A Powerful Global Test Statistic for Functional Statistical Inference

Jingwen Zhang,[1] **Joseph Ibrahim,**[1] **Tengfei Li,**[2] **Hongtu Zhu**[1]
[1]Department of Biostatistics, UNC Gillings School of Global Public Health
[2]Biomedical Research Imaging Center, UNC School of Medicine
University of North Carolina at Chapel Hill
jingwenz@live.unc.edu, ibrahim@bios.unc.edu
tengfeili2006@gmail.com, htzhu@email.unc.edu

Abstract

We consider the problem of performing an association test between functional data and scalar variables in a varying coefficient model setting. We propose a functional projection regression model and an associated global test statistic to aggregate relatively weak signals across the domain of functional data, while reducing the dimension. An optimal functional projection direction is selected to maximize signal-to-noise ratio with ridge penalty. Theoretically, we systematically study the asymptotic distribution of the global test statistic and provide a strategy to adaptively select the optimal tuning parameter. We use simulations to show that the proposed test outperforms all existing state-of-the-art methods in functional statistical inference. Finally, we apply the proposed testing method to the genome-wide association analysis of imaging genetic data in UK Biobank dataset.

Introduction

Functional regression modeling with a functional response $y(s), s \in \mathcal{S}$ and multivariate covariates $\boldsymbol{x} \in \mathbb{R}^p$ is a powerful statistical tool in modern high-dimensional inference, with wide applications in various biomedical studies. Specifically, functional responses frequently arise in medical imaging, computational biology and computer vision, and have been widely used to characterize brain structure and function, such as cortical complexity and white matter microstructure (Grenander and Miller 2007; Miller and Qiu 2009; Kendall et al. 2009; Srivastava and Klassen 2016; Smith et al. 2006; Smith and Nichols 2009; Huang et al. 2017). In imaging genetic studies, our primary problem of interest is to identify genetic variants ($\boldsymbol{x}$) associated with functional phenotypic variation ($y(s)$), which may ultimately lead to discoveries of risk genes contributing to neuropsychiatric and neurological disorders.

Suppose that we observe a functional response $y_i(s)$ and a set of clinical variables (e.g., age, genetic markers, and gender) $\boldsymbol{x}_i \in \mathbb{R}^p$ for n unrelated subjects. Without loss of generality, we assume $\mathcal{S} = [0, S]$ for a positive scalar S. Throughout this paper, we consider n independent observations $(y_i(s), \boldsymbol{x}_i)$ and a varying coefficient model given by

$$y_i(s) = \boldsymbol{x}_i^T \boldsymbol{\beta}(s) + \eta_i(s) + e_i(s), \quad (1)$$

where $\boldsymbol{\beta}(s)$ is a $p \times 1$ vector of functional coefficients, $\eta_i(s)$ is a function of random effect that characterizes subject-specific spatial variation, and $e_i(s)$ represents measurement error. It is assumed that $\eta_i(s)$ and $e_i(s)$ are mutually independent and identical copies of $\text{SP}\{0, \Sigma_\eta(s, s')\}$ and $\text{SP}\{0, \sigma_e^2(s) I(s = s')\}$, respectively, where $\text{SP}(\mu, \Sigma)$ denotes a stochastic process with mean function $\mu(s)$ and covariance function $\Sigma(s, s')$, and $I(\cdot)$ is the indicator function of an event. Many hypothesis testing problems of interest, such as comparison across groups, can often be formulated as a global testing problem across $\mathcal{S}$, which is given by,

$$\begin{aligned} H_0 : &\quad \mathbf{C}\boldsymbol{\beta}(s) = \boldsymbol{b}_0(s) \ \ \forall s \in \mathcal{S}, \\ H_1 : &\quad \mathbf{C}\boldsymbol{\beta}(s) \neq \boldsymbol{b}_0(s) \ \ \exists s \in \mathcal{S}, \end{aligned} \quad (2)$$

in which $\mathbf{C}$ is an $r \times p$ matrix and $\boldsymbol{b}_0(s)$ is an $r \times 1$ matrix. Without loss of generality, we center the covariates, standardize the responses, and assume $\text{rank}(\mathbf{C}) = r = 1$ and $\boldsymbol{b}_0(s) = \mathbf{0}$.

The key problem is how to design a powerful global test statistic that can efficiently aggregate weak signals across $\mathcal{S}$, while achieving high statistical power for testing problem (2). To the best of our knowledge, such problem has not been fully solved yet. We focus on a specific setting that all components in $\boldsymbol{\beta}(s)$ lie in an infinite-dimensional functional space, but p is relatively small. Existing testing methods are not powerful enough to detect moderate or weak signals due to two major challenges, (i) infinite-dimensional functional parameters and (ii) complicated covariance structure $\Sigma_\eta(s, s')$. Popular pooled global test statistics are to conduct univariate analysis at each sample grid point of $\mathcal{S}$ and then combine their results (Zhu, Li, and Kong 2012; Huang et al. 2017). However, since most of such tests ignore the correlation structure of $y_i(s)$, they may suffer from severe power loss in the presence of high correlation. Moreover, testing at each grid point individually in the mass univariate analysis requires a substantial penalty of controlling for multiplicity. The Hotelling's T^2 type test is also not well-defined for our problem of interest, since the sample estimate of Σ_η is usually non-invertible. Although dimension reduction techniques, such as principal component analysis (PCA), can be applied to reduce the dimension of functional response, most of the methods ignore the variation of covariates and their associations with responses. Thus, such methods are sub-optimal for our prob-

lem. Finally, some recent developments in regularization methods, such as multiple task learning, do not provide a post-inference tool, e.g., p-values, (Sun, Ji, and Ye 2016; Trevor, Tibshirani, and Friedman 2009).

The proposed method has three major contributions given as follows:

- A novel functional projection regression model and its associated global test statistic are introduced to aggregate relatively weak signals across $\mathcal{S}$, while reducing the dimension of functional data. An optimal functional projection direction is calculated by maximizing statistical power with ridge penalty.
- The asymptotic distribution of the global test statistic is studied systematically under both null and alternative hypotheses and a data-driven strategy is provided to adaptively select the optimal tuning parameter.
- Numerical simulations show that the proposed test outperforms all existing state-of-the-art methods in functional statistical inference.

The rest of the paper is organized as follows. We first introduce a functional projection model and the associated global test statistic for testing problem (2). Next, we derive the asymptotic distribution of the test statistic under both null and alternative hypotheses. Finally we use numerical simulations and a real data example to examine the finite sample performance of the global test statistic and conclude with some remarks.

Method

Functional Projection Regression Model

We propose a functional projection regression model as follows. Specifically, let $\omega(s)$ be a weight function in $\mathcal{L}_2(\mathcal{S})$, we define the following pseudo response $y_{w,i}$, which is a projected variable onto the functional direction $\omega(s)$ given as

$$y_{w,i} \triangleq \boldsymbol{x}_i^T \boldsymbol{\beta}_w + \eta_{w,i}(s), \tag{3}$$

where $\boldsymbol{\beta}_w = \int_{\mathcal{S}} \boldsymbol{\beta}(s)\omega(s)ds$, and $\eta_{w,i} = \int_{\mathcal{S}} \eta_i(s)\omega(s)ds$. The term associated with $e_i(s)$ in (3) would converges to 0 in probability through local kernel smoothing and therefore is asymptotically ignorable. The projected model (3) transforms the functional data into a univariate variable $y_{w,i}$. Let $\widehat{\boldsymbol{\beta}}_w$ and $\widehat{\Sigma}_\eta(s,s')$ be the estimates of $\boldsymbol{\beta}_w$ and $\Sigma_\eta(s,s')$ respectively, then a standard Wald-type statistic for testing problem (2) can be given by

$$T_n(\omega) = \frac{\widehat{\boldsymbol{\beta}}_w^T \mathbf{C}^T[\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T]^{-1}\mathbf{C}\widehat{\boldsymbol{\beta}}_w}{\iint \widehat{\Sigma}_\eta(s,s')\omega(s)\omega(s')dsds'}. \tag{4}$$

For a given projection direction of $\omega(s)$, we need to estimate $\widehat{\boldsymbol{\beta}}(s)$ and $\widehat{\Sigma}_\eta(s,s')$ in order to calculate $T_n(\omega)$. In real data, functional responses $\{y_i(s)\}_{i=1}^n$ are usually observed on a set of M discrete sample points on $\mathcal{S}$, which is denoted as $\widehat{\mathcal{S}} = \{s_1, \cdots, s_m, \cdots, s_M\}$. To estimate $\boldsymbol{\beta}(s)$, we use a weighted least square (WLS) method based on the local polynomial kernel (LPK) smoothing technique (Fan and Gijbels 2018; Zhu, Li, and Kong 2012). Let $K(s)$ be a predetermined smoothing kernel on $[-1,1]$ and let $K_h(s) = h^{-1}K(s/h)$ for bandwidth h. A smooth estimate of $\boldsymbol{\beta}(s)$ can be given as the minimizers of the following weighted least square function given by

$$\widehat{\boldsymbol{\beta}}_{h_1}(s) = \underset{\boldsymbol{\beta}(s)}{\operatorname{argmin}} \sum_{i=1}^n K_{h_1}[y_i(s), \boldsymbol{x}_i^T\boldsymbol{\beta}(s)|\widehat{\mathcal{S}}], \tag{5}$$

where $K_{h_1}[y_i(s), \boldsymbol{x}_i^T\boldsymbol{\beta}(s)|\widehat{\mathcal{S}}]$ is defined as

$$\sum_{m=1}^M [y_i(s_m) - \boldsymbol{x}_i^T\boldsymbol{\beta}(s)]^2 K_{h_1}(s_m - s). \tag{6}$$

Similarly, each random function $\eta_i(s)$ can also be estimated by

$$\widehat{\eta}_{i,h_2}(s) = \underset{\eta_i(s)}{\operatorname{argmin}} K_{h_2}[y_i(s) - \boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}_{h_1}(s), \eta_i(s)|\widehat{\mathcal{S}}], \tag{7}$$

where $K_{h_2}[y_i(s) - \boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}_{h_1}(s), \eta_i(s)|\widehat{\mathcal{S}}]$ is defined as

$$\sum_{m=1}^M [y_i(s_m) - \boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}_{h_1}(s_m) - \eta_i(s)]^2 K_{h_2}(s_m - s). \tag{8}$$

With $\{\widehat{\eta}_{i,h_2}(s)\}_{i=1}^n$, we can obtain a consistent estimate of $\boldsymbol{\Sigma}_\eta(s,s')$ as

$$\widehat{\boldsymbol{\Sigma}}_\eta(s,s') = \frac{1}{n}\sum_{i=1}^n \widehat{\eta}_{i,h_2}(s)\widehat{\eta}_{i,h_2}(s'). \tag{9}$$

Finally, we address the problem of determining $\omega(s)$ in order to achieve optimal power. Specifically, we consider the signal-to-noise ratio of test statistics $T_n(\omega)$, which dominates the asymptotic power, as follows:

$$L_1(\omega) = \frac{\boldsymbol{\beta}_w^T\mathbf{C}^T[\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T]^{-1}\mathbf{C}\boldsymbol{\beta}_w}{\iint \boldsymbol{\Sigma}_\eta(s,s')\omega(s)\omega(s')dsds'}. \tag{10}$$

An optimal projection direction would be the maximizer of $L_1(\omega)$. However, when plugging in the estimates of $\boldsymbol{\beta}(s)$ and $\boldsymbol{\Sigma}_\eta(s,s')$, directly maximizing $L_1(\omega)$ can be an ill-conditioned problem. The eigenvalues of $\widehat{\boldsymbol{\Sigma}}_\eta(s,s')$ usually decrease to zero very fast and the maximum value of $L_1(w)$ tend to be ∞. To solve this issue, we add a ridge penalty term to the denominator in (10). Given a positive tuning parameter λ, the optimal projection direction $\widehat{\omega}_\lambda(s)$ can be estimated as

$$\underset{\omega(s)}{\operatorname{argmax}} \frac{\widehat{\boldsymbol{\beta}}_{w,h_1}^T\mathbf{C}^T[\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T]^{-1}\mathbf{C}\widehat{\boldsymbol{\beta}}_{w,h_1}}{\iint \widehat{\boldsymbol{\Sigma}}_\eta(s,s')\omega(s)\omega(s')dsds' + \lambda\|\omega(s)\|_2^2}, \tag{11}$$

where $\|\omega(s)\|_2^2 = \int_{\mathcal{S}} \omega^2(s)ds$ is the L_2-norm.

For a given λ, we calculate $\widehat{\omega}_\lambda(\cdot)$ as follows. Let $\{\widehat{\tau}_k\}_{k=1}^{+\infty}$ be the eigenvalues of $\widehat{\Sigma}_\eta(s,s')$ in a decreasing order and let $\{\widehat{\phi}_k(s)\}_{k=1}^{+\infty}$ be the corresponding eigenfunctions. We further assume that $\omega(s) \in span\{\phi_k(s)\}_{k=1}^{+\infty}$ such that

$\omega(s) = \sum_{k=1}^{+\infty} w_k \phi_k(s)$. In that case, we can seek solution $\widehat{\omega}_\lambda(s)$ in the space spanned by $\{\widehat{\phi}_k(s)\}_{k=1}^{+\infty}$, which is given by

$$\begin{aligned} \widehat{\mathbf{w}}_\lambda &= (\widehat{w}_{1,\lambda}, \cdots, \widehat{w}_{k,\lambda}, \cdots) \\ &= \underset{w_1,\cdots,w_k,\cdots}{\operatorname{argmax}} \frac{[\sum_{k=1}^{+\infty} \widehat{d}_{k,h_1} w_k]^2}{\sum_{k=1}^{+\infty} w_k^2(\widehat{\tau}_k + \lambda)}, \end{aligned} \tag{12}$$

where

$$\begin{aligned} \widehat{w}_{k,\lambda} &= \int_0^S \widehat{\omega}_\lambda(s)\widehat{\phi}_k(s)ds \text{ and} \\ \widehat{d}_{k,h_1} &= \int_0^S \mathbf{C}\widehat{\boldsymbol{\beta}}_{h_1}(s)\widehat{\phi}_k(s)ds \end{aligned}$$

are the projections of functional direction $\widehat{\omega}_\lambda(s)$ and estimated signal $\mathbf{C}\widehat{\boldsymbol{\beta}}_{h_1}(s)$ on the estimated eigenfunctions $\{\widehat{\phi}_k(s)\}_{k=1}^{\infty}$ respectively. The solutions to (12) can be explicitly expressed as,

$$\widehat{w}_{k,\lambda} = \widehat{d}_{k,h_1}/(\widehat{\tau}_k + \lambda). \tag{13}$$

Finally, we obtain a global test statistic based on the optimal projection direction $\widehat{\omega}_\lambda(s) = \sum_{k=1}^{+\infty} \widehat{w}_{k,\lambda}\widehat{\phi}_k(s)$ as follows:

$$T_n(\widehat{\omega}_\lambda) = \frac{(\sum_{k=1}^{+\infty} \widehat{d}_{k,h_1}\widehat{w}_{k,\lambda})^2}{[\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T]\sum_{k=1}^{+\infty}\widehat{\tau}_k\widehat{w}_{k,\lambda}^2}. \tag{14}$$

An unsolved question is how to choose the tuning parameter λ, which will be answered in Section 3. To approximate the distribution of $T_n(\widehat{\omega}_\lambda)$ under H_0, we adopt a wild-boostrap procedure described as follows.

Algorithm 1.

(a) Fit the varying coefficient model under the null hypothesis and get the estimate of $\widehat{\beta}_0(s)$ and $\{\widehat{\eta}_{i,0}(s), \widehat{e}_{i,0}(s)\}_{i=1}^n$.

(b) For $g = 1, \cdots, G$, generate independent random numbers $\nu_i^{(g)}$ and $\nu_i^{(g)}(s_m)$ from $N(0,1)$, and the wild bootstrap sample on each grid point can be calculated as

$$\widehat{y}_i^{(g)}(s_m) = \widehat{\beta}_0(s_m)^T\mathbf{x}_i + \nu_i^{(g)}\widehat{\eta}_{i,0}(s_m) + \nu_i^{(g)}(s_m)\widehat{e}_{i,0}(s_m).$$

(c) Repeat the testing procedure and obtain G samples of $T_n(\widehat{\omega}_\lambda^{(g)})$ under the null hypothesis.

(d) The p-value is approximated by

$$p = G^{-1}\sum_{g=1}^{G} I\{T_n(\widehat{w}_\lambda) \geq T_n(\widehat{\omega}_\lambda^{(g)})\}.$$

Approximation of the null distribution requires repeated calculation of the estimation-test procedure by G times, and G should be large enough to guarantee accuracy.

Theoretical Result

In this section, we study the asymptotic distribution of the proposed test statistic $T_n(\widehat{\omega}_\lambda)$ for fixed λ and consider the problem of determining the tuning parameter λ for optimally testing (2).

Assumptions

Throughout the paper, the following assumptions are used to facilitate the technical details. Some of the assumptions might be weakened but the current version simplifies the proof.

Assumption 1. *Smoothing kernel $K(u)$ is a symmetric positive function with compact support $[-1,1]$ and upper bound c_1. Moreover, $K(u)$ has continuous first order derivative satisfying $\sup_u |\dot{K}(u)| < c_2 < +\infty$.*

Assumption 2. *Variable of interest $\boldsymbol{x}_i$ are identically and independently distributed variables with mean μ_x and positive definite covariance $\boldsymbol{\Sigma}_x$. And $\|\boldsymbol{x}_i\|_\infty < c_3 < +\infty$.*

Assumption 3. *Sample grid point set $\widehat{\mathcal{S}}$ is composed of M equidistant points on $[0, S]$.*

Assumption 4. *Fixed effects $\boldsymbol{\beta}(s)$ are continuous functions in $C^1[0,S]$ with universally bounded first order derivatives, i.e., $\sup_s \|\dot{\boldsymbol{\beta}}(s)\|_\infty < c_4 < +\infty$.*

Assumption 5. *Random functions $\{\eta_i(s)\}_{i=1}^n$ are identically and independently distributed copies from gaussian process and the sample path has continuous first-order derivative on $[0,S]$. Additionally, it is assumed that $\dot{\eta}_i(s)$ is from gaussian process and its covariance function has continuous first-order derivatives, i.e., $\boldsymbol{\Sigma}_{\dot{\eta}}(s,t) \in C^1[0,S]^2$.*

Assumption 6. *Error terms $\{e_i(s)\}_{i=1}^n$ are from a universally upper bounded process, that is, $\sup_s |e_i(s)| < c_7 < +\infty$.*

Assumption 7. *Let $\Sigma_\eta(s,s') = \sum_{k=1}^{+\infty}\tau_k\phi_k(s)\phi_k(s')$ be the spectral expansion of $\Sigma_\eta(s,s')$, in which $\tau_1 > \cdots > \tau_k > \cdots \geq 0$ are eigenvalues in decreasing order. It is assumed that all eigenvalues have simple multiplicity that satisfy*

$$\min_k \min_{j\neq k} |\tau_j - \tau_k|/\tau_k > \epsilon_0 > 0.$$

Let $\{\lambda_n\}$ be a sequence of tuning parameters that satisfy $\lambda_n \to 0$ as $n \to +\infty$, we further assume that one of the two conditions holds,

(i) $\{\tau_k\}_{k=1}^{+\infty}$ follows polynomial decay rate, i.e. $\tau_k \asymp k^{-r}$ with $r > 1$, it is assumed that $\lambda_n^{1-\frac{1}{r}}Mh_1 \to +\infty$ and $\lambda_n^{3-\frac{1}{r}}\min\{h_1^{-2}, h_2^{-2}, Mh_2, n/\log n\} \to +\infty$.

(ii) $\{\tau_k\}_{k=1}^{+\infty}$ follows exponential decay rate, i.e. $\tau_k \asymp \alpha^{-k}$ with $\alpha > 1$, it is assumed that $Mh_1\lambda_n \log\lambda_n^{-1} \to +\infty$ and $\lambda_n^3 \log\lambda_n^{-1}\min\{h_1^{-2}, h_2^{-2}, Mh_2, n/\log n\} \to +\infty$.

Assumption 8 (Local Alternative Hypothesis)**.** *A sequence of local alternative hypotheses $H_{1n}: \mathbf{C}\boldsymbol{\beta}(s) = n^{-1/2}d_0(s)$ is satisfied as $n \to +\infty$, where $d_0(s) \in C^1[0,S] \cap span\{\phi_k(s)\}_{k=1}^{+\infty}$.*

Assumptions 1-6 are standard conditions in functional data analysis (Zhu, Li, and Kong 2012; Hall, Müller, and Wang 2006; Li and Hsing 2010), which are required to guarantee that the estimates of $\boldsymbol{\beta}(s)$ and $\Sigma_\eta(s,s')$ are consistent. Assumption 7 is required in order to specify the bound of tuning parameter λ_n for two different decay rates of $\{\tau_k\}_{k=1}^{+\infty}$. These two decay rates are the most commonly used conditions in the literature of functional data analysis (Qu and Wang 2017;

Yuan and Cai 2010; Wang and Ruppert 2015). Here, we only consider distinct eigenvalues. And the distance between one eigenvalue and any other eigenvalues can not be too large compared to itself. Conclusions for multiplicity greater than one could be reached, yet is beyond the discussion of this paper. Assumption 8 specified a sequence of local alternative hypotheses from which we will derive the asymptotic power.

Main Theoretical Results

We present the key results below according to different decay rates of $\{\tau_k\}_{k=1}^{+\infty}$. The proof of the theorem is given in https://app.box.com/v/aaai19PFGT.

Theorem 1. *When Assumptions 1 - 6 and 7(i) (or 7(ii)) hold, as $n, M \to +\infty$, $T_n(\widehat{\omega}_{\lambda_n})$ has the following asymptotic normal distribution under the null hypothesis,*

$$T_n(\widehat{\omega}_{\lambda_n}) \quad \xrightarrow{d} \quad N\{\mu_0, \sigma_0^2\}, \tag{15}$$

where $\xrightarrow{d}$ denotes convergence in distribution and μ_0 and σ_0^2 are given by,

$$\mu_0 = \frac{a_1^2}{a_2} \quad \textit{and} \quad \sigma_0^2 = \frac{8a_1^2}{a_2} + \frac{2a_1^4 a_4}{a_2^4} - \frac{8a_1^3 a_3}{a_2^3}, \tag{16}$$

in which a_1, a_2, a_3, and a_4 are defined as,

$$\begin{aligned} a_1 &= \sum_{k=1}^{+\infty} \frac{\tau_k}{\tau_k + \lambda_n}, \quad a_2 = \sum_{k=1}^{+\infty} (\frac{\tau_k}{\tau_k + \lambda_n})^2, \\ a_3 &= \sum_{k=1}^{+\infty} (\frac{\tau_k}{\tau_k + \lambda_n})^3, \ \ a_4 = \sum_{k=1}^{+\infty} (\frac{\tau_k}{\tau_k + \lambda_n})^4. \end{aligned}$$

When Assumptions 1 - 6 and 7(i) (or 7(ii)) hold, and the local alternative hypothesis specified by Assumption 8 is satisfied, $T_n(\widehat{\omega}_{\lambda_n})$ has the following asymptotic normal distribution given by

$$T_n(\widehat{\omega}_{\lambda_n}) \quad \xrightarrow{d} \quad N\{\mu_1, \sigma_1^2\}, \tag{17}$$

where μ_1 and σ_1^2 are defined as,

$$\begin{aligned} \mu_1 &= \frac{(a_1 + d_1)^2}{a_2 + d_2}, \\ \sigma_1^2 &= \frac{8(a_1 + d_1)^2 (a_2 + 2d_2)}{(a_2 + d_2)^2} \\ &+ \frac{2(a_1 + d_1)^4 (a_4 + 2d_4)}{(a_2 + d_2)^4} \\ &- \frac{8(a_1 + d_1)^3 (a_3 + 2d_3)}{(a_2 + d_2)^3}. \end{aligned} \tag{18}$$

In the above equation, d_1, d_2, d_3, and d_4 are defined as

$$\begin{aligned} d_1 &= \sum_{k=1}^{+\infty} \frac{\delta_{k,0}^2}{\sigma_c^2(\tau_k + \lambda_n)}, \ \ d_2 = \sum_{k=1}^{+\infty} \frac{\tau_k \delta_{k,0}^2}{\sigma_c^2(\tau_k + \lambda_n)^2}, \\ d_3 &= \sum_{k=1}^{+\infty} \frac{\tau_k^2 \delta_{k,0}^2}{\sigma_c^2(\tau_k + \lambda_n)^3}, \ \ d_4 = \sum_{k=1}^{+\infty} \frac{\tau_k^3 \delta_{k,0}^2}{\sigma_c^2(\tau_k + \lambda_n)^4}, \end{aligned}$$

where $\delta_{0,k} = \int_0^S d_0(s)\phi_k(s)ds$ and $\sigma_c^2 = \mathbf{C}\mathbf{\Sigma}_x^{-1}\mathbf{C}^T$.

Theorem 1 establishes the asymptotic distribution of the proposed test statistics for polynomial and exponential decay rates under both null and alternative hypotheses. It inspires a data-driven criterion to adaptively select tuning parameter λ_n in order to achieve optimal power. Specifically, we choose

$$\widehat{\lambda}_n = \text{argmax}[\widehat{\mu}_1/\widehat{\sigma}_1 - \widehat{\mu}_0/\widehat{\sigma}_0], \tag{19}$$

in which $\widehat{\mu}_0, \widehat{\mu}_1, \widehat{\sigma}_0, \widehat{\sigma}_1$ are calculated by plugging in their corresponding estimates. Intuitively, (19) tends to maximize the asymptotic power of $T_n(\widehat{\omega}_{\lambda_n})$ under alternative hypothesis.

Numerical Simulation

Setup

In this section, we use numerical simulations to evaluate the finite-sample performance of the proposed global test statistic. Data was generated from the following model

$$y_i(s) = \beta_0(s) + x_{i,1}\beta_1(s) + \eta_i(s) + e_i(s), \ \ i = 1, \cdots n,$$

where $x_{i,1} \sim N(0, 1)$. We set $n = 200$ and $S = 1$ and put the number of grid points $M = 100$ evenly in $[0, 1]$. Our hypothesis of interest is to test the following problem:

$$\begin{aligned} H_0 &: \beta_1(s) = 0, \forall s \in [0, 1], \\ H_1 &: \beta_1(s) \neq 0, \exists s \in [0, 1]. \end{aligned}$$

In this experiment, we simulated $\beta_1(s)$ as a spatially heterogenous function under the alternative hypothesis. Other model parameters were estimated from the UK Biobank dataset introduced in Application Section. We considered two decay rates of $\{\tau_k\}_{k=1}^{+\infty}$ including a polynomial decay rate with $\tau_k = k^{-3/2}$ and an exponential decay rate given by $\tau_k = 0.75^k$. The signal-to-noise ratios under alternative hypothesis are shown in Figure 1 (a)-(b) and the structure of the covariance functions are presented in Fig 1 (c)-(d). As can be seen, responses in the polynomial decay case have stronger spatial correlation than those in the exponential decay case. For the choice of the tuning parameter, we considered both fixed quantities where $\log \lambda_n$ takes values from $[-2, 0]$ with an equal increment of 0.1 (PFGT-λ_n) and an optimal $\widehat{\lambda}_n$ selected by (19) in each run (PFGT-optimal). As a comparison, we considered two state-of-the-art methods in functional statistical inference, including FADTTS (Zhu, Li, and Kong 2012) and FLMtest (Zhang 2011). In each scenario, 1,000 simulation replicates were generated to evaluate type I and type II error rates respectively. To calculate p-values, $G = 1,000$ wild-bootstrap samples were generated in each run.

Results

Simulation results are summarized in Figure 2. The rejection rates under null hypotheses are shown in Figure 2 (a) - (b). As can be seen, FADTTS controls type I error rates well. Although our global test has slightly inflated false positive rate as λ_n is relatively large, but the optimal $\widehat{\lambda}_n$ does not have such problem. For FLMtest, type I error is slightly inflated. Under the alternative hypothesis, PFGT substantially outperforms FADTTS and FLMtest for both the polynomial

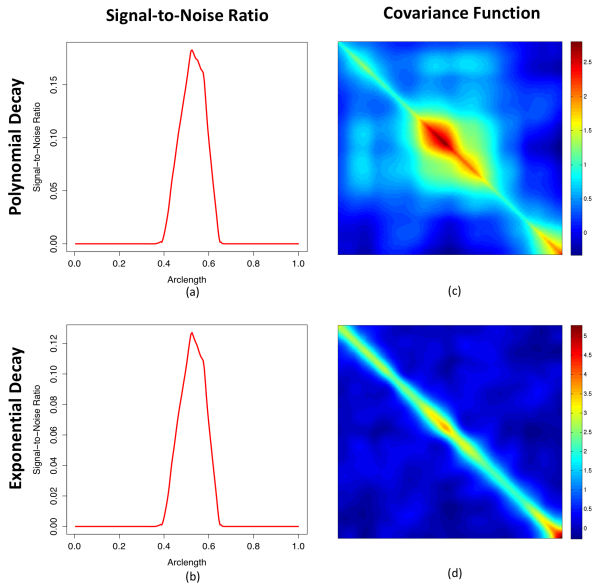

Figure 1: Simulation settings: Panels (a)-(b) demonstrate the signal-to-noise ratios under alternative hypothesis. Panels (c)-(d) visualize the covariance function of simulated responses along the curve.

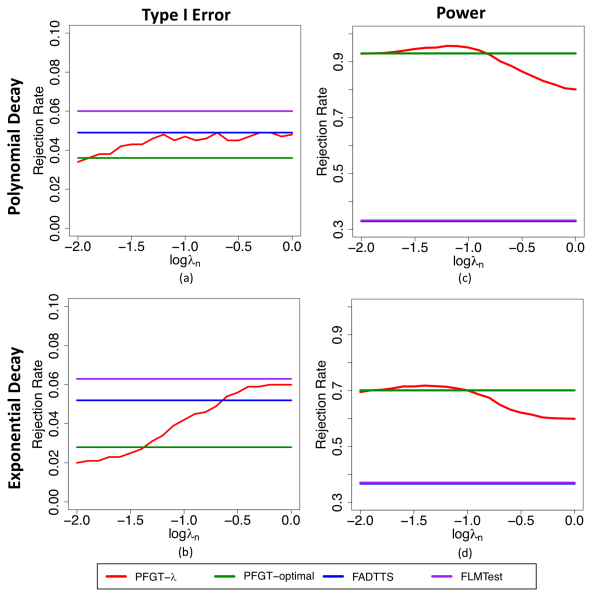

Figure 2: Simulation results: Panels (a)-(b) present the type I error for PFGT-λ_n, PFGT-optimal, FADTTS and FLMTest. Panels (c)-(d) present the power under alternative hypothesis.

decay rate and the exponential decay rate. In addition, the proposed test using adaptive tuning parameter $\widehat{\lambda}_n$ achieves almost the best performance given by PFGT using fixed λ_n, which indicates the effectiveness of the data-driven strategy to choose λ_n. Moreover, PFGT shows larger improvement in the polynomial decay case than in the exponential decay case compared with FADTTS and FLMtest. It is suggested that the proposed method has better performance in the presence of stronger spatial correlation.

Application: UK Biobank Data Analysis

UK Biobank Study

UK Biobank is a large-scale cohort in the United Kingdom designed to investigate the influences of genetic susceptibility, environmental exposures and lifestyle factors to a wide range of health-related outcomes and disorders in middle aged and elderly population. In this section, we perform a genome-wide association analysis on the functional neuroimaging phenotypes from this study.

Diffusion weighted images (DWI) were acquired for 8751 subjects in total. We ran the TBSS-ENIGMA pipeline (McMahon and Thompson 2017) on DWIs with the FSL tool set (Jenkinson et al. 2012) to perform quality control and registration. The ENIGMA skeleton was then projected onto the registered FA images and FA statistics on 26,334 voxels from 21 regions of interest (ROIs) were obtained. The primary phenotype of interest is the distributional density of voxel-wise FA statistics of the whole brain. As the density function is constrained by the normalization condition, we applied a log quantile density transformation introduced in (Petersen and Müller 2016) and took the output as the functional phenotypes for further analysis.

The Affymetrix Axiom platform was used to genotype 8057 subjects from the full population with imaging data, which resulted in a set of 784,256 single-nucleotide polymorphism (SNPs). The genotype data were preprocessed by standard quality control steps and SNPs with minor allele frequency (MAF) less than 1% were removed. Eventually, 459,588 SNPs were remained in the dataset for further analysis.

Statistical Analysis and Results

Our problem of interest is to perform a genome-wide association analysis on the log quantile curve of the whole brain FA measure. We fitted model (1) with covariates including an intercept term, a specific SNP, age, gender, and the top 5 genetic principal components. We developed a computationally efficient strategy to approximate the p-values of SNPs with different MAFs. For each MAF category, we generated 10,000 bootstrap samples and adopted a mixed chis-quare approximation (Zhang 2005) to approximate the null distribution of the test statistic. The histograms and the QQ-plots for a fixed λ_n are, respectively, presented in Figure 3 and Figure 4 as an example. Other λ_n choices showed very similar pattern. The mixed chi-square approximation works reasonably well for a wide range of MAFs. To obtain a single p-value, we chose the optimal λ_n from (19) for each SNP.

We present the Manhattan plot and the QQ plot of the GWAS results in Figure 5. The top 10 loci along with their p-values are summarized in Table 1. As can be seen, no genome-wide significant marker (p-value $< 1.08 \times 10^{-7}$) is observed. Additionally, five loci exceed the suggestive genome-wide association threshold (p-value $< 5 \times 10^{-6}$). Among the top genes, CAMK2N1 plays an important role in long-term potentiation, which is a process closely related to learning and memory (Lisman, Schulman, and Cline 2002).

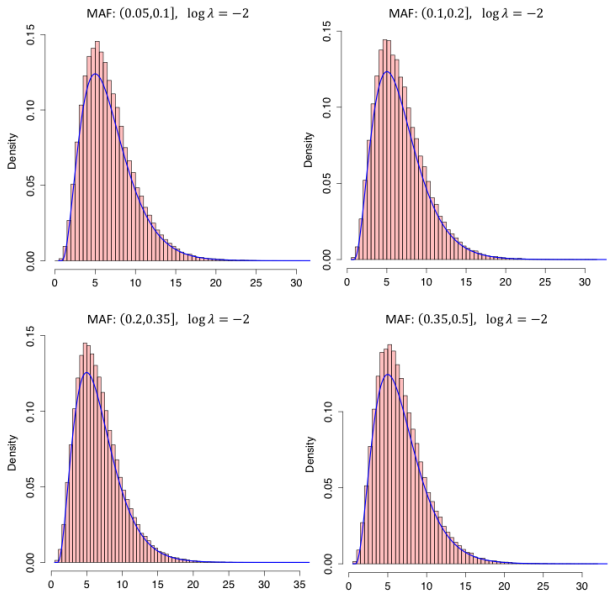

Figure 3: Histograms of wild bootstrap statistics for different MAF intervals with fixed $\lambda_n = 10^{-2}$, along with their density approximations by mixed chis-quare distribution.

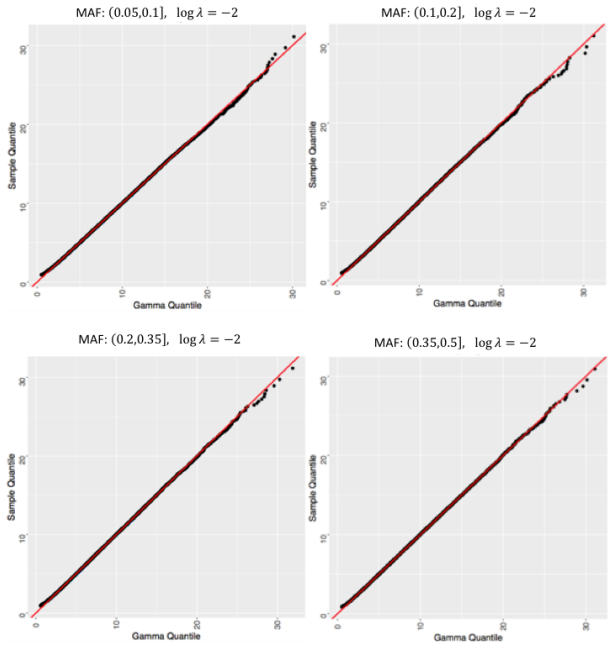

Figure 4: QQ Plots of wild bootstrap statistics for different MAF intervals for fixed $\lambda_n = 10^{-2}$.

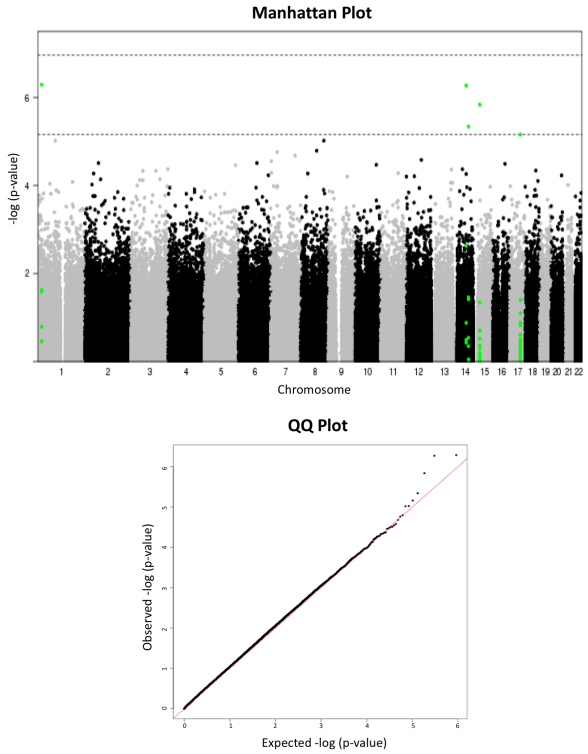

Figure 5: Visualization of GWAS results: Manhattan Plot and QQ Plot of the p-values of 450,899 SNPs.

ZFP36L1, CEP128, HAS2 and EVI5 are risk genes implicated by certain neurodegenerative diseases (Yuan et al. 2013; Perga et al. 2015; Mowry et al. 2013; Le-Niculescu et al. 2009). MSI2 gene is known to be related to the proliferation and maintenance of stem cells in the central nervous system (Sakakibara et al. 2001).

Conclusion

We proposed a powerful functional global testing framework (PFGT) to perform statistical inference on the varying coefficient model. The asymptotic distribution of the test statistic has been systematically studied and we provided a strategy to adaptively select the optimal tuning parameter in order to maximize the testing power.

As a continuation of this paper, it is interesting and important to investigate optimal testing procedures for other statistical inference problems of parametric and nonparametric models using dimension reduction techiniques and power maximization framework, for example, inference on the transformed measurements (Zhou et al. 2016), test of distributional differences (Jitkrittum et al. 2016), test of independence (Heller and Heller 2016; Sen et al. 2017), test of goodness-of-fit (Jitkrittum et al. 2017) and many others (Liu and Coull 2017; Cecchi and Hegde 2017).

Table 1: Top 10 SNPs from GWAS and their nearest genes

SNP	Chr	p-value	Gene
rs6663450	1	5.15E-07	CAMK2N1
rs11158764	14	5.37E-07	ZFP36L1
rs2339157	15	1.45E-06	FMN1
rs143406098	14	3.87E-06	CEP128
rs17821769	17	4.93E-06	MSI2
rs79320696	8	9.47E-06	HAS2
rs72722496	1	9.61E-06	EVI5
rs893282	8	1.61E-05	RALYL
rs73086843	7	1.74E-05	HERPUD2
s55783991	7	2.10E-05	CPA4

Acknowledgments

Dr. Zhu's research was partially supported by the National Institutes of Health grants MH086633 and MH116527.

References

Cecchi, F., and Hegde, N. 2017. Adaptive active hypothesis testing under limited information. In *Advances in Neural Information Processing Systems*, 4038–4046.

Fan, J., and Gijbels, I. 2018. *Local polynomial modelling and its applications: monographs on statistics and applied probability*, volume 66. Routledge.

Grenander, U., and Miller, M. I. 2007. *Pattern Theory From Representation to Inference*. Oxford University Press.

Hall, P.; Müller, H. G.; and Wang, J. L. 2006. Properties of principal component methods for functional and longitudinal data analysis. *The Annals of Statistics* 1493–1517.

Heller, R., and Heller, Y. 2016. Multivariate tests of association based on univariate tests. In *Advances in Neural Information Processing Systems*, 208–216.

Huang, C.; Thompson, P.; Wang, Y.; Yu, Y.; Zhang, J.; Kong, D.; Colen, R. R.; Knickmeyer, R. C.; and Zhu, H. 2017. Fgwas: Functional genome wide association analysis. *NeuroImage* 159:107–121.

Jenkinson, M.; Beckmann, C. F.; Behrens, T. E.; Woolrich, M. W.; and Smith, S. M. 2012. Fsl. *NeuroImage* 62(2):782–790.

Jitkrittum, W.; Szabó, Z.; Chwialkowski, K. P.; and Gretton, A. 2016. Interpretable distribution features with maximum testing power. In *Advances in Neural Information Processing Systems*, 181–189.

Jitkrittum, W.; Xu, W.; Szabó, Z.; Fukumizu, K.; and Gretton, A. 2017. A linear-time kernel goodness-of-fit test. In *Advances in Neural Information Processing Systems*, 261–270.

Kendall, D. G.; Barden, D.; Carne, T. K.; and Le, H. 2009. *Shape and shape theory*, volume 500. John Wiley & Sons.

Le-Niculescu, H.; Patel, S. D.; Bhat, M.; Kuczenski, R.; Faraone, S. V.; Tsuang, M. T.; McMahon, F. J.; Schork, N. J.; Nurnberger Jr, J. I.; and Niculescu Iii, A. B. 2009. Convergent functional genomics of genome-wide association data for bipolar disorder: Comprehensive identification of candidate genes, pathways and mechanisms. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* 150(2):155–181.

Li, Y., and Hsing, T. 2010. Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *The Annals of Statistics* 38(6):3321–3351.

Lisman, J.; Schulman, H.; and Cline, H. 2002. The molecular basis of camkii function in synaptic and behavioural memory. *Nature Reviews Neuroscience* 3(3):175.

Liu, J., and Coull, B. 2017. Robust hypothesis test for nonlinear effect with gaussian processes. In *Advances in Neural Information Processing Systems*, 795–803.

McMahon, M. A. B., and Thompson, P. M. 2017. Enhancing neuroimaging genetics through meta analysis: global collaborations in psychiatry by the enigma consortium. *European Neuropsychopharmacology* 27:S715.

Miller, M. I., and Qiu, A. 2009. The emerging discipline of computational functional anatomy. *NeuroImage* 45:S16–S39.

Mowry, E. M.; Carey, R. F.; Blasco, M. R.; Pelletier, J.; Duquette, P.; Villoslada, P.; Malikova, I.; Roger, E.; Kinkel, R. P.; and McDonald, J. 2013. Multiple sclerosis susceptibility genes: associations with relapse severity and recovery. *PloS one* 8(10):e75416.

Perga, S.; Montarolo, F.; Martire, S.; Berchialla, P.; Malucchi, S.; and Bertolotto, A. 2015. Anti-inflammatory genes associated with multiple sclerosis: a gene expression study. *Journal of neuroimmunology* 279:75–78.

Petersen, A., and Müller, H. G. 2016. Functional data analysis for density functions by transformation to a hilbert space. *The Annals of Statistics* 44(1):183–218.

Qu, S., and Wang, X. 2017. Optimal global test for functional regression. *arXiv preprint arXiv:1710.02269*.

Sakakibara, S. I.; Nakamura, Y.; Satoh, H.; and Okano, H. 2001. Rna-binding protein musashi2: developmentally regulated expression in neural precursor cells and subpopulations of neurons in mammalian cns. *Journal of Neuroscience* 21(20):8091–8107.

Sen, R.; Suresh, A. T.; Shanmugam, K.; Dimakis, A. G.; and Shakkottai, S. 2017. Model-powered conditional independence test. In *Advances in Neural Information Processing Systems*, 2955–2965.

Smith, S. M., and Nichols, T. E. 2009. Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage* 44:83–98.

Smith, S. M.; Jenkinson, M.; Johansen-Berg, H.; Rueckert, D.; Nichols, T. E.; Mackay, C. E.; Watkins, K. E.; Ciccarelli, O.; Cader, M. Z.; Matthews, P. M.; and Behrens, T. E. 2006. Tract-based spatial statistics: voxelwise analysis of multi-subject diffusion data. *NeuroImage* 31:1487–1505.

Srivastava, A., and Klassen, E. P. 2016. *Functional and shape data analysis*. Springer.

Sun, L.; Ji, S.; and Ye, J. 2016. *Multi-label dimensionality reduction*. Chapman and Hall/CRC.

Trevor, H.; Tibshirani, R.; and Friedman, J. H. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd)*. Hoboken, New Jersey.: Springer.

Wang, X., and Ruppert, D. 2015. Optimal prediction in an additive functional model. *Statistica Sinica* 567–589.

Yuan, M., and Cai, T. T. 2010. A reproducing kernel hilbert space approach to functional linear regression. *The Annals of Statistics* 38(6):3412–3444.

Yuan, Y.; Tang, B.; Yu, R.; Li, K.; Lv, Z.; Yan, X.; and Guo, J. 2013. Marginal association between snp rs2046571 of the has2 gene and parkinson's disease in the chinese female population. *Neuroscience letters* 552:58–61.

Zhang, J. T. 2005. Approximate and asymptotic distributions of chi-squared–type mixtures with applications. *Journal of the American Statistical Association* 100(469):273–285.

Zhang, J. T. 2011. Statistical inferences for linear models with functional responses. *Statistica Sinica* 1431–1451.

Zhou, H.; Ithapu, V. K.; Ravi, S. N.; Singh, V.; Wahba, G.; and Johnson, S. C. 2016. Hypothesis testing in unsupervised domain adaptation with applications in alzheimer's disease. In *Advances in Neural Information Processing Systems*, 2496.

Zhu, H.; Li, R.; and Kong, L. 2012. Multivariate varying coefficient model for functional responses. *The Annals of Statistics* 40(5):2634.