the minimax distinguishable rate $d_n^{\dagger} \asymp n^{-2m/(4m+1)}$. Thus, we demonstrate that our proposed Wald type test is minimax optimal. We remark that Corollary 9 still holds when extending $\mathcal{H}^{\langle 1 \rangle}$ as a standard Sobolev space.

4.3. The Choice of Regularization Parameter

Different from the classical "bias-variance" tradeoff in optimal nonparametric estimation, Theorem 8 states that the optimal nonparametric testing for Equation (3) can be achieved by another type of tradeoff between the squared bias of the estimator and the standard deviation of the test statistic.

Such intrinsic difference further leads to different orders of optimal regularization parameters: as shown in Corollary 9, 10, the optimal λ is chosen as the order of $n^{-\frac{4m}{4m+1}}$; while as the order of $n^{-\frac{2m}{2m+1}}$ for optimal estimation (Gu, 2013).

In practice, cross validation method is often used as a tuning procedure for nonparametric estimation based on penalized loss functions (Golub et al., 1979). Raskutti et al. (2014) proposed another data-dependent algorithmic regularization technique, that is, choosing an early stopping rule for an iterative algorithm to avoid over-fitting in nonparametric estimation. Both of the above approaches are optimal for estimation but suboptimal for testing. There has few theoretically justified tuning procedure for obtaining optimal testing in nonparametric inference. One related work we are aware currently is Liu and Cheng (2018), under which they developed a data-dependent early stopping regularization rule from an algorithmic perspective for testing $f = 0$ in nonparametric regression model $Y = f(X) + \epsilon$. The total step size determined via the early stopping rule in gradient descent algorithm plays the same role with $1/\lambda$ in the penalized regularization, to avoid over-fitting. However, a data-adaptive choice of the regularization parameter λ is still lacking for nonparametric inference in Equation (3) under the penalization regularization.

We propose a data-adaptive method to choose λ with testing optimality guarantee based on Theorem 8. In practice, we can choose the optimal smoothing parameter λ^* satisfying

$$\lambda^* = \min\big\{\lambda \mid \lambda < \sigma_{n,\lambda}\big\}, \tag{28}$$

where $\sigma_{n,\lambda}$ can be explicitly calculated based on the observed data by the expression defined in Theorem 7, i.e., $\sigma_{n,\lambda}^2 = 2\sigma^4 Tr(\Delta^2)/(ns)^2$, with $\Delta = M^{-1}K_{11}^2 M^{-1}$.

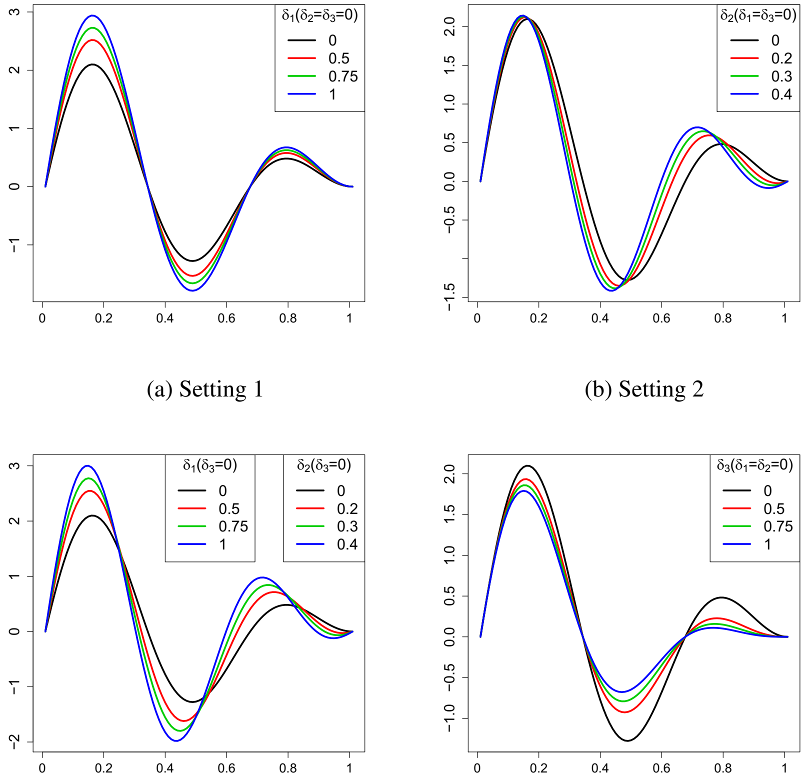
The above criterion in Equation (28) in choosing λ is a data-dependent rule that produces a minimax-optimal nonparametric testing method. Based on the Rademacher complexity, $\sigma_{n,\lambda} \asymp \frac{\sigma^2}{ns}\sqrt{\sum_{i=1}^n \min\{1, \widehat{\mu}_i/\lambda\}}$. That is, the rule in Equation (28) depends on the eigenvalues of the kernel matrix, especially the first few leading eigenvalues. There are many efficient methods to compute the top eigenvalues fast (Drineas and Mahoney, 2005; Ma and Belkin, 2017). As a future work, we can also introduce the randomly projected kernel methods to accelerate the computing time.

5. Simulation Study

To assess the performance of our proposed test, we carried out extensive analyses on simulated data sets. We compared our approach with F-test (SSF) (Ma et al., 2009), parallelism trend test (PTT) (Degras et al., 2011) and a random permutation test with 500 permutations. In the three methods, permutation test can be used as a benchmark because it can closely approximate null distribution when the number of permutations is adequate. However, the permutation test is computationally intensive, especially for calculating the Kullback-Leibler distance under the null and alternative hypothesis for SSANOVA model (Gu, 2004).

5.1. Empirical Power Analysis

We illustrate the empirical power performance of our proposed test through four well-designed examples. In all four examples, we generated 100 to 1000 observations with an increment of 100 observations in each simulation for both case and control groups in Equation (1), where $x_i^{\langle 1\rangle} \overset{iid}{\sim} U(0,1)$ and $\epsilon_{ij} \overset{iid}{\sim} N(0,1)$. Each example was repeated 500 times for power and other comparisons. To make the simulation more close to the reality, we considered two types of nonparallel patterns between $f(x^{\langle 1\rangle},1)$ and $f(x^{\langle 1\rangle},0)$: magnitude and frequency. These two kinds of nonparallel patterns are often observed in real applications. For example, the hypermethylated DNA regions, i.e., regions with low methylation levels, are related to transcriptional silencing which plays an important role in cancer development; the frequency differences are often related to different brain functions between the neurodisease and control groups in fMRI studies. In the first four examples, we consider the



(a) Setting 1

(b) Setting 2

(c) Setting 3

(d) Setting 4

Figure 3: Plotted here are functions of the control group (solid line) and case group (dashed, dotted and dot-dash lines) with four types of nonparallel patterns: magnitude differences only (Setting 1), frequency differences only (Setting 2), both magnitude and frequency differences (Setting 3), and magnitude dynamic differences (Setting 4).

following function in Equation (1),

$$f(x^{\langle 1\rangle},x^{\langle 2\rangle})=\begin{cases}2.5\sin(3\pi x^{\langle 1\rangle})(1-x^{\langle 1\rangle}) & \text{if } x^{\langle 2\rangle}=0, \text{i.e., control}\\ (2.5+\delta_1)\sin((3+\delta_2)\pi x^{\langle 1\rangle})(1-x^{\langle 1\rangle})^{(1+\delta_3)} & \text{if } x^{\langle 2\rangle}=1, \text{i.e., case}\end{cases} \tag{29}$$

where δ_1, δ_2 and δ_3 control the magnitude of nonparallelism between the null hypothesis and the alternative hypothesis in Equation (18). In general, varying δ_1, δ_2 and δ_3 give rise to different distinguishable rates d_ns. The larger the δ_1, δ_2 and δ_3 are, the larger the d_n is. To illustrate how the testing power is affected by different δ's, as shown in Figure 3, we considered the following four settings. **Setting 1:**

Case and control have constant magnitude differences ($\delta_1 = 0.50, 0.75, 1.00$ and $\delta_2, \delta_3 = 0.00$); **Setting 2:** Case and control have frequency differences ($\delta_2 = 0.20, 0.30, 0.40$ and $\delta_1, \delta_3 = 0.00$); **Setting 3:** Both magnitude and frequency are different ($\delta_1, \delta_2 = (0.50, 0.20), (0.75, 0.30), (1.00, 0.40)$ and $\delta_3 = 0.00$); **Setting 4:** Case and control have non-constant magnitude differences ($\delta_1, \delta_2 = 0.00$ and $\delta_3 = 0.50, 0.75, 1.00$). The corresponding functions $f(x^{\langle 1 \rangle}, 0)$ and $f(x^{\langle 1 \rangle}, 1)$ are shown in Figure 3.

		Sample Size									
		100	200	300	400	500	600	700	800	900	1000
$\delta_1 = 0.50$	Proposed	0.17	0.33	0.49	0.59	0.69	0.75	0.86	0.91	0.92	0.96
	Permutation	0.19	0.38	0.53	0.60	0.62	0.76	0.80	0.88	0.94	0.97
	SSF	0.02	0.09	0.11	0.16	0.26	0.28	0.36	0.54	0.58	0.72
	PTT	0.05	0.06	0.05	0.1	0.11	0.1	0.14	0.21	0.11	0.17
$\delta_1 = 0.75$	Proposed	0.37	0.67	0.90	0.93	0.97	0.98	1.00	1.00	1.00	1.00
	Permutation	0.38	0.66	0.81	0.90	0.96	0.99	0.99	1.00	1.00	1.00
	SSF	0.04	0.21	0.37	0.50	0.81	0.86	0.91	0.96	0.97	0.98
	PTT	0.09	0.14	0.15	0.33	0.38	0.36	0.47	0.55	0.44	0.54
$\delta_1 = 1.00$	Proposed	0.61	0.92	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Permutation	0.57	0.89	0.95	0.99	1.00	1.00	1.00	1.00	1.00	1.00
	SSF	0.14	0.48	0.79	0.90	0.97	0.99	1.00	1.00	1.00	1.00
	PTT	0.08	0.23	0.42	0.43	0.54	0.62	0.77	0.77	0.79	0.85

Table 1: Table lists the empirical power of our proposed test and permutation test for Setting 1 with $\delta_1 = 0.50, 0.75, 1.00$, $\delta_2 = \delta_3 = 0.00$ and sample size ranging from 100 to 1000.

		Sample Size									
		100	200	300	400	500	600	700	800	900	1000
$\delta_2 = 0.20$	Proposed	0.28	0.46	0.66	0.79	0.86	0.95	0.95	0.97	0.98	0.99
	Permutation	0.27	0.43	0.59	0.74	0.86	0.94	0.94	0.98	1.00	1.00
	SSF	0.02	0.05	0.21	0.32	0.48	0.62	0.79	0.84	0.88	0.95
	PTT	0.04	0.03	0.04	0.08	0.11	0.14	0.12	0.09	0.16	0.26
$\delta_2 = 0.30$	Proposed	0.40	0.63	0.81	0.94	0.96	0.99	0.99	1.00	1.00	1.00
	Permutation	0.36	0.64	0.79	0.89	0.97	0.98	0.99	1.00	1.00	1.00
	SSF	0.03	0.13	0.35	0.52	0.72	0.85	0.91	0.97	0.99	1.00
	PTT	0.03	0.08	0.09	0.15	0.31	0.23	0.28	0.4	0.35	0.4
$\delta_2 = 0.40$	Proposed	0.73	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Permutation	0.78	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	SSF	0.24	0.74	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00
	PTT	0.11	0.16	0.18	0.38	0.39	0.52	0.56	0.59	0.81	0.89

Table 2: Table lists the empirical power of our proposed test and permutation test for Setting 2 with $\delta_2 = 0.20, 0.30, 0.40$, $\delta_1 = \delta_3 = 0.00$ and sample size ranging from 100 to 1000.

The empirical powers of our proposed Wald-type test, permutation test, SSF test and PTT test are summarized in Tables 1-2 for Settings 1-2. For Setting 1, as shown in Table 1, the empirical

		Sample Size									
		100	200	300	400	500	600	700	800	900	1000
$\delta_1 = 0.50$	Proposed	0.35	0.51	0.74	0.86	0.91	0.95	0.97	0.98	1.00	1.00
$\delta_2 = 0.20$	SSF	0.04	0.15	0.29	0.41	0.57	0.72	0.85	0.89	0.91	0.96
	PTT	0.03	0.07	0.07	0.08	0.08	0.06	0.15	0.19	0.21	0.2
$\delta_1 = 0.75$	Proposed	0.42	0.70	0.86	0.96	0.99	1.00	1.00	1.00	1.00	1.00
$\delta_2 = 0.30$	SSF	0.05	0.26	0.46	0.64	0.79	0.93	0.94	0.95	1.00	1.00
	PTT	0.04	0.07	0.11	0.15	0.19	0.23	0.31	0.29	0.43	0.46
$\delta_1 = 1.00$	Proposed	0.72	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
$\delta_2 = 0.40$	SSF	0.25	0.72	0.97	0.99	1.00	1.00	1.00	1.00	1.00	1.00
	PTT	0.11	0.19	0.22	0.32	0.52	0.5	0.64	0.61	0.73	0.69

Table 3: Table lists the empirical power of our proposed test and permutation test for Setting 3 with $\delta_1, \delta_2 = (0.50, 0.20), (0.75, 0.30), (1.00, 0.40)$, $\delta_3 = 0$ and sample size ranging from 100 to 1000.

		Sample Size									
		100	200	300	400	500	600	700	800	900	1000
$\delta_3 = 0.50$	Proposed	0.15	0.33	0.47	0.58	0.66	0.75	0.83	0.88	0.89	0.94
	SSF	0.01	0.04	0.07	0.16	0.18	0.28	0.35	0.47	0.57	0.64
	PTT	0.06	0.03	0.08	0.09	0.09	0.14	0.07	0.13	0.08	0.13
$\delta_3 = 0.75$	Proposed	0.35	0.61	0.73	0.84	0.92	0.95	0.99	1.00	1.00	1.00
	SSF	0.03	0.12	0.18	0.34	0.56	0.70	0.83	0.86	0.96	0.96
	PTT	0.01	0.07	0.06	0.07	0.09	0.12	0.13	0.18	0.18	0.24
$\delta_3 = 1.00$	Proposed	0.42	0.70	0.85	0.95	0.99	0.99	1.00	1.00	1.00	1.00
	SSF	0.07	0.20	0.52	0.76	0.82	0.92	0.98	0.98	1.00	1.00
	PTT	0.09	0.04	0.08	0.10	0.18	0.18	0.21	0.24	0.25	0.28

Table 4: Table lists the empirical power of our proposed test and permutation test for Setting 4 with $\delta_3 = 0.50, 0.75, 1.00$, $\delta_1 = \delta_2 = 0.00$ and sample size ranging from 100 to 1000.

power of our test increases rapidly as sample size increases, and approaches to 1 even for the smallest magnitude ($\delta_1 = 0.50$). The empirical powers of the proposed test are comparable with that of the permutation test. In contrast, the empirical powers of SSF and PTT increase slower than our proposed test. For the weak signal scenario, i.e., $\delta_1 = 0.50$, the proposed test has significantly gain of power under different sample sizes. For the strong signal scenario, i.e, $\delta = 1.00$, our proposed test is significantly more powerful than SSF and PTT when sample size is less than 500. For Setting 2, as shown in Table 2, the empirical power of our proposed test converges to 1 as the sample size increases for all three cases with $\delta_2 = 0.20$, 0.30 and 0.40. In contrast, the empirical power of SSF and PTT converges to 1 slower than the proposed test.

For Settings 3 and 4, we only included the empirical results for our proposed test and SSF test due to the extremely high computational cost of the permutation test. As shown in Table 6, it takes more than 150 hours to complete the permutation test for one setting. For Setting 3, we simulated the signal with differences in both scale and frequency across case and control groups. The empirical powers of the simulation with different distinguishable parameters are listed in Table 3. The empirical powers of our proposed test and SSF increase for all the three cases with $\delta_1, \delta_2 = 0.20, 0.30, 0.40$. The empirical power of PTT also increases, but with a much slower pattern. When the sample size is small and signal strength is weak, our proposed test has significant gain of power compared to the SSF and PTT test. For Setting 4, there is a nonlinear magnitude difference along the $x^{\langle 1\rangle}$ between the two groups. As shown in Table 4, the empirical power of SSF test converges to one slower than the proposed test and is lower than 0.65 for the least distinguishable case.

5.2. Empirical Size Analysis

To examine the approximation of significance levels, we generated data from a new setting **Setting 5**. We kept the function form of control group the same as Equation (5.4) and only added a parallel shift over the control function as the function of the case group, i.e., the model does not include the nonparallel patterns. In particular,

$$f(x^{\langle 1\rangle}, x^{\langle 2\rangle}) = 2.5\sin(3\pi x^{\langle 1\rangle})(1 - x^{\langle 1\rangle}) + \delta_4 I_{\{x^{\langle 2\rangle}=1\}},$$

where δ_4 was set to be 0, 0.5 and 1 to characterize different level parallel difference in the two groups. We generated data from Equation (1) with function f specified in Setting 5. The rest of parameters were set the same as before.

		Sample Size									
		100	200	300	400	500	600	700	800	900	1000
$\delta_4 = 0.00$	Proposed	0.04	0.07	0.06	0.06	0.05	0.06	0.06	0.07	0.06	0.05
	Permutation	0.04	0.08	0.05	0.08	0.06	0.05	0.06	0.04	0.07	0.06
	SSF	0.06	0.11	0.03	0.08	0.08	0.03	0.07	0.09	0.07	0.03
	PTT	0.03	0.05	0.02	0.02	0.03	0.02	0.12	0.09	0.08	0.06
$\delta_4 = 0.50$	Proposed	0.06	0.05	0.05	0.06	0.06	0.05	0.06	0.04	0.05	0.06
	Permutation	0.07	0.04	0.05	0.06	0.08	0.09	0.04	0.03	0.05	0.04
	SSF	0.06	0.05	0.07	0.06	0.07	0.08	0.07	0.04	0.04	0.07
	PTT	0.02	0.02	0.03	0.03	0.07	0.04	0.06	0.07	0.06	0.04
$\delta_4 = 1.00$	Proposed	0.07	0.06	0.07	0.06	0.05	0.05	0.06	0.06	0.06	0.05
	Permutation	0.04	0.06	0.03	0.05	0.05	0.04	0.03	0.02	0.04	0.04
	SSF	0.07	0.07	0.08	0.06	0.04	0.07	0.06	0.09	0.07	0.04
	PTT	0.03	0.04	0.03	0.05	0.05	0.04	0.06	0.05	0.05	0.08

Table 5: Table lists the empirical sizes of the proposed test, permutation test, SSF, and PTT for $\delta_4 = 0.00, 0.50, 1.00$ and sample size ranging from 100 to 1000.

Table 5 lists the empirical sizes of our proposed test, permutation test, SSF test, and PTT under Setting 5. We varied δ_4 from 0.00 to 1.00 to model different magnitudes of the main effect. The empirical size of our proposed test approaches to 0.05 as the sample size increases for different values of δ_4. The empirical size of SSF test is fluctuating from 0.03 to 0.1. The inaccurate size of the SSF test may be attributed to the fact that the degrees of freedom of the SSF test is very roughly approximated by the rounding value of the trace of the smoothing matrix. The empirical size of PTT test is fluctuating from 0.02 to 0.12.

5.3. Computation Time

As shown in Tables 1 and 2, our purposed test achieves the power similar to the permutation test. Next, we compared the computation time of our proposed test and permutation test for 500 replicated samples. We conducted the comparison on a computer workstation with core Intel i7 8700k CPU and 32 Gb RAM. In Table 6, we reported the computational time in Setting 1 with $\delta_1 = 0.5$ and sample size ranging from 100 to 1000. As shown in Table 6, our proposed test is consistently faster than the permutation test. Our proposed test is nearly 263×faster than the permutation test when the sample size is 1000. Note that the computational time is more than 42 hours when the sample size is 1000 for running 500 test. In practice, the huge computational cost limits the application of the permutation test in many large scale studies involving large sample size and multiple tests.

	Sample Size									
	100	200	300	400	500	600	700	800	900	1000
Proposed	0.01	0.03	0.04	0.06	0.07	0.09	0.10	0.12	0.14	0.16
Permutation	3.22	6.14	9.29	13.29	17.93	22.26	26.74	31.26	36.57	42.23

Table 6: Table lists computational time (in hour) of running the simulation with 500 replications for our proposed test and the permutation test.

5.4. Simulation Studies with Correlated Noise

We established **Setting 6** to evaluate the performance of the proposed test when the noises are correlated. In this example, we generated 100 to 1000 observations with an increment of 100 observations in each simulation for both case and control groups in Equation (1). We considered $x_i^{\langle 1 \rangle}, i = 1, \ldots, n$ are evenly distributed in $[0, 1]$. We generated two correlated noise vector $(\epsilon_{11}, \ldots, \epsilon_{n1})$ and $(\epsilon_{12}, \ldots, \epsilon_{n2})$ i.i.d. from $N(0, \Sigma)$ where Σ is autoregressive, i.e., each of its element $\sigma_{ii'} = \rho^{|i-i'|}$ with $\rho = 0.5$. We generated the signal $Y_{ij} = f(x_i^{\langle 1 \rangle}, x_j^{\langle 2 \rangle}) + \epsilon_{ij}$ where f is defined in Equation (5.4) with $\delta_1 = 0.00, 0.50, 0.75, 1.00$ and $\delta_2, \delta_3 = 0.00$, that is,

$$f(x^{\langle 1 \rangle}, x^{\langle 2 \rangle}) = \begin{cases} 2.5 \sin(3\pi x^{\langle 1 \rangle})(1 - x^{\langle 1 \rangle}) & \text{if } x^{\langle 2 \rangle} = 0, \\ (2.5 + \delta_1) \sin(3\pi x^{\langle 1 \rangle})(1 - x^{\langle 1 \rangle}) & \text{if } x^{\langle 2 \rangle} = 1. \end{cases}$$

We set the significance level as 0.05 and repeated 500 times for evaluating the empirical size and power.

As shown in Table 7, when $\delta_1 = 0.00$, the size of our proposed method concentrates around $0.05 - 0.07$, while the sizes of SSF and PTT are fluctuating from 0.02 to 0.16. When $\delta_1 > 0.00$, compared with SSF and PTT, the power of our proposed method has the highest performance, and approaches to 1 as δ_1 increases.

5.5. Simulation Studies with Non-smooth Cases

We evaluate the robustness of the proposed method when the smoothness assumption is invalid. We established **Setting 7** to test the performance of the proposed test for the cases with non-smooth trends. In this setting, we generated 100 to 1000 observations with an increment of 100 observations in each simulation for both case and control groups in model (1). We considered $x_i^{\langle 1 \rangle}, i = 1, \ldots, n$,

		Sample Size									
		100	200	300	400	500	600	700	800	900	1000
$\delta_1 = 0.00$	Proposed	0.08	0.04	0.06	0.06	0.06	0.08	0.10	0.07	0.06	0.07
	SSF	0.08	0.06	0.06	0.09	0.10	0.05	0.09	0.14	0.08	0.06
	PTT	0.02	0.1	0.04	0.08	0.11	0.05	0.16	0.12	0.13	0.06
$\delta_1 = 0.50$	Proposed	0.21	0.33	0.48	0.57	0.73	0.73	0.82	0.91	0.94	0.96
	SSF	0.01	0.05	0.10	0.17	0.29	0.32	0.46	0.48	0.63	0.72
	PTT	0.13	0.22	0.35	0.50	0.51	0.53	0.72	0.73	0.78	0.86
$\delta_1 = 0.75$	Proposed	0.66	0.89	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	SSF	0.13	0.48	0.74	0.89	0.98	1.00	0.99	1.00	1.00	1.00
	PTT	0.16	0.32	0.41	0.43	0.66	0.67	0.73	0.85	0.85	0.89
$\delta_1 = 1.00$	Proposed	0.93	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	SSF	0.47	0.93	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	PTT	0.16	0.41	0.55	0.62	0.69	0.85	0.86	0.90	0.90	0.95

Table 7: Table lists the empirical size ($\delta_1 = 0$) and power ($\delta_1 = 0.50, 0.75, 1.00$) of our proposed test, SSF and PTT for Setting 6 with $\delta_2 = \delta_3 = 0.00$ and sample size ranging from 100 to 1000.

are evenly distributed in $[0, 1]$ and $\epsilon_{ij} \overset{iid}{\sim} N(0, 1)$. We generated the signal $Y_{ij} = f(x_i^{\langle 1\rangle}, x_j^{\langle 2\rangle}) + \epsilon_{ij}$ with f defined as

$$f(x^{\langle 1\rangle}, x^{\langle 2\rangle}) = 2.5\sin(2\pi x^{\langle 1\rangle}) I_{\{x^{\langle 1\rangle}\in(0,0.5)\}} + (1 + \delta_5 I_{\{x^{\langle 2\rangle}=1\}})(x-1) I_{\{x^{\langle 1\rangle}\in[0.5,0)\}}$$

which is shown in Figure 4. This curve is non-differentiable at $x^{\langle 1\rangle} = 0.5$ which is a change point from nonlinear to linear trend. We set the significance level as 0.05 and repeated 500 times to evaluate the empirical size and power.

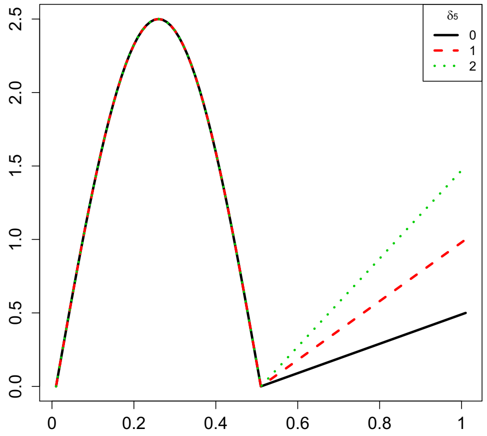


Figure 4: Solid line with $\delta_5 = 0$: function of the control group; dashed and dotted lines with $\delta_5 = 1, 2$: the case group for Setting 7.

As shown in Table 8, when $\delta_5 = 0.00$, the empirical size of our proposed method concentrates around 0.05. The empirical size of our proposed method is slightly inflated compared with SSF and PTT. When $\delta_5 = 1, 2$, compared with SSF and PTT, the power of our proposed method has the highest performance, and approaches to 1 as n increases.

		Sample Size									
		100	200	300	400	500	600	700	800	900	1000
$\delta_5 = 0.00$	Proposed	0.08	0.04	0.06	0.06	0.06	0.08	0.10	0.07	0.06	0.07
	SSF	0.01	0.01	0.00	0.00	0.00	0.01	0.01	0.01	0.00	0.00
	PTT	0.02	0.06	0.04	0.08	0.03	0.05	0.03	0.06	0.03	0.03
$\delta_5 = 1.00$	Proposed	0.21	0.33	0.48	0.57	0.73	0.73	0.82	0.91	0.94	0.96
	SSF	0.03	0.02	0.06	0.07	0.09	0.15	0.23	0.29	0.34	0.37
	PTT	0.01	0.05	0.01	0.02	0.04	0.02	0.04	0.04	0.08	0.06
$\delta_5 = 2.00$	Proposed	0.66	0.89	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	SSF	0.07	0.24	0.46	0.63	0.76	0.87	0.91	0.98	0.97	0.99
	PTT	0.02	0.04	0.02	0.09	0.03	0.06	0.07	0.10	0.06	0.06

Table 8: Table lists the empirical size ($\delta_5 = 0$) and power ($\delta_5 = 1.00, 2.00$) of our proposed test, SSF and PTT for Setting 7.

6. Real Data Examples

We apply the technique to analyze two real data sets: DNA methylation in chronic lymphocytic leukemia and neuroimaging of Alzheimer's Disease using fMRI.

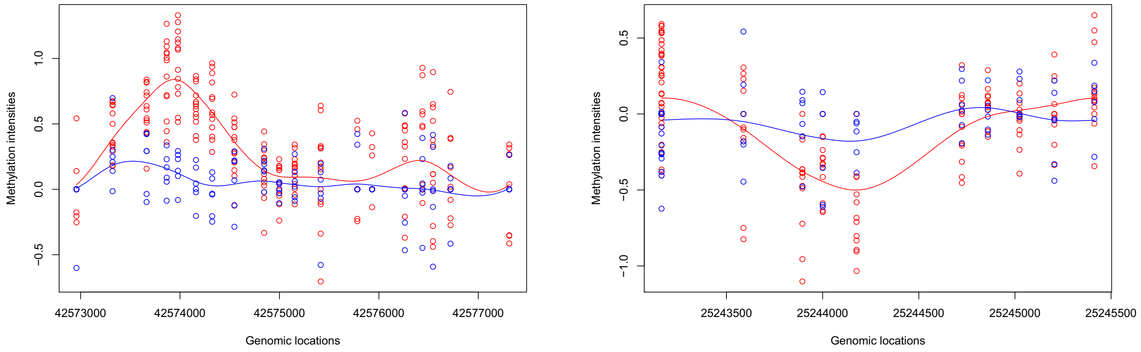
6.1. DNA Methylation in Chronic Lymphocytic Leukemia

Recently, Filarsky et al. (2016) reported a DNA methylation study for chronic lymphocytic leukemia (CLL) patients. In the study, the DNA samples were extracted from CD19+ cells from 12 CLL patients and B cells from 6 normal subjects. The DNA methylation is profiled by the whole-genome tiling array technique. The goal is to identify differentially methylated regions (DMRs), i.e., the genome regions that have significantly different methylation levels, between CLL patients and normal subjects.

To achieve this goal, we compiled the DNA methylation intensities within the -3.8 to $+1.8$ kb of transcription start sites (TSS) for each gene. We used the M-value suggested by Irizarry et al. (2008) as methylation level at each site and as our response variable. In particular, the data consists of $(Y_{ij}, x_i^{\langle 1\rangle}, x_j^{\langle 2\rangle})$, where Y_{ij} is the methylation level at the ith genome location $x_i^{\langle 1\rangle}$ of the jth subject in group $x_j^{\langle 2\rangle}$, which equals to 1 if the jth subject is in the case group and equals to 0 if the jth subject is in the control group. We fit the model in Equation (1) with SSANOVA decomposition in Equation (2) to the data.

We applied the proposed hypothesis testing on 10383 regions. Through controlling FDR < 0.01 using Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995), we selected 613 DMRs. We conducted gene ontology analysis on the 613 genes corresponding 613 identified DMRs using the GSEA (Subramanian et al., 2005). Among these genes, 79 genes participate the lipid metabolic process, which plays an important role in the development of CLL (Pallasch et al., 2008). This biological process contributes to apoptosis resistance in CLL cells. Furthermore, 78 and 61 genes participate the immune related biological processes: “Immune system process” and “Regulation of immune system process” respectively. The observation indicates that the aberrant DNA methylation has the potential impact on the immune system.

Our Wald-type test, even after FDR control, yields p-values that are as small as 10^{-9}. Consequently, it is very difficult to compare our test with the permutation test with only hundreds or thousands of permutations. Thus, we only compared our proposed test with permutation test (based on 500 permutation) for regions with p-values larger than 0.05 the averaged difference between our test and permutation test is 0.012.



(a) MTA3

(b) DNMT3A

Figure 5: The promoter regions of two genes, (a) MTA3 and (b) DNMT3A. The horizontal axis is the genomic location and the y axis is the M-value representing the methylation levels. The red and blue lines are the fitted curves for the case and control groups respectively.

We highlighted two DMRs with significant nonparallel patterns in Figure 5. The focal hypermethylation at genome locations 42574000 and 42576500 are observed on the promoter region of gene MTA3. It was reported in (Bilban et al., 2006) that MTA3 signaling pathway is a potential bio-marker for CLL and shows significantly altered gene expression. Our test also identified that the methylation levels between CLL patients and normal subjects, of MTA3 gene have significant difference, which has potential prognostic value. In the promoter region of DNMT3, we observed significant hypomethylation at genome location 25244500. DNMT3 is a family of DNA methyltransferases that could methylate hemimethylated and unmethylated CpG sites at the same rate (Okano et al., 1998). Since the global hypomethylation is observed, the aberrant methylation levels of this DNA methylatransferase may have influence on this global trend.

6.2. Neuroimaging of Alzheimer’s Disease using fMRI

Alzheimer’s disease (AD) is one of the most commonly known neurology disease characterized with neurodegeneration and cognitive decline (Rombouts et al., 2005; Wang et al., 2006). Despite the prevalence of AD, there are no cure or preventive methods available due to the lack of a complete understanding of the mechanisms that contribute to AD pathophysiology. Discovering aberrant neural network of AD will fundamentally advance the scientific understanding of this disease.

In this study, we analyzed the data that was collected by Alzheimer's Disease Neuroimaging Initiative (ADNI) [2], in which the resting-state fMRI signals of 60 normal/early-mild-cognitive-impairment subjects (control group) and 50 AD/late-mild-cognitive-impairment subjects (AD group) were collected from $256 \times 256 \times 170$ voxels for 140 consecutive time points with equal time intervals of 30ms. The fMRI signals for each subject were preprocessed using fMRI Expert Analysis Tool (FEAT) (Smith et al., 2004) for skull-stripping, motion correction, slice timing correction, temporal filtering, spatial smoothing and registration to standard space (MNI152_T1_2mm model) so that signals from all subjects can be considered as from the same engineered brain template. Sixty-nine brain-region-of-interests (ROI) that are defined by Harvard-Oxford-Atlas (http://fsl.fmrib.ox .ac.uk/fsl/fslwiki/Atlases) was extracted by automatic regional labeling approach using the refined fMRI data. For each ROI, we consider model (1) with SSANOVA decomposition in Equation (2), where Y_{ij} records the average blood-oxygen-level (Huettel et al., 2004) of the brain region for subject j measured at the $x_i^{\langle 1 \rangle}$ time point. As the blood-oxygen-level can accurately quantify the corresponding brain activity, we can detect abnormal AD related brain activity. Testing problem in Equation (18) is equivalent to testing whether the brain activities of a given ROI have different temporal patterns in case and control groups.

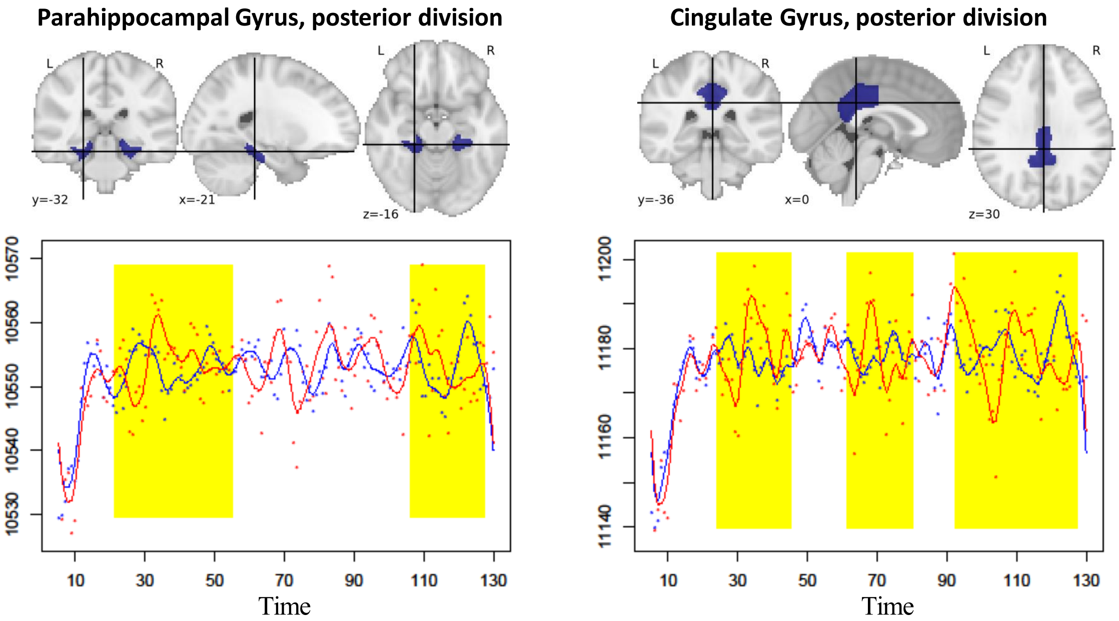


Figure 6: Plotted here are blood-oxygen-levels of *parahippocampal gyrus* (left) and *cingulate gyrus* (right) for control group (blue) and AD group (red) observed at 140 time points. Physical locations of either ROIs on frontal, axial and lateral sides are illustrated on the top of each panel.

[2] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at `http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf`

Seven cortical regions *parahippocampal gyrus*, *cingulate gyrus*, *inferior temporal gyrus*, *postcentral gyrus*, *juxtapositional lobule cortex*, *precuneous cortex*, *central opercular cortex* and one sub-cortical region *right thalamus* with significantly different temporal patterns were identified using our test with the false discovery rate controlled at 5% using Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). Among the eight ROIs, *parahippocampal gyrus* and *cingulate gyrus* have been shown clinically to be risk factors for AD. As demonstrated in Echávarri et al. (2011) and Kesslak et al. (1991), *parahippocampal gyrus* of AD patients have significant atrophy. Meanwhile, *cingulate gyrus* was also found to be AD related (Scheff et al., 2015) due to its extensive connectivity with multiple different cortical areas, especially areas involved with learning and memory. In Figure 6, we plotted frontal, axial, and lateral views and corresponding temporal patterns of *parahippocampal gyrus* and *cingulate gyrus*. The temporal regions with significant difference between AD/late-mild-cognitive-impairment subjects (red line) and normal/early-mild-cognitive-impairment subjects (blue line) are highlighted. As clearly demonstrated in lower left panel of Figure 6, the first highlighted area of *parahippocampal gyrus* has a significant reversed pattern between case group and control group. The second highlighted area shows the reduced levels for the AD group. For *cingulate gyrus*, the highlighted regions in the right panel of Figure 6 show clearly larger magnitude for the AD groups. This difference was also observed via fMRI in a visual encoding memory task (Rami et al., 2012). Both of the two experiments suggest that the difference may change the memory function.

7. Discussion

The hypothesis testing in SSANOVA is a very challenge problem. In this paper, we develop a Wald-type test for testing the significance of the nonparallelism in a two-way SSANOVA model. The optimality of the proposed test is justified by the minimax distinguishable rate. The extensive empirical studies suggest that the proposed test has a superior performance over existing methods. Although we only discuss the test of the significance of the nonparallelism in a two-way SSANOVA model, the test on a higher order SSANOVA model can be developed parallel to our framework.

Acknowledgments

PM was funded in part by NSF DMS-1440037, 1438957, 1925066 and NIH 1R01GM122080-01. WZ was funded in part by NSF DMS-1440038, 1903226, 1925066 and NIH 1R01GM113242-01. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics.The

Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Appendix A. Proof of Main Results

In this section, we present main proofs of the theorems and lemmas in the main text.

A.1. Notation Clarification

We rewrite (16) as

$$\hat{\boldsymbol{f}}_{11} = K_{11}M^{-1}(I_{ns} - S(S^TM^{-1}S)^{-1}S^TM^{-1})\mathbf{y},$$

where

$$S = \begin{bmatrix} I_n \otimes \mathbf{1}_{w_0} & 0 \\ 0 & I_n \otimes \mathbf{1}_{w_1} \end{bmatrix} \begin{bmatrix} \mathbf{1}_n & \mathbf{1}_n \\ \mathbf{1}_n & 0 \end{bmatrix}, K_{11} = \frac{1}{2}\begin{bmatrix} K_1^{\langle 1\rangle} & -K_1^{\langle 1\rangle} \\ -K_1^{\langle 1\rangle} & K_1^{\langle 1\rangle} \end{bmatrix}$$

and

$$M = \begin{bmatrix} I_n \otimes \mathbf{1}_{w_0} & 0 \\ 0 & I_n \otimes \mathbf{1}_{w_1} \end{bmatrix} \left(\frac{\theta_{10}}{2}\begin{bmatrix} K_1^{\langle 1\rangle} & K_1^{\langle 1\rangle} \\ K_1^{\langle 1\rangle} & K_1^{\langle 1\rangle} \end{bmatrix} + \frac{\theta_{11}}{2}\begin{bmatrix} K_1^{\langle 1\rangle} & -K_1^{\langle 1\rangle} \\ -K_1^{\langle 1\rangle} & K_1^{\langle 1\rangle} \end{bmatrix} + \lambda I_{2n} \right)$$
$$\begin{bmatrix} I_n \otimes \mathbf{1}_{w_0}^T & 0 \\ 0 & I_n \otimes \mathbf{1}_{w_1}^T \end{bmatrix},$$

$K_1^{\langle 1\rangle}$ is the kernel matrix of $\mathcal{H}_1^{\langle 1\rangle}$ with (i,i')th entry as $\frac{1}{n}\mathcal{K}_1^{\langle 1\rangle}(x_i^{\langle 1\rangle}, x_{i'}^{\langle 1\rangle})$, w_0 is the number of subjects in control group, w_1 is the number of subjects in case group, and $\otimes$ denotes the Kronecker product. Based on Chapter A.3 in Gu (2013), we set $\theta_{10}^{-1} \propto \mathrm{Tr}(K_{10})$ and $\theta_{11}^{-1} \propto \mathrm{Tr}(K_{11})$ with $\theta_{10} + \theta_{11} = 1$.

In the following theoretical derivation, we only focus on the case with $s = 2$, i.e. $w_0 = w_1 = 1$. If we have $s > 2$ subjects, the proof can be easily generalized to this situation by replacing Equation (15) by the penalized weighted least squares; see Section 3.2.4 in Gu (2013).

A.2. Proofs for Section 3

A.2.1. Preliminary

We identify a sequence model that is equivalent to our nonparametric model (1) with SSANOVA decomposition in Equation (2). Let $\{\rho_i, \phi_i\}_{i=1}^{\infty}$ be pairs of eigenvalue and eigenfunction in $\mathcal{H}^{\langle 1\rangle}$ and $\{\nu_j, \psi_j\}_{j=1}^{2}$ be pairs of eigenvalue and eigenfunction in $\mathcal{H}^{\langle 2\rangle}$. In the tensor product space $\mathcal{H} = \mathcal{H}^{\langle 1\rangle} \otimes \mathcal{H}^{\langle 2\rangle}$, as shown in Lin (2000), eigenvalues and eigenfunctions are $\{\mu_i\nu_j, \phi_i\psi_j\}_{i=1,\dots,\infty,\, j=1,2}$. Model (1) is equivalent to a sequence model

$$z_{ij} = \theta_{ij} + \omega_{ij}, \tag{30}$$

where $\theta_{ij} = \frac{1}{2}\sum_{x^{\langle 2\rangle}=0}^{1}\int_{\mathcal{X}_1} f(x_i^{\langle 1\rangle}, x_j^{\langle 2\rangle})\phi_i(x^{\langle 1\rangle})\psi_j(x^{\langle 2\rangle})d\omega(x^{\langle 1\rangle})$ are the basis expansion coefficients, the random noise ω_{ij} is mean zero and variance σ^2/n. The space $\mathcal{E} = \{f \in \mathcal{H} : ||f||_{\mathcal{H}} < 1\}$ in Equation (30) is equivalent to $\mathcal{E} = \{\sum_{i=1}^{\infty}\sum_{j=1}^{2}\frac{\theta_{ij}^2}{(\mu_i\nu_j)} \leq 1\}$. The hypothesis in Equation (18) is equivalent to the hypothesis

$$H_0 : \theta_{i2} = 0 \text{ for } i = 2, \dots, n.$$

Let $\boldsymbol{\theta}_{11} = (\theta_{22}, \theta_{32}, \dots, \theta_{n2})^T$, and $\mathcal{E}_{11} = \{\boldsymbol{\theta}_{11} \mid \sum_{i=2}^{n}\frac{\theta_{i2}^2}{(\mu_i\nu_2)} \leq 1\}$. Consider a local alternative $H_{1n} : \boldsymbol{\theta}_{11} \in \mathcal{E}_{11}$ with $\|\boldsymbol{\theta}_{11}\|_2 \geq d_n$, where d_n represents a generic distinguishable rate. The total error of a generic testing rule ϕ_n under distinguishable rate d_n can be rewritten as

$$pseudo.risk(\phi_n, d_n) = \mathbb{E}_{H_0}\{\phi_n | H_0 \text{ is true}\} + \sup_{\substack{\boldsymbol{\theta}_{11} \in \mathcal{E}_{11} \\ \|\boldsymbol{\theta}_{11}\|_2 \geq d_n}} \mathbb{E}\{1 - \phi_n | H_1 \text{ is true}\}. \tag{31}$$

Equation (31) is consistent with the testing error defined by Ingster (1993), Wei and Wainwright (2020). For the simplicity of description, we order the axis length $\{(\mu_i \nu_2)\}_{i=2}^{\infty}$ from the smallest to the largest as $\{\rho_p\}_{p=1}^{\infty}$. Next we introduce a lemma to give a low bound of the minimum pseudo risk.

Lemma 11 *For every set C and probability measure Q supported on $C \cap \mathbb{B}^c(d_n)$, we have*

$$\inf_{\phi_n} pseudo.risk(\phi_n, d_n) \geq 1 - \frac{1}{2}\sqrt{\mathbb{E}_{\eta,\eta'} \exp(\frac{\langle \eta, \eta' \rangle}{\sigma^2}) - 1}$$

where $\mathbb{E}_{\eta,\eta'}$ denotes expectation with respect to an i.i.d. pair η, $\eta' \sim \mathbb{Q}$

The proof of this lemma directly follows Lemma 3 in Wei and Wainwright (2020).

A.2.2. PROOF OF LEMMA 4

Proof As shown in Lemma 11, we have

$$\inf_{\phi_n} pseudo.risk(\phi_n, d_n) \geq 1 - \frac{1}{2}\sqrt{\mathbb{E}_{\eta,\eta'} \exp(\frac{\langle \eta, \eta' \rangle}{\sigma^2}) - 1} \tag{32}$$

Next we show that if $\delta^2 \leq \frac{\sqrt{k_B(\delta)}\sigma^2}{4}$, we have the last term in Equation (32) larger than $1/2$. Let $\theta_b = \frac{\delta}{\sqrt{k}} \sum_{i=1}^{k} b_i e_i$ where e_i is the standard basis vector with ith coordinate as one. We consider $\mathcal{Q}$ as the uniform distribution on $\{\theta_b, b \in \{-1, 1\}^k\}$. The expectation in the last term of Equation (32) can be written as

$$\begin{aligned}
\mathbb{E}_{n\eta,\eta'} \exp(\frac{n\langle \eta, \eta' \rangle}{\sigma^2}) =& \frac{1}{2^k} \sum_{b,b'} \exp(\frac{n\theta_b^T \theta_{b'}}{\sigma^2}) = \frac{1}{2^k} \sum_{b,b'} \exp(\frac{n\delta^2 \sum_{i=1}^{k} b_i b_i'}{k\sigma^2}) \\
=& \frac{1}{2^k} (\exp(\frac{n\delta^2}{k\sigma^2}) + \exp(-\frac{n\delta^2}{k\sigma^2}))^k \\
\overset{(i)}{\leq}& (1 + \frac{n^2\delta^4}{k^2\sigma^4})^k \\
\overset{(ii)}{\leq}& \exp(\frac{n^2\delta^4}{k\sigma^4}),
\end{aligned}$$

where (i) is due to that $\frac{1}{2}(\exp(x) + \exp(-x)) \leq 1 + x^2$ for $|x| \leq 1/2$ and (ii) is due to that $1 + x \leq e^x$. Thus for any $\delta^4 \leq \frac{k\sigma^4}{16n^2}$, we have

$$\inf_{\phi_n} pseudo.risk(\phi_n, d_n) \geq 1 - \frac{1}{2}\sqrt{e^{1/16} - 1} \geq 1/2.$$

By the definition of r_B, we have $pseudo.risk(\phi_n, d_n) > 1/2$ for all $d_n \leq r_B$. ■

A.2.3. Proof of Lemma 5

Proof We show that $b_{k,2}(\mathcal{E}_{11})$ is bounded below by $\sqrt{\rho_{k+1}}$. It is sufficient to show that $\mathcal{E}_{11}$ contains a l_2 ball centered at $f_{11} = 0$ with radius $\sqrt{\rho_{k+1}}$. For any $v \in \mathcal{E}_{11}$ with $||v||_2 \leq \sqrt{\rho_{k+1}}$, we have

$$b_{2,k} \overset{(i)}{\leq} \sum_{i=1}^{k+1} \frac{v_i^2}{\rho_i} \overset{(ii)}{\leq} \frac{1}{\mu_{k+1}} \sum_{i=1}^{k+1} v_i^2,$$

where inequality (i) holds by set the $(k+1)$-dimensional subspace spaned by the eigenvectors corresponding to the first $(k+1)$ largest eigenvalues; inequality (ii) holds by the decreasing order of the eigenvalues, i.e., $\rho_1 \geq \rho_2 \geq \ldots \rho_{k+1}$.

Recall that the definition of the Bernstein lower critical dimension is $k_B(\delta) = \arg\max_k \{b_{k-1,2}^2(\mathcal{E}_{11}) \geq \delta^2\}$, we have

$$k_B(\delta) \geq \arg\max_k \{\sqrt{\rho_k} \geq \delta\}.$$

■

A.2.4. Proof of Theorem 6

Proof By Lemme 4, we have

$$d_n \leq \sup\{\delta : k_B(\delta) \geq 16n^2\delta^4\}.$$

We plug in the lower bound of $k_B(\delta)$ in Lemma 5. Then we have

$$d_n \leq \sup\{\delta : \arg\max_k \{\sqrt{\rho_k} \geq \delta\} \geq 16n^2\delta^4\}. \tag{33}$$

The eigenvalues have polynomial decay rate i.e., $\rho_p \asymp p^{-2m}$, and consequently, $\arg\max_k\{\sqrt{\rho_k} \geq \delta\} \asymp \delta^{-1/m}$. Plugging this into Equation (33), it is easy to see that the supremum on the right hand side has an order $n^{-\frac{2m}{4m+1}}$. Proof is thus completed. ■

A.3. Proof of Theorem 7

Before deriving the proof of Theorem 7, we first state Lemma 12, Lemma 13, Lemma 14, and Lemma 15, which are used in the proof of Theorem 7. The proof of these auxiliary lemmas is referred to the Supplementary.

A.3.1. Some Auxiliary Lemmas

Lemma 12 shows the projection of $\boldsymbol{f}_{10}$ on $\mathcal{H}_{11} \cap \mathcal{H}_{model}$ is zero. This result indicates our test statistic does not depend on the nuisance parameter $\boldsymbol{f}_{10}$.

Lemma 12 *The quantity, $K_{11}M^{-1}(I_n - S(S^TM^{-1}S)^{-1}S^TM^{-1})\boldsymbol{f}_{10}$, equals to zero.*

The next two lemmas show the equivalence of τ_λ and $\hat{\tau}_\lambda$ under the quasi-uniform design and uniform design.

Lemma 13 *If $x^{\langle 1\rangle}$ follows the quasi-uniform random design, for any $\lambda = \frac{1}{n^{1-c}}$, $m > 3/2$, and any $\delta, c > 0$, we have*

$$P(\hat{\tau}_\lambda \asymp \tau_\lambda) \geq 1 - (n^{\frac{2}{2m-1}-2\delta} + n^{\frac{1}{2m-1}}) \exp\{-cn^{\frac{2m-3}{2m-1}+2\delta}\},$$

where $\tau_\lambda = \max\{i \mid \mu_i \geq \lambda\}$ and $\hat{\tau}_\lambda = max\{i \mid \hat{\mu}_i \geq \lambda\}$.

Lemma 14 *If $x^{\langle 1\rangle}$ follows the uniform fixed design condition, for $m > 1/2$ and $\lambda > 0$, we have*

$$\hat{\tau}_\lambda \asymp \tau_\lambda.$$

In the following lemma, we bound $\mathrm{Tr}(\Delta)$ by a function of $\hat{\tau}_\lambda$. This result is essential in deriving the asymptotic distribution of $T_{n,\lambda}$.

Lemma 15 *For $\Delta = M^{-1}K_{11}^2M^{-1}$ defined in Theorem 7, we have*

$$\frac{4\hat{\tau}_\lambda}{9} \leq \mathrm{Tr}(\Delta) \leq \frac{4}{(1-\theta_d)^2}(\hat{\tau}_\lambda + \frac{1}{2\lambda}\sum_{i=\hat{\tau}_\lambda+1}^{n} \hat{\mu}_i). \tag{34}$$

A.3.2. PROOF OF THEOREM 7

Proof For simplicity, we suppose $\sigma^2 = 1$. We define the three terms on the right-hand side of Equation (26) as T_1, T_2 and T_3, i.e.,

$$\begin{aligned}
T_1 &= \frac{1}{n}\boldsymbol{\epsilon}^T\Delta\boldsymbol{\epsilon},\\
T_2 &= \frac{1}{n}\boldsymbol{\epsilon}^T M^{-1}S(S^TM^{-1}S)^{-1}S^T\Delta S(S^TM^{-1}S)^{-1}S^TM^{-1}\boldsymbol{\epsilon},\\
T_3 &= \frac{1}{n}\boldsymbol{\epsilon}^T M^{-1}S(S^TM^{-1}S)^{-1}S^T\Delta\boldsymbol{\epsilon}.
\end{aligned}$$

We now show T_2 and T_3 are in smaller order compared to T_1. First, we analyze the second term T_2 in Equation (26). We have

$$\begin{aligned}
\mathbb{E}[T_2] =& \frac{1}{n}\mathbb{E}[\boldsymbol{\epsilon}^T M^{-1}S(S^TM^{-1}S)^{-1}S^T\Delta S(S^TM^{-1}S)^{-1}S^TM^{-1}\boldsymbol{\epsilon}]\\
=& \frac{1}{n}\,\mathrm{Tr}(M^{-1}S(S^TM^{-1}S)^{-1}S^T\Delta S(S^TM^{-1}S)^{-1}S^TM^{-1})\\
\leq& \frac{2}{n}\lambda_{\max}(\Delta)\lambda_{\max}(M^{-1}S(S^TM^{-1}S)^{-1}S^TS(S^TM^{-1}S)^{-1}S^TM^{-1})\\
\leq& \frac{2}{n}\lambda_{\max}(\Delta),
\end{aligned}$$

where $\lambda_{\max}$ denotes the largest eigenvalue. Since all eigenvalues of Δ are less than 1, we have $\mathbb{E}[T_2] \leq \frac{2}{n}$. Analogously, we can derive the variance inequality of T_2. Combining the results together and using the Chebyshev inequality, we have

$$T_2 = \mathcal{O}_p(\frac{1}{n}). \tag{35}$$

Second, we analyze the third term T_3 in Equation (26). We apply the Cauchy-Schwarz inequality and have

$$|T_3| \leq \sqrt{T_2}\sqrt{T_1}. \tag{36}$$

Finally, we derive the magnitude of T_1. We first consider the testing consistency of T_1 conditional on X. Denote $\mathbb{E}_\epsilon$ as the expectation with respect to ϵ, and define $\mathbb{V}ar_\epsilon$ as the variance with respect to ϵ. Note that

$$\mathbb{E}_\epsilon[\boldsymbol{\epsilon}^T \Delta \boldsymbol{\epsilon}] = \mathrm{Tr}(\Delta), \quad \mathbb{V}ar_\epsilon[\boldsymbol{\epsilon}^T \Delta \boldsymbol{\epsilon}] = 2\,\mathrm{Tr}(\Delta^2).$$

Let $Z = (\boldsymbol{\epsilon}^T \Delta \boldsymbol{\epsilon} - \mathrm{Tr}(\Delta))/\sqrt{2\,\mathrm{Tr}(\Delta^2)}$ and $t \in (-1/2, 1/2)$. Then the log-characteristic function of Z can be written as

$$\begin{aligned}
&\log \mathbb{E}_\epsilon[\exp(itZ)] \\
=&\log \mathbb{E}_\epsilon[\exp(it\boldsymbol{\epsilon}^T \Delta \boldsymbol{\epsilon}/\sqrt{2\,\mathrm{Tr}(\Delta^2)})] - it\,\mathrm{Tr}(\Delta)/\sqrt{2\,\mathrm{Tr}(\Delta^2)} \\
=&-\frac{1}{2}\log\det\{I_{2n} - 2it\Delta/\sqrt{2\,\mathrm{Tr}(\Delta^2)}\} - it\,\mathrm{Tr}(\Delta)/\sqrt{2\,\mathrm{Tr}(\Delta^2)}.
\end{aligned} \tag{37}$$

Through Taylor expansion, one has

$$\begin{aligned}
&-\frac{1}{2}\log\det\{I_{2n} - 2it\Delta/\sqrt{2\,\mathrm{Tr}(\Delta^2)}\} \\
=&it\frac{\mathrm{Tr}(\Delta)}{\sqrt{2\,\mathrm{Tr}(\Delta^2)}} - t^2\frac{\mathrm{Tr}(\Delta^2)}{2\,\mathrm{Tr}(\Delta^2)} + \mathcal{O}(t^3\frac{\mathrm{Tr}(\Delta^3)}{[\mathrm{Tr}(\Delta^2)]^{3/2}}).
\end{aligned} \tag{38}$$

Combining Equations (37) and (38), we have

$$\log \mathbb{E}_\epsilon[\exp(itZ)] = -\frac{t^2}{2} + \mathcal{O}(t^3\frac{\mathrm{Tr}(\Delta^3)}{[\mathrm{Tr}(\Delta^2)]^{3/2}}). \tag{39}$$

Since all eigenvalues of Δ are less than 1, we have $\frac{\mathrm{Tr}(\Delta^3)}{\mathrm{Tr}(\Delta^2)} \leq 1$. Analogous to (S.11), we have

$$\mathrm{Tr}(\Delta^2) \geq \frac{16}{81}\hat{\tau}_\lambda. \tag{40}$$

Under the quasi-uniform design, we have $\mathrm{Tr}(\Delta^2) \to \infty$ as $\lambda \to 0$ with probability approaching 1 by Lemma 13 and Equation (40). Hence, the second term on the right-hand side of Equation (39) is $o_p(1)$. We thus conclude that

$$\mathbb{E}_\epsilon[\exp(itZ)] \xrightarrow{P} \exp(-\frac{t^2}{2}).$$

Next, we show that

$$\mathbb{E}[\exp(itZ)] = \mathbb{E}_X\left[\mathbb{E}_\epsilon[\exp(itZ)]\right] \to \exp(-t^2/2)$$

for $t \in (-\frac{1}{2}, \frac{1}{2})$. If not, there exists a subsequence of r.v $X_{nk}^{\langle 1\rangle}$, such that for $\forall \varepsilon > 0$, $|\mathbb{E}_{X_{nk}^{\langle 1\rangle}}\mathbb{E}_\epsilon \exp(itZ) - \exp(-t^2/2)| > \varepsilon$. On the other hand, since $\mathbb{E}_\epsilon \exp(itZ(X_{nk}^{\langle 1\rangle})) \xrightarrow{P} \exp(-t^2/2)$, which is bounded, there exists a sub-sub sequence $\{X_{n_{kl}}^{\langle 1\rangle}\}$, such that $\mathbb{E}_\epsilon \exp(itZ(X_{n_{kl}}^{\langle 1\rangle})) \xrightarrow{a.s} \exp(-t^2/2)$. Then by dominate convergence theorem, $\mathbb{E}_{X_{n_{kl}}^{\langle 1\rangle}}\mathbb{E}_\epsilon \exp(itZ) \to \exp(-t^2/2)$, which is a contradiction. Under the uniform design, we can easily obtain $\mathbb{E}[\exp(itZ)] \to \exp(-\frac{t^2}{2})$ by Lemma 14 and Equation (40).

Thus Z is asymptotically normally distributed, and

$$\frac{T_1 - \mathrm{Tr}(\Delta)/n}{\sqrt{2\,\mathrm{Tr}(\Delta^2)/n^2}} \xrightarrow{d} N(0,1). \tag{41}$$

Combining (35), (36) and (41), the theorem follows. ■

A.4. Proof of Theorem 8

Proof Under the alternative hypothesis, the statistic $T_{n,\lambda}$ in Equation (26) can be decomposed into three terms as follows

$$T_{n,\lambda} = \frac{1}{n}||H\boldsymbol{\epsilon}||_2^2 + \frac{1}{n}||H\boldsymbol{f}_{11}||_2^2 + \frac{2}{n}\boldsymbol{f}_{11}^T H^T H\boldsymbol{\epsilon}. \tag{42}$$

where $H = \theta_{11}K_{11}M^{-1}(I - S(S^TM^{-1}S)^{-1}S^TM^{-1})$. Let $W_1 = \frac{1}{n}||H\boldsymbol{\epsilon}||_2^2$, $W_2 = \frac{1}{n}||H\boldsymbol{f}_{11}||_2^2$, and $W_3 = \frac{2}{n}\boldsymbol{f}_{11}^T H^T H\boldsymbol{\epsilon}$ denote corresponding three terms on the right-hand side of Equation (42).

We now derive a lower bound for W_2. By Lemma S.1, we have

$$\begin{aligned}\frac{1}{n}||H\boldsymbol{f}_{11} - \boldsymbol{f}_{11}||_2^2 &\leq \frac{1}{n}||H\boldsymbol{f}_{11} - \boldsymbol{f}_{11}||_2^2 + \frac{1}{n}||H\boldsymbol{f}_{10} - \boldsymbol{f}_{10}||_2^2 \\ &= \frac{1}{n}||H\boldsymbol{f}_{10} + H\boldsymbol{f}_{11} - \boldsymbol{f}_{10} - \boldsymbol{f}_{11}||_2^2 \\ &= ||\tilde{g}^* - g^*||_n^2 \leq c\lambda. \end{aligned}\tag{43}$$

Let $c' = \sqrt{c}$, we consider the distinguishable rate

$$\frac{1}{n}||\boldsymbol{f}_{11}||_2^2 = ||f_{11}||_n^2 > c'^2 d_n^2 = c(\lambda + \sigma_{n,\lambda}). \tag{44}$$

where the inequality is satisfied since $||\cdot||_n$ dominates $||\cdot||_2$ by Lemma S.2. The lower bound of W_2 is thus,

$$W_2 = \frac{1}{n}||H\boldsymbol{f}_{11}||_2^2 = \frac{1}{n}||\boldsymbol{f}_{11}||_2^2 - \frac{1}{n}||\boldsymbol{f}_{11} - H\boldsymbol{f}_{11}||_2^2 \geq cd_n^2 - c\lambda \geq c\sigma_{n,\lambda}. \tag{45}$$

where the first inequality is obtained by (43) and the second inequality is obtained through plugging in Equation (44).

For the third term W_3, it is seen that $\mathbb{E}W_3 = 0$. It is easy to verify that the eigenvalues of HH^T are all less than 1. Moreover,

$$\begin{aligned}\mathbb{E}W_3^2 =& \frac{4}{n^2}\mathbb{E}[\boldsymbol{f}_{11}^T H^T H\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T H^T H\boldsymbol{f}_{11}] = \frac{4}{n^2}(H\boldsymbol{f}_{11})^T HH^T(H\boldsymbol{f}_{11}) \\ \leq& \frac{4}{n^2}(H\boldsymbol{f}_{11})^T(H\boldsymbol{f}_{11}) = \frac{4}{n}W_2.\end{aligned}$$

By the Chebyshev's inequality, for any $\epsilon > 0$, we have

$$\mathbb{P}(|W_3| \geq \frac{2\epsilon^{-\frac{1}{2}}W_2^{\frac{1}{2}}}{\sqrt{n}}) \leq \frac{n\mathbb{E}W_3^2}{4\epsilon^{-1}W_2} \leq \epsilon.$$

Consequently, there exists an n_0, for any $n > n_0$, we have

$$\mathbb{P}\{|W_3| > \frac{1}{2}W_2\} \leq \mathbb{P}(|W_3| \geq \frac{2\epsilon^{-\frac{1}{2}}W_2^{\frac{1}{2}}}{\sqrt{n}}) \leq \epsilon. \tag{46}$$

Now, we are ready to prove our theorem. By the triangle inequality, we have

$$\begin{aligned} |\frac{W_1 - \mu_{n,\lambda}}{\sigma_{n,\lambda}} + \frac{W_2 + W_3}{\sigma_{n,\lambda}}| \geq & |\frac{W_2 + W_3}{\sigma_{n,\lambda}}| - |\frac{W_1 - \mu_{n,\lambda}}{\sigma_{n,\lambda}}| & (47) \\ \geq & |\frac{W_2}{\sigma_{n,\lambda}}| - |\frac{W_3}{\sigma_{n,\lambda}}| - |\frac{W_1 - \mu_{n,\lambda}}{\sigma_{n,\lambda}}|. & (48) \end{aligned}$$

If $\frac{|W_1 - \mu_{n,\lambda}|}{\sigma_{n,\lambda}} \leq C_\epsilon$, and $|W_3| \leq \frac{1}{2}W_2$ hold, in view of (47) and Equation (45), we have

$$|\frac{W_1 - \mu_{n,\lambda}}{\sigma_{n,\lambda}} + \frac{W_2 + W_3}{\sigma_{n,\lambda}}| \geq \frac{1}{2}c - C_\epsilon.$$

Noting that W_1 is identical to Equation (26), by Theorem 7, we have $\frac{|W_1 - \mu_{n,\lambda}|}{\sigma_{n,\lambda}} = \mathcal{O}_p(1)$. That is for any $C_\epsilon > 0$, there exists an s, for any $n > s$, we have

$$\mathbb{P}(\frac{|W_1 - \mu_{n,\lambda}|}{\sigma_{n,\lambda}} > C_\epsilon) \leq \epsilon. \tag{49}$$

Setting $c \geq 2(C_\epsilon + z_{1-\frac{\alpha}{2}})$ and $N = \max(n_0, s)$, for any $n > N$, we have

$$\begin{aligned} \mathbb{P}(\phi_{n,\lambda} = 1) =& \mathbb{P}\{\frac{|W_1 + W_2 + W_3 - \mu_{n,\lambda}|}{\sigma_{n,\lambda}} \geq z_{1-\frac{\alpha}{2}}\} \\ \geq& \mathbb{P}\{\frac{|W_1 - \mu_{n,\lambda}|}{\sigma_{n,\lambda}} \leq C_\epsilon, |W_3| \leq \frac{1}{2}W_2\} \\ \geq& 1 - \mathbb{P}\{\frac{|W_1 - \mu_{n,\lambda}|}{\sigma_{n,\lambda}} > C_\epsilon\} - \mathbb{P}\{|W_3| > \frac{1}{2}W_2\} \\ \geq& 1 - 2\epsilon, \end{aligned}$$

where the second inequality is due to the Boole's inequality (Casella and Berger, 2002) and the last inequality is obtained by combining Equation (45) and Equation (49). Thus, we have

$$\sup_{H_1^*} \mathbb{E}(1 - \phi_{n,\lambda}|H_1^* \text{ is true}) < \delta,$$

where $H_1^* = \{f \mid f \in \mathcal{H}_{model}^\infty \text{ and } ||f_{11}||_2 \geq C_\delta\sqrt{\lambda + \sigma_{n,\lambda}} \triangleq d_n\}$.

■

References

Milton Abramowitz and Irene A Stegun. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*. National Bureau of Standards, Washington, DC., 1964.

Ahmed Alaoui and Michael W Mahoney. Fast randomized kernel ridge regression with statistical guarantees. In *Advances in Neural Information Processing Systems 28*, pages 775–783. 2015.

Peter L Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *Annals of Statistics*, 33(4):1497–1537, 2005.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

Martin Bilban, Daniel Heintel, Theresa Scharl, Thomas Woelfel, Michael M Auer, Edit Porpaczy, Birgit Kainz, Alexander Krober, Vincent J Carey, Medhat Shehata, C Zielinski, W Pickl, S Stilgenbauer, A Gaiger, O Wagner, U Jager, and German CLL Study Group. Deregulated expression of fat and muscle genes in b-cell chronic lymphocytic leukemia with high lipoprotein lipase expression. *Leukemia*, 20(6):1080–1088, 2006.

Mikio L Braun. Accurate error bounds for the eigenvalues of the kernel matrix. *Journal of Machine Learning Research*, 7(Nov):2303–2328, 2006.

George Casella and Roger L Berger. *Statistical inference*. Duxbury Pacific Grove, CA, 2nd edition, 2002.

David Degras, Zhiwei Xu, Ting Zhang, and Wei Biao Wu. Testing for parallelism among trends in multiple time series. *IEEE Transactions on Signal Processing*, 60(3):1087–1097, 2011.

Petros Drineas and Michael W Mahoney. On the nyström method for approximating a gram matrix for improved kernel-based learning. *Journal of Machine Learning Research*, 6(Dec):2153–2175, 2005.

C Echávarri, P Aalten, HBM Uylings, HIL Jacobs, PJ Visser, EHBM Gronenschild, FRJ Verhey, and S Burgmans. Atrophy in the parahippocampal gyrus as an early biomarker of alzheimer's disease. *Brain Structure and Function*, 215(3-4):265–271, 2011.

Paulus Petrus Bernardus Eggermont and Vincent N LaRiccia. *Maximum penalized likelihood estimation*, volume II. Springer, 2001.

Jianqing Fan and Jian Zhang. Sieve empirical likelihood ratio tests for nonparametric functions. *Ann. Statist.*, 32(5):1858–1907, 10 2004.

Jianqing Fan, Chunming Zhang, and Jian Zhang. Generalized likelihood ratio statistics and wilks phenomenon. *Ann. Statist.*, 29(1):153–193, 02 2001.

Katharina Filarsky, Angela Garding, Natalia Becker, Christine Wolf, Manuela Zucknick, Rainer Claus, Dieter Weichenhan, Christoph Plass, Hartmut Döhner, Stephan Stilgenbauer, Peter Lichter, and Daniel Mertens. Krüppel-like factor 4 (klf4) inactivation in chronic lymphocytic leukemia correlates with promoter dna-methylation and can be reversed by inhibition of notch signaling. *Haematologica*, 101(6):249, 2016.

Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models*. Cambridge University Press, 2015.

Gene H Golub, Michael Heath, and Grace Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979.

Chong Gu. Model diagnostics for smoothing spline ANOVA models. *Canadian Journal of Statistics*, 32(4):347–358, 2004.

Chong Gu. *Smoothing spline ANOVA models*. Springer, 2nd edition, 2013.

Kasper D Hansen, Benjamin Langmead, and Rafael A Irizarry. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biology*, 13(10):R83, 2012.

Scott A Huettel, Allen W Song, and Gregory McCarthy. *Functional magnetic resonance imaging*. Sinauer Associates Sunderland, 2004.

Yuri Ingster and Irina A Suslina. *Nonparametric goodness-of-fit testing under Gaussian models*. Springer Science & Business Media, 2012.

Yuri I Ingster. Asymptotically minimax hypothesis testing for nonparametric alternatives. i, ii, iii. *Math. Methods Statist*, 2(2):85–114, 1993.

Rafael A Irizarry, Christine Ladd-Acosta, Benilton Carvalho, Hao Wu, Sheri A Brandenburg, Jeffrey A Jeddeloh, Bo Wen, and Andrew P Feinberg. Comprehensive high-throughput arrays for relative methylation (charm). *Genome Research*, 18(5):780–790, 2008.

J Patrick Kesslak, Orhan Nalcioglu, and Carl W Cotman. Quantification of magnetic resonance scans for hippocampal and parahippocampal atrophy in alzheimer's disease. *Neurology*, 41(1):51–51, 1991.

Young-Ju Kim and Chong Gu. Smoothing spline gaussian regression: more scalable computation via efficient approximation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(2):337–356, 2004.

Yi Lin. Tensor product space ANOVA models. *Annals of Statistics*, 28(3):734–755, 2000.

Anna Liu and Yuedong Wang. Hypothesis testing in smoothing spline models. *Journal of Statistical Computation and Simulation*, 74(8):581–597, 2004.

Meimei Liu and Guang Cheng. Early stopping for nonparametric testing. In *Advances in Neural Information Processing Systems*, pages 3985–3994, 2018.

Meimei Liu, Zuofeng Shang, and Guang Cheng. Sharp theoretical analysis for nonparametric testing under random projection. In *Conference on Learning Theory*, pages 2175–2209, 2019.

Qiang Liu, Jason Lee, and Michael Jordan. A kernelized stein discrepancy for goodness-of-fit tests. In *International Conference on Machine Learning*, pages 276–284, 2016.

Ping Ma, Wenxuan Zhong, and Jun S Liu. Identifying differentially expressed genes in time course microarray data. *Statistics in Biosciences*, 1(2):144, 2009.

Ping Ma, Michael W Mahoney, and Bin Yu. A statistical perspective on algorithmic leveraging. *Journal of Machine Learning Research*, 16(1):861–911, 2015.

Siyuan Ma and Mikhail Belkin. Diving into the shallows: a computational perspective on large-scale shallow learning. In *Advances in Neural Information Processing Systems*, pages 3778–3787, 2017.

Axel Munk and Holger Dette. Nonparametric comparison of several regression functions: exact and asymptotic theory. *Annals of Statistics*, 26(6):2339–2368, 1998.

Thomas E Nichols and Andrew P Holmes. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human Brain Mapping*, 15(1):1–25, 2002.

Masaki Okano, Shaoping Xie, and En Li. Cloning and characterization of a family of novel mammalian dna (cytosine-5) methyltransferases. *Nature Genetics*, 19(3):219, 1998.

William W Orrison, Jeffrey Lewine, John Sanders, and Michael F Hartshorne. *Functional brain imaging*. Elsevier Health Sciences, 2017.

CP Pallasch, J Schwamb, S Königs, A Schulz, S Debey, D Kofler, JL Schultze, M Hallek, A Ultsch, and CM Wendtner. Targeting lipid metabolism by the lipoprotein lipase inhibitor orlistat results in apoptosis of b-cell chronic lymphocytic leukemia cells. *Leukemia*, 22(3):585–592, 2008.

Allan Pinkus. *N-widths in Approximation Theory*, volume 7. Springer Science & Business Media, 2012.

Lorena Rami, Roser Sala-Llonch, Cristina Solé-Padullés, Juan Fortea, Jaume Olives, Albert Lladó, Cleofe Peña-Gómez, Mircea Balasa, Bea Bosch, Anna Antonell, R. Sanchez-Valle, D. Bartrés-Faz, and J. L. Molinuevo. Distinct functional activity of the precuneus and posterior cingulate cortex during encoding in the preclinical stage of alzheimer's disease. *Journal of Alzheimer's Disease*, 31 (3):517–526, 2012.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Early stopping and non-parametric regression: an optimal data-dependent stopping rule. *Journal of Machine Learning Research*, 15(1):335–366, 2014.

Serge ARB Rombouts, Frederik Barkhof, Rutger Goekoop, Cornelis J Stam, and Philip Scheltens. Altered resting state networks in mild cognitive impairment and mild alzheimer's disease: an fmri study. *Human Brain Mapping*, 26(4):231–239, 2005.

Stephen W Scheff, Douglas A Price, Mubeen A Ansari, Kelly N Roberts, Frederick A Schmitt, Milos D Ikonomovic, and Elliott J Mufson. Synaptic change in the posterior cingulate gyrus in the progression of alzheimer's disease. *Journal of Alzheimer's Disease*, 43(3):1073–1090, 2015.

Bernhard Schölkopf, Ralf Herbrich, and Alex J Smola. A generalized representer theorem. In *International Conference on Computational Learning Theory*, pages 416–426. Springer, 2001.

Dirk Schübeler. Function and information content of dna methylation. *Nature*, 517(7534):321, 2015.

Zuofeng Shang and Guang Cheng. Local and global asymptotic inference in smoothing spline models. *Annals of Statistics*, 41(5):2608–2638, 2013.

Zuofeng Shang and Guang Cheng. Computational limits of a distributed algorithm for smoothing spline. *Journal of Machine Learning Research*, 18(1):3809–3845, 2017.

Xiaotong Shen, Hsin-Cheng Huang, and Noel Cressie. Nonparametric hypothesis testing for a spatial signal. *Journal of the American Statistical Association*, 97(460):1122–1140, 2002.

Stephen M Smith, Mark Jenkinson, Mark W Woolrich, Christian F Beckmann, Timothy EJ Behrens, Heidi Johansen-Berg, Peter R Bannister, Marilena De Luca, Ivana Drobnjak, David E Flitney, RK Niazy, J Saunders, J Vickers, Y Zhang, N De Stefano, JM Brady, and PM Matthews. Advances in functional and structural mr image analysis and implementation as fsl. *Neuroimage*, 23: S208–S219, 2004.

Dirk Stach, Oliver J Schmitz, Stephan Stilgenbauer, Axel Benner, Hartmut DoÈhner, Manfred Wiessler, and Frank Lyko. Capillary electrophoretic analysis of genomic DNA methylation levels. *Nucleic Acids Research*, 31(2):e2–e2, 2003.

Lars Sthle and Svante Wold. Analysis of variance (anova). *Chemometrics and Intelligent Laboratory Systems*, 6(4):259–272, 1989.

John D Storey, Wenzhong Xiao, Jeffrey T Leek, Ronald G Tompkins, and Ronald W Davis. Significance analysis of time course microarray experiments. *Proceedings of the National Academy of Sciences*, 102(36):12837–12842, 2005.

Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, , , and Jill P Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43): 15545–15550, 2005.

Mehrdad Vossoughi, SMT Ayatollahi, Mina Towhidi, and Seyyed Taghi Heydari. A distribution-free test of parallelism for two-sample repeated measurements. *Statistical Methodology*, 30:31–44, 2016.

Grace Wahba. *Spline models for observational data*. SIAM, 1990.

Grace Wahba, Yuedong Wang, Chong Gu, Ronald Klein, Barbara Klein, et al. Smoothing spline anova for exponential families, with application to the wisconsin epidemiological study of diabetic retinopathy: the 1994 neyman memorial lecture. *Annals of Statistics*, 23(6):1865–1895, 1995.

Liang Wang, Yufeng Zang, Yong He, Meng Liang, Xinqing Zhang, Lixia Tian, Tao Wu, Tianzi Jiang, and Kuncheng Li. Changes in hippocampal connectivity in the early stages of alzheimer's disease: evidence from resting state fmri. *Neuroimage*, 31(2):496–504, 2006.

Yazhen Wang. Change curve estimation via wavelets. *Journal of the American Statistical Association*, 93(441):163–172, 1998.

Yuedong Wang. *Smoothing splines: methods and applications*. CRC Press, 2011.

Yuting Wei and Martin J Wainwright. The local geometry of testing in ellipses: Tight control via localized kolmogorov widths. *IEEE Transactions on Information Theory*, 2020.

Simon N Wood. Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):95–114, 2003.

Yun Yang, Mert Pilanci, Martin J Wainwright, et al. Randomized sketches for kernels: Fast and optimal nonparametric regression. *Annals of Statistics*, 45(3):991–1023, 2017.

Supplementary

Supplement to
Minimax Nonparametric Parallelism Test

This document contains some auxiliary lemmas, the proofs of Corollary 9 and Corollary 10 as well as the proofs of Lemma 12, Lemma 13, Lemma 14, and Lemma 15 stated in Appendix.

- Section S.1 includes some auxiliary lemmas in proving Theorem 8.
- Section S.2 includes the proof of Corollary 9 and Corollary 10.
- Section S.3 includes the proof of Lemma 12, Lemma 13, Lemma 14, Lemma 15, Lemma S.1, Lemma S.2, and Lemma S.4.

S.1. Some Auxiliary Lemmas in Proving Theorem 8

We first introduce several notations and lemmas and then start the main proof of Theorem 8.
Let $g^* = f_{10} + f_{11}$ and its estimator as

$$\tilde{\boldsymbol{g}}^* = R[M^{-1} - M^{-1}S(S^TM^{-1}S)^{-1}S^TM^{-1}]\boldsymbol{g}^*. \tag{S.1}$$

Lemma S.1 *If $\|f\|_{\mathcal{H}} < 1$ for any $f \in \mathcal{H}$, as $n \to \infty$, $\lambda \to 0$ and $\lambda \geq n^{-1}$, we have*

$$||\tilde{g}^* - g^*||_n^2 \leq c\lambda,$$

where c is a constant, $||\cdot||_n$ is the empirical norm.

In the following Lemma S.2, we discuss the relationship between the empirical norm and L_2 norm. Recall the definition of empirical norm and L_2 norm are as follows:

$$||f||_n^2 = \frac{1}{2n}\sum_{j=1}^{2}\sum_{i=1}^{n} f^2(x_i^{\langle 1\rangle}, x_j^{\langle 2\rangle}) \quad \text{and} \quad ||f||_2^2 = \sum_{x^{\langle 2\rangle}=0}^{1}\int_0^1 f^2(x^{\langle 1\rangle}, x^{\langle 2\rangle})d\omega_1.$$

Lemma S.2 *Under the quasi-uniform design or the uniform design, for $f : \mathcal{X}_1 \times \mathcal{X}_2 \to \mathbb{R}$ and a positive constant c, we have*

$$||f||_2 \leq c||f||_n,$$

i.e. the empirical norm of f dominates the L_2 norm.

S.2. Proofs of Corollary 9 and Corollary 10

In order to find the optimal distinguishable rate, we need to bound the tail sum of the eigenvalues of the empirical kernel matrix. We state the following two lemmas which give upper bounds for the tail sum of the eigenvalues of the empirical kernel matrix under the quasi-uniform design and the uniform design respectively.

Lemma S.3 *(Liu et al. (2019)) If $1/n < \lambda \to 0$ and the quasi-uniform design is satisfied, then with probability at least $1 - 4e^{-\tau_\lambda}$,*

$$\sum_{i=\hat{\tau}_\lambda+1}^{n} \hat{\mu}_i \le C\tau_\lambda \mu_{\tau_\lambda},$$

where $C > 0$ is an absolute constant.

Lemma S.4 *If $\lambda > 0$ and the uniform design is satisfied, we have*

$$\sum_{i=\hat{\tau}_\lambda+1}^{n} \hat{\mu}_i \le C\tau_\lambda \mu_{\tau_\lambda},$$

where $C > 0$ is an absolute constant.

Now we start the main proof of Corollary 9 and Corollary 10. The distinguishable rate is

$$d_n = \sqrt{\lambda + \sigma_{n,\lambda}},$$

where $\sigma_{n,\lambda}^2 = 2\theta_{11}^4 \sigma^4 \operatorname{Tr}(\Delta^2)/n^2$. We now derive the order of $\sigma_{n,\lambda}^2$. Since the eigenvalues of Δ are less than 1, and by Lemma 15, we have

$$\operatorname{Tr}(\Delta^2) \le \operatorname{Tr}(\Delta) \le \frac{4}{(1-\theta_d)^2}(\hat{\tau}_\lambda + \frac{1}{2\lambda}\sum_{i=\hat{\tau}_\lambda+1}^{n} \hat{\mu}_i).$$

Proof Under the quasi-uniform design, applying Lemma A.1, we have

$$\operatorname{Tr}(\Delta^2) \lesssim \frac{4}{(1-\theta_d)^2}(\hat{\tau}_\lambda + \frac{1}{2\lambda}\lambda\tau_\lambda)$$

with probability at least $1 - 4e^{-\tau_\lambda}$. Combining the lower bound of $\operatorname{Tr}(\Delta^2)$ in Equation (40) and Lemma S.3, we have

$$\operatorname{Tr}(\Delta^2) \asymp \tau_\lambda, \tag{S.2}$$

with probability at least $1 - 4e^{-\tau_\lambda} - (n^{\frac{2}{2m-1}-2\epsilon} + n^{\frac{1}{2m-1}})\exp\{-cn^{\frac{2m-3}{2m-1}+2\epsilon}\}$. Similarly, we have Equation (S.2) satisfied under the uniform design by applying Lemma 14 and Lemma S.4.

Using Equation (S.2), we have

$$\sigma_{n,\lambda}^2 \asymp \lambda^{-\frac{1}{2m}} n^{-2} \asymp \tau_\lambda n^{-2}. \tag{S.3}$$

By the Cauchy-Schwartz inequality, the distinguishable rate $d_n = \sqrt{\lambda + \sigma_{n,\lambda}}$ is minimized when $\lambda \asymp \sigma_{n,\lambda}$, i.e.,

$$\lambda \asymp n^{-4m/(4m+1)}.$$

Thus we have the minimum distinguishable rate

$$d_n^* \asymp n^{-2m/(4m+1)}.$$

By Lemma S.2, this optimal distinguishable rate is achieved in the sense of L_2 norm. ■

S.3. Proofs of Auxiliary Results

S.3.1. Proof of Lemma 12

Proof Write matrix R as

$$R = \theta_{01}K_{01} + \theta_{11}K_{11} = \frac{1}{2}\begin{bmatrix} K_1^{\langle 1\rangle} & \theta_d K_1^{\langle 1\rangle} \\ \theta_d K_1^{\langle 1\rangle} & K_1^{\langle 1\rangle} \end{bmatrix},$$

where $\theta_d = \theta_{01} - \theta_{11}$. The inverse of M can be written as

$$\begin{aligned} M^{-1} &= \begin{bmatrix} \frac{1}{2}K_1^{\langle 1\rangle} + \lambda I_n & \frac{\theta_d}{2}K_1^{\langle 1\rangle} \\ \frac{\theta_d}{2}K_1^{\langle 1\rangle} & \frac{1}{2}K_1^{\langle 1\rangle} + \lambda I_n \end{bmatrix}^{-1} \triangleq \begin{bmatrix} A & B \\ B & A \end{bmatrix}^{-1} \\ &= \begin{bmatrix} A^{-1} + A^{-1}B(A - BA^{-1}B)^{-1}BA^{-1} & -A^{-1}B(A - BA^{-1}B)^{-1} \\ -A^{-1}B(A - BA^{-1}B)^{-1} & (A - BA^{-1}B)^{-1} \end{bmatrix}, \end{aligned}$$

where $A = \frac{1}{2}K_1^{\langle 1\rangle} + \lambda I_n$, $B = \frac{\theta_d}{2}K$, and I_n denotes the $n \times n$ identity matrix. Note that S is a $2n \times 2$ matrix defined as $S = (\mathbf{1}_n, \mathbf{1}_n)^T$. We thus have

$$S^T M^{-1} S = \begin{bmatrix} a & b \\ b & c \end{bmatrix},$$

where

$$\begin{aligned} a &= \mathbf{1}^T A^{-1}\mathbf{1} + \mathbf{1}^T A^{-1}B(A - BA^{-1}B)^{-1}BA^{-1}\mathbf{1} + 2b - c, \\ b &= -\mathbf{1}^T BA^{-1}(A - BA^{-1}B)^{-1}\mathbf{1} + \mathbf{1}^T(A - BA^{-1}B)^{-1}\mathbf{1}, \\ c &= \mathbf{1}^T(A - BA^{-1}B)^{-1}\mathbf{1}. \end{aligned}$$

Consequently,

$$\begin{aligned} S(S^T M^{-1} S)^{-1} S^T &= \frac{1}{ac - b^2}\begin{bmatrix} c\mathbf{1}\mathbf{1}^T & (c-b)\mathbf{1}\mathbf{1}^T \\ (c-b)\mathbf{1}\mathbf{1}^T & (a+c-2b)\mathbf{1}\mathbf{1}^T \end{bmatrix} \\ &= \frac{1}{ac - b^2}\begin{bmatrix} c\mathbf{1}\mathbf{1}^T & (c-b)\mathbf{1}\mathbf{1}^T \\ (c-b)\mathbf{1}\mathbf{1}^T & c\mathbf{1}\mathbf{1}^T \end{bmatrix}, \end{aligned}$$

where the second equality holds by the fact $a - 2b = 0$ and Woodbury matrix identity.

Note that $\boldsymbol{f}_{10} = (f_{10}(x_1^{\langle 1\rangle}), \ldots, f_{10}(x_n^{\langle 1\rangle}), f_{10}(x_1^{\langle 1\rangle}), \ldots, f_{10}(x_n^{\langle 1\rangle}))$. Let

$$(f_{10}(x_1^{\langle 1\rangle}), \ldots, f_{10}(x_n^{\langle 1\rangle}) \triangleq \mathbf{h}^T.$$

Therefore, we have

$$\begin{aligned} &K_{11}M^{-1}(I_n - S(S^T M^{-1} S)^{-1} S^T M^{-1})\boldsymbol{f}_{10} \\ =&\frac{1}{2}\begin{bmatrix} K_1^{\langle 1\rangle} & -K_1^{\langle 1\rangle} \\ -K_1^{\langle 1\rangle} & K_1^{\langle 1\rangle} \end{bmatrix}(M^{-1} - \frac{1}{ac - b^2}M^{-1}\begin{bmatrix} c\mathbf{1}\mathbf{1}^T & (c-b)\mathbf{1}\mathbf{1}^T \\ (c-b)\mathbf{1}\mathbf{1}^T & c\mathbf{1}\mathbf{1}^T \end{bmatrix}M^{-1})\begin{bmatrix} \mathbf{h} \\ \mathbf{h} \end{bmatrix}. \end{aligned}$$

Since both M^{-1} and $\frac{1}{ac-b^2}M^{-1}\begin{bmatrix} c\mathbf{1}\mathbf{1}^T & (c-b)\mathbf{1}\mathbf{1}^T \\ (c-b)\mathbf{1}\mathbf{1}^T & c\mathbf{1}\mathbf{1}^T \end{bmatrix} M^{-1})$ are symmetric matrices and their diagnonal entries are identical, we have

$$\begin{bmatrix} \mathbf{h}^* \\ \mathbf{h}^* \end{bmatrix} \triangleq (M^{-1} - \frac{1}{ac-b^2}M^{-1}\begin{bmatrix} c\mathbf{1}\mathbf{1}^T & (c-b)\mathbf{1}\mathbf{1}^T \\ (c-b)\mathbf{1}\mathbf{1}^T & c\mathbf{1}\mathbf{1}^T \end{bmatrix} M^{-1}) \begin{bmatrix} \mathbf{h} \\ \mathbf{h} \end{bmatrix}.$$

Simple algebra yields $K_{11}M^{-1}(I_n - S(S^TM^{-1}S)^{-1}S^TM^{-1})\boldsymbol{f}_{10} = 0$. ■

S.3.2. PROOF OF LEMMA 13

Proof Under the quasi-uniform design, $X_1^{\langle 1\rangle}$,..., $X_n^{\langle n\rangle}$ are i.i.d with distribution $\omega^{\langle 1\rangle}$. Therefore, by Theorem 3 in Braun (2006), for $1 \le i \le n$ and $i \le r \le n$, simple algebra yields

$$\mathbb{P}(|\hat{\mu}_i - \mu_i| \le c_m\mu_i + \mu_r + \Lambda_r) \ge 1 - r(r+1)\exp\{-\frac{nc_m^2}{2C^4r^2}\},$$

where $\Lambda_r = \sum_{i=r+1}^{\infty}\mu_i$, C is an absolute constant, and c_m is a constant depending solely on m. Since the eigenvalue μ_i has the polynomial decay rate i^{-2m}, we have

$$\Lambda_r \asymp \sum_{i=r+1}^{\infty} i^{-2m}.$$

For $m > 1/2$,

$$\sum_{i=r+1}^{\infty} i^{-2m} \le \int_r^{\infty} x^{-2m}dx = \frac{r^{1-2m}}{2m-1} = \mathcal{O}(r^{1-2m}).$$

Let $r = n^{1/(2m-1)-\epsilon}$, we have $\Lambda_r + \mu_r = \mathcal{O}(n^{2\epsilon m-1-\epsilon}) = o(\mu_i)$ for $i = 1, \dots, n^{1/2m-\epsilon}$. Next, we have, for any $i = 1, \dots, n^{\frac{1}{2m}-\epsilon}$, the empirical eigenvalue $\hat{\mu}_i$ satisfies

$$|\hat{\mu}_i - \mu_i| \le c_m\mu_i,$$

with probability at least

$$1 - (n^{\frac{2}{2m-1}-2\epsilon} + n^{\frac{1}{2m-1}})\exp\{-cn^{\frac{2m-3}{2m-1}+2\epsilon}\}, \tag{S.4}$$

where $c = \frac{c_m^2}{2C^4}$, c_m is a constant only related to m, and M is an absolute constant. To ensure the probability in Equation (S.4) goes to 1, we further require $m > 3/2$. Thus, for $\lambda > 1/n$ and $m > 3/2$,

$$\hat{\tau}_\lambda \asymp \tau_\lambda$$

with probability at least $1 - (n^{\frac{2}{2m-1}-2\epsilon} + n^{\frac{1}{2m-1}})\exp\{-cn^{\frac{2m-3}{2m-1}+2\epsilon}\}$. ■

S.3.3. Proof of Lemma 14

Proof Considering $\mathcal{H}^{\langle 1\rangle}$ as the homogeneous Sobolev space, the kernel function $\mathcal{K}_1^{\langle 1\rangle}$ can be explicitly written as

$$\mathcal{K}_1^{\langle 1\rangle}(x,y)=2\sum_{i=1}^{\infty}\frac{cos(2\pi k(x-y))}{(2\pi k)^{2m}}.$$

Under the uniform design, we have the $X_1^{\langle 1\rangle},\ldots,X_n^{\langle 1\rangle}$ evenly distributed on $[0,1]$. Without loss of generality, we assume that $X_1^{\langle 1\rangle}<\cdots<X_n^{\langle 1\rangle}$. Therefore, the ii'th entry of kernel matrix $K_1^{\langle 1\rangle}$ is $\mathcal{K}_1^{\langle 1\rangle}(x_i^{\langle 1\rangle},x_{i'}^{\langle 1\rangle})$ which is a symmetric circulant matrix of order n (Shang and Cheng, 2017) with eigenvalues

$$\hat{\mu}_i^*=\begin{cases}\sum_{k=1}^{\infty}\frac{1}{[2\pi(kn-i)]^{2m}}+\sum_{k=0}^{\infty}\frac{1}{[2\pi(kn+i)]^{2m}} & \text{if } 1\le i\le n-1\\ 2\sum_{k=1}^{\infty}\frac{1}{(2\pi kn)^{2m}} & \text{if } i=n\end{cases}. \tag{S.5}$$

Note that $\hat{\mu}_i^*$ is a re-arrangement of $\hat{\mu}_i$. When $m>1/2$, simple calculation yields

$$\begin{aligned}\frac{1}{[2\pi(n-i)]^{2m}}+\frac{1}{(2\pi i)^{2m}}+2\underline{c}_m(2\pi n)^{-2m}\le&\hat{\mu}_i^*\\ \le&\frac{1}{[2\pi(n-i)]^{2m}}+\frac{1}{(2\pi i)^{2m}}+2\bar{c}_m(2\pi n)^{-2m},\end{aligned} \tag{S.6}$$

for $i=1,\ldots,n-1$, and

$$\hat{\mu}_n^*=2\bar{c}_m(2\pi n)^{-2m},$$

where $\underline{c}_m:=\sum_{k=1}^{\infty}k^{-2m}$, and $\bar{c}_m=\sum_{k=2}^{\infty}k^{-2m}$. By Equation (S.6), we have $\hat{\mu}_i^*\asymp\mu_i$ for $1\le i\le\frac{n}{2}$ and $\hat{\mu}_i^*\asymp\mu_{n-i}$ for $\frac{n}{2}\le i\le n$. Since $\{\hat{\mu}\}_{i=1}^n$ are obtained by ordering $\{\hat{\mu}_i^*\}_{i=1}^n$ decreasingly, we have $\mu_i\asymp\hat{\mu}_i$, and consequently,

$$\tau_\lambda\asymp\hat{\tau}_\lambda$$

for any $\lambda>0$. ■

S.3.4. Proof of Lemma 15

Proof

Note that the kernel matrix $K_1^{\langle 1\rangle}$ in Equation (A.1) has the spectral decomposition $K_1^{\langle 1\rangle}=UDU^T$, where the eigenvector matrix U is a $n\times n$ unitary matrix and the eigenvalue matrix $D=Diag\{\hat{\mu}_i\}$ is a diagonal matrix with eigenvalues $\hat{\mu}_1\ge\hat{\mu}_2\ge\cdots\ge\hat{\mu}_n$. Correspondingly, we have the following decomposition,

$$K_{11}=\frac{1}{2}\begin{bmatrix}U&0\\0&U\end{bmatrix}\begin{bmatrix}D&-D\\-D&D\end{bmatrix}\begin{bmatrix}U^T&0\\0&U^T\end{bmatrix},$$

$$M=\frac{1}{2}\begin{bmatrix}U&0\\0&U\end{bmatrix}\begin{bmatrix}D+2\lambda I_n&\theta_dD\\\theta_dD&D+2\lambda I_n\end{bmatrix}\begin{bmatrix}U^T&0\\0&U^T\end{bmatrix},$$

where I_n is the $n \times n$ identity matrix, and $\theta_d = \theta_{01} - \theta_{11}$. Letting $E = D + 2\lambda I_n = Diag\{\hat{\mu}_i + 2\lambda\}$ and $F = \theta_d D = Diag\{\theta_d \hat{\mu}_i\}$, we have

$$K_{11} M^{-1} = \begin{bmatrix} U & 0 \\ 0 & U \end{bmatrix} \begin{bmatrix} D & -D \\ -D & D \end{bmatrix} \begin{bmatrix} E & F \\ F & E \end{bmatrix}^{-1} \begin{bmatrix} U^T & 0 \\ 0 & U^T \end{bmatrix}.$$

Using the inverse of block matrix, we have

$$\begin{bmatrix} D & -D \\ -D & D \end{bmatrix} \begin{bmatrix} E & F \\ F & E \end{bmatrix}^{-1} \triangleq \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}$$

where

$$\begin{aligned} V_{11} &= DE^{-1} + (D + DE^{-1}F)(E - FE^{-1}F)^{-1} FE^{-1}, \\ V_{12} &= -(DE^{-1}F + D)(E - FE^{-1}F)^{-1}, \\ V_{21} &= -V_{12}, && \text{(S.7)} \\ V_{22} &= -V_{11}. && \text{(S.8)} \end{aligned}$$

Consequently,

$$\Delta = M^{-1} K_{11}^2 M^{-1} = \begin{bmatrix} U & 0 \\ 0 & U \end{bmatrix} \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}^T \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix} \begin{bmatrix} U^T & 0 \\ 0 & U^T \end{bmatrix}.$$

We thus have

$$\begin{aligned} \mathrm{Tr}(\Delta) = \mathrm{Tr}(M^{-1} K_{11}^2 M^{-1}) &= \mathrm{Tr}(\begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}^T \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}) \\ &= \mathrm{Tr} \begin{bmatrix} V_{11}^T V_{11} + V_{21}^T V_{21} & V_{11}^T V_{12} + V_{21}^T V_{22} \\ V_{12}^T V_{11} + V_{22}^T V_{21} & V_{12}^T V_{12} + V_{22}^T V_{22} \end{bmatrix}. \qquad \text{(S.9)} \end{aligned}$$

By Equation (S.7) and Equation (S.8), we have

$$V_{11}^T V_{11} + V_{21}^T V_{21} = V_{12}^T V_{12} + V_{22}^T V_{22}.$$

Simple algebra yields

$$V_{11}^T V_{11} + V_{21}^T V_{21} = 2 V_{11}^T V_{11}.$$

Therefore, we have

$$\mathrm{Tr}(\Delta) = 4\, \mathrm{Tr}(V_{11}^T V_{11}). \qquad \text{(S.10)}$$

Notice that D, E, F are diagonal matrices, we have

$$\mathrm{Tr}(\Delta) = 4\, \mathrm{Tr}(V_{11}^T V_{11}) \geq 4\, \mathrm{Tr}\,(D^2 E^{-2}).$$

Since

$$D^2 E^{-2} = Diag\{\frac{\hat{\mu}_i^2}{(\hat{\mu}_i + 2\lambda)^2}\},$$

we have

$$\mathrm{Tr}(\Delta) \geq 4 \sum_{i=1}^{n} \frac{\hat{\mu}_i^2}{(\hat{\mu}_i + 2\lambda)^2} \geq 4 \sum_{i=1}^{\hat{r}_\lambda} \frac{\hat{\mu}_i^2}{(\hat{\mu}_i + 2\lambda)^2}, \qquad \text{(S.11)}$$

where $\hat{\tau}_\lambda$ is the effective dimension for kernel matrix K_{11}. For the any $i < \hat{\tau}_\lambda$, we have $\frac{\hat{\mu}_i}{\hat{\mu}_i+2\lambda} > \frac{1}{3}$. Thus we have

$$\mathrm{Tr}(\Delta) \geq \frac{4}{9}\hat{\tau}_\lambda.$$

Now we shall prove the upper bound for $\mathrm{Tr}(\Delta)$. Since $\mathrm{Tr}(\Delta)$ has the expression in Equation (S.10), we expand V_{11} as

$$V_{11} = DE^{-1} + DE^{-1}(F(E-FE^{-1}F)^{-1}FE^{-1} + (E-FE^{-1})^{-1}F).$$

The ith diagonal entry of $F(E-FE^{-1}F)^{-1}FE^{-1}$ is

$$\begin{aligned} Diag_i(F(E-FE^{-1}F)^{-1}FE^{-1}) &= \frac{\theta_d^2\hat{\mu}_i^2}{(\hat{\mu}_i + 2\lambda - \frac{\theta_d^2\hat{\mu}_i^2}{\hat{\mu}_i+2\lambda})(\hat{\mu}_i+2\lambda)} \\ &\leq \frac{\theta_d^2}{1-\theta_d^2}, \end{aligned} \tag{S.12}$$

and the ith diagonal entry of $(E-FE^{-1})^{-1}F$ is

$$Diag_i((E-FE^{-1})^{-1}F) = \frac{\theta_d\hat{\mu}}{\hat{\mu}_i + 2\lambda - \frac{\theta_d^2\hat{\mu}_i^2}{\hat{\mu}_i+2\lambda}} \leq \frac{\theta_d}{1-\theta_d^2}. \tag{S.13}$$

Combining Equation (S.12) and Equation (S.13), we have the ith diagonal entry of V_{11}

$$Diag_i(V_{11}) \leq (1 + \frac{\theta_d^2}{1-\theta_d^2} + \frac{\theta_d}{1-\theta_d^2}Diag_i(DE^{-1}) = \frac{1}{1-\theta_d}Diag_i(DE^{-1}).$$

Since the lower diagonal block of DE^{-1} is identical to the upper diagonal block, we only need to bound the trace of DE^{-1}. We have

$$\begin{aligned} \mathrm{Tr}(D^2E^{-2}) =& \sum_{i=1}^{\hat{\tau}_\lambda} \frac{\hat{\mu}_i^2}{(\hat{\mu}_i+2\lambda)^2} + \sum_{i=\hat{\tau}_\lambda+1}^{n} \frac{\hat{\mu}_i^2}{(\hat{\mu}_i+2\lambda)^2} \\ \leq& \sum_{i=1}^{\hat{\tau}_\lambda} \frac{\hat{\mu}_i}{\hat{\mu}_i+2\lambda} + \sum_{i=\hat{\tau}_\lambda+1}^{n} \frac{\hat{\mu}_i}{\hat{\mu}_i+2\lambda} \\ \leq& \hat{\tau}_\lambda + \frac{1}{2\lambda}\sum_{i=\hat{\tau}_\lambda+1}^{n} \hat{\mu}_i. \end{aligned}$$

Thus we have $\mathrm{Tr}(\Delta) \leq \frac{4}{(1-\theta_d)^2}(\hat{\tau}_\lambda + \frac{1}{2\lambda}\sum_{i=\hat{\tau}_\lambda+1}^{n}\hat{\mu}_i)$. ■

S.3.5. PROOF OF LEMMA S.1

Proof By the functional decomposition in Equation (10), we have

$$||f_{10} + f_{11}||^2_{\mathcal{H}_{10}\oplus\mathcal{H}_{11}} \leq ||f||^2_{\mathcal{H}} < 1. \tag{S.14}$$

For any function g in $\mathcal{H}_{10} \oplus \mathcal{H}_{11}$, we write $g = \xi^T \tilde{\mathbf{c}} + \zeta(\cdot)$, where $\zeta(\cdot) \in \mathcal{H}_{10} \oplus \mathcal{H}_{11}$ is orthogonal to ξ. Moreover,

$$\begin{aligned} ||f_{10} + f_{11}||^2_{\mathcal{H}_{10}\oplus\mathcal{H}_{11}} =& ||\xi^T \tilde{\mathbf{c}}||^2_{\mathcal{H}_{10}\oplus\mathcal{H}_{11}} + ||\zeta(\cdot)||^2_{\mathcal{H}_{10}\oplus\mathcal{H}_{11}} \\ \geq& n\tilde{\mathbf{c}}^T R \tilde{\mathbf{c}} = \frac{1}{n}(n\tilde{\mathbf{c}}^T R)R^{-1}(nR\tilde{\mathbf{c}}) \\ =& \frac{1}{n}\boldsymbol{g}^{*T} R^{-1} \boldsymbol{g}^*. \end{aligned} \tag{S.15}$$

Combining Equation (S.14) and Equation (S.15), we have

$$\frac{1}{n}\boldsymbol{g}^{*T} R^{-1} \boldsymbol{g}^* < 1. \tag{S.16}$$

By Equation (S.1), we have

$$\begin{aligned} ||\tilde{g}^* - g^*||^2_n = & \frac{1}{n}||\boldsymbol{g}^* - RM^{-1}\boldsymbol{g}^* + RM^{-1}S(S^T M^{-1} S)^{-1} S^T M^{-1} \boldsymbol{g}^*||^2_2 \\ & = \frac{1}{n}\boldsymbol{g}^{*T}(I - RM^{-1})^2 \boldsymbol{g}^* + \frac{1}{n}||RM^{-1}S(S^T M^{-1} S^T)^{-1} S^T M^{-1}\boldsymbol{g}^*||^2_2. \end{aligned}$$

Noting that $M = R + \lambda I_n$, the eigenvalues of $I_n - R(R + \lambda I_n)^{-1}$ are all smaller than 1, and the rank of $RM^{-1}S(S^T M^{-1} S^T)^{-1} S^T M^{-1}$ is 2, we have

$$\begin{aligned} ||\tilde{g}^* - g^*||^2_n \leq & \frac{1}{n}\boldsymbol{g}^*(I - R(R + \lambda I)^{-1})\boldsymbol{g}^* + \mathcal{O}(\frac{1}{n}) \\ \leq & \lambda + \mathcal{O}(\frac{1}{n}), \end{aligned}$$

where the last inequality holds by applying Woodbury matrix identity,

$$(R + \lambda I_n)^{-1} = R^{-1} - R^{-1}(\frac{1}{\lambda}I_n + R^{-1})^{-1} R^{-1} \geq R^{-1} - \lambda R^{-2},$$

and Equation (S.16). The proof is thus completed. ■

S.3.6. PROOF OF LEMMA S.2

Proof Under the quasi-uniform design, Theorem 3.1 of Eggermont and LaRiccia (2001)(page 384, Eggermont and LaRiccia (2001)) implies that $|| \cdot ||_{\omega^{\langle 1 \rangle} mh}$ norm is equivalent to $|| \cdot ||_n$ for any fixed $x^{\langle 2 \rangle}$. The $|| \cdot ||_{\omega^{\langle 1 \rangle} hm}$, which is defined as $||f(x^{\langle 1 \rangle}, 0)||^2_{\omega^{\langle 1 \rangle} mh} = ||f(x^{\langle 1 \rangle}, 0)||^2_{L_2(\omega^{\langle 1 \rangle})} + h^{2m}||f(x^{\langle 1 \rangle}, 0)^{(m)}||^2$, trivially dominates the $|| \cdot ||_{L_2(\omega^{\langle 1 \rangle})}$ norm. Since $x^{\langle 2 \rangle}$ can only take the values 0 or 1, we have that $|| \cdot ||_n$ dominates $|| \cdot ||_2$, i.e. there exists a positive constant c such that $||f||_2 \leq c||f||_n$.

Under the uniform design, Lemma 2.27 in Eggermont and LaRiccia (2001) states that $|| \cdot ||_n$ dominates $|| \cdot ||_2$ for $x^{\langle 1 \rangle}$.

■

S.3.7. PROOF OF LEMMA S.4

Proof Under the uniform design, the empirical eigenvalues could be calculated by Equation (S.5). By the definition of $\hat{\tau}_\lambda$, we have

$$\sum_{i=\hat{\tau}_\lambda+1}^{n} \hat{\mu}_i = \sum_{\{i|\hat{\mu}_i^*<\lambda\}} \hat{\mu}_i^*. \tag{S.17}$$

Since the population eigenvalues are $\{(2\pi i)^{-2m}\}_{i=1}^{\infty}$, we calculate the population efficient dimension as $\tau_\lambda = (\lambda)^{-1/2m}/2\pi$. By the inequalities Equation (S.6), we have $\hat{\mu}_i^* \geq \lambda$ for $i = 1, \ldots, \tau_\lambda$ or $i = n - \tau_\lambda, \ldots, n$. We can bound the term in Equation (S.17)

$$\sum_{\{i|\hat{\mu}_i^*<\lambda\}} \hat{\mu}_i^* \leq \sum_{i=\tau_\lambda}^{n-\tau_\lambda} \hat{\mu}_i^*.$$

By the upper bound of $\hat{\mu}_i^*$ given in Equation (S.6), we have

$$\sum_{i=\tau_\lambda}^{n-\tau_\lambda} \hat{\mu}_i^* \leq C\tau_\lambda \mu_{\tau_\lambda},$$

which completes the proof. ■