Statistica Sinica Preprint No: SS-2022-0309	
Title	Hypotheses Testing of Functional Principal Components
Manuscript ID	SS-2022-0309
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202022.0309
Complete List of Authors	Zening Song, Lijian Yang and Yuanyuan Zhang
Corresponding Authors	Lijian Yang
E-mails	yanglijian@tsinghua.edu.cn
Notice: Accepted version subject to English editing.	

Statistica Sinica

HYPOTHESES TESTING OF FUNCTIONAL PRINCIPAL COMPONENTS

Zening Song[1], Lijian Yang[1*], Yuanyuan Zhang[2]

[1] *Tsinghua University,* [2] *Soochow University*

Abstract: We propose a test for the hypothesis that the standardized functional principal components (FPCs) of a functional data equal a given set of orthonormal basis (e.g., the Fourier basis). Using estimates of individual trajectories that satisfy certain approximation conditions, a chi-square type statistic is constructed and shown to be oracally efficient under the null hypothesis in the sense that its limiting distribution is the same as an infeasible statistic using all trajectories, known by "oracle". The null limiting distribution is an infinite Gaussian quadratic form, and a consistent estimator of its quantile is obtained. A test statistic based on the chi-square type statistic and approximate quantile of the Gaussian quadratic form is shown to be both of the nominal asymptotic significance level and asymptotically correct. It is further shown that B-spline trajectory estimates meet the required approximation conditions. Simulation studies illustrate superior finite sample performance of the proposed testing procedure. For the EEG (ElectroEncephalogram) data, the proposed procedure has confirmed an interesting discovery that the centered EEG data is generated from a small number of elements of the standard Fourier basis.

Key words and phrases: B-spline, ElectroEncephalogram, Functional principal components, Gaussian quadratic form, Oracle efficiency.

1. Introduction

Functional data analysis (FDA) encompasses the analysis of data that come in the form of functions, see Ramsay and Sliverman (2002, 2005) for exploratory tools, Ferraty and Vieu (2006) for Banach/Hilbert space approach to FDA, and Hsing and Eubank (2015) for data-driven theory and methods of FDA.

A raw functional data set consists of observations $\{Y_{ij}, 1 \leq i \leq n, 1 \leq j \leq N\}$, where Y_{ij} is the observation at the j-th measurement point j/N of a random curve $\eta_i(\cdot)$ with $N \to \infty$. For the i-th subject, $i = 1, 2, ..., n$, its sample path $(Y_{ij}, j/N), j = 1, \cdots, N$ is a noisy realization of the latent continuous time stochastic process $\eta_i(\cdot)$ in the sense that

$$Y_{ij} = \eta_i(j/N) + \sigma_i(j/N)\varepsilon_{ij}, 1 \leq i \leq n, 1 \leq j \leq N.$$

The stochastic processes $\eta_i(\cdot)$ are called trajectories of the i-th subject, $1 \leq i \leq n$, and are i.i.d. copies of a canonical stochastic process $\eta(x), x \in [0,1]$ which is square-integrable continuous, i.e., $\eta(\cdot) \in \mathcal{C}[0,1]$ almost surely and $\mathbb{E}\int_{[0,1]} \eta^2(x)dx < +\infty$. The terms $\sigma_i(j/N)\varepsilon_{ij}$ are measurement errors, in which $\{\varepsilon_{ij}\}_{i=1,j=1}^{n,N}$ are i.i.d. noises with mean 0, variance 1, and $\sigma_i(\cdot)$ are

standard deviation functions of i-th subject.

According to Bosq (2000), the $\mathcal{C}[0,1]$-valued random variable $\eta(\cdot)$ has mean $m(\cdot) \in \mathcal{C}[0,1]$ and covariance $G(\cdot,\cdot) \in \mathcal{C}[0,1]^2$ where $m(x) \equiv \mathbb{E}\eta(x), x \in [0,1]$ and $G(x,x') \equiv \mathrm{Cov}\{\eta(x),\eta(x')\}$, $x,x' \in [0,1]$. According to Mercer Lemma, there exist eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$, $\sum_{k=1}^{\infty}\lambda_k < \infty$, with corresponding eigenfunctions $\{\psi_k\}_{k=1}^{\infty}$ of $G(\cdot,\cdot)$, the latter being an orthonormal basis of $\mathcal{L}^2[0,1]$, such that $G(x,x') \equiv \sum_{k=1}^{\infty}\lambda_k\psi_k(x)\psi_k(x')$ and $\int G(x,x')\psi_k(x')dx' = \lambda_k\psi_k(x)$. For each $k \in \mathbb{N}_+$, let $I_k = \{k' \in \mathbb{N}_+ | \lambda_{k'} = \lambda_k\}$, then $\min I_k \leq k \leq \max I_k$. If $\lambda_k > 0$, the cardinality $\#(I_k) = \max I_k - \min I_k + 1$ of I_k is finite, since the integral operator defined by $G(x,x')$ is compact. The linear space of functions spanned by $\{\psi_{k'}\}_{k'\in I_k}$ is the eigen subspace Ψ_k with dimension $\#(I_k)$, corresponding to eigenvalue λ_k of multiplicity $\#(I_k)$. The Mercer expansion of $G(\cdot,\cdot)$ in terms of $\lambda_k, \psi_k, k \in \mathbb{N}_+$ is unique up to orthogonal transformation of $\{\psi_{k'}, k' \in I_k\}$ within each eigenspace Ψ_k.

The standard process $\eta(\cdot)$ then allows the Karhunen-Loève (K-L) expansion $\eta(\cdot) = m(\cdot) + \sum_{k=1}^{\infty}\xi_k\phi_k(\cdot)$ according to Theorem 1.5 of Bosq (2000), in which the rescaled eigenfunctions, $\{\phi_k\}_{k=1}^{\infty}$, called functional principal components (FPCs) satisfy

$$\phi_k(\cdot) = \sqrt{\lambda_k}\psi_k(\cdot), k \geq 1, \tag{1.1}$$

and the random coefficients $\{\xi_k\}_{k=1}^{\infty}$, uncorrelated with mean 0 and variance 1. The i-th trajectory $\eta_i(\cdot)$ is decomposed as

$$\eta_i(\cdot) = m(\cdot) + \xi_i(\cdot), \xi_i(\cdot) = \sum_{k=1}^{\infty} \xi_{ik}\phi_k(\cdot), \tag{1.2}$$

in which the $\mathcal{C}[0,1]$-valued random variable $\xi_i(\cdot)$ is a small-scale variation of x with $\mathbb{E}\xi_i(\cdot) \equiv 0$ and covariance $G(x,x') \equiv \mathbb{E}\{\xi_i(x)\xi_i(x')\}$, $x, x' \in [0,1]$. The random coefficients $\{\xi_{ik}\}_{k=1}^{\infty}, i = 1, ..., n$, are i.i.d. copies of $\{\xi_k\}_{k=1}^{\infty}$, and are called FPC scores. The raw functional data can then be written as

$$Y_{ij} = m(j/N) + \sum_{k=1}^{\infty} \xi_{ik}\phi_k(j/N) + \sigma_i(j/N)\varepsilon_{ij}, 1 \le i \le n, 1 \le j \le N, \tag{1.3}$$

where the infinite series converges absolutely almost surely. Denote also the rescaled FPC scores

$$\zeta_{ik} = \int \xi_i(x)\psi_k(x)\,dx, \tag{1.4}$$

which by (1.1) and (1.2) satisfy

$$\begin{aligned} \zeta_{ik} &= \int \left(\sum_{k'=1}^{\infty} \xi_{ik'}\phi_{k'}(x)\right)\psi_k(x)\,dx = \sqrt{\lambda_k}\xi_{ik}, \\ \zeta_{ik}\psi_k(\cdot) &= \xi_{ik}\phi_k(\cdot), 1 \le i \le n, k \in \mathbb{N}_+. \end{aligned} \tag{1.5}$$

For convenience, the orthonormal eigenfunctions $\{\psi_k(\cdot)\}_{k=1}^{\infty}$ are called canonical FPCs.

Just as mathematical statistics textbooks treat extensively first and second moments estimation, the estimation of mean function $m(\cdot)$ and

covariance function $G(\cdot,\cdot)$ are also essential first steps of FDA, and have been adequately addressed over the last decade. In particular, simultaneous confidence regions are constructed for $m(\cdot)$ in Degras (2011), Cao et al. (2012), Ma et al. (2012), Zheng et al. (2014), Gu et al. (2014), Cai et al. (2020), Li and Yang (2023), Huang et al (2022); and for $G(\cdot,\cdot)$ in Cao et al. (2016), Wang et al. (2020), Zhong and Yang (2023).

The covariance function $G(x,x')$ is intricately composed of eigenvalues $\{\lambda_k\}_{k=1}^{\infty}$ and FPCs $\{\phi_k(\cdot)\}_{k=1}^{\infty}$, all of which are unknown parameters not directly estimable. Likewise, the FPC scores $\{\xi_{ik}, 1 \leq i \leq n, k \in \mathbb{N}_+\}$ are well-defined mathematical objects, but unobservable to the data-handling statistician.

Data analytical tools for computing FPCs and FPC scores are collectively referred to as Functional Principal Components Analysis (FPCA), a simplifying preliminary step for many interesting applications involving trajectories $\{\eta_i(\cdot)\}_{i=1}^{n}$ as independent variables, see Hall and Hosseini-Nasab (2006), Aue et al. (2015) and Shang (2017). Typically FPCA first estimates FPCs and eigenvalues as eigenfunctions and eigenvalues of some estimated $G(\cdot,\cdot)$, and subsequently FPC scores, see Ramsay and Sliverman (2005), Horváth and Kokoszka (2012), Shang (2014), Zhang et al. (2020), and Huang et al (2021). Rigorous inference for functional regression models

remains difficult if FPC scores estimated from eigen equations are used as predictor variables in place of the true ones, because the differences between the true and estimated FPC scores are of order $n^{-1/2}$ only implicitly. Under the special circumstance that the FPCs are known a priori, we have established in (S.20) explicit form of the differences between the true and estimated FPC scores, which could be useful in developing inferential tools for functional regression models.

If the canonical FPCs $\{\psi_k(\cdot)\}_{k=1}^{\infty}$ were "known" a priori as $\{\psi_{0,k}(\cdot)\}_{k=1}^{\infty}$, then rescaled FPC scores $\{\zeta_{ik}, 1 \le i \le n, k \in \mathbb{N}_+\}$ in (1.4) can be estimated by method of moments as

$$\hat{\zeta}_{ik} = \int \hat{\xi}_i(x)\,\psi_{0,k}(x)\,dx, 1 \le i \le n, 1 \le k < \infty, \tag{1.6}$$

where $\left\{\hat{\xi}_i(\cdot)\right\}_{i=1}^{n}$ are some good estimators of centered trajectories $\{\xi_i(\cdot)\}_{i=1}^{n}$. Estimators of eigenvalues and covariance function are also explicit

$$\hat{\lambda}_k = n^{-1}\sum_{i=1}^{n}\hat{\zeta}_{ik}^2, \tag{1.7}$$

$$\hat{G}(x,x') = n^{-1}\sum_{i=1}^{n}\hat{\xi}_i(x)\,\hat{\xi}_i(x'). \tag{1.8}$$

As $\left\{\psi_{0,k}(x)\,\psi_{0,k'}(x')\right\}_{k,k'=1}^{\infty}$ is an orthonormal basis of $\mathcal{L}^2[0,1]^2$, $G(x,x')$

has the following expansion with coordinates $C_{kk'}, k, k' \in \mathbb{N}_+$

$$G(x,x') \equiv \sum_{k,k'=1}^{\infty} C_{kk'} \psi_{0,k}(x) \psi_{0,k'}(x'),$$

$$C_{kk'} \equiv \int G(x,x') \psi_{0,k}(x) \psi_{0,k'}(x') dx dx', k, k' \in \mathbb{N}_+. \tag{1.9}$$

If $\{\psi_k(\cdot)\}_{k=1}^{\infty}$ and $\{\psi_{0,k}(\cdot)\}_{k=1}^{\infty}$ are the "same" set subject to permutation of eigenspaces $\{\Psi_k\}_{k\in\mathbb{N}_+}$ and orthogonal transformation within each eigenspace Ψ_k, then exist $\lambda_{0,k} \geq 0, k \in \mathbb{N}_+, \sum_{k=1}^{\infty} \lambda_{0,k} < \infty$ such that $G(x,x') \equiv \sum_{k=1}^{\infty} \lambda_{0,k} \psi_{0,k}(x) \psi_{0,k}(x')$, in other words, all off-diagonal coordinates $C_{kk'}$ $(k \neq k')$ are 0. Since $C_{kk'} \equiv C_{k'k}, k, k' \in \mathbb{N}_+$, one can test hypotheses

$$\begin{aligned} H_0 &: C_{kk'} \equiv 0, \forall k < k' \in \mathbb{N}_+ \\ H_1 &: \exists k < k' \in \mathbb{N}_+, C_{kk'} \neq 0 \end{aligned} \tag{1.10}$$

Formulation of (1.10) is motivated by studies of ElectroEncephalogram (EEG) data in Li and Yang (2023) and Zhong and Yang (2023). Both have observed trigonometric shape trajectories, with explicit and sparse Fourier expansions of mean $m(\cdot)$ and covariance $G(\cdot,\cdot)$ accepted by using simultaneous confidence regions. Similar phenomenon had also been noticed in studies of Event Related Potentials (ERP) data. The present work goes deeper to directly test canonical FPCs at the more fundamental level.

The paper is organised as follows. Section 2 states main theoretical results on a hypothesis test for the canonical FPCs, including asymptotic

significance level and asymptotic correctness of chi-square type statistics, both "infeasible"and two-step data-driven, and that all requirements for these asymptotics to hold are met by B-spline trajectory estimates. Procedures to implement the proposed test are given in Section 3. Section 4 contains some simulation findings and an empirical study of EEG data is in Section 5. All technical proofs are in the Supplement.

2. Main Results

2.1 Asymptotic properties

To better formulate the hypotheses in (1.10), define the following Hilbert space of infinite real matrices with the usual Frobenius norm

$$\begin{aligned}\mathcal{H} \;=\; \Big\{ & (a_{kk'})_{1\le k,k'<\infty}, a_{kk'} \in \mathbb{R} : \\ & \left\|(a_{kk'})_{1\le k,k'<\infty}\right\|_{\mathcal{H}} = \left(\sum_{1\le k,k'<\infty} a_{kk'}^2\right)^{1/2} < \infty \Big\}.\end{aligned}$$

A natural orthonormal basis of $\mathcal{H}$ consists of coordinate vectors $(\mathbf{e}_{kk'})_{1\le k,k'<\infty}$, where $\mathbf{e}_{kk'}$ is a vector with $a_{kk'} = 1$ and all other elements 0. Denote the subspace of upper triangle matrices

$$\mathcal{H}_{\mathrm{UT}} = \Big\{(a_{kk'})_{1\le k,k'<\infty} \in \mathcal{H} : a_{kk'} \equiv 0, 1 \le k' \le k < \infty\Big\}$$

with corresponding orthogonal projection operator $\mathcal{P}_{\mathrm{UT}}$:

$$\mathcal{P}_{\mathrm{UT}}\left(a_{kk'}\right)_{1\leq k,k'<\infty}=\left(a_{kk'}\right)_{1\leq k<k'<\infty}. \tag{2.11}$$

Relative to the orthonormal basis $\left\{\psi_{0,k}\left(x\right)\psi_{0,k'}\left(x'\right)\right\}_{k,k'=1}^{\infty}$ of $\mathcal{L}^2\left[0,1\right]^2$, there is a natural isometry Π between $\mathcal{H}$ and $\mathcal{L}^2\left[0,1\right]^2$:

$$\begin{aligned}\Pi\left\{\left(a_{kk'}\right)_{1\leq k,k'<\infty}\right\}&=\sum_{1\leq k,k'<\infty}a_{kk'}\psi_{0,k}\left(x\right)\psi_{0,k'}\left(x'\right),\left(a_{kk'}\right)_{1\leq k,k'<\infty}\in\mathcal{H}\\ \Pi^{-1}\left(\Theta\right)&=\left(\int\Theta\left(x,x'\right)\psi_{0,k}\left(x\right)\psi_{0,k'}\left(x'\right)dxdx'\right)_{1\leq k,k'<\infty},\Theta\in\mathcal{L}^2\left[0,1\right]^2.\end{aligned} \tag{2.12}$$

As (1.9) entails that $\Pi^{-1}\left(G\right)=\left(C_{kk'}\right)_{1\leq k,k'<\infty}$, so H_0 in (1.10) is equivalent to $\mathcal{P}_{\mathrm{UT}}\Pi^{-1}\left(G\right)=\left(0\right)_{1\leq k<k'<\infty}$. The hypotheses are therefore reformulated in terms of the Hilbert space parameter $\mathcal{P}_{\mathrm{UT}}\Pi^{-1}\left(G\right)$, with projection operator $\mathcal{P}_{\mathrm{UT}}$ and isometry Π^{-1} defined in (2.11)) and (2.12:

$$\begin{aligned}H_0&:\ \mathcal{P}_{\mathrm{UT}}\Pi^{-1}\left(G\right)=\left(0\right)_{1\leq k<k'<\infty},\text{ or }\left\|\mathcal{P}_{\mathrm{UT}}\Pi^{-1}\left(G\right)\right\|_{\mathcal{H}}^2=0,\\ H_1&:\ \mathcal{P}_{\mathrm{UT}}\Pi^{-1}\left(G\right)\neq\left(0\right)_{1\leq k<k'<\infty},\text{ or }\left\|\mathcal{P}_{\mathrm{UT}}\Pi^{-1}\left(G\right)\right\|_{\mathcal{H}}^2>0.\end{aligned} \tag{2.13}$$

Under H_0, by permuting eigen subspaces Ψ_k and applying orthogonal transformations, one may assume that $\psi_k\left(\cdot\right)\equiv\psi_{0,k}\left(\cdot\right),k\in\mathbb{N}_+$.

An infeasible estimator of the covariance function $G\left(x,x'\right)$ is

$$\tilde{G}\left(x,x'\right)\equiv n^{-1}\sum_{i=1}^{n}\xi_i\left(x\right)\xi_i\left(x'\right).$$

One may write by (1.2) and (1.5)

$$\tilde{G}(x,x') = \sum\nolimits_{k,k'=1}^{\infty} n^{-1} \sum_{i=1}^{n} \zeta_{ik}\zeta_{ik'}\psi_k(x)\,\psi_{k'}(x'). \tag{2.14}$$

The coordinates of $\tilde{G}(x,x')$ relative to $\left\{\psi_{0,k}(x)\,\psi_{0,k'}(x')\right\}_{k,k'=1}^{\infty}$ are

$$\begin{aligned} Z_{kk'} &= Z_{k'k} = \int \tilde{G}(x,x')\,\psi_{0,k}(x)\,\psi_{0,k'}(x')\,dxdx' \\ &= n^{-1}\sum_{i=1}^{n}\sum\nolimits_{k_1,k_2=1}^{\infty} \zeta_{ik_1}\zeta_{ik_2}u_{k_1k_2,kk'}, k,k' \in \mathbb{N}_+, \end{aligned} \tag{2.15}$$

in which the inner products

$$u_{k_1k_2,kk'} = \int \psi_{k_1}(x)\,\psi_{k_2}(x')\,\psi_{0,k}(x)\,\psi_{0,k'}(x')\,dxdx', \tag{2.16}$$

satisfy for $k_1,k_2,k_3,k_4,k,k',k'',k''' \in \mathbb{N}_+$

$$\begin{aligned} \sum\nolimits_{k_1,k_2=1}^{\infty} u_{k_1k_2,kk'}u_{k_1k_2,k''k'''} &= \left\langle \psi_{0,k}\psi_{0,k'}, \psi_{0,k''}\psi_{0,k'''}\right\rangle = \delta_{kk''}\delta_{k'k'''}, \\ \sum\nolimits_{k,k'=1}^{\infty} u_{k_1k_2,kk'}u_{k_3k_4,kk'} &= \left\langle \psi_{0,k_1}\psi_{0,k_2}, \psi_{0,k_3}\psi_{0,k_4}\right\rangle = \delta_{k_1k_3}\delta_{k_2k_4}, \end{aligned} \tag{2.17}$$

where the Kronecker indices $\delta_{kk'} = 1$ for $k = k'$ and 0 for $k \neq k'$. Thus if one defines an operator $\mathbf{U}: \mathcal{H} \to \mathcal{H}$ by

$$\mathbf{U}\left(a_{kk'}\right)_{1\le k,k'<\infty} = \left(\sum\nolimits_{1\le k,k'<\infty}^{\infty} u_{k_1k_2,kk'}a_{kk'}\right)_{1\le k_1,k_2<\infty}, \tag{2.18}$$

then $\mathbf{U}$ is unitary, its corresponding infinite orthogonal matrix transforms orthonormal basis $\left\{\psi_{k_1}(x)\,\psi_{k_2}(x')\right\}_{k_1,k_2=1}^{\infty}$ to $\left\{\psi_{0,k}(x)\,\psi_{0,k'}(x')\right\}_{k,k'=1}^{\infty}$. Under H_0, $\mathbf{U} = \mathbf{I}$, the identity operator.

The infeasible estimator then can be written as

$$\begin{aligned}\tilde{G}(x,x') &= \sum_{k,k'=1}^{\infty} Z_{kk'}\psi_{0,k}(x)\psi_{0,k'}(x'),\\ (Z_{kk'})_{1\le k,k'<\infty} &= \Pi^{-1}\left(\tilde{G}\right), (Z_{kk'})_{1\le k<k'<\infty} = \mathcal{P}_{\mathrm{UT}}\Pi^{-1}\left(\tilde{G}\right), \end{aligned} \quad (2.19)$$

so to determine if $\|\mathcal{P}_{\mathrm{UT}}\Pi^{-1}(G)\|_{\mathcal{H}}^2 = 0$ as in H_0 of (2.13), one defines the following chi-square type statistic $\tilde{S}_n$, the larger value of which favors H_1:

$$\tilde{S}_n = n\sum_{1\le k<k'<\infty} Z_{kk'}^2 = n\left\|\mathcal{P}_{\mathrm{UT}}\Pi^{-1}\left(\tilde{G}\right)\right\|_{\mathcal{H}}^2 = \left\|\mathcal{P}_{\mathrm{UT}}\left(n^{-1/2}\sum_{i=1}^{n}\mathbf{X}_i\right)\right\|_{\mathcal{H}}^2, \quad (2.20)$$

in which

$$(Z_{kk'})_{1\le k,k'<\infty} = n^{-1}\sum_{i=1}^{n}\mathbf{X}_i, \quad (2.21)$$

$$\mathbf{X}_i = \left(\sum_{k_1,k_2=1}^{\infty}\zeta_{ik_1}\zeta_{ik_2}u_{k_1k_2,kk'}\right)_{1\le k,k'<\infty} = \mathbf{U}\left(\zeta_{ik_1}\zeta_{ik_2}\right)_{1\le k_1,k_2<\infty}, 1\le i\le n, \quad (2.22)$$

the infinite matrices $\mathbf{X}_i$ being i.i.d. $\mathcal{H}$-valued with mean $\boldsymbol{\mu}_{\mathbf{X}}\in\mathcal{H}$ given in (2.26) of Theorem 1, and $\mathbf{U}$ being the unitary operator in (2.18). Denote also i.i.d. variables

$$\mathbf{Y}_i = \left(\zeta_{ik_1}\zeta_{ik_2} - \lambda_{k_1}\delta_{k_1k_2}\right)_{1\le k_1,k_2<\infty} = \mathbf{U}^{-1}\left(\mathbf{X}_i - \boldsymbol{\mu}_{\mathbf{X}}\right), \quad (2.23)$$

the covariance operator $\mathbf{C}_{\mathbf{Y}}$ of $\mathbf{Y}_i$ and $\mathbf{C}_{\mathbf{X}}$ of $\mathbf{X}_i$ satisfy

$$\mathbf{C}_{\mathbf{Y}}(\mathbf{x}) = \mathbf{U}^{-1}\mathbf{C}_{\mathbf{X}}\mathbf{U}(\mathbf{x}), \forall\mathbf{x}\in\mathcal{H}. \quad (2.24)$$

Finally, define an infinite Gaussian quadratic form

$$S = \sum_{1 \le k < k' < \infty} \lambda_k \lambda_{k'} \chi^2_{kk'}(1), \tag{2.25}$$

where $\chi^2_{kk'}(1)$ are independent chi-square variables of degree of freedom 1. The infinite series in (2.25) converges absolutely almost surely since $\mathbb{E}S = \sum_{1 \le k < k' < \infty} \lambda_k \lambda_{k'} < \left(\sum_{1 \le k < \infty} \lambda_k \right)^2 < \infty$.

The following assumption is needed for asymptotics of $\tilde{S}_n$.

(A1) The FPC scores $\{\xi_{ik}\}_{i \ge 1, k \ge 1}$ are independent over $k \ge 1$ and i.i.d. over $i \ge 1$. The number of distinct distributions for all FPC scores $\{\xi_{1k}\}_{k \ge 1}$ is finite, and $\max_{1 \le k < \infty} \mathbb{E}\xi^4_{1k} < \infty$.

The independence condition in (A1) is common in existing works on functional data analysis, see Cao et al. (2012), Ma et al. (2012), Gu et al. (2014), Zheng et al. (2014), Wang et al. (2020). Each of the FPC scores $\{\xi_{1k}\}_{k \ge 1}$ may have its own probability distribution, but the number of distinct distributions must be finite. For example, $\xi_{11}, \xi_{13} \sim N(0,1)$, $\xi_{12} \sim t_{(10)}/\sqrt{1.25}$, $\xi_{14} \sim U\left(-\sqrt{3}, \sqrt{3}\right)$ for Case 2 in Section 4, the distributions of $\xi_{1k}, k > 4$ can be all set to $N(0,1)$ as $\lambda_k \equiv 0, k > 4$.

Theorem 1. *Under Assumption (A1),* $\{\mathbf{X}_i\}_{i=1}^n$ *in (2.22) are i.i.d.* $\mathcal{H}$-

valued random variables, with

$$\mathbb{E}\mathbf{X}_i = \boldsymbol{\mu}_{\mathbf{X}} = \left(\sum_{k_1=1}^{\infty} \lambda_{k_1} u_{k_1 k_1, kk'}\right)_{1\le k,k'<\infty}, \tag{2.26}$$

$$\mathbb{E}\left\|\mathbf{X}_i\right\|_{\mathcal{H}}^2 = \sum_{k=1}^{\infty} \lambda_k^2 \left(\mathbb{E}\xi_k^4 - 1\right) + \left(\sum_{k=1}^{\infty} \lambda_k\right)^2 < \infty. \tag{2.27}$$

As $n \to \infty$, there is an $\mathcal{H}$-valued normal variable $\mathcal{N}$ such that

$$n^{1/2}\left\{\left(Z_{kk'}\right)_{1\le k,k'<\infty} - \boldsymbol{\mu}_{\mathbf{X}}\right\} = n^{-1/2}\sum_{i=1}^{n}\left(\mathbf{X}_i - \boldsymbol{\mu}_{\mathbf{X}}\right) \xrightarrow{D} \mathcal{N} \sim N\left(\mathbf{0}, \mathbf{C}_{\mathbf{X}}\right), \tag{2.28}$$

which, under H_0 in (2.13), becomes the following special case

$$n^{1/2}\left(Z_{kk} - \lambda_k, Z_{kk'}\right)_{1\le k\ne k'<\infty} = n^{-1/2}\sum_{i=1}^{n}\mathbf{Y}_i \xrightarrow{D} \mathcal{N} \sim N\left(\mathbf{0}, \mathbf{C}_{\mathbf{Y}}\right), \tag{2.29}$$

$$\mathbf{C}_{\mathbf{Y}}\left(\mathbf{e}_{kk}\right) = \lambda_k^2\left(\mathbb{E}\xi_k^4 - 1\right)\mathbf{e}_{kk}, \mathbf{C}_{\mathbf{Y}}\left(\mathbf{e}_{kk'}\right) = \lambda_k\lambda_{k'}\left(\mathbf{e}_{kk'} + \mathbf{e}_{k'k}\right), 1 \le k \ne k' < \infty. \tag{2.30}$$

Consequently, under H_0, with S as in (2.25)

$$\tilde{S}_n = \left\|\mathcal{P}_{\mathrm{UT}}\left(n^{-1/2}\sum_{i=1}^{n}\mathbf{Y}_i\right)\right\|_{\mathcal{H}}^2 \xrightarrow{D} \left\|\mathcal{P}_{\mathrm{UT}}\left(\mathcal{N}\right)\right\|_{\mathcal{H}}^2 \overset{D}{=} S.$$

Lemma S.2 in Supplement stipulates that the distribution function $F_S(q) = \mathbb{P}[S \le q]$ of quadratic form S in (2.25) is continuous and strictly increasing, so the inverse function F_S^{-1} is well-defined. For any $\alpha \in (0,1)$, the $(1-\alpha)$-th quantile $Q_{1-\alpha}$ of S is the unique q that solves $F_S(q) = 1-\alpha$:

$$Q_{1-\alpha} = F_S^{-1}(1-\alpha).$$

Under H_0, $Z_{kk'}$ in (2.15) and (2.21) has the following simpler expression

$$Z_{kk'} = n^{-1}\sum_{i=1}^{n}\zeta_{ik}\zeta_{ik'} = n^{-1}\sum_{i=1}^{n}\sqrt{\lambda_k}\sqrt{\lambda_{k'}}\xi_{ik}\xi_{ik'}. \tag{2.31}$$

Since $\{\zeta_{ik}, 1 \le i \le n, k \in \mathbb{N}_+\}$ are unobservable, $\{Z_{kk'}\}_{k\neq k'}$ and $\tilde{S}_n$ are all "infeasible". Substituting ζ_{ik} with $\hat{\zeta}_{ik}$ in (1.6), yields the following feasible replicas of $Z_{kk'}$ in (2.31):

$$\hat{Z}_{kk'} = n^{-1}\sum_{i=1}^{n}\hat{\zeta}_{ik}\hat{\zeta}_{ik'}, \left(\hat{Z}_{kk'}\right)_{1\le k<k'<\infty} = \mathcal{P}_{\mathrm{UT}}\Pi^{-1}\left(\hat{G}\right). \tag{2.32}$$

Using $\hat{Z}_{kk'}$ in (2.32), a feasible statistic $\hat{S}_n$ is defined to mimic $\tilde{S}_n$ in (2.20)

$$\hat{S}_n = n\sum_{1\le k<k'\le\kappa_n}\hat{Z}_{kk'}^2, \tag{2.33}$$

where truncation indices $\kappa_n \in \mathbb{N}_+$ satisfy

$$\kappa_n \to \infty, \kappa_n^2 n^{-1/2}\log^{3/2} n \to 0. \tag{2.34}$$

In what follows, for function $\varphi(\cdot)$ defined on $[0,1]$, denote $\|\varphi\|_\infty = \sup_{x\in[0,1]}|\varphi(x)|$, and $\varphi^{(q)}(\cdot)$ its q-th order derivative if it exists. For $q \in \mathbb{N}, \mu \in (0,1]$, denote the (q,μ) Hölder seminorm of function φ as

$$\|\varphi\|_{q,\mu} = \sup_{x,x'\in[0,1],x\neq x'}\left|\frac{\varphi^{(q)}(x) - \varphi^{(q)}(x')}{|x-x'|^\mu}\right|$$

and the space of functions with finite (q,μ) Hölder seminorm as $\mathcal{C}^{(q,\mu)}[0,1] = \left\{\varphi \mid \|\varphi\|_{q,\mu} < +\infty\right\}$. As a special case, $\mathcal{C}^{(0,1)}[0,1]$ is the space of Lipschitz continuous functions.

(B1) The FPCs $\phi_k(\cdot) \in \mathcal{C}^{(0,1)}[0,1]$ with $\sum_{k=1}^{\infty} \|\phi_k\|_{\infty} + \sum_{k=1}^{\infty} \|\phi_k\|_{0,1} < +\infty$.

(B2) The trajectory estimates $\left\{\hat{\xi}_i(\cdot)\right\}_{i=1}^{n}$ used in (1.6) satisfy

$$\max_{1\le i\le n} \left\| \hat{\xi}_i(\cdot) - \xi_i(\cdot) + n^{-1} \sum_{i'=1}^{n} \xi_{i'}(\cdot) \right\|_{\infty} = \mathcal{O}_{a.s.}\left(\rho_{n,N}\right), \tag{2.35}$$

where $\left\{\rho_{n,N}\right\}_{n=1}^{\infty}$ are such that $\rho_{n,N} > 0, \kappa_n^2 n^{1/2} \rho_{n,N} \log^{1/2} n \to 0$ as $n \to \infty$ for some $\{\kappa_n\}_{n=1}^{\infty}$ satisfying (2.34).

Collective boundedness and Lipschitz bounded smoothness of principal components in Assumption (B1) are necessary for $\mathcal{C}[0,1]$ Central Limit Theorem of $n^{-1}\sum_{i'=1}^{n} \xi_{i'}(\cdot)$, see Lemma S.6 in Supplement.

Propositions 1 and 2 in Supplement lead to the following theorem.

Theorem 2. *Under Assumptions (A1), (B1)-(B2) and H_0 in (2.13), as $n \to \infty$, $\hat{S}_n$ in (2.33) is oracally efficient, i.e., $\hat{S}_n - \tilde{S}_n \to_p 0$. Hence*

$$\sup_{\alpha\in(0,1)} \left| \mathbb{P}\left[\tilde{S}_n > Q_{1-\alpha}\right] - \alpha \right| \to 0, \ \sup_{\alpha\in(0,1)} \left| \mathbb{P}\left[\hat{S}_n > Q_{1-\alpha}\right] - \alpha \right| \to 0.$$

Using eigenvalue estimates $\hat{\lambda}_k$ in (1.7), define an approximation of S:

$$\bar{S}_n = \sum_{1\le k<k'\le \kappa_n} \hat{\lambda}_k \hat{\lambda}_{k'} \chi_{kk'}^2(1), \tag{2.36}$$

with $(1-\alpha)$-th quantile denoted as $\hat{Q}_{1-\alpha}$. The following theorem provides full justification to use $\hat{Q}_{1-\alpha}$ in place of $Q_{1-\alpha}$ so one can define a test statistic

$$T_n = I_{\left\{\hat{S}_n > \hat{Q}_{1-\alpha}\right\}}, \tag{2.37}$$

with the rule of rejecting H_0 if and only if $T_n = 1$.

Theorem 3. *Under Assumptions (A1), (B1)-(B2), and under H_0 in (2.13), as $n \to \infty$, the finite approximation $\bar{S}_n$ in (2.36) converges to S in probability, i.e., $\bar{S}_n - S = o_p(1)$. Consequently, for any $\alpha \in (0,1)$, $\hat{Q}_{1-\alpha} - Q_{1-\alpha} = o_p(1)$ and*

$$\mathbb{P}(T_n = 1) = \mathbb{P}\left(\hat{S}_n > \hat{Q}_{1-\alpha}\right) \to \alpha, \mathbb{P}\left(\tilde{S}_n > \hat{Q}_{1-\alpha}\right) \to \alpha.$$

Theorem 3 provides that the asymptotic significance level is α for both the data-driven test $T_n = I_{\left\{\hat{S}_n > \hat{Q}_{1-\alpha}\right\}}$ and the infeasible $I_{\left\{\tilde{S}_n > \hat{Q}_{1-\alpha}\right\}}$, Theorem 2 the other infeasible $I_{\left\{\tilde{S}_n > Q_{1-\alpha}\right\}}$ and $I_{\left\{\hat{S}_n > Q_{1-\alpha}\right\}}$.

We establish next asymptotic consistency of test T_n in (2.37).

Theorem 4. *Under Assumptions (A1), (B1)-(B2) and H_1 in (2.13), there exist $k_1 < k_2 \in \mathbb{N}_+, C_{k_1k_2} \neq 0$, where $C_{k_1k_2}$ is given in (1.9). As $n \to \infty$,*

$$\min\left\{\tilde{S}_n, \hat{S}_n\right\} \geq n\hat{Z}^2_{k_1k_2} = nC^2_{k_1k_2} + \mathcal{O}_p\left(n^{1/2}\right),$$

$$\mathbb{P}(T_n = 1) = \mathbb{P}\left[\hat{S}_n > \hat{Q}_{1-\alpha}\right] \to 1,$$

$$\min\left\{\mathbb{P}\left[\tilde{S}_n > Q_{1-\alpha}\right], \mathbb{P}\left[\hat{S}_n > Q_{1-\alpha}\right], \mathbb{P}\left[\tilde{S}_n > \hat{Q}_{1-\alpha}\right]\right\} \to 1.$$

Theorem 4 reveals that under alternative H_1 in (2.13), the data-driven test T_n in (2.37), is consistent, along with the infeasible $I_{\left\{\tilde{S}_n > \hat{Q}_{1-\alpha}\right\}}, I_{\left\{\tilde{S}_n > Q_{1-\alpha}\right\}}$ and $I_{\left\{\hat{S}_n > Q_{1-\alpha}\right\}}$.

2.2 B-spline estimation

Theorems 2, 3, and 4 depend on a high level Assumption (B2) involving trajectory estimates $\left\{\hat{\xi}_i(\cdot)\right\}_{i=1}^{n}$ in (1.6). In this section it is shown that B-spline trajectory estimates meet Assumption (B2).

To define splines, the interval $[0,1]$ is divided into (J_s+1) equal subintervals $I_J = [Jh, (J+1)h), 0 \le J \le J_s - 1, I_{J_s} = [J_s h, 1]$ with length $h = 1/(J_s+1)$. For positive integer p, let $\mathcal{H}^{(p-2)} = \mathcal{H}^{(p-2)}[0,1]$ be the space of functions that are $(p-2)$ times continuously differentiable on $[0,1]$, polynomials of degree $(p-1)$ on subintervals $I_J, 0 \le J \le J_s$. Denote by $\{B_{J,p}(\cdot), 1 \le J \le J_s + p\}$ the p-th order B-spline basis of $\mathcal{H}^{(p-2)}$ (de Boor (2001)), $\mathcal{H}^{(p-2)} = \left\{\sum_{J=1}^{J_s+p} \lambda_{J,p} B_{J,p}(\cdot) \mid \lambda_{J,p} \in \mathbb{R}\right\}$.

Latent trajectories $\eta_i(\cdot)$ are estimated via B-spline for each subject i

$$\hat{\eta}_i(\cdot) = \underset{g(\cdot)\in\mathcal{H}^{(p-2)}}{\operatorname{argmin}} \sum_{j=1}^{N} \{Y_{ij} - g(j/N)\}^2, 1 \le i \le n. \tag{2.38}$$

B-spline estimates of mean $m(\cdot)$ and centered trajectories $\xi_i(\cdot)$ are:

$$\hat{m}(\cdot) = n^{-1} \sum_{i=1}^{n} \hat{\eta}_i(\cdot), \tag{2.39}$$

$$\hat{\xi}_i(\cdot) = \hat{\eta}_i(\cdot) - \hat{m}(\cdot), 1 \le i \le n, \tag{2.40}$$

with $\hat{\eta}_i(\cdot)$ defined in (2.38). The B-spline estimates $\hat{\xi}_i(\cdot)$ in (2.40) is then used for estimating rescaled FPC scores in (1.6), as well as covariance function in (1.8).

The following constraints are listed as constants ν, q, μ, etc., appear sequentially in Assumptions (C1)-(C5)

$$\nu \in (0,1], q \in \mathbb{N}^+, \mu \in (0,1], p^* = q + \mu, \tag{2.41}$$

$$\theta \in \left(0, \min\left(\frac{2p^*}{2p^*+1}, \nu\right)\right), \tag{2.42}$$

$$\beta_2 \in \left(0, \min\left\{\frac{1}{2}, \nu - \frac{\theta}{2}, 1 - \frac{\theta(p^*+1)}{2p^*}\right\}\right), \tag{2.43}$$

$$r_1 > \max\left(6, \frac{4\theta}{2p^*(1-\theta)-\theta}, \frac{4\theta}{2p^*(1-\beta_2-\theta/2)-\theta}\right), \tag{2.44}$$

$$\max\left(1-\nu, \left(\frac{2}{r_1}+\frac{1}{2}\right)\frac{\theta}{p^*}\right) < \gamma < \min\left(1-\theta, 1-\beta_2-\frac{\theta}{2}\right). \tag{2.45}$$

Elementary algebra shows that (2.42) is needed for (2.43) to hold, while both (2.42) and (2.43) are needed for (2.44). One also verifies that (2.42), (2.43) and (2.44) together ensure the existence of γ that satisfies (2.45).

The above (2.42), (2.43), (2.44), (2.45) enable the following assumptions.

(C1) The standard deviation functions $\sigma_i(\cdot) \in \mathcal{C}^{(0,\nu)}[0,1]$ for ν in (2.41), $\max_{1\le i\le n}\|\sigma_i\|_\infty \le C_\sigma$, $\max_{1\le i\le n}\|\sigma_i\|_{0,\nu} \le C_\sigma$ for $0 < C_\sigma < \infty$.

(C2) The FPCs $\phi_k(\cdot) \in \mathcal{C}^{(q,\mu)}[0,1]$ for integer q and constant μ in (2.41) with $\sum_{k=1}^{\infty}\|\phi_k\|_{q,\mu} < +\infty$.

(C3) As $n \to \infty, N = N(n) \to \infty, n = O(N^\theta)$ for θ in (2.42).

(C4) The i.i.d. noises $\{\varepsilon_{ij}\}_{i\ge1,j\ge1}$ satisfy $\mathbb{E}\varepsilon_{11}^2 < \infty$. There are i.i.d. $N(0,1)$

variables $\{U_{ij,\varepsilon}\}_{i=1,j=1}^{n,N}$ such that

$$\mathbb{P}\left\{\max_{1\le i\le n}\max_{1\le t\le N}\left|\sum_{j=1}^{t}\varepsilon_{ij}-\sum_{j=1}^{t}U_{ij,\varepsilon}\right|>N^{\beta_2}\right\}<C_\varepsilon N^{-\gamma_2}$$

for constants $C_\varepsilon\in(0,+\infty)$, $\gamma_2\in(1,+\infty)$ and β_2 in (2.43). For r_1 in (2.44), $\max_{1\le k<\infty}\mathbb{E}\left|\xi_{1k}\right|^{r_1}<\infty$.

(C5) The spline order $p\ge p^*$, the number of interior knots $J_s=N^\gamma d_N$ with γ in (2.45) and $d_N+d_N^{-1}=\mathcal{O}\left(\log^\tau N\right)$ as $N\to\infty$ for some $\tau>0$.

Uniform boundedness and Hölder continuity for the standard deviation functions $\sigma_i(\cdot)$ in Assumption (C1) are both common for spline smoothing, see Wang et al. (2020), Li and Yang (2023) and Zhong and Yang (2023). Allowing $\sigma_i(\cdot)$ for each subject i and not imposing any smoothness condition on the mean function $m(\cdot)$ are new features which substantially enhance applicability of our proposed method. The collective (q,μ)-Hölder bounded smoothness of principal components in Assumption (C2) is for bias reduction. Assumption (C3) requires that the number N of observations per curve grows in sync with the sample size n, not slower than $n^{1/\theta}$. The probability inequalities in Assumption (C4) provides Gaussian partial sum strong approximation of measurement errors $\{\varepsilon_{ij}\}_{i\ge1,j\ge1}$. The high level Assumption (C4) can be ensured by the elementary Assumption (C4') below together with Assumption (C3), the proof of which is in the

Supplement. The requirement for the number of knots of the splines is stated in Assumption (C5), which aims to modulate smoothness of B-spline estimator by that of FPCs.

(C4') There exist $r_2 > (2+\theta)/\beta_2$ for θ in (2.42) and β_2 in (2.43), such that $\mathbb{E}|\varepsilon_{11}|^{r_2} < \infty$. For r_1 in (2.44), $\max_{1\le k<\infty} \mathbb{E}|\xi_{1k}|^{r_1} < \infty$.

Remark 1. The above assumptions are mild and are satisfied in various practical situations. One simple and reasonable setting for parameters $q, \mu, \nu, \theta, p, \gamma$ can be the following: $q+\mu = p^* = 4$, $\nu = 1$, $\theta < 8/9$ (e.g., 0.6), $p = 4$ (cubic spline), $\gamma = 0.2$. These constants are used as implementation defaults in Section 3, together with $d_N \asymp \log\log N$.

The next crucial theorem ensures the feasibility of Assumption (B2).

Theorem 5. *Under Assumptions (A1), (B1), (C1)-(C5), the B-spline trajectory estimates* $\left\{\hat{\xi}_i(\cdot)\right\}_{i=1}^n$ *in (2.40) satisfy Assumption (B2) with*

$$\rho_{n,N} = J_s^{-p^*}(n\log n)^{2/r_1} + N^{-1/2}J_s^{1/2}\log^{1/2} N + J_s N^{\beta_2-1}. \tag{2.46}$$

3. Implementation

This section details how the test is performed. All trajectories are estimated by cubic splines, i.e., $p = 4$. The smoothness order (q, μ) of eigenfunctions

$\phi_k(\cdot)$ is taken as $(3,1)$ or $(4,0)$ by default. The number of knots for B-spline smoothing $J_s = [cN^\gamma \log\log N]$ is recommended with constant c, where $[a]$ denotes the integer part of a. The default values $\gamma = 0.2$ and $c = 2$ are adequate. These B-spline trajectory estimates satisfy Assumption (B2) if one takes the number of FPCs for test statistic $\kappa_n = [c_1 \log n] + c_2$, the default values are $c_1 = 3/2, c_2 = 0$. Then $\hat{S}_n$ is computed according to (2.33) and T_n (2.37).

To obtain $\hat{Q}_{1-\alpha}$, one generates $\hat{\tau}_b = \sum_{1\le k<k'\le\kappa_n} \hat{\lambda}_k \hat{\lambda}_{k'} \chi_{kk',b}$ where $\chi_{kk',b}$ are i.i.d. central chi-square variables with degree of freedom 1, $1 \le k < k' < \kappa_n$, $b = 1, \ldots, b_M$, b_M is a preset large integer with default value 1000. Then $\hat{Q}_{1-\alpha}$ is taken as $(1-\alpha)$-th sample quantile of $\{\hat{\tau}_b\}_{b=1}^{b_M}$.

4. Simulation

Two candidate sets $\{\psi_{0,k}(\cdot)\}_{k=1}^{\infty}$ of canonical FPCs are used in this section:

(a) FPCs of Ornstein-Uhlenbeck (O-U) process: for $k \in \mathbb{N}_+$,

$$\psi_{\mathrm{OU},k}(x) = \left\{1/2 + \left(1+\omega_k^2\right)^{-1}\right\}^{-1/2} \sin\left\{\omega_k (x - 1/2) + k\pi/2\right\}, \quad (4.47)$$

where the ω_k's are the positive roots of $\tan\omega = -2\omega(1-\omega^2)^{-1}$ arranged in ascending order;

(b) the Fourier basis: for $l \in \mathbb{N}_+$,

$$\psi_{\mathrm{F},1}(x) \equiv 1, \psi_{\mathrm{F},2l}(x) \equiv \sqrt{2}\cos(2l\pi x), \psi_{\mathrm{F},2l+1}(x) \equiv \sqrt{2}\sin(2l\pi x). \quad (4.48)$$

Data are generated from model

$$Y_{ij} = m\left(\frac{j}{N}\right) + \sum_{k=1}^{\kappa} \xi_{ik}\sqrt{\lambda_k}\psi_k\left(\frac{j}{N}\right) + \sigma\epsilon_{ij}, 1 \le j \le N, 1 \le i \le n,$$

with $\sigma = 0.3, n = 150, 250, 400, 600, N = 200, 500, 1000, 2000, \alpha = 0.01, 0.05, 0.1, 0.2$. Noises $\epsilon_{ij} \sim N(0,1), i, j \in \mathbb{N}_+$. Each combination of (n, N, α) is replicated 1000 times.

Case 1: $m(x) = 10 - \sin(2\pi x), \kappa = 2, (\lambda_1, \lambda_2) = (2, 1/2)$, $\psi_1(x) = \psi_{\mathrm{OU},1}(x)$, $\psi_2(x) = \psi_{\mathrm{OU},2}(x)$. FPC scores $\xi_{i1} \sim N(0,1), \xi_{i2} \sim t_{(10)}/\sqrt{1.25}, i \in \mathbb{N}_+$.

Case 2: $m(x) = 10 + \sin(3\pi x)$, $\kappa = 4$, $(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = (4, 2, 1, 1/2)$, $\psi_1(x) = \psi_{\mathrm{F},3}(x), \psi_2(x) = \psi_{\mathrm{F},2}(x), \psi_3(x) = \psi_{\mathrm{F},5}(x), \psi_4(x) = \psi_{\mathrm{F},8}(x)$. $\xi_{i1}, \xi_{i3} \sim N(0,1)$, $\xi_{i2} \sim t_{(10)}/\sqrt{1.25}$, $\xi_{i4} \sim U\left(-\sqrt{3}, \sqrt{3}\right), i \in \mathbb{N}_+$.

Under the null hypothesis, i.e., (a) for Case 1 and (b) for Case 2, Table 1 shows that the rejection frequency approaches the nominal significance level α as n increases. Under the alternative hypothesis, i.e, (a) for Case 2 and (b) for Case 1, the rejection frequency is found to equal 1 for all combinations, thus the test is clearly consistent.

Table 1: Rejection frequency under null hypothesis

	(a) for Case 1				(b) for Case 2			
(n, N)	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.20$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.20$
$(150, 200)$	0.015	0.059	0.111	0.216	0.015	0.056	0.105	0.209
$(250, 500)$	0.009	0.056	0.105	0.207	0.008	0.046	0.104	0.197
$(400, 1000)$	0.007	0.053	0.109	0.196	0.008	0.052	0.098	0.203
$(600, 2000)$	0.011	0.050	0.099	0.197	0.012	0.051	0.102	0.201

5. Real Data Analysis

To further illustrate the testing procedure, an ElectroEncephalogram (EEG) data is studied. EEG is known for containing a great deal of information about the function of the brain. The data used consists of 142 people with EEG signals recorded from 32 scalp locations at 1000Hz sample rate. The mid 200 signals of each person at the 10-th scalp location are used, so the data is functional of form (1.3), with $n = 142$, $N = 200$. The null hypothesis is that the canonical FPCs of this EEG data are a finite subset of standard Fourier basis in (4.48) subject to permutation.

The default $\kappa_n = [c_1 \log n] + c_2$ with $c_1 = 3/2, c_2 = 0$ has yielded $\kappa_n = 7$. For $\hat{G}(x, x')$ defined in (1.8), the largest κ_n estimated eigenvalues are

$$\left(\hat{\lambda}_k\right)_{1 \le k \le 7} = (40.658, 9.049, 7.023, 4.482, 2.468, 1.331, 0.990),$$

with corresponding canonical FPCs $\{\psi_{0,k}(x)\}_{1\le k\le 7}$

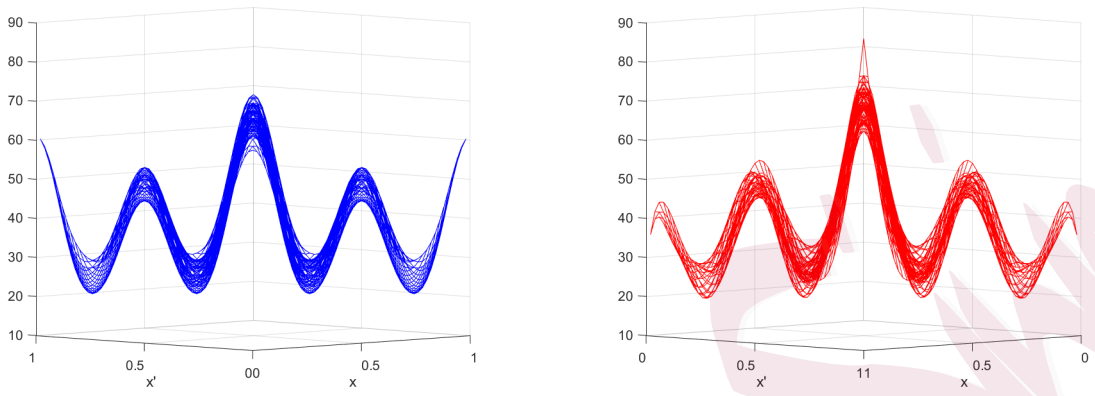
$$1, \sqrt{2}\sin(4\pi x), \sqrt{2}\cos(4\pi x), \sqrt{2}\sin(2\pi x),$$
$$\sqrt{2}\cos(2\pi x), \sqrt{2}\sin(6\pi x), \sqrt{2}\sin(8\pi x).$$

One then obtains $\hat{S}_n = 754.778$ according to (2.33), and the lowest confidence level empirical quantile $\hat{Q}_{1-\alpha}$ greater than $\hat{S}_n$ is $\hat{Q}_{0.2552} = 754.930$. So the null hypothesis is retained with p-value $= 0.7448$.

The estimated covariance function $\hat{G}(x, x')$ defined by (1.8) is in fact well approximated by $\hat{G}_0(x, x') \equiv \sum_{k=1}^{7} \hat{\lambda}_k \psi_{0,k}(x) \psi_{0,k}(x')$, with coefficient of determination $R^2 = 0.892$. Graphically, Figure 1 (a) depicts $\hat{G}_0(x, x')$, which appears to be a faithful representation of the estimated covariance function $\hat{G}(x, x')$ in Figure 1 (b).

Figure 2 shows for 4 randomly selected participants the raw EEG data $Y_{ij}, 1 \le j \le 200$ (crosses), spline estimated trajectories $\hat{\eta}_i(j/200), 1 \le j \le 200$ (solid), and null trajectories $\hat{m}(j/200) + \sum_{k=1}^{7} \hat{\zeta}_{ik} \psi_{0,k}(j/200), 1 \le j \le 200$ (dashed). The coefficients of determination of spline trajectories and null hypothesis trajectories against the 4 raw data segments are $(0.992, 0.911)$, $(0.983, 0.919)$, $(0.982, 0.927)$ and $(0.994, 0.931)$ respectively. This further validates that for this particular EEG data, the Fourier canonical FPCs are rather appropriate.

We have also tested the EEG data against the O-U FPCs in (4.47) as



(a) (b)

Figure 1: (a) The covariance function $\hat{G}_0(x, x')$ under null hypothesis in Section 5; (b) the estimated covariance function $\hat{G}(x, x')$ defined in (1.8).

canonical FPCs. Having obtained $\hat{S}_n = 2687.381$ and $\hat{Q}_{0.95} = 2588.731$, the null hypothesis is rejected with p-value< 0.05.

6. Conclusions

A chi-square type statistic is constructed via estimates of individual trajectories to test specifications of FPCs in functional data. Limiting distribution of the statistic under the null hypothesis equals an infinite Gaussian quadratic form whose quantiles are estimated consistently. The data-driven test has correct significance level under null hypothesis and is consistent under alternative if trajectory estimates satisfy some constraints, which are met by B-spline estimates. Numerical experiments demonstrate

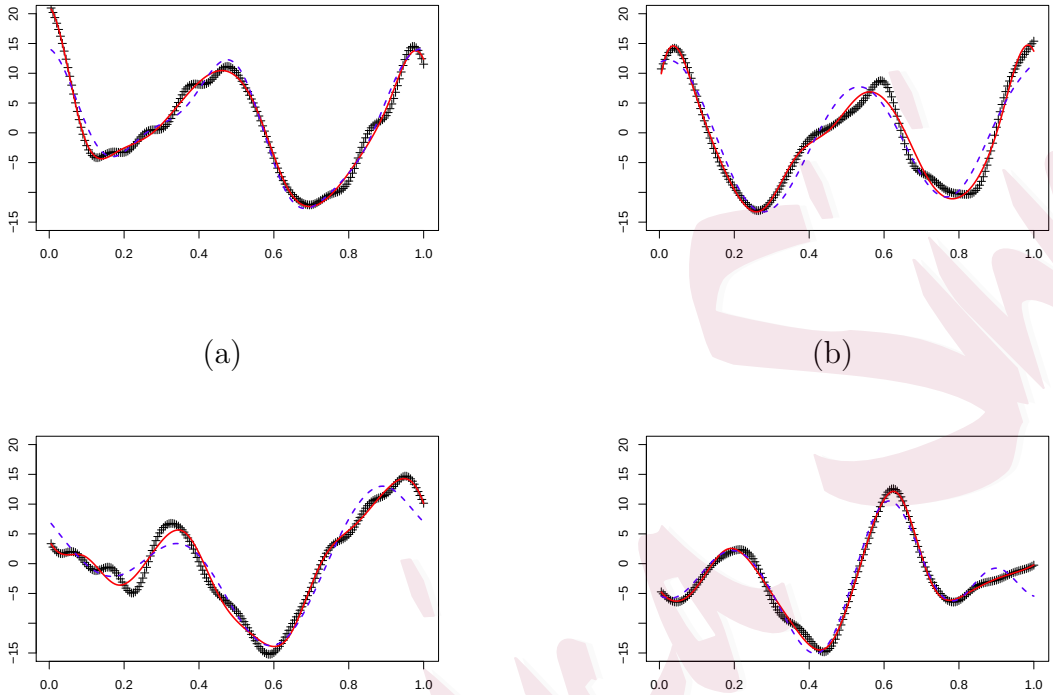


Figure 2: Randomly selected segments of raw EEG data (crosses), spline estimated trajectories (solid) and null hypothesis trajectories (dashed).

excellent performance of the test corroborating the asymptotic theory. For one EEG data, strong evidence points to canonical FPCs as a small set of standard Fourier basis. The proposed test is expected to be widely applicable in various scientific fields by simplifying functional data models with validated simple sets of FPCs.

Further research may reveal that other trajectory estimates based on

wavelet or local polynomial also satisfy Assumption (B2), and can be used to formulate tests with desirable properties in Theorems 3 and 4. It is also feasible to extend our results to functional data recorded over irregular grid, albeit with messier algebra. Similar tests may also be constructed for temporally dependent functional data, such as the functional moving average (FMA) in Li and Yang (2023) and Zhong and Yang (2023).

Supplementary Materials

The online supplement contains detailed proofs of technical results.

Acknowledgements

This research is partially supported by National Natural Science Foundation of China award 12171269. The insightful comments from Associate Editor and two Referees have led to major improvement of the paper.

References

Aue, A., Nourinho, D.D. and Hörmann, S. (2015). On the prediction of stationary functional time series. *Journal of the American Statistical Association* **110**, 378-392.

Bosq, D. (2000). *Linear Processes in Function Spaces: Theory and Applications*. Springer-Verlag, New York.

REFERENCES

Cai, L., Li, L., Huang, S., Ma, L. and Yang, L. (2020). Oracally efficient estimation for dense functional data with holiday effects. *TEST* **29**, 282-306.

Cao, G., Wang, L., Li, Y. and Yang, L. (2016). Oracle-efficient confidence envelopes for covariance functions in dense functional data. *Statistica Sinica* **26**, 359-383.

Cao, G., Yang L. and Todem, D. (2012). Simultaneous inference for the mean function based on dense functional data. *Journal of Nonparametric Statistics* **24**, 359-377.

de Boor, C. (2001). *A Practical Guide to Splines*. Springer-Verlag, New York.

Degras, D. A. (2011). Simultaneous confidence bands for nonparametric regression with functional data. *Statistica Sinica* **21**, 1735-1765.

Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. Springer-Verlag, New York.

Gu, L., Wang, L., Härdle, W. and Yang, L. (2014). A simultaneous confidence corridor for varying coefficient regression with sparse functional data. *TEST* **23**, 806-843.

Hall, P. and Hosseini-Nasab, M. (2006). On properties of functional principal components analysis. *Journal of the Royal Statistical Society: Series B* **68**, 109-126.

Horváth, L. and Kokoszka, P. (2012). *Inference for Functional Data with Applications*. Springer-Verlag, New York.

Hsing, T. and Eubank, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. Wiley, Chichester.

REFERENCES

Huang K., Chen, D., Wang, F. and Yang, L. (2021). Prediction of dispositional dialectical thinking from resting-state electroencephalography. *Brain and Behavior* **11**, e2327.

Huang, K., Zheng, S. and Yang, L. (2022). Inference for dependent error functional data with application to event related potentials. *TEST* **31**, 1100-1120.

Li, J. and Yang, L. (2023). Statistical inference for functional time series. *Statistica Sinica* **33**, 519-549.

Ma. S., Yang, L. and Carroll, R. (2012). A simultaneous confidence band for sparse longitudinal regression. *Statistica Sinica* **22**, 95-122.

Ramsay, J. and Sliverman, B. (2002). *Applied Functional Data Analysis: Methods and Case Studies*. Springer-Verlag, New York.

Ramsay, J. and Sliverman, B. (2005). *Functional Data Analysis*. Springer-Verlag, New York.

Shang, H. L. (2014). A survey of functional principal component analysis. *AStA Advances in Statistical Analysis* **98**, 121-142.

Shang, H. L. (2017). Functional time series forecasting with dynamic updating: An application to intraday particulate matter concentration. *Econometrics and Statistics* **1**, 184-200.

Wang, J., Cao, G., Wang, L. and Yang, L. (2020). Simultaneous confidence band for stationary covariance function of dense functional data. *Journal of Multivariate Analysis* **176**, 104584.

Zhang, Y., Wang, C., Wu, F., Huang, K., Yang, L. and Ji, L. (2020). Prediction of working memory ability based on EEG by functional data analysis. *Journal of Neuroscience*

Methods **333**, 108552.

Zheng, S., Yang, L. and Härdle, W. (2014). A smooth simultaneous confidence corridor for the mean of sparse functional data. *Journal of the American Statistical Association* **109**, 661-673.

Zhong, C. and Yang, L. (2023). Statistical inference for functional time series: autocovariance function. *Statistica Sinica* **in press** DOI: 10.5705/ss.202021.0121.

Center for Statistical Science & Department of Industrial Engineering, Tsinghua University, Beijing 100084, China

E-mail: szn18@mails.tsinghua.edu.cn

Center for Statistical Science & Department of Industrial Engineering, Tsinghua University, Beijing 100084, China

E-mail: yanglijian@tsinghua.edu.cn

School of Mathematical Sciences, Soochow University, Suzhou 215006, China

E-mail: zhangyy@suda.edu.cn