Received: 23 August 2023 | Revised: 26 October 2023 | Accepted: 14 November 2023 | Published online: 17 November 2023

RESEARCH ARTICLE

Journal of Data Science and Intelligent Systems
2024, Vol. 2(3) 181–190
DOI: 10.47852/bonviewJDSIS32021578

BON VIEW PUBLISHING

Bootstrap Methods for Canonical Correlation Analysis of Functional Data

Haoyu Yu[1] and Lihong Wang[1,*]

[1]*Department of Mathematics, Nanjing University, China*

Abstract: The bootstrap method is a very general resampling procedure for investigating the distributional property of statistics. In this paper, we present two bootstrap methods with the aim of studying the functional canonical components for functional data. The bootstrap I method constructs the bootstrap replications by resampling from the raw data, while the bootstrap II algorithm samples with replacement from the principal component scores. Simulation studies are conducted to examine the performance of the proposed bootstrap methods. The method is also applied to the motion analysis dataset, which consists of the angles formed by the hip and knee of each of 39 children over each child's gait cycle. Numerical simulations and real data analysis show the good performance of both bootstrap methods for functional canonical correlation analysis. Moreover, as measured by the mean error and mean squared error, the bootstrap II algorithm performs better in approximating sample canonical components than the bootstrap I method.

Keywords: bootstrap method, canonical correlation analysis, functional data, functional principal component, resampling

1. Introduction

The development of data storage and computing technology has facilitated the diversity of data type. The diversified data types, such as text, graphics, and images, are gradually changing the traditional statistical analysis and data mining methods. Functional data, as a special kind of infinite dimensional data, have been attracting a growing attention, see, for example, Cuevas [1], Ferraty and Romain [2], Goia and Vieu [3], Horváth and Kokoszka [4], Hsing and Eubank [5], and Ramsay and Silverman [6].

In recent years, statistical methods applied to finite dimensional data have been widely extended to functional data. These algorithms together with Hilbert space analysis support the theory and practice of functional data analysis (FDA). Classical canonical correlation analysis (CCA) is one of the important algorithms in statistical inference to measure the strength of the overall relationships between two random vectors. It has been widely used in image recognition, machine learning, pattern recognition, and so on. Therefore, as indicated in Horváth and Kokoszka [4], it is necessary to extend CCA to its functional context.

In the FDA, a main task is to make inference about the probability distribution of statistics, such as the estimators of the functional canonical correlations, from a set of realized samples. However, in practice, one may only be able to obtain relatively small samples, especially in FDA. This makes the limiting distributional properties of the statistics inapplicable. When such a problem arises, resampling methodology turns out to be the only practical alternative [7, 8]. The bootstrap technique introduced by Efron [9] and Efron and Tibshirani [10] is a useful resampling tool for investigating the distributional property of statistics with small samples. In contrast to multivariate data analysis, there is comparably less work that has been done on bootstrapping functional data.

Charkaborty and Panaretos [11] constructed a bootstrap implementation of a test procedure for the rank of covariance operator of functional data. Chen and Pun [12] examined bootstrap methods to construct a generalized KPSS test for functional time series. Chowdhury and Chaudhuri [13] developed a bootstrap implementation for functional analysis of variance. Cuevas et al. [7] presented the bootstrap confidence bands of functional parameters with several resampling methods for functional data. Kim and Lim [14] proposed a classification method based on bootstrap aggregating for sparse functional data. Paparoditis and Sapatinas [15] considered bootstrap-based testing of equality of mean functions for functional data. Paparoditis and Shang [16] proposed a bootstrap procedure for constructing prediction bands of stationary functional time series. Poskitt and Sengarapillai [17] and Shang [18] proposed bootstrap procedures by randomly sampling with replacement from the functional principal component (FPC) scores.

For the multivariate CCA, Fan and Wang [19] and Lee [20] developed the bootstrap methods for the estimation of canonical correlations. But the existing bootstrap methods used in traditional multivariate CCA cannot be directly applied to the functional (infinitely dimensional) CCA. For the multivariate case, the standard bootstrap resampling does not involve functional principal component analysis (FPCA), while for functional data one needs to calculate the FPC scores. Therefore, new bootstrap methods based on FPCA are necessary. However, to our knowledge, no study has explored the bootstrapping for functional CCA. In this paper, with the aim of investigating the sample functional canonical components (FCCs) without increasing the sample size, we present two bootstrap methods for mimicking the behavior of sample FCC between two random functions. It is expected that the bootstrap methods will receive increasing popularity in functional CCA, where the object of interest is on the distribution of functional estimators.

*Corresponding author: Lihong Wang, Department of Mathematics, Nanjing University, China. Email: lhwang@nju.edu.cn

The main contribution of this paper is to propose two techniques for bootstrapping the estimators of the canonical components for functional data. Based on the bootstrapped estimators, one can make inference about the distribution of the sample FCC, the confidence intervals, and various hypothesis tests of the FCCs. The bootstrap procedures are addressed in Section 2. Section 3 provides a series of numerical simulations and real data analysis to evaluate the accuracy of the proposed bootstrap methods.

2. The Bootstrap Procedure

We consider square integrable random functions $X \in H_1$ and $Y \in H_2$, where $H_1 = L^2(T_1)$ and $H_2 = L^2(T_2)$ are two L^2 Hilbert spaces with the inner product $\langle x, y\rangle = \int x(t)y(t)dt$ which generates the norm $\|\cdot\|$, T_1 and T_2 are, possibly different, subsets of a Euclidean space. Using the notations in Chapter 4.2 of Horváth and Kokoszka [4], we define the FCC as follows:

Let $\mu_X(t) = E[X(t)]$, $\mu_Y(t) = E[Y(t)]$,

$$c_{11}(t,s) = E[(X(t) - \mu_X(t))(X(s) - \mu_X(s))],$$

$$c_{12}(t,s) = E[(X(t) - \mu_X(t))(Y(s) - \mu_Y(s))],$$

$$c_{21}(t,s) = E[(Y(t) - \mu_Y(t))(X(s) - \mu_X(s))],$$

$$c_{22}(t,s) = E[(Y(t) - \mu_Y(t))(Y(s) - \mu_Y(s))],$$

and

$$C_{11}(x)(t) = \int_{T_1} c_{11}(t,s)x(s)ds = E[\langle X - \mu_X, x\rangle(X(t) - \mu_X(t))],$$

$$C_{12}(y)(t) = \int_{T_2} c_{12}(t,s)y(s)d = E[\langle Y - \mu_Y, y\rangle(X(t) - \mu_X(t))],$$

$$C_{21}(x)(t) = \int_{T_1} c_{21}(t,s)x(s)ds = E[\langle X - \mu_X, x\rangle(Y(t) - \mu_Y(t))],$$

$$C_{22}(y)(t) = \int_{T_2} c_{22}(t,s)y(s)ds = E[\langle Y - \mu_Y, y\rangle(Y(t) - \mu_Y(t))].$$

Then the kth canonical components are defined as $(\rho_k, a_k, b_k, A_k, B_k)$ with $A_k = \langle a_k, X\rangle$ and $B_k = \langle b_k, Y\rangle$, where ρ_k is the kth canonical correlation, a_k and b_k are the associated weight functions, if there exist, by

$$\begin{aligned}\rho_k &= \langle a_k, C_{12}(b_k)\rangle \\ &= \sup\{\langle a, C_{12}(b)\rangle : a \in H_1, b \in H_2, \langle a, C_{11}(a)\rangle = 1, \langle b, C_{22}(b)\rangle = 1\}\end{aligned}$$

subject to the conditions, for $k > 1$,

$$\begin{aligned}\langle a_k, C_{11}(a_j)\rangle = \langle a_k, C_{12}(b_j)\rangle = \langle b_k, C_{22}(b_j)\rangle = \langle b_k, C_{21}(a_j)\rangle \\ = 0, \quad j < k.\end{aligned}$$

To estimate the theoretical FCC, one needs to derive the sample FCC by using a sample of pairs of functions $(X_1, Y_1), \cdots, (X_N, Y_N)$. The sample covariance operator of the random function X is denoted by

$$\hat{C}_{11}(x) = \tfrac{1}{N}\sum_{n=1}^{N}\langle X_n - \hat{\mu}_X, x\rangle(X_n - \hat{\mu}_X), \quad x \in H_1,$$

where $\hat{\mu}_X(t) = N^{-1}\sum_{n=1}^{N} X_n(t)$.

$\hat{\lambda}_i$ and $\hat{v}_i$ denote the eigenvalues and eigenfunctions of the sample covariance operator $\hat{C}_{11}$, and analogously $\hat{\gamma}_j$ and $\hat{u}_j$ define the sample covariance operator $\hat{C}_{22}$ of the function Y. The numbers p and q are determined such that $\sum_{i=1}^{p}\hat{\lambda}_i$ and $\sum_{j=1}^{q}\hat{\gamma}_j$ explain the required proportion of the total variance. Then, the FPC scores are computed

$$\hat{\xi}_{in} = \langle X_n - \hat{\mu}_X, \hat{v}_i\rangle, i = 1, \cdots, p, \quad \hat{\zeta}_{jn} = \langle Y_n - \hat{\mu}_Y, \hat{u}_j\rangle, j = 1, \ldots, q.$$

Let $\hat{\xi}_n = (\hat{\xi}_{1n}, \cdots, \hat{\xi}_{pn})^T$, $\hat{\zeta}_n = (\hat{\zeta}_{1n}, \cdots, \hat{\zeta}_{qn})^T$. Based on the pairs $(\hat{\xi}_1, \hat{\zeta}_1), \ldots, (\hat{\xi}_N, \hat{\zeta}_N)$, the original functional CCA can be reduced to the multivariate sample CCA. He et al. [21] state that the usual properties of canonical correlations and canonical weights known from multivariate analysis can extend to the functional canonical analysis.

Setting $m = \min(p, q)$, we obtain the multivariate sample canonical components $(\hat{\rho}_k, \hat{\mathbf{a}}_k, \hat{\mathbf{b}}_k)$, $k = 1, \cdots, m$, where

$$\begin{aligned}\hat{\rho}_k &= \hat{\mathbf{a}}_k^T\hat{\mathbf{C}}_{12}\hat{\mathbf{b}}_k \\ &= \max\{\mathbf{a}^T\hat{\mathbf{C}}_{12}\mathbf{b} : \mathbf{a} \in R^p, \mathbf{b} \in R^q, \mathbf{a}^T\hat{\mathbf{C}}_{11}\mathbf{a} = 1, \mathbf{b}^T\hat{\mathbf{C}}_{22}\mathbf{b} = 1\}\end{aligned}$$

subject to the conditions

$$\hat{\mathbf{a}}_k^T\hat{\mathbf{C}}_{11}\hat{\mathbf{a}}_j = \hat{\mathbf{a}}_k^T\hat{\mathbf{C}}_{12}\hat{\mathbf{b}}_j = \hat{\mathbf{b}}_k^T\hat{\mathbf{C}}_{22}\hat{\mathbf{b}}_j = \hat{\mathbf{b}}_k^T\hat{\mathbf{C}}_{21}\hat{\mathbf{a}}_j = 0, \quad j < k \quad for \quad k > 1,$$

where

$$\hat{\mathbf{C}}_{11} = \frac{1}{N-1}\sum_{n=1}^{N}\hat{\xi}_n\hat{\xi}_n^T, \quad \hat{\mathbf{C}}_{12} = \frac{1}{N-1}\sum_{n=1}^{N}\hat{\xi}_n\hat{\zeta}_n^T,$$

$$\hat{\mathbf{C}}_{21} = \frac{1}{N-1}\sum_{n=1}^{N}\hat{\zeta}_n\hat{\xi}_n^T, \quad \hat{\mathbf{C}}_{22} = \frac{1}{N-1}\sum_{n=1}^{N}\hat{\zeta}_n\hat{\zeta}_n^T.$$

Finally, the estimators of FCC are defined as $\left(\hat{\rho}_k, \hat{a}_k, \hat{b}_k, \hat{\mathbf{A}}_k, \hat{\mathbf{B}}_k\right)$, where

$$\hat{a}_k = \hat{\mathbf{a}}_k^T\left[\hat{v}_1, \ldots, \hat{v}_p\right]^T, \quad \hat{b}_k = \hat{\mathbf{b}}_k^T\left[\hat{u}_1, \ldots, \hat{u}_p\right]^T,$$

$$\hat{\mathbf{A}}_k = \left[\langle\hat{a}_k, X_1\rangle, \cdots, \langle\hat{a}_k, X_N\rangle\right]^T, \quad \hat{\mathbf{B}}_k = \left[\langle\hat{b}_k, Y_1\rangle, \cdots, \langle\hat{b}_k, Y_N\rangle\right]^T.$$

In order to study the distributional property of the sample FCC, we introduce two bootstrap algorithms. The first one (bootstrap I) is a direct method, which is similar to the bootstrap method of the traditional CCA, where we resample from the raw data and construct the bootstrap replications. This algorithm intuitively reflects the idea of bootstrap resampling, but one needs to calculate the FPC for each replication, which increases computation load. Since $\hat{\mathbf{A}}_k = [\langle\hat{a}_k, X_1\rangle, \cdots, \langle\hat{a}_k, X_N\rangle]^T$ and $\hat{\mathbf{B}}_k = [\langle\hat{b}_k, Y_1\rangle, \cdots, \langle\hat{b}_k, Y_N\rangle]^T$, we only generate the bootstrap realizations for $\left(\hat{\rho}_k, \hat{a}_k, \hat{b}_k\right)$.

The bootstrap I algorithm is as follows:

(1) Generate bootstrap replication $\aleph^* = \{(X_1^*, Y_1^*), \cdots, (X_N^*, Y_N^*)\}$ by taking i.i.d random draw from the observations $\{(X_1, Y_1), \cdots, (X_N, Y_N)\}$;
(2) Calculate the eigenfunctions $\hat{v}_i^*$ and $\hat{u}_j^*$ of the sample covariance operators of the X and Y components of $\aleph^*$, respectively, and compute the corresponding FPC scores

$$\hat{\xi}_n^* = [\langle X_n^* - \hat{\mu}_X^*, \hat{v}_1^*\rangle, \cdots, \langle X_n^* - \hat{\mu}_X^*, \hat{v}_p^*\rangle]^T,$$

$$\hat{\xi}_n^* = [\langle Y_n^* - \hat{\mu}_Y^*, \hat{\mu}_1^* \rangle, \cdots, \langle Y_n^* - \hat{\mu}_Y^*, \hat{\mu}_q^* \rangle]^T, \quad n = 1, \cdots, N.$$

Let $\Im^* = \{(\hat{\xi}_1^*, \hat{\zeta}_1^*), \cdots, (\hat{\xi}_N^*, \hat{\zeta}_N^*)\}$;

(3) Construct the sample canonical components $\left(\hat{\rho}_k^*, \hat{a}_k^*, \hat{b}_k^*\right)$ of pairs $\Im^*$.

Now we introduce the bootstrap II algorithm. The basic idea is to bootstrap a set of sample FCC by randomly sampling principal component scores. This idea was also adopted by Poskitt and Sengarapillai [17] and Shang [18]. The advantage of this method is that the principal component is only calculated once from the original sample, and there is no additional computational cost. The bootstrap II algorithm is as follows.

(1) Calculate the eigenfunctions $\hat{v}_i$ and $\hat{u}_j$ of the sample covariance operators of the X and Y components of the raw data $\{(X_1, Y_1), \cdots, (X_N, Y_N)\}$ respectively, and compute the corresponding FPC scores $\varsigma_n = (\hat{\xi}_n, \hat{\zeta}_n)$, $n = 1, \cdots, N$. Let $\Im = \{\varsigma_1, \cdots, \varsigma_N\}$;
(2) Generate bootstrap replication $\Im^{**}$ by taking i.i.d random draw from $\Im$, where $\Im^{**} = \{\varsigma_1^{**}, \cdots, \varsigma_N^{**}\} = \{(\hat{\xi}_1^{**}, \hat{\zeta}_1^{**}), \cdots, (\hat{\xi}_N^{**}, \hat{\zeta}_N^{**})\}$;
(3) Construct the sample canonical components $\left(\hat{\rho}_k^{**}, \hat{a}_k^{**}, \hat{b}_k^{**}\right)$ of pairs $\Im^{**}$.

3. Numerical Simulations and Empirical Studies

In this section, we evaluate the performance of two bootstrap methodologies via simulation. We also apply the proposed methods for an empirical study of the motion analysis dataset, which consists of the angles formed by the hip and knee of each of 39 children over each child's gait cycle.

3.1. Numerical simulations

In this subsection we concentrate on the bootstrap accuracy for the sample FCC through simulation studies. The performance of two bootstrap methods is evaluated and compared based on the difference between the original sample FCC and bootstrapped sample FCC. We calculate the mean error (ME) and mean squared error (MSE) to measure such a difference, which are given by

$$ME = \frac{1}{MB}\sum_{i=1}^{M}\sum_{b=1}^{B}\left(\hat{\rho}_{k,b,i}^* - \hat{\rho}_{k,i}\right),$$

$$MSE = \frac{1}{MB}\sum_{i=1}^{M}\sum_{b=1}^{B}\left(\hat{\rho}_{k,b,i}^* - \hat{\rho}_{k,i}\right)^2,$$

where M represents the total number of simulation runs, and B represents the total number of bootstrap replications.

First, we generate data from the following model with sample size $N = 50$:

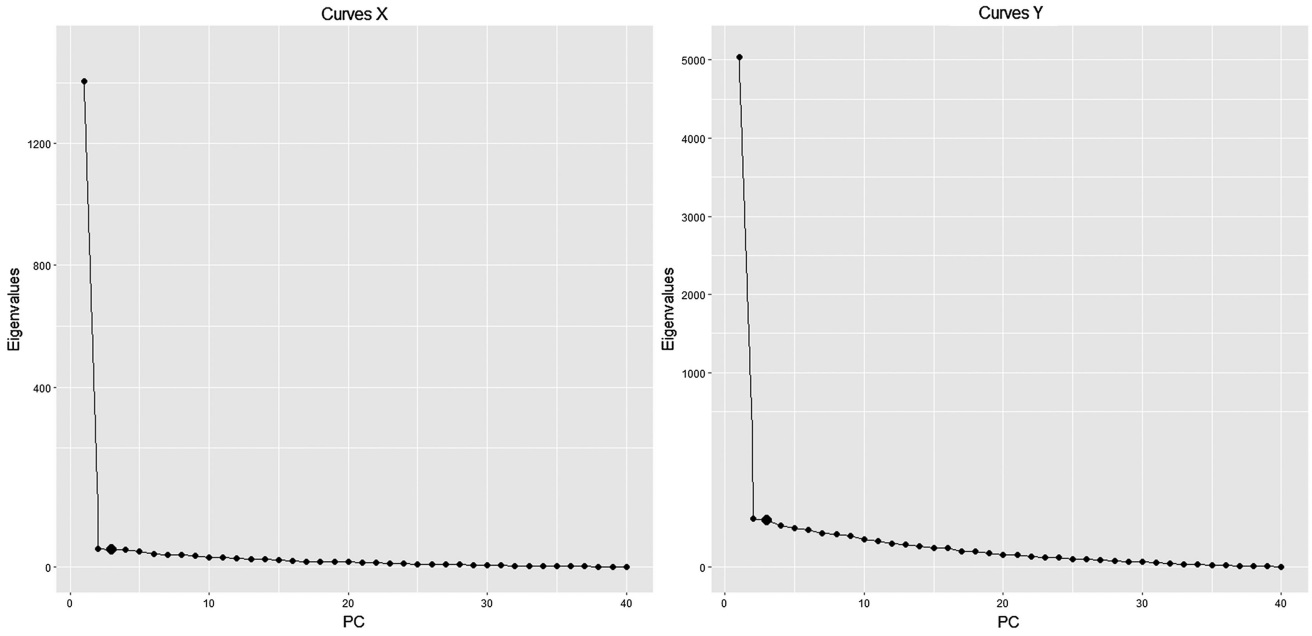
$$X(t) = t + \varepsilon(t), \quad Y(t) = \sqrt{t} + \varepsilon(t) + \eta(t), \quad t \in [0, 1], \qquad (1)$$

where $\varepsilon(t)$ and $\eta(t)$ are mutually independent standard Brownian motions.

The FPCA for $X(t)$ and $Y(t)$ indicates that the first two principal components explain 99.9% of the cumulative percentage of total variance. Therefore, we choose $p = q = 2$. Another way to pick the numbers of the principal components is to use the scree test, which is a graphical method first proposed by Cattell [22]. To apply the scree method, one plots the successive eigenvalues, see Figure 1, and finds the place where the smooth decrease of eigenvalues appears to level off. This provides the number of most important principal components. Figure 1 suggests the same p and q values, $p = q = 2$, and then $m = 2$. That is, we calculate the first two sample FCCs.

With $M = 100$ replications, we first compute $\hat{\rho}_{k,i}$ for $k = 1, 2$, $i = 1, \cdots, M$. Next, setting $B = 5000$, we use the two bootstrap methods to calculate the first two sample functional canonical correlation coefficients $\hat{\rho}_{k,b,i}^*$ and $\hat{\rho}_{k,b,i}^{**}$, $k = 1, 2$, $b = 1, \cdots, B$, $i = 1, \cdots, M$.

Figure 1
The scree plot for model (1)



"ME," "MSE," and the average of $\hat{\rho}_{k,i}$, $\hat{\rho}^{*}_{k,b,i}$, $\hat{\rho}^{**}_{k,b,i}$ over 100 simulation repetitions and 5000 bootstrap repetitions are reported in Table 1. Figures 2 and 3 illustrate the histograms and the estimated densities of $\hat{\rho}_k$, $\hat{\rho}^{*}_k$, $\hat{\rho}^{**}_k$ and of the differences $\hat{\rho}^{*}_k - \hat{\rho}_k$ and $\hat{\rho}^{**}_k - \hat{\rho}_k$. Figure 4 depicts the plots of the first two pairs of canonical weight functions from one simulation run based on the raw sample, bootstrap I algorithm, and bootstrap II algorithm, respectively.

It can be seen that both bootstrap I and II methods perform well in approximating the FCCs. The bootstrapped canonical correlation coefficient $\hat{\rho}^{*}_k$ (or $\hat{\rho}^{**}_k$) has high accuracy in estimating the original sample FCC. The maximum relative error $|\hat{\rho}^{*}_k - \hat{\rho}_k|/|\hat{\rho}_k|$ is only 2.46%, while the bootstrap estimates of the second canonical correlation have a larger error than the estimates of the first canonical

Table 1
ME and MSE between the raw sample FCCs and bootstrap sample FCCs based on 100 sample replications and 5000 bootstrap replications for model (1)

FCC	Raw sample	Bootstrap I	Bootstrap II
$\hat{\rho}_1$	0.729487	0.729254	0.728938
ME	–	−2.33e-4	−5.49e-4
MSE	–	3.4514e-5	2.1822e-5
$\hat{\rho}_2$	0.397995	0.388201	0.399369
ME	–	−9.7936e-3	1.3738e-3
MSE	–	8.5613e-4	1.4807e-4
Time (ms)	9	328	34

Figure 2
The histograms of the raw sample FCCs and bootstrap sample FCCs for model (1)

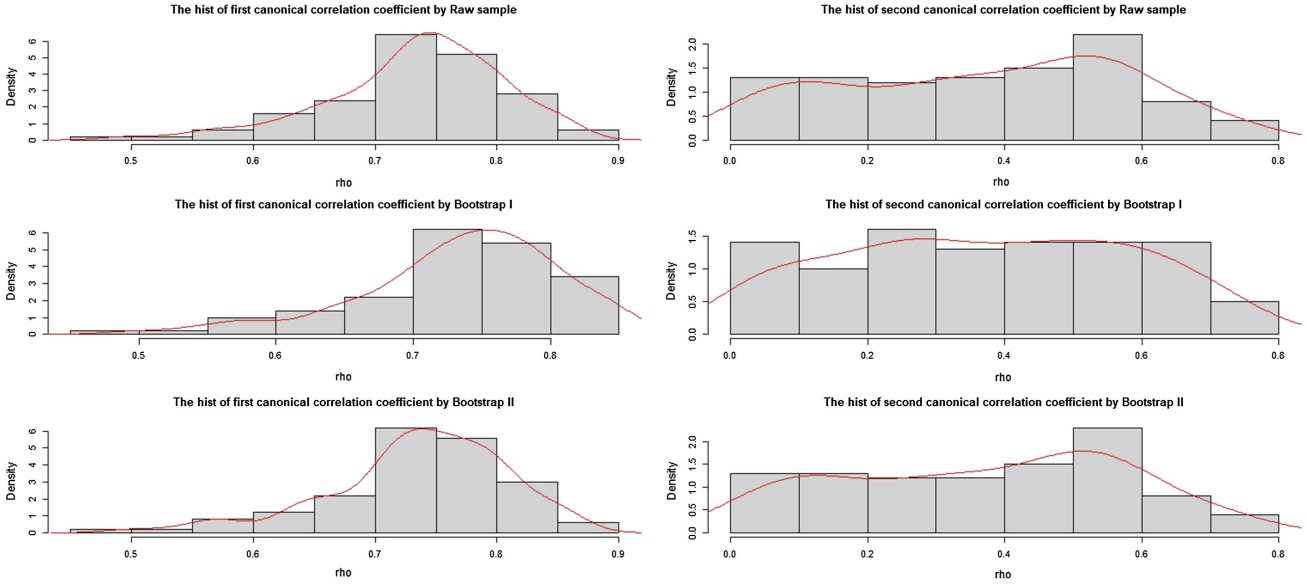
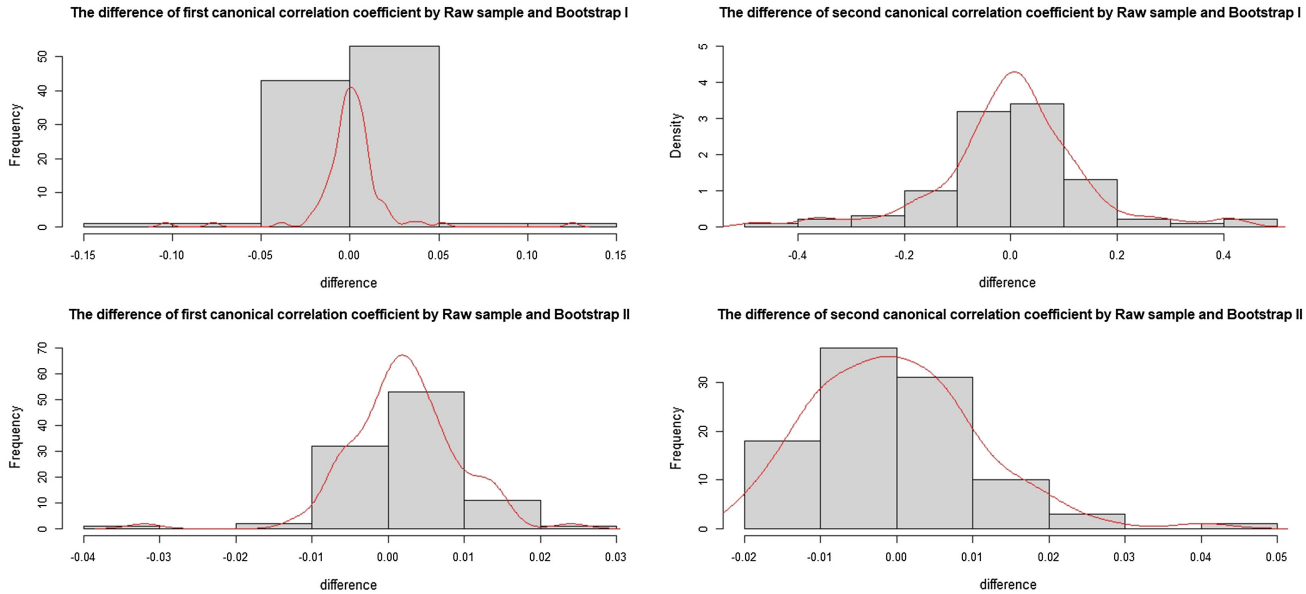


Figure 3
The histograms of the differences between raw sample FCCs and bootstrap sample FCCs for model (1)



correlation. In general, the errors, especially MSEs of bootstrap II method, are smaller than those of bootstrap I method. This indicates that bootstrap II is superior to bootstrap I in estimating the FCC.

From Table 1, we see that the average running time of bootstrap I for each simulation replication is 328 ms, while the average time of bootstrap II is 34 ms. The time complexity of bootstrap II is significantly lower than that of bootstrap I.

Figure 2 implies that the FCC based on bootstrap sampling, especially using bootstrap II procedure, approximately has the same distribution as the original sample FCC. Moreover, the histograms and densities suggest that the sample first canonical correlation coefficient approximately has a normal distribution, and the differences between raw sample FCCs and bootstrap sample FCCs shown in Figure 3 are also approximately normally distributed with zero mean.

In terms of the canonical weight function estimation, both bootstrap methods have good performance. It is observed from Figure 4 that the bootstrapped canonical weight functions are very similar to the original ones, with the same trend and peaks. In comparison, bootstrap II procedure performs better than bootstrap I method. Moreover, the computation cost of bootstrap II is lower than that of bootstrap I.

Next, we consider the model suggested in He et al. [23]. The processes are defined as follows:

$$X(t) = \sum_{i=1}^{21} \varepsilon_i \theta_i(t), \quad Y(t) = \sum_{i=1}^{21} \eta_i \theta_i(t), \quad t \in [0, D], \tag{2}$$

where $D = 50$, $\{\theta_i(t)\}$ is the Fourier basis on $[0, D]$, with $\theta_1(t) = \sqrt{1/D}$, $\theta_2(t) = \sqrt{2/D}\sin((t - D/2)2\pi/D)$, $\theta_3(t) = \sqrt{2/D}\cos((t - D/2)2\pi/D)$, $\cdots$, $\theta_{20}(t) = \sqrt{2/D}\sin((t - D/2)20\pi/D)$, $\theta_{21}(t) = \sqrt{2/D}\cos((t - D/2)20\pi/D)$, and $\varepsilon = \{\varepsilon_i\}$ and $\eta = \{\eta_i\}$ are Gaussian vectors with covariance matrices $Cov(\varepsilon) = R_{11}$, $Cov(\eta) = R_{22}$ and $Cov(\varepsilon, \eta) = R_{12}$, where R_{11}, R_{12}, and R_{22} are diagonal matrices with $diag(R_{11}) = diag(R_{22}) = \{10 \times 0.75^i\}$, $i = 0, 0, 0, 1, 2, \cdots, 18$, $diag(R_{12}) = \{7, 3, 1, 0, 0, \cdots, 0\}$.

Then, direct calculation yields that the canonical correlations for X and Y are $\rho_1 = 0.7$, $\rho_2 = 0.3$, $\rho_3 = 0.1$, $\rho_4 = \rho_5 = \cdots = \rho_{21} = 0$, and the canonical weight functions are $a_k(t) = b_k(t) = \sqrt{0.1}\theta_k(t)$ for $k = 1, 2, 3$, and 0 otherwise.

We generate 50 pairs of X and Y and compute the sample FCCs and the bootstrapped FCCs using the same procedures as the previous example.

Table 2 lists the means, MEs, and MSEs for estimates of the first three canonical correlations obtained by using the raw samples and the bootstrap samples based on 5000 bootstrap replications over 100 simulation runs, which shows that both bootstrap algorithms provide fairly accurate estimates for the first three canonical correlations.

Table 2
ME and MSE between the raw sample FCCs and bootstrap sample FCCs based on 100 sample replications and 5000 bootstrap replications for model (2)

FCC	Raw sample	Bootstrap I	Bootstrap II
$\hat{\rho}_1$	0.691549	0.691697	0.692023
ME	–	−1.48e-4	4.746e-4
MSE	–	4.8709e-5	2.8425e-5
$\hat{\rho}_2$	0.297705	0.297846	0.298742
ME	–	−1.406e-4	1.037e-3
MSE	–	4.0311e-5	3.1642e-5
$\hat{\rho}_3$	0.100058	0.099947	0.099947
ME	–	−1.11e-4	−1.11e-4
MSE	–	3.2068e-5	1.9355e-5
Time (ms)	30	282	69

In Figure 5, again one finds that the sample canonical correlation coefficients are approximately normally distributed with mean equal to the true coefficients, and the bootstrapped FCC approximately has the same distribution as the raw sample FCC. Figure 6 displays the pointwise errors of first three pairs of the estimated canonical weight functions by two bootstrap methods.

Figure 4
The first two pairs of canonical weight functions based on the raw sample, bootstrap I algorithm, and bootstrap II algorithm for model (1)

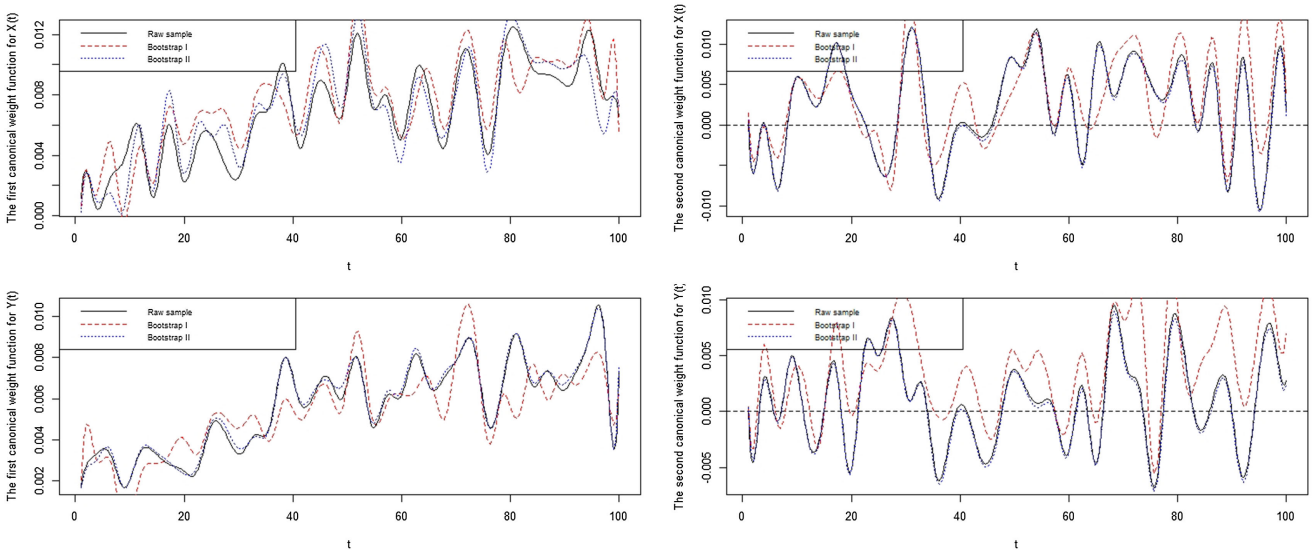


Figure 5
The histograms of the raw sample FCCs and bootstrap sample FCCs for model (2)

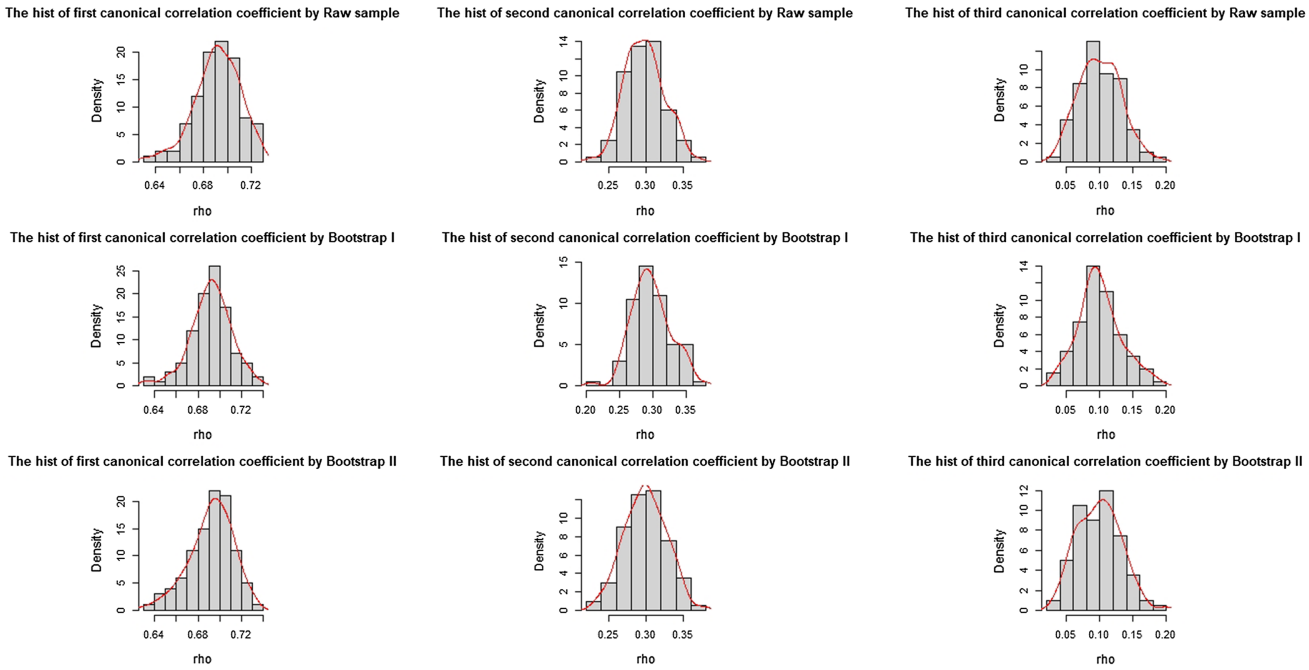
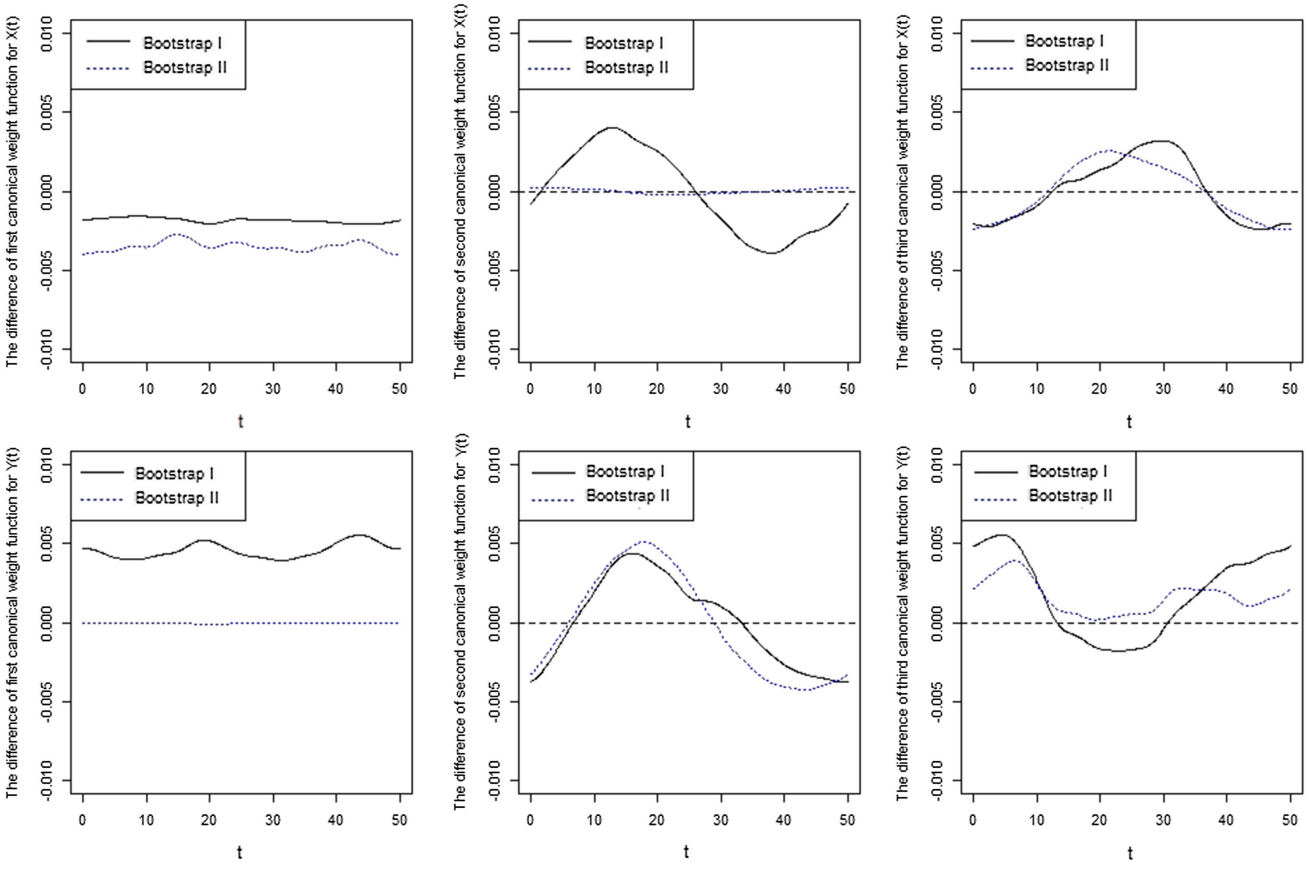


Figure 6
The errors of first three pairs of the estimated canonical weight functions using two bootstrap methods for model (2)



Overall, the errors are very small, while bootstrap II method produces more accurate estimators than bootstrap I in general. As indicated in Section 2, the reason why bootstrap II outperforms bootstrap I is that bootstrap I estimates the FPC scores for each replication, which not only increases computational load but also causes additional error, while bootstrap II algorithm only

Figure 7
The angles formed by the hip and by the knee as 39 children go through a gait cycle

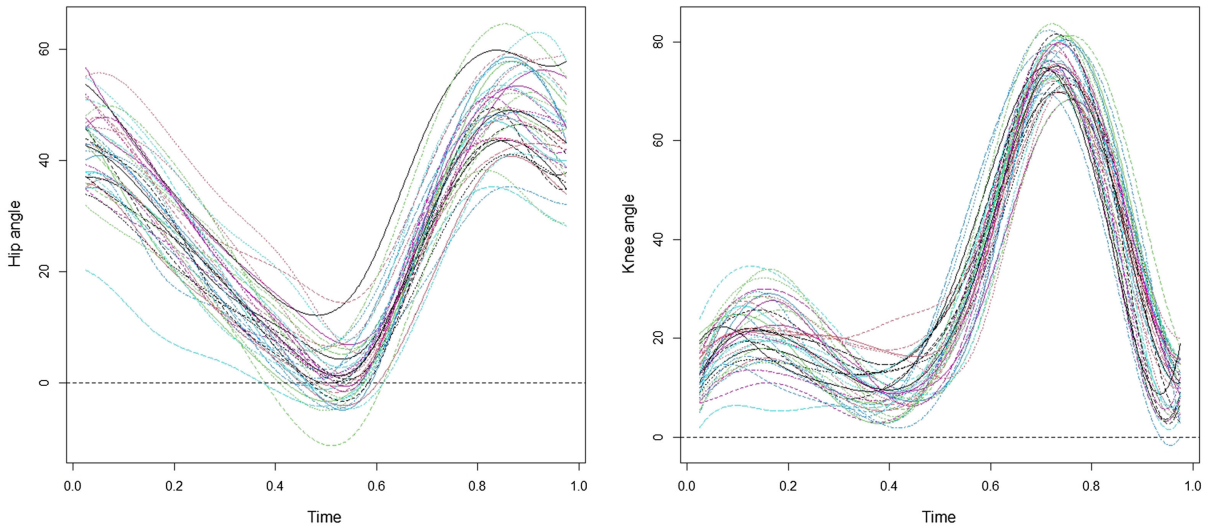
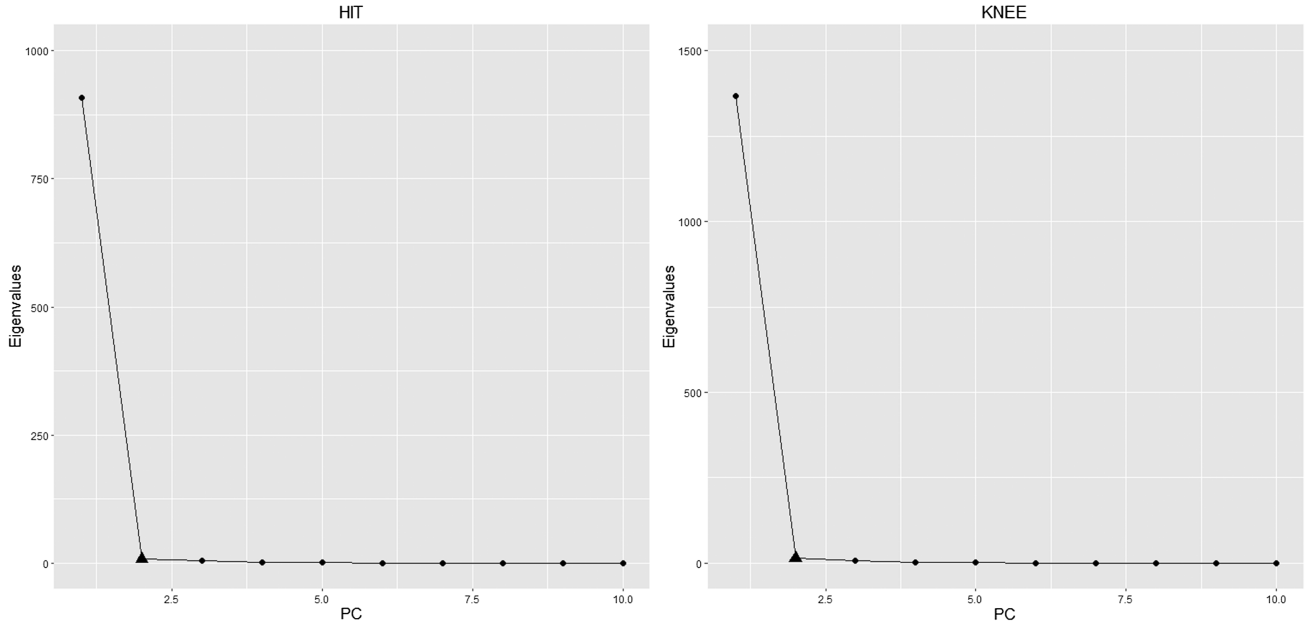


Figure 8
The scree plot for the gait data



calculates the FPC once from the original sample with no additional error and computational cost.

Based on these limited simulation studies, our recommendation is to use bootstrap II, that is, the bootstrap method of randomly sampling with replacement from the principal component scores for the functional CCA, because it performs better in approximating the sample FCCs and the associated weight functions with the additional benefit that it is computationally less expensive.

3.2. Empirical example

In this subsection, the proposed bootstrap methods are applied to the CCA of hip and knee angle while walking. In the process of walking, the angle formed by the hip joint and the ground is called the hip angle, and the angle formed by the knee joint and the ground is called the knee angle. The dataset "gait" contains the hip and knee angles of each of 39 children over each child's gait cycle. Considering the continuity of the movement process, it is obviously reasonable to connect the hip angles and knee angles of each child into curves. Time is measured in terms of the individual gait cycle, so that every curve is given for values of t in $[0, 1]$. See Ramsay and Silverman [6] for full details.

$X_n(t)$ and $Y_n(t)$, $n = 1, \cdots, 39$, denote the curves of hip angle and knee angle on the interval $[0, 1]$, which are shown in Figure 7.

It is found that the first two empirical FPCs of the hip angle and knee angle samples account for more than 99% of the total variation explained. The scree plot in Figure 8 also suggests selecting the first

Table 3
ME and MSE between the raw sample FCCs and bootstrap sample FCCs based on 5000 bootstrap replications for the gait data

FCC	Raw sample	Bootstrap I	Bootstrap II
$\hat{\rho}_1$	0.71148	0.711639	0.712672
ME	–	1.581e-4	1.1916e-3
MSE	–	1.4467e-5	2.3977e-5
$\hat{\rho}_2$	0.161603	0.158594	0.161345
ME	–	−3.0095e-3	−2.588e-4
MSE	–	3.7014e-4	1.8619e-5
Time (ms)	11	94	27

two FPCs. Thus, we set $p = q = 2$ and calculate the first two sample FCCs. The canonical correlations are $\hat{\rho}_1 = 0.71148$ and $\hat{\rho}_2 = 0.161603$.

Next we use two bootstrap methods to measure the canonical correlation between $X(t)$ and $Y(t)$. Table 3 presents the empirical mean of the bootstrapped canonical correlations with $B = 5000$ bootstrap replications. "ME" and "MSE", representing the difference between the original canonical correlations and the bootstrapped ones, are also shown in Table 3. We see that the FCCs based on both bootstrap methods are very close to the original FCCs. The maximum relative error of canonical correlations for bootstrap I is 1.86%, and that of bootstrap II is only 0.26%. Generally speaking, the performance of bootstrap II is better than that of bootstrap I.

Figure 9
The canonical weight vectors for the gait data

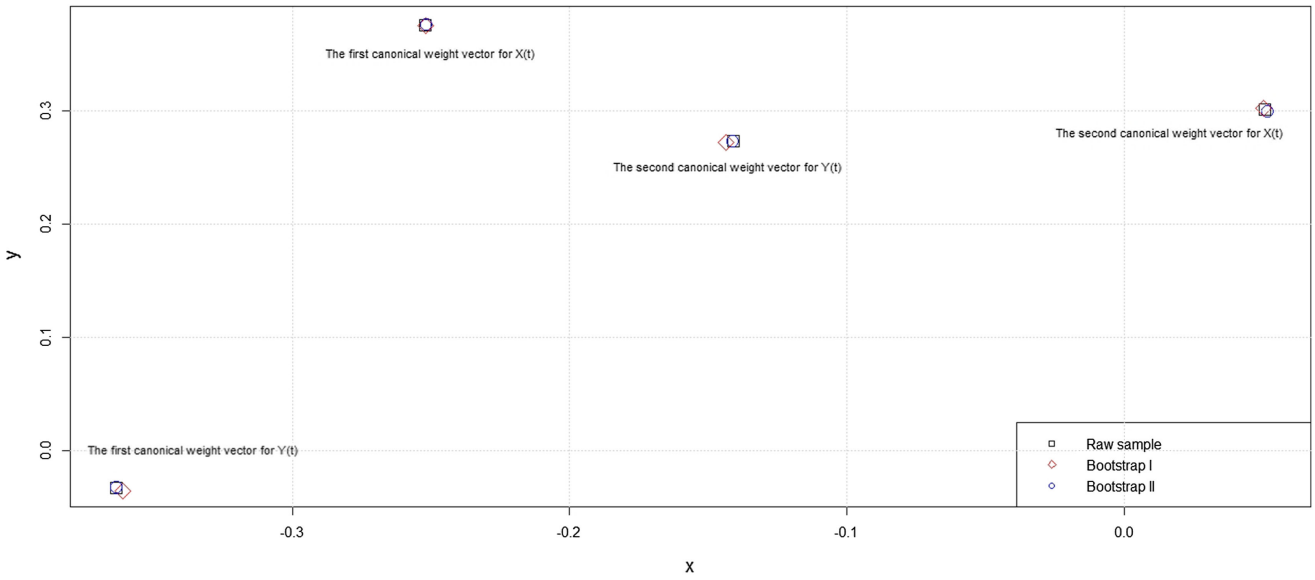
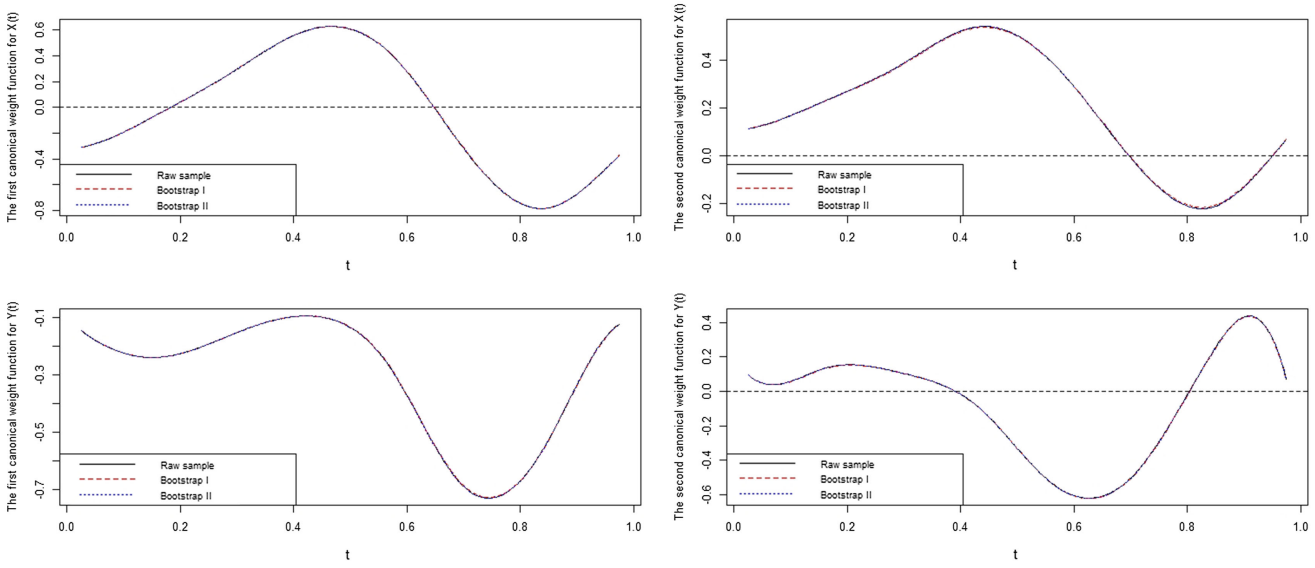


Figure 10
The canonical weight functions for the gait data



However, for the first canonical correlation, bootstrap I method has more accuracy than bootstrap II, which may be due to the possible noise in the gait data. As discussed in Horváth and Kokoszka [4], FPCA not only allows us to reduce the dimension of infinitely dimensional functional data to a small finite dimension in an optimal way but also is capable of eliminating possible noise in the unknown data generating process. Since bootstrap I repeats the FPCA in each bootstrapping resample, bootstrap I may have more accuracy than bootstrap II in practice where there is possible noise in data, but at a much higher computational cost based on the running times shown in Table 3.

Figure 9 displays the associated first two pairs of canonical weight vectors $\hat{\mathbf{a}}_k$ and $\hat{\mathbf{b}}_k$, $k = 1, 2$, on the two-dimensional plane. The weight vectors are divided into four clusters on the two-dimensional plane, representing the two weight vectors of the hip angle and knee angle, respectively. As shown in Figure 9, the weight vectors obtained by the two bootstrap methods are very close to those calculated by the raw sample, especially by the bootstrap II method.

Furthermore, Figure 10 shows the canonical weight functions obtained by the raw data and by using the proposed two bootstrap methods. Again we see that the weight functions calculated by the bootstrap methods are almost the same as those calculated by the raw sample, with similar trend and peak valley. The simulation evidence reveals that both bootstrap methods perform well in approximating the behavior of the sample canonical correlations and the associated weight functions.

4. Conclusions

The bootstrap schemes for the CCA of functional data are considered in this paper. Two bootstrap methods are proposed to estimate the FCCs. These procedures are then applied to the simulated data and a dataset in empirical example. As measured by the ME and MSE, the bootstrap II method that samples with replacement from the estimated principal component scores performs better in approximating sample canonical components than the bootstrap I method of resampling from the raw data. For bootstrap I algorithm, which is similar to the bootstrap method of the traditional CCA, one needs to calculate the FPC for each replication, which not only increases computation load but also causes additional error due to the estimation of the FPC scores in each replication. But, for bootstrap II algorithm, the main advantage is that the principal component is only calculated once from the original sample, so there is no additional error and computational cost. Therefore, in general the bootstrap method that samples with replacement from the estimated principal component scores is better to approximate the sample canonical components.

However, bootstrap II has its own limitations. Since bootstrap II resamples from the FPC scores of the raw data, the FPCs should be fully representative of the original sample. That is, the choice of the number of principal components is important. We suggest to choose the number p for which the cumulative percentage of total variance (CPV) explained by the first p components exceeds at least 95%. Other ways, such as the scree plot, should also be used in conjunction with the CPV method.

Overall, from the simulation results, we see that the distribution of the bootstrapped estimator is approximately the same as that of the original estimator. This ensures that the proposed bootstrap methods can be applied to investigate distributional property of sample FCCs in functional data. In practical applications, the proposed bootstrap methods can be used to make inference about the sample functional canonical correlations, such as the estimation of the distribution of the sample FCC, the construction of confidence intervals, and the implementation of hypothesis tests of the FCCs. However, the theoretical investigation on the asymptotic equivalence of the distributions of the raw sample and bootstrap sample canonical correlations is not yet available. This topic will be pursued in the future research.

Funding Support

This work was supported by National Natural Science Foundation of China (grant number 11671194).

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

Data available on request from the corresponding author upon reasonable request.

References

[1] Cuevas, A. (2014). A partial overview of the theory of statistics with functional data. *Journal of Statistical Planning and Inference*, *147*, 1–23. https://doi.org/10.1016/j.jspi.2013.04.002

[2] Ferraty, F., & Romain, Y. (2011). *The Oxford handbook of functional data analysis*. UK: Oxford University Press.

[3] Goia, A., & Vieu, P. (2016). An introduction to recent advances in high/infinite dimensional statistics. *Journal of Multivariate Analysis*, *146*, 1–6. https://doi.org/10.1016/j.jmva.2015.12.001

[4] Horváth, L., & Kokoszka, P. (2012). *Inference for functional data with applications*. USA: Springer.

[5] Hsing, T., & Eubank, R. (2015). *Theoretical foundations of functional data analysis, with an introduction to linear operators*. USA: Wiley.

[6] Ramsay, J. O., & Silverman, B. W. (2005). *Functional data analysis*. USA: Springer.

[7] Cuevas, A., Febrero, M., & Fraiman, R. (2006). On the use of the bootstrap for estimating functions with functional data. *Computational Statistics & Data Analysis*, *51*(2), 1063–1074. https://doi.org/10.1016/j.csda.2005.10.012

[8] McMurry, T., & Politis, D. N. (2011). Resampling methods for functional data. In F. Ferraty & Y. Romain (Eds.), *The Oxford handbook of functional data analysis*. Oxford University Press.

[9] Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, *7*(1), 1–26.

[10] Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. UK: Taylor & Francis.

[11] Charkaborty, A., & Panaretos, V. M. (2022). Testing for the rank of a covariance operator. *Annals of Statistics*, *50*(6), 3510–3537. https://doi.org/10.1214/22-AOS2238

[12] Chen, Y., & Pun, C. S. (2019). A bootstrap-based KPSS test for functional time series. *Journal of Multivariate Analysis*, *174*, 104535. https://doi.org/10.1016/j.jmva.2019.104535

[13] Chowdhury, J., & Chaudhuri, P. (2022). Multi-sample comparison using spatial signs for infinite dimensional data. *Electronic Journal of Statistics*, *16*(2), 4636–4678. https://doi.org/10.1214/22-EJS2054

[14] Kim, H., & Lim, Y. (2022). Bootstrap aggregated classification for sparse functional data. *Journal of Applied Statistics*, *49*(8), 2052–2063. https://doi.org/10.1080/02664763.2021.1889997

[15] Paparoditis, E., & Sapatinas, T. (2016). Bootstrap-based testing of equality of mean functions or equality of covariance operators for functional data. *Biometrika*, *103*(3), 727–733. https://doi.org/10.1093/biomet/asw033

[16] Paparoditis, E., & Shang, H. L. (2023). Bootstrap prediction bands for functional time series. *Journal of the American Statistical Association*, *118*(542), 972–986. https://doi.org/10.1080/01621459.2021.1963262

[17] Poskitt, D. S., & Sengarapillai, A. (2013). Description length and dimensionality reduction in functional data analysis. *Computational Statistics & Data Analysis*, *58*, 98–113. https://doi.org/10.1016/j.csda.2011.03.018

[18] Shang, H. L. (2015). Resampling techniques for estimating the distribution of descriptive statistics of functional data. *Communications in Statistics – Simulation and Computation*, *44*(3), 614–635. https://doi.org/10.1080/03610918.2013.788703

[19] Fan, Z., & Wang, L. (1996). Comparability of jackknife and bootstrap results: An investigation for a case of canonical correlation analysis. *The Journal of Experimental Education*, *64*(2), 173–189. https://doi.org/10.1080/00220973.1996.9943802

[20] Lee, H. (2007). Canonical correlation analysis using small number of samples. *Communications in Statistics – Simulation and Computation*, *36*(5), 973–985. https://doi.org/10.1080/03610910701539443

[21] He, G., Müller, H., & Wang, J. (2003). Functional canonical analysis for square integrable stochastic processes. *Journal of Multivariate Analysis*, *85*(1), 54–77. https://doi.org/10.1016/S0047-259X(02)00056-8

[22] Cattell, R. B. (1966). The Scree test for the number of factors. *Multivariate Behavioral Research*, *1*(2), 245–276. https://doi.org/10.1207/s15327906mbr0102_10

[23] He, G., Müller, H., & Wang, J. (2004). Methods of canonical analysis for functional data. *Journal of Statistical Planning and Inference*, *122*(1–2), 141–159. https://doi.org/10.1016/j.jspi.2003.06.003

How to Cite: Yu, H., & Wang, L. (2024). Bootstrap Methods for Canonical Correlation Analysis of Functional Data. *Journal of Data Science and Intelligent Systems, 2*(3), 181–190. https://doi.org/10.47852/bonviewJDSIS32021578