International Journal of Mathematics, Statistics and Operations Research
Vol. 6; Number 1; 2026, pp. 115-129
https://doi.org/10.47509/IJMSOR.2026.v06i01.06

ARF INDIA
Academic Open Access Publishing
www.arfjournals.com

Two-sample tests for sparse functional data

YANG, Qingyue and WANG, Lihong[a]

[a] *School of Mathematics, Nanjing University, Nanjing, China*

ABSTRACT
In this article, we consider the two-sample testing for sparse functional data, which is widely used in functional data analysis and longitudinal data analysis. Due to irregular observations and small sample sizes, the testing problem for sparse functional data is challenging and complicated. We propose several non-parametric testing methods using Cramér-von Mises test statistics and Energy statistics based on conditional functional principal component analysis. The numerical simulation results show that the developed procedures have good performance in controlling two types of errors. Finally, empirical analysis is conducted on the two-sample tests for a carbon monoxide concentration dataset.

KEYWORDS
Cramér-von Mises statistics, conditional principal components analysis, energy statistics, sparse functional data, two-sample tests

Preamble

Functional data analysis (FDA) is a central topic of modern statistical science and has been receiving growing attention from both researchers and practitioners. The applications of FDA across various industries are numerous, from environmental science, finance, food science, to neuromedicine, see, e.g., King et al. [8], Peltier et al. [12], Tian et al. [18], and Yu et al. [22].

Functional data can be categorized into dense and sparse designs based on measurement density. Most existing research has focused on dense data, where observing time points per subject are densely and regularly distributed, allowing smooth curves to be easily obtained through parametric methods. However, sparse functional data involves only limited amount of irregularly-spaced measurements for each subject. This type of data arises frequently in many real world applications, particularly in longitudinal studies.

The two-sample testing is a fundamental problem in FDA. For two-sample test of dense functional data, Benko et al. [3] decomposed the two-sample distribution test into tests for the mean function, characteristic function, and eigenvalues based on principal component analysis. Corain et al. [5] employed permutation and combination-based tests and time-to-time analysis. Fan and Lin [6] proposed the adaptive Neyman test and the wavelet thresholding tests for comparing two sets of curves. Hall and

CONTACT Author[a]. Email:lhwang@nju.edu.cn

Article History
Received: 24 April 2026; Revised: 20 May 2026; Accepted: 27 May 2026; Published: 30 June 2026

To cite this paper
YANG, Qingyue and WANG, Lihong (2026). Two-sample tests for sparse functional data. *International Journal of Mathematics, Statistics and Operations Research.* 6(1), 115-129.

Keilegom [7] extended the Cramér-von Mises tests from the multivariate case to the functional case, directly computing the empirical distribution function for smoothed functional curves and using the bootstrap method to determine critical values. Krzyśko and Smaga [9] used Energy statistics for conformity of distributions based on a basis function representation of functional data. However, these approaches for dense data are not directly applicable to sparse data.

Regarding the study of sparse data, Ma et al. [10] presented simultaneous confidence bands for the mean function in sparse functional data via a piecewise-constant spline smoothing approach. Müller [11] provided a comprehensive overview on sparse data analysis. Pomann et al. [14] extended the traditional two-sample Anderson-Darling test statistic to sparse functional data by using marginal functional principal component analysis. Staicu et al. [16] introduced a global pseudo likelihood ratio test which can be applied to test of the mean function in sparsely observed functional data. Wang [19] developed a two-sample test for detecting difference among mean functions for sparse and irregularly-observed data based on the conditional principal component analysis. Yao et al. [21] proposed the conditional principal component analysis method for estimating the functional principal component score, mean function, and covariance function. In this article, we further study the two-sample testing problem for sparse functional data and focus on testing the null hypothesis that two samples of curves with sparse and irregular observations have the same underlying distribution.

Assume that $X_1(\cdot)$ and $X_2(\cdot)$ are independent square-integrable random processes defined on T, where T is a compact interval. Suppose we have observed independent samples $\{X_{1i}(\cdot)\}_{i=1}^{n_1}$ and $\{X_{2i}(\cdot)\}_{i=1}^{n_2}$ from two populations $X_1(\cdot)$ and $X_2(\cdot)$, respectively, among which the i-th sample has N_{ri} observations X_{rij} at possibly irregular time points $\{T_{rij}\}_{1\leq i\leq n_r, 1\leq j\leq N_{ri}}$ for $r = 1, 2$, and the grid of points for each subject is sparse, i.e. N_{1i} and N_{2i} are as small as few observations.

For the two-sample inference problem, the primary focus is to compare the distributions of the two samples, i.e., testing the following hypothesis:

$$H_0 : X_1(\cdot) \stackrel{d}{=} X_2(\cdot) \longleftrightarrow H_A : X_1(\cdot) \stackrel{d}{\neq} X_2(\cdot). \tag{1}$$

The detailed testing procedures are introduced in the next section. In section 2, numerical simulations and empirical analysis are provided to compare the proposed two-sample tests for sparse functional data. The simulation results demonstrate the good performance of the proposed methods in identifying the difference between the two populations.

1. Two-sample testing for sparse data

Let the sample sizes n_1 and n_2 tend to infinity, with $\lim_{n_1,n_2\to\infty} n_1/(n_1+n_2) = p \in (0, 1)$. Denote by $X(t)$ the mixture process of $X_1(t)$ and $X_2(t)$, taking values $X_1(t)$ and $X_2(t)$ with probabilities p and $1-p$, respectively. Let $\mu(t)$ be the mean function of $X(t)$ and $G(s, t)$ the covariance function. By Mercer's theorem, $G(s, t) = \sum_{k\geq 1} \lambda_k \phi_k(s)\phi_k(t)$, where $\{\phi_k(\cdot), k = 1, 2, \cdots, \infty\}$ are the eigenfunctions of $G(s, t)$ forming an orthonormal basis of $L^2(T)$, corresponding to the non-increasing eigenvalue sequence

$\{\lambda_k, k = 1, 2, \cdots, \infty\}$. It follows from the Karhunen-Loéve (K-L) expansion that $X(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_k \phi_k(t)$, where $\xi_k = \int_T (X(t) - \mu(t))\phi_k(t)dt$ are the functional principal component (FPC) scores, which are random variables with mean 0 and covariance λ_k. Truncating the infinite series in the K-L expansion yields $X^K(t) = \mu(t) + \sum_{k=1}^{K} \xi_k \phi_k(t)$. Then $X^K(t)$ converges uniformly to $X(t)$ in the L^2 norm as $K \to \infty$.

Define

$$X_1^K(t) = \mu(t) + \sum_{k=1}^{K} \xi_{1k}\phi_k(t),\ X_2^K(t) = \mu(t) + \sum_{k=1}^{K} \xi_{2k}\phi_k(t),$$

where

$$\xi_{1k} = \int_T (X_1(t) - \mu(t))\phi_k(t)dt,\ \xi_{2k} = \int_T (X_2(t) - \mu(t))\phi_k(t)dt. \tag{2}$$

Obviously, $X_1^K(t)$ and $X_2^K(t)$ converge to $X_1(t)$ and $X_2(t)$ as $K \to \infty$, respectively. Let $\xi_{r\cdot} = (\xi_{r1}, \xi_{r2}, \cdots, \xi_{rK})^T$, $r = 1, 2$. Then testing (1) is asymptotically equivalent to a multi-dimensional two-sample testing

$$H_0^K : \xi_{1\cdot} \stackrel{d}{=} \xi_{2\cdot} \longleftrightarrow H_A^K : \xi_{1\cdot} \stackrel{d}{\neq} \xi_{2\cdot}. \tag{3}$$

In order to choose the number K of eigenfunctions that provide a reasonable approximation to the infinite-dimensional process, one may adapt AIC type criteria or use the cross-validation score based on the one-curve-leave-out prediction error. The details can be found in Section 2.5 of Yao et al. [21].

Given random subjects $\{X_{1i}(\cdot)\}_{i=1}^{n_1}$ and $\{X_{2i}(\cdot)\}_{i=1}^{n_2}$, by (2), the FPC scores ξ_{1ik} and ξ_{2ik} for two samples are defined as

$$\xi_{1ik} = \int_T (X_{1i}(t) - \mu(t))\phi_k(t)dt,\ \xi_{2ik} = \int_T (X_{2i}(t) - \mu(t))\phi_k(t)dt. \tag{4}$$

As discussed in Yao et al. [21], for dense functional data, the FPC scores in (4) have traditionally been estimated by numerical integration which works well when the density of the grid of measurements for each subject is sufficiently large. However, for sparse functional data, this method will not provide reasonable approximations to FPC scores, for example when one has only two observations per subject. Therefore, Yao et al. [21] developed the Principal Components Analysis through Conditional Expectation (PACE) method to estimate the FPC scores.

In order to obtain the consistent estimators of the FPC scores, we consider the pooled sample $\{(T_{1ij}, X_{1ij}), j = 1, \cdots, N_{1i}\}_{i=1}^{n_1} \bigcup \{(T_{2ij}, X_{2ij}), j = 1, \cdots, N_{2i}\}_{i=1}^{n_2}$, where $\{(T_{ij}, Y_{ij}), j = 1, \cdots, N_i\}_{i=1}^{n}$ are the generic functional observations in the pooled set. The measurements $\{Y_{ij}, j = 1, \cdots, N_i\}$ are viewed as the observations of the samples $\{Y_i(t)\}_{i=1}^{n}$ of the mixture process $X(t)$ at time points $\{T_{ij}, j = 1, \cdots, N_i\}$ and are contaminated with errors. Specifically, the observation model is given by

$$Y_{ij} = Y_i(T_{ij}) + \epsilon_{ij},$$

where $\{\epsilon_{ij}\}$ are independent and identically distributed (i.i.d.) noises with mean 0 and

variance σ^2, and are assumed to be independent of $X(t)$.

First, we review the local linear scatterplot smoother for $\mu(t)$, which can be obtained from the entire data ensemble by minimizing

$$\sum_{i=1}^{n}\sum_{j=1}^{N_i}\kappa_1\left(\frac{T_{ij}-t}{h_\mu}\right)\left\{Y_{ij}-\beta_0-\beta_1(T_{ij}-t)\right\}^2 \tag{5}$$

with respect to β_0, β_1. The estimate of $\mu(t)$ is then $\hat{\mu}(t)=\hat{\beta}_0(t)$. In a similar way, the local linear surface smoother for $G(s,t)$ is defined by minimizing

$$\sum_{i=1}^{n}\sum_{1\le j\ne l\le N_i}\kappa_2\left(\frac{T_{ij}-s}{h_G},\frac{T_{il}-t}{h_G}\right)\left\{\widehat{\varepsilon}_{ij}\widehat{\varepsilon}_{il}-\beta_0-\beta_{11}(T_{ij}-s)-\beta_{12}(T_{il}-t)\right\}^2 \tag{6}$$

with regard to β_0, β_{11}, β_{12}, where $\widehat{\varepsilon}_{ij}=Y_{ij}-\hat{\mu}(T_{ij})$, κ_1, κ_2 are nonnegative univariate and bivariate kernel functions, and h_μ, h_G are bandwidths. This yields the estimate $\widehat{G}(s,t)=\hat{\beta}_0(s,t)$.

Then, as described in Rice and Silverman [15], the estimates of eigenfunctions and eigenvalues obtained from (5) and (6) correspond to the solutions $\hat{\phi}_k$ and $\hat{\lambda}_k$ of the eigen-equations

$$\int_T \widehat{G}(s,t)\hat{\phi}_k(s)ds=\hat{\lambda}_k\hat{\phi}_k(t).$$

Let $\boldsymbol{Y}_i=(Y_{i1},Y_{i2},\cdots,Y_{iN_i})^T$. Denote the covariance matrix of $\boldsymbol{Y}_i$ as G_i, where its (j,l) entry is $(G_i)_{j,l}=G(T_{ij},T_{il})+\sigma^2\delta_{jl}$ with $\delta_{jl}=1$ if $j=l$ and 0 if $j\ne l$. $(G_i)_{j,j}$ can be estimated by $\widehat{(G_i)}_{j,j}=\widehat{G}(T_{ij},T_{ij})+\hat{\sigma}^2$, where $\hat{\sigma}^2$ can be obtained using the method introduced in Section 2.2 of Yao et al. [21]. One can also estimate $(G_i)_{j,j}$ directly by local linear smoother, i.e., $\widehat{(G_i)}_{j,j}=\widehat{a}_1(T_{ij})$, where

$$(\widehat{a}_1,\widehat{b}_1)=\arg\min_{a_1,b_1}\sum_{i=1}^{n}\sum_{j=1}^{N_i}\kappa_1\left(\frac{T_{ij}-t}{h'_G}\right)\left\{\widehat{\varepsilon}_{ij}^2-a_1-b_1\left(T_{ij}-t\right)\right\}^2.$$

Let $\boldsymbol{X}_{ri}=(X_{ri1},X_{ri2},\cdots,X_{riN_{ri}})^T$, $\hat{\boldsymbol{\mu}}_{ri}=(\hat{\mu}(T_{ri1}),\hat{\mu}(T_{ri2}),\cdots,\hat{\mu}(T_{riN_{ri}}))^T$, $\widehat{\boldsymbol{\phi}}_{rik}=(\hat{\phi}_k(T_{ri1}),\ \hat{\phi}_k(T_{ri2}),...,\hat{\phi}_k(T_{riN_{ri}}))^T$, $r=1,2$. Using PACE method based on the best linear prediction as in Yao et al. [21], the consistent estimators of the FPC scores are obtained by

$$\hat{\xi}_{rik}=\hat{\lambda}_k\widehat{\boldsymbol{\phi}}_{rik}^T\widehat{G}_i^{-1}(\boldsymbol{X}_{ri}-\hat{\boldsymbol{\mu}}_{ri}). \tag{7}$$

Based on the estimated FPC scores in (7), we introduce six test statistics for the hypothesis testing (3). Let $\hat{\xi}_{ri\cdot}=(\hat{\xi}_{ri1},\hat{\xi}_{ri2},...,\hat{\xi}_{riK})^T$, $i=1,2,\cdots,n_r$, $r=1,2$. Let $\hat{F}_r(x)$ be the estimator of the empirical distribution function of $\xi_{r\cdot}$ obtained from the "samples" $\{\hat{\xi}_{ri\cdot}\}_{i=1}^{n_r}$. Denote $\hat{F}(x)=\frac{1}{n}\{n_1\hat{F}_1(x)+n_2\hat{F}_2(x)\}$ the mixed empirical distribution function.

Three Cramer-von Mises type of test statistics (see, e.g. Anderson[1], Pettitt[13],

Watson[20]) are defined as

$$W^2 = \frac{n_1 n_2}{n^2} \sum_{r=1}^{2} \sum_{i=1}^{n_r} (\hat{F}_1(\hat{\xi}_{ri\cdot}) - \hat{F}_2(\hat{\xi}_{ri\cdot}))^2,$$

$$AD^2 = \frac{n_1 n_2}{n^2} \sum_{r=1}^{2} \sum_{i=1}^{n_r} \frac{(\hat{F}_1(\hat{\xi}_{ri\cdot}) - \hat{F}_2(\hat{\xi}_{ri\cdot}))^2}{\hat{F}(\hat{\xi}_{ri\cdot})(1 - \hat{F}(\hat{\xi}_{ri\cdot}))},$$

$$U^2 = \frac{n_1 n_2}{n^2} \sum_{r=1}^{2} \sum_{i=1}^{n_r} \left\{ \hat{F}_1(\hat{\xi}_{ri\cdot}) - \hat{F}_2(\hat{\xi}_{ri\cdot}) - \frac{1}{n} \sum_{r=1}^{2} \sum_{i=1}^{n_r} (\hat{F}_1(\hat{\xi}_{ri\cdot}) - \hat{F}_2(\hat{\xi}_{ri\cdot})) \right\}^2.$$

Under the null hypothesis, the test statistics should be close to 0. Therefore, when the statistic is greater than the upper α quantile (i.e., the critical value), the null hypothesis is rejected. The exact null distributions of W^2, AD^2 and U^2 are difficult to determine, an appropriate approximation is needed. One can use permutation tests or bootstrap procedures to approximate the finite sample p-values at a given significance level α.

Next we construct three test statistics using Energy statistics introduced in Aslan and Zech[2], Chen et al. [4], and Székely and Rizzo [17]. Denote $\|\cdot\|$ the Euclidean norm. The Energy test statistics are defined as

$$E_{sr} = \frac{2}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \|\hat{\xi}_{1i\cdot} - \hat{\xi}_{2j\cdot}\| - \frac{1}{n_1^2} \sum_{i,j=1}^{n_1} \|\hat{\xi}_{1i\cdot} - \hat{\xi}_{1j\cdot}\| - \frac{1}{n_2^2} \sum_{i,j=1}^{n_2} \|\hat{\xi}_{2i\cdot} - \hat{\xi}_{2j\cdot}\|,$$

$$E_{az} = \frac{2}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \log \|\hat{\xi}_{1i\cdot} - \hat{\xi}_{2j\cdot}\| - \frac{1}{n_1^2} \sum_{i,j=1}^{n_1} \log \|\hat{\xi}_{1i\cdot} - \hat{\xi}_{1j\cdot}\| - \frac{1}{n_2^2} \sum_{i,j=1}^{n_2} \log \|\hat{\xi}_{2i\cdot} - \hat{\xi}_{2j\cdot}\|,$$

$$E_{exp} = \frac{1}{n_1^2} \sum_{i,j=1}^{n_1} e^{-\|\hat{\xi}_{1i\cdot} - \hat{\xi}_{1j\cdot}\|} + \frac{1}{n_2^2} \sum_{i,j=1}^{n_2} e^{-\|\hat{\xi}_{2i\cdot} - \hat{\xi}_{2j\cdot}\|} - \frac{2}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} e^{-\|\hat{\xi}_{1i\cdot} - \hat{\xi}_{2j\cdot}\|}.$$

Again, to obtain the critical values or p values of the tests at a given significance level, one can use permutation tests to approximate the null distribution through shuffling procedures. Note that Energy test statistics measure the distance between two characteristic functions and have rotational invariance, while Cramér-von Mises test statistics compute the distance between distribution functions and do not have the property of rotational invariance. This makes Energy statistics more adaptable in the multi-dimensional testing problems. We will compare the performance of these statistics via simulations in next section.

2. Numerical simulations and empirical studies

In this section, we evaluate the performance of the proposed approaches via simulation. We also apply the proposed test for an empirical analysis of the carbon monoxide concentration data in an Italian city. All computations are performed using Matlab R2024a.

2.1. Numerical simulations

In this subsection we conduct simulation studies to illustrate and compare the effectiveness of the proposed six test statistics. Specifically, we generate samples of size $(n_1, n_2) = (50, 50), (50, 100), (100, 100), (100, 150)$, respectively, as follows.

$$X_{rij} = \mu_r(T_{rij}) + \sum_{k=1}^{K} \xi_{rik}\phi_{rk}(T_{rij}) + \epsilon_{rij}, \quad i = 1, ..., n_r, \ j = 1, \cdots, N_{ri}, \ r = 1, 2, \quad (8)$$

where the errors ϵ_{rij} are i.i.d. from $N(0, 1)$, T_{rij} follow a uniform distribution on $[0, 10]$ with N_{ri} from a discrete uniform distribution on $[2, 10]$, ξ_{rik} are independently sampled from $N(0, \lambda_k)$ with λ_k chosen from a discrete uniform distribution on $[5, 15]$. For eigenfunctions ϕ_{rk}, we use three types of basis functions: Fourier basis, B-Spline basis and Polynomial basis. We set $K = 3, 5$ and the significance level $\alpha = 0.05$. For the smoothing steps, univariate and bivariate Epanechnikov kernel functions are used. The number of eigenfunctions in each simulation run is chosen by the AIC criterion and the criterion based on the percentage of the cumulative explained variance (90%).

We consider five types of setting:

1 (null hypothesis). $\mu_1(t) = \mu_2(t) = 0$.

2 (different means). $\mu_1(t) = \ln(t)$, $\mu_2(t) = \ln(t) + \delta t$, where $\delta = 0.1, 0.2, \cdots, 1$.

3 (different variances). $\mu_1(t) = \mu_2(t) = 0$, ξ_{rik} are independently sampled from $N(0, \lambda_{rk})$ with $\lambda_{1k} = 1 + k$ and $\lambda_{2k} = \lambda_{1k} + \delta$, where $\delta = 1, 2, \cdots, 9$.

4 (continuous process and mean shift). Data are generated by Brownian motion, i.e., data process satisfies $dX(t) = \mu(t)dt + \tilde{\sigma}dW(t)$. We set $\tilde{\sigma} = 0.3$, $\mu_1(t) = 0$, $\mu_2(t) = \delta$, where $\delta = 0, 0.02, \cdots, 0.08$, and randomly select 2 to 10 points from each subject to generate sparse data.

5 (continuous process and variance shift). Data are generated by Brownian motion. We set $\mu_1(t) = \mu_2(t) = 0$, $\tilde{\sigma}_1 = 0.3$, $\tilde{\sigma}_2 = 0.3 + \delta$, where $\delta = 0, 0.1, \cdots, 0.4$, and randomly select 2 to 10 points from each subject to generate sparse data.

Now we report the comparison of the performance of six tests W^2 (KS), AD^2 (AD), U^2 (Watson), E_{sr} (Energy(sr)), E_{az} (Energy(az)) and E_{exp} (Energy(exp)). Table 1 summarises the simulation results obtained for setting 1 with three types of basis functions and different values of n_1, n_2 and K, when the null hypothesis holds. The empirical size is approximated by an average over 1000 simulation runs for each scenario. Table 2 compares the mean squared error (MSE) between the empirical size and the nominal significance level.

Tables 1 and 2 reveal that, the empirical sizes of all test statistics are relatively stable around 0.05, except for the Energy statistic E_{exp}, which exhibits several empirical size values much smaller than $\alpha = 0.05$ in the case of Polynomial basis. Overall, all test statistics are not sensitive to the choice of the type of basis functions, the number of eigenfunctions and sample size. From Table 2, it can be seen that the Energy statistic E_{sr} has best performance, followed by the Cramér-von Mises test statistic AD^2.

Next, we consider the alternative hypothesis. In a similar fashion, the computation results under setting 2 are presented in Table 3. Figure 1 depicts the empirical powers of six test statistics using Fourier basis with $n_1 = n_2 = 100$ for different δ values, respectively.

The simulation results in Table 3 and Figure 1 indicate that the Energy statistics E_{sr} and E_{az} outperform the other tests, especially E_{sr}, which shows the highest powers in all cases. Among the Cramér-von Mises test statistics, AD^2 statistic performs the best, but still lower than E_{sr} and E_{az}. The performance of Energy statistics E_{exp} is

Table 1. The empirical sizes for various scenarios under setting 1

Sample size	(n_1, n_2)=(50,50)		(n_1, n_2)=(50,100)		(n_1, n_2)=(100,100)		(n_1, n_2)=(100,150)	
K	3	5	3	5	3	5	3	5
Basis				Fourier				
KS	0.042	0.043	0.042	0.045	0.046	0.049	0.050	0.048
AD	0.049	0.040	0.045	0.039	0.055	0.062	0.057	0.050
Watson	0.049	0.042	0.050	0.056	0.046	0.044	0.045	0.060
Energy(sr)	0.045	0.055	0.050	0.045	0.054	0.053	0.054	0.043
Energy(az)	0.046	0.046	0.045	0.041	0.052	0.053	0.049	0.038
Energy(exp)	0.051	0.044	0.046	0.046	0.054	0.044	0.048	0.053
Basis				B-spline				
KS	0.044	0.040	0.031	0.035	0.049	0.040	0.053	0.056
AD	0.063	0.046	0.044	0.041	0.055	0.048	0.051	0.060
Watson	0.046	0.030	0.043	0.036	0.055	0.056	0.055	0.051
Energy(sr)	0.064	0.047	0.042	0.041	0.048	0.045	0.047	0.051
Energy(az)	0.069	0.037	0.049	0.034	0.046	0.053	0.053	0.049
Energy(exp)	0.046	0.048	0.054	0.048	0.047	0.044	0.053	0.058
Basis				Polynomial				
KS	0.037	0.049	0.056	0.063	0.062	0.057	0.046	0.047
AD	0.045	0.059	0.048	0.056	0.052	0.050	0.046	0.043
Watson	0.044	0.052	0.061	0.071	0.058	0.053	0.048	0.057
Energy(sr)	0.044	0.060	0.045	0.059	0.051	0.051	0.044	0.046
Energy(az)	0.043	0.066	0.056	0.069	0.056	0.050	0.045	0.045
Energy(exp)	0.053	0.016	0.056	0.040	0.036	0.023	0.042	0.042

the worst. The reason is that in exponential form, the value of the statistic becomes very small under negative exponentiation even if the distance between two samples is very large, resulting in poor testing performance.

Furthermore, the simulation results under setting 3 are displayed in Table 4 and Figure 2 which exhibit the empirical powers of six test statistics using Fourier basis with $n_1 = n_2 = 100$ for different δ, respectively. It can be seen that the Energy statistic E_{az} is the most powerful among the six statistics, followed by the Cramér-von Mises statistic U^2. In addition, statistics E_{sr} and AD^2 also have satisfactory performance, while the statistics W^2 and E_{exp} have relatively low powers especially for small δ.

Generally speaking, the above simulation evidence suggests that Energy statistics E_{sr} and E_{az} should be a good choice for two-sample testing of functional sparse data. The test statistic AD^2 is also quite powerful and is worth recommending.

Regarding setting 4, comparing results in Table 5 and Figure 3, we find that the three Energy statistics perform better than Cramér-von Mises statistics, among them E_{sr} has the best performance under the alternative hypothesis. In addition, the statistic

Table 2. MSEs of the empirical sizes under setting 1

	Fourier	B-spline	Polynomial	Overall
KS	2.23E-04	8.68E-04	5.93E-04	1.68E-03
AD	4.65E-04	4.32E-04	**2.15E-04**	**1.11E-03**
Watson	2.78E-04	7.48E-04	7.28E-04	1.75E-03
Energy(sr)	**1.65E-04**	**3.89E-04**	**2.96E-04**	**8.50E-04**
Energy(az)	2.96E-04	8.22E-04	7.88E-04	1.91E-03
Energy(exp)	**1.34E-04**	**1.58E-04**	2.35E-03	2.65E-03

Table 3. The empirical powers of six test statistics for different δ values under setting 2

δ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
KS	5.1%	13.2%	30.1%	48.0%	72.9%	91.0%	96.8%	99.5%	100.0%	100.0%
AD	8.9%	15.1%	36.4%	60.6%	85.4%	97.1%	99.0%	100.0%	100.0%	100.0%
Watson	4.4%	6.6%	15.5%	22.7%	42.0%	63.3%	81.1%	91.5%	97.7%	99.6%
Energy(sr)	9.5%	23.2%	50.7%	75.9%	93.6%	98.8%	100.0%	100.0%	100.0%	100.0%
Energy(az)	7.9%	21.0%	43.0%	68.1%	88.9%	97.6%	99.5%	100.0%	100.0%	100.0%
Energy(exp)	5.3%	4.3%	5.2%	6.0%	9.3%	9.7%	11.3%	13.0%	17.7%	21.8%

Table 4. The empirical powers of six test statistics for different δ values under setting 3

δ	1	2	3	4	5	6	7	8	9
KS	5.7%	9.8%	17.0%	27.6%	42.7%	59.8%	72.0%	83.7%	90.7%
AD	7.6%	16.6%	30.9%	55.3%	80.6%	92.8%	97.7%	99.8%	100.0%
Watson	10.9%	40.6%	76.9%	94.0%	98.9%	100.0%	100.0%	100.0%	100.0%
Energy(sr)	8.9%	26.6%	65.4%	92.2%	99.1%	100.0%	100.0%	100.0%	100.0%
Energy(az)	12.8%	50.9%	89.7%	99.4%	99.9%	100.0%	100.0%	100.0%	100.0%
Energy(exp)	8.0%	19.7%	37.1%	54.0%	73.0%	83.8%	88.8%	93.7%	96.6%

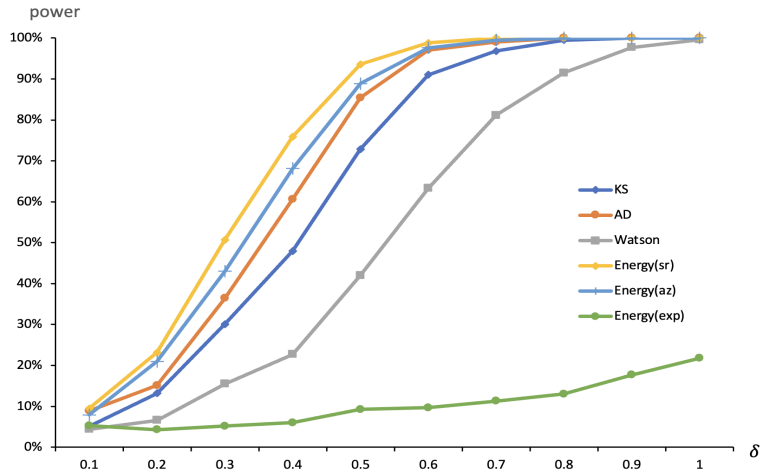

Figure 1. The empirical powers of six test statistics for different δ values under setting 2

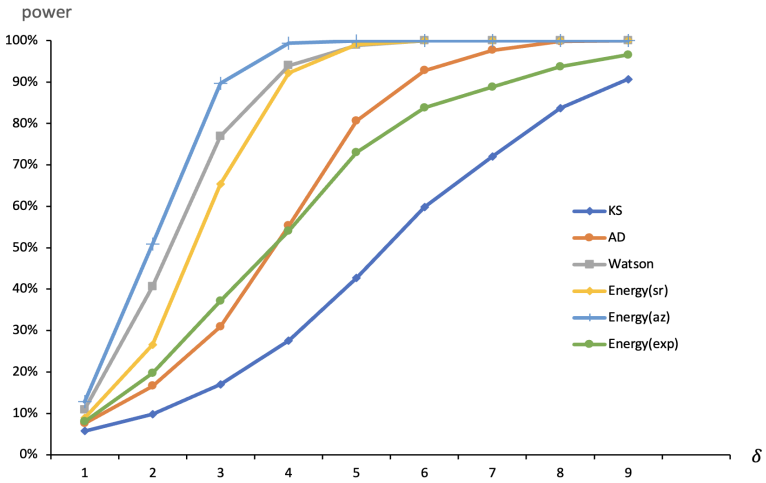

Figure 2. The empirical powers of six test statistics for different δ values under setting 3

AD^2 outperforms the other two Cramér-von Mises statistics for $\delta \geq 0.04$, while the statistic U^2 performs the worst. However, when the null hypothesis holds or for very small δ, the performance of the statistics U^2, E_{az} and E_{exp} is superior to the other three statistics. In Figure 3, we can see more clearly the performance of each statistic.

Moreover, Table 6 and Figure 4 show that the statistic U^2 performs the best under setting 5. E_{az} and E_{exp} have similar performance and are second only to U^2. The performance of E_{sr} and AD^2 is also very similar with satisfactory powers, while the statistic W^2 performs the worst.

Overall, the ranking of the six statistics is summarized in Table 7 based on the simulations conducted under settings 2-5 where the null hypothesis does not hold. One can conclude that the Energy statistics E_{sr} and E_{az} have good performance in various situations. Next is AD^2, which shows a moderate level of testing performance compared to the other statistics. The statistic W^2 performs poorly in almost all scenarios. The performance of E_{exp} is unstable with low powers in discrete cases, but performs well

under continuous settings. The statistic U^2 also has obvious characteristics: it has better performance in detecting difference in variances, especially for Brownian motion setting, while it is insensitive to difference in means.

Table 5. The empirical powers of six test statistics for different δ values under setting 4

δ	0	0.02	0.04	0.06	0.08
KS	1.5%	1.8%	19.4%	81.0%	99.4%
AD	1.9%	2.2%	30.5%	98.7%	100.0%
Watson	5.9%	6.4%	11.5%	37.9%	80.9%
Energy(sr)	0.4%	4.2%	91.7%	100.0%	100.0%
Energy(az)	3.9%	7.2%	57.6%	99.8%	100.0%
Energy(exp)	4.1%	7.1%	51.9%	99.5%	100.0%

Table 6. The empirical powers of six test statistics for different δ values under setting 5

δ	0	0.1	0.2	0.3	0.4
KS	1.5%	7.9%	32.7%	72.7%	94.1%
AD	1.9%	13.7%	65.4%	97.6%	99.9%
Watson	5.9%	57.9%	97.5%	99.9%	100.0%
Energy(sr)	0.4%	15.9%	68.6%	95.5%	99.8%
Energy(az)	3.9%	29.9%	85.7%	98.8%	100.0%
Energy(exp)	4.1%	31.0%	85.6%	98.7%	100.0%

Table 7. Summary of the performance of six statistics

Settings	Data generation	The empirical powers: from the highest (left) to the lowest (right)					
2, 4	Fourier Basis	Energy(sr)	Energy(az)	AD	KS	Watson	Energy(exp)
	Brownian Motion	Energy(sr)	Energy(az)	Energy(exp)	AD	KS	Watson
3, 5	Fourier Basis	Energy(az)	Watson	Energy(sr)	AD	Energy(exp)	KS
	Fourier Basis	Watson	Energy(exp)	Energy(az)	Energy(sr)	AD	KS

2.2. Empirical study

In this subsection we apply the proposed test statistics to a real dataset. The dataset is the urban air quality data, which is available at https://archive.ics.uci.edu/dataset/360/air+quality#.

The dataset was collected in an Italian city, spanning from March 2004 to April 2005, recorded hourly. It contains 9,358 records, including hourly average concentrations of carbon monoxide (CO), benzene, and nitrogen oxides.

We mainly consider the CO concentration data. The complete data should have 24 hours of records per day, but there are many missing values in the original data. The missing degree of CO concentration reaches 18%, which makes it applicable for sparse

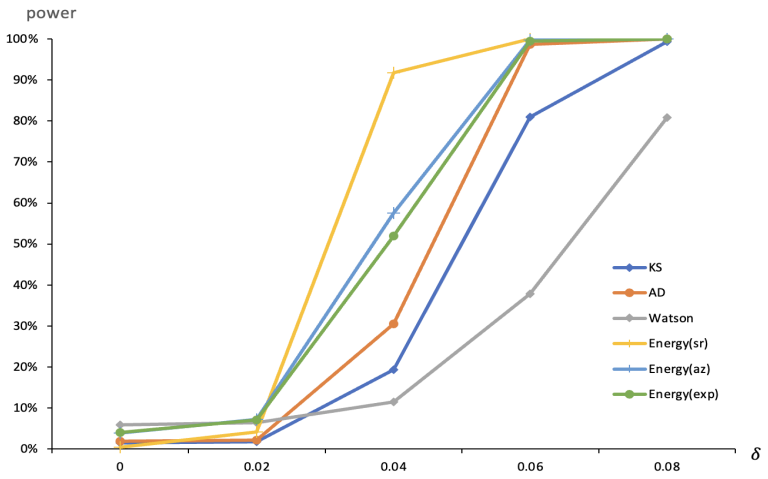

Figure 3. The empirical sizes and powers of six test statistics for different δ values under setting 4

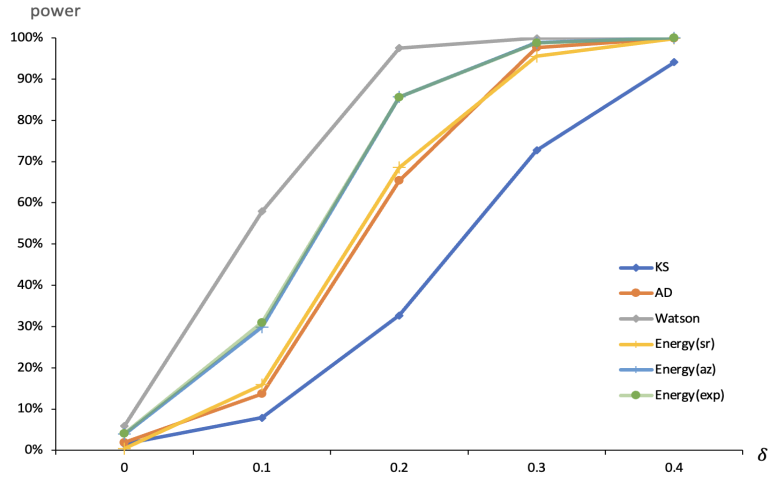

Figure 4. The empirical sizes and powers of six test statistics for different δ values under setting 5

functional data analysis. One can view the daily CO concentration as a function on $[0, 24]$ and convert the data into 355 functional data subjects. There are 254 workdays and 101 weekends in total, based on which we divide the data into two samples.

Figure 5 displays the plots of CO concentration data on workdays and weekends, respectively. We also calculate the first three functional principal component eigenfunctions based on the entire data and the corresponding principal component scores for two samples, as shown in Figure 6. It can be seen that the CO concentration on weekends is slightly lower than that on workdays. Two distinct peaks can be observed on workdays, corresponding to the rush hours. The peak values over the weekend are relatively scattered and there is no significant peak.

Now we use the six two-sample testing methods to analyze whether there are significant differences in the distributions of CO concentration on weekdays and weekends. The results are shown in Table 8. It can be seen that five tests reject the null hypothesis with small p-values, especially the Cramér-von Mises test statistics have significant

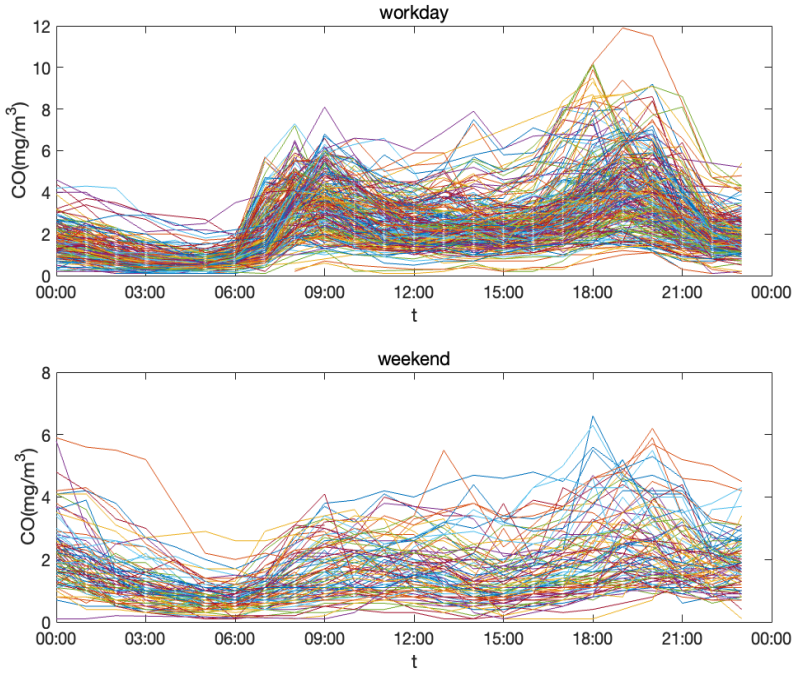

Figure 5. CO concentration: workday (upper) and weekend (lower)

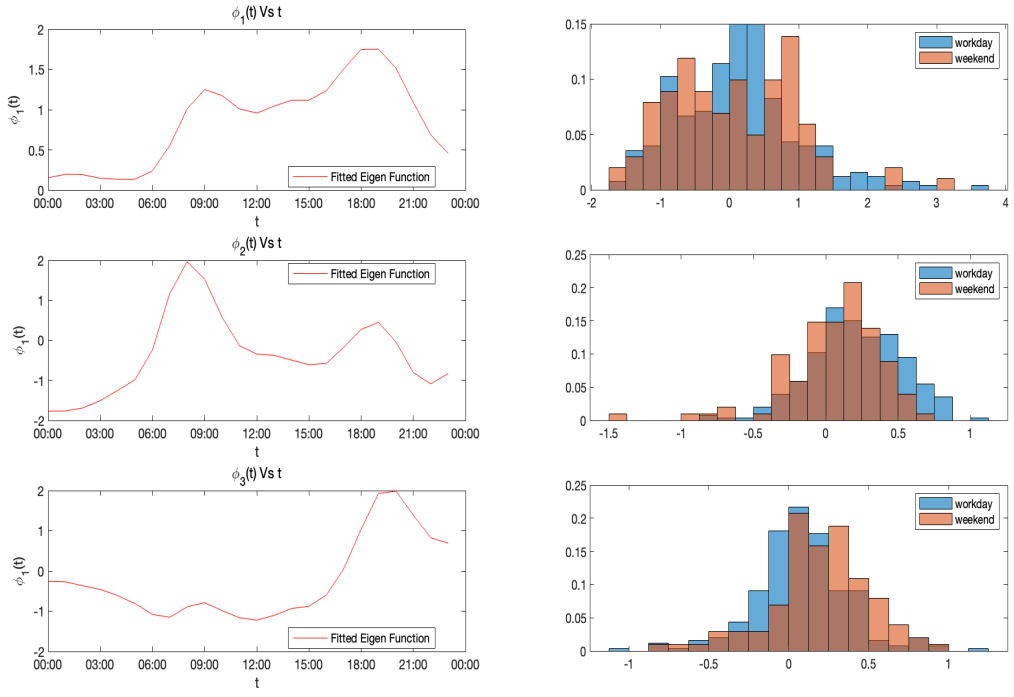

Figure 6. The first three eigenfunctions (left) and the corresponding FPC scores for two samples (right)

small p-values. The Energy statistic E_{exp} is infeasible, as the value of the statistic is negative infinity. Overall, one can reject the null hypothesis and conclude that the daily CO concentration has different distributions on weekdays and weekends.

Table 8. Two-sample tests for the CO concentration data

Test	p-value	statistic value	Whether to reject H_0
KS	9.08E-11	0.4177	Y
AD	2.07E-28	22.8314	Y
Watson	8.75E-07	0.8530	Y
Energy(sr)	0.001	1.0610	Y
Energy(az)	-	-Inf	-
Energy(exp)	0.001	-0.0002	Y

We further conduct the principal component analysis and estimate the mean functions for two samples, respectively. Figure 7 shows the fitted mean functions of two samples, and their first three eigenfunctions. It is obvious that the two samples have different mean functions. Moreover, the CO concentration fluctuates significantly during the weekdays, while the changes are more gradual over the weekend.

Table 9. Testing results under different degrees of sparsity

Sparsity	25%		30%		50%	
Test	p-value	reject H_0	p-value	reject H_0	p-value	reject H_0
KS	1.22E-06	Y	2.01E-06	Y	1.57E-05	Y
AD	3.42E-18	Y	9.50E-18	Y	1.18E-20	Y
Watson	2.89E-06	Y	3.64E-06	Y	1.87E-04	Y
Energy(sr)	0.001	Y	0.001	Y	0.002	Y
Energy(az)	0.001	Y	0.007	Y	0.002	Y

In order to explore the impact of sparsity on the testing results, we artificially construct sparse data in this example. We define the proportion of missing data as data sparsity. The sparsity of the original data is 18%. We randomly remove some data to increase the sparsity to 25%, 40%, and 50%. The corresponding test results are listed in Table 9. One can see that even when the sparsity reaches 50%, the five test

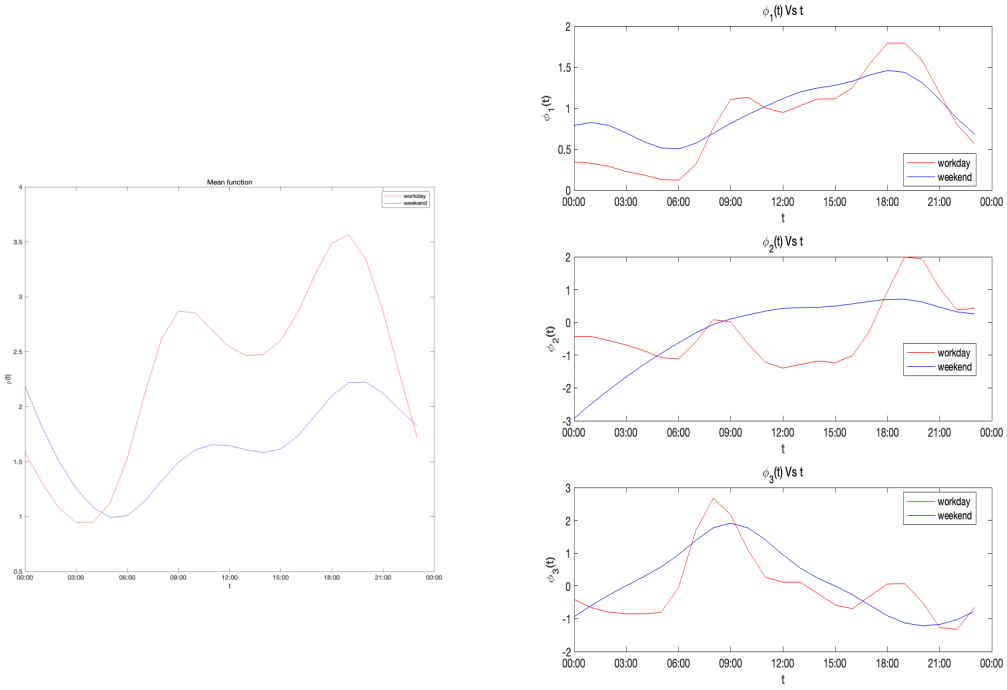

Figure 7. The fitted mean functions (left) and the first three eigenfunctions (right) for two samples

statistics still consistently reject the null hypothesis. In general, the above computation results show that the proposed two-sample tests for sparse functional data are reliable in practice.

3. Conclusion

In this paper, three Cramer-von Mises statistics and three Energy statistics are proposed for two-sample distribution test of sparse functional data. The test statistics are constructed based on the non-parametric local linear smoothing method and the conditional principal component analysis, which are powerful tools for studying sparse functional data. We conduct simulation studies and real data analysis. The results indicate that the proposed methods have good performance in controlling both the first and second types of errors.

References

[1] T. W. Anderson, On the distribution of the two-sample Cramér-Von Mises criterion, Ann. Math. Statist., 33(3), 1148–1159 (1962).

[2] B. Aslan and G. Zech, Statistical energy as a tool for binning-free, multivariate goodness-of-fit tests, two-sample comparison and unfolding, Nucl. Instrum. Methods Phys. Res. A. 537(3), 626–636 (2005).

[3] M. Benko, W. Härdle, and A. Kneip, Common functional principal components, Ann. Statist., 37(1), 1–34 (2009).
[4] F. Chen, S. G. Meintanis, and L. Zhu, On some characterizations and multidimensional criteria for testing homogeneity, symmetry and independence, J. Multivariate Anal., 173, 125–144 (2019).
[5] L. Corain, V. B. Melas, A. Pepelyshev, and L. Salmaso, New insights on permutation approach for hypothesis testing on functional data, Adv. Data Anal. Classif., 8(3), 339–356 (2014).
[6] J. Fan and S. Lin, Test of significance when data are curves, J. Amer. Statist. Assoc., 93(443), 1007–1021 (1998).
[7] P. Hall and I. Keilegom, Two-sample tests in functional data analysis starting from discrete data, Statist. Sinica, 17(4), 1511–1531 (2007).
[8] M. C. King, A. M. Staicu, J. M. Davis, B. J. Reich, and B. Eder, A functional data analysis of spatiotemporal trends and variation in fine particulate matter, Atmos. Environ., 184, 233– 243 (2018).
[9] M. L. Krzyśko and L. Smaga, Two-sample tests for functional data using characteristic functions, Austrian J. Stat., 50, 53–64 (2021).
[10] S. Ma, L. Yang, and R. J. Carroll, A simultaneous confidence band for sparse longitudinal regression, Statist. Sinica, 22(1), 95–122 (2012).
[11] H. G. Müller, Functional modeling and classification of longitudinal data. Scand. J. Statist., 32(2), 223–240 (2005).
[12] C. Peltier, M. Visalli, P. Schlich, and H. Cardot, Analyzing temporal dominance of sensations data with categorical functional data analysis, Food Qual. Prefer., 109, 104893 (2023).
[13] A. N. Pettitt, Two-sample Cramér-Von Mises type rank statistics, J. R. Stat. Soc. Ser. B Stat. Methodol., 41(1), 46–53 (1979).
[14] G. M. Pomann, A. M. Staicu, and S. Ghosh, A two-sample distribution-free test for functional data with application to a diffusion tensor imaging study of multiple sclerosis, J. Roy. Statist. Soc. Ser. C, 65, 395–414 (2016).
[15] J. A. Rice and B. W. Silverman, Estimating the mean and covariance structure nonparametrically when the data are curves, J. R. Stat. Soc. Ser. B Stat. Methodol., 53(1), 233–243 (1991).
[16] A. M. Staicu, Y. Li, C. M. Crainiceanu, and D. Ruppert, Likelihood ratio tests for dependent data with applications to longitudinal and functional data analysis, Scand. J. Statist., 41(4), 932–949 (2014).
[17] G. J. Székely and M. L. Rizzo, Energy statistics: A class of statistics based on distances, J. Statist. Plann. Inference, 143(8), 1249–1272 (2013).
[18] S. Tian, S. Li, and Q. Gu, Measurement and contagion modelling of systemic risk in China's financial sectors: Evidence for functional data analysis and complex network. Int. Rev. Financ. Anal., 90, 102913 (2023).
[19] Q. Wang, Two-sample inference for sparse functional data, Electron. J. Stat., 15(1), 1395–1423 (2021).
[20] G. S. Watson, Goodness-of-fit tests on a circle, Biometrika, 48(1/2), 109–114 (1961).
[21] F. Yao, H. G. Müller, and J. L. Wang, Functional data analysis for sparse longitudinal data, J. Amer. Statist. Assoc., 100(470), 577–590 (2005).
[22] D. Yu, L. Kong, and I. Mizera, Partial functional linear quantile regression for neuroimaging data analysis, Neurocomputing, 195, 74–87 (2016).