Contrastive Functional Principal Components Analysis

Eric Zhang, Didong Li

Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA.
eyzhang@unc.edu, didongli@unc.edu

Abstract

As functional data assumes a central role in contemporary data analysis, the search for meaningful dimension reduction becomes critical due to its inherent infinite-dimensional structure. Traditional methods, such as Functional Principal Component Analysis (FPCA), adeptly explore the overarching structures within the functional data. However, these methods may not sufficiently identify low-dimensional representations that are specific or enriched in a foreground dataset (case or treatment group) relative to a background dataset (control group). This limitation becomes critical in scenarios where the foreground dataset, such as a specific treatment group in biomedical applications, contains unique patterns or trends that are not as pronounced in the background dataset. Addressing this gap, we propose Contrastive Functional Principal Component Analysis (CFPCA), a method designed to spotlight low-dimensional structures unique to or enriched in the foreground dataset relative to the background counterpart. We supplement our method with theoretical guarantees on CFPCA estimates supported by multiple simulations. Through a series of applications, CFPCA successfully identifies these foreground-specific structures, thereby revealing distinct patterns and trends that traditional FPCA overlooks.

Code — https://github.com/ezhang1218/CFPCA

Introduction

Recent advancements in technology have enabled continuous or intermittent observations of data over time intervals, both examples of functional data (Kokoszka and Reimherr 2017). An essential characteristic distinguishing functional data from its multivariate counterpart is smoothness: observed values exhibit a smooth variation that can, theoretically, be observed indefinitely (Ramsay 1997), and exists at any given moment (Hörmann and Kokoszka 2012). While functional data provides a wealth of valuable information, it also poses a number of challenges in theory, method, and computation due to its intrinsically infinite-dimensional nature. Thus, dimension reduction is a natural strategy for better visualization and interpretation of the data (Wang, Chiou, and Müller 2016).

Principal component analysis (PCA) is a well-established dimension reduction method for finite high-dimensional data (Joliffe and Morgan 1992), widely used for data exploration, visualization, and downstream analysis (Hotelling 1933). Functional principal component analysis (FPCA) extends PCA to functional data, uncovering a set of basis functions – termed functional principal components (FPCs)–that best approximate the data curves through linear combinations (Ramsay 1997). FPCA converts infinite dimensional data into a finite-dimensional vector of scores, where the first two (or carefully chosen two) PCs can be visualized in two dimensions (Wang, Chiou, and Müller 2016). In addition to dimension reduction, FPCA also serves as a useful tool for data modeling, clustering, and density estimation (Shang 2014; Hall 2010; Wang, Chiou, and Müller 2016), providing consistent estimates of basis functions and the FPC scores (Hall and Hosseini-Nasab 2006; Yao, Müller, and Wang 2005). FPCA has become a popular tool in functional data analysis, finding applications across various fields including biomechanics, medicine, geophysics, and finance (Jolliffe and Cadima 2016).

In this paper, we consider a specific type of functional data consisting of two groups, referred to as the foreground, case, or treatment group, and another known as the background group or control group. The goal is to identify low dimensional structures that are unique to or enriched in the foreground group. Such scenarios are common in various scientific fields. For example, in Electrocardiogram (ECG) data, the foreground group could be patients with heart abnormalities such as myocardial infarction, and the background could be healthy patients with normal heartbeats. In this situation, we are interested in finding unique structures and patterns in patients with heart abnormalities. In a financial setting, we could be interested in trends specific to the stock price of technology-based companies such as Meta or Google relative to non-technology based companies such as JP Morgan or Exxon Mobil.

Often, the foreground and background groups exhibit substantial similarities in behavior and trends, making it difficult to identify patterns that are unique to one group. Previous research has compared and estimated mean functions (Fan and Lin 1998; Cai and Yuan 2011), proposed two-sample tests and ANOVA tests to determine mean differences (Hall and Van Keilegom 2007), and examined differences in distributions and covariance functions of functional data (Panaretos, Kraus, and Maddocks 2010; Boente,

Rodriguez, and Sued 2011). While these methods are able to detect these differences between functions, they do not sufficiently characterize unique characteristics of one group over the other.

Existing contrastive models for finite dimensional data, such as contrastive latent variable model (CLVM, Severson, Ghosh, and Ng (2019)), contrastive principal components analysis (cPCA, Abid et al. (2017)), and probablistic component analysis (PCPCA, Li, Jones, and Engelhardt (2020)), successfully identify low dimensional subspaces that captures information specific to the foreground group, eliminating common and less interesting information shared by two groups. However, to our knowledge, no method has yet adapted this approach to functional data.

In this paper, we introduce Contrastive Functional Principal Component Analysis (CFPCA), a novel method designed to detect trends specific to a foreground group in comparison to a background group in functional data. This approach integrates key concepts from CLVM and cPCA and, while initially applied in a time series context, can be generalized to other functional data types. From this point, we use these two terms, functional data, and time series, interchangeably.

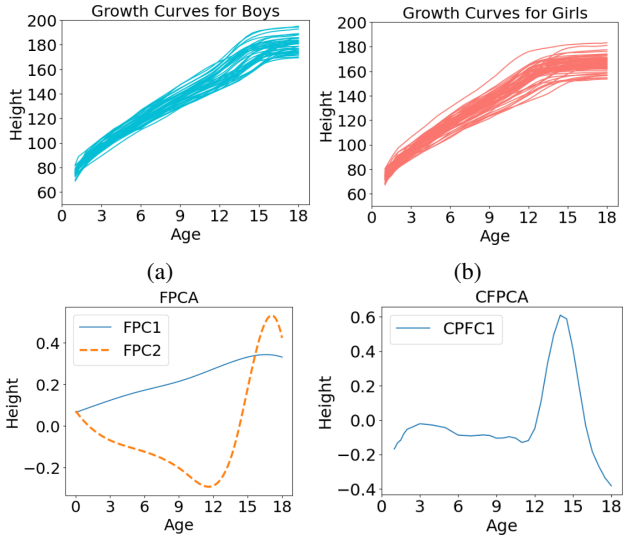


Figure 1: A toy dataset from a Berkeley growth study. (a) Height progression of boys. (b) Height progression of girls. (c) The first two FPCs. (d) The first CFPC.

As an illustrative example, Fig. 1 examines a benchmark dataset from the Berkeley growth study, tracking the heights of boys and girls as a function of age from 0 to 18 (Tuddenham and Snyder 1954). In Fig. 1c, the first FPC represents the mean progression of height, while the second indicates a divergence around age 12, a period of differential growth between boys and girls. However, FPCA fails to clearly delineate this difference. In contrast, CFPC1, with boys as the foreground and girls as the background, shows a steep ascent around age 12, highlighting the faster growth rate in boys due to the later onset of puberty compared to girls (Rogol, Roemmich, and Clark 2002). This example illustrates how CFPCA successfully unveils unique information specific to the foreground.

This paper is organized as follows. In the Method section, we introduce the CFPCA algorithm. In Theory section, we provide asymptotic theory on the estimated CFPCs. In Simulation, we validate the theory with four different simulations. The Application section presents results from applying CFPCA to multiple datasets. Finally, in Discussion, we summarize our findings and discuss potential future directions for research. All proofs and additional experimental details are available in the same GitHub repository as the code.

Method

Let $c(t,s) = \text{Cov}(u(t), u(s))$ be the covariance function of a random curve $u(t) \in L^2(\mathbb{R})$, the space of real-valued square-integrable functions. Its covariance operator, denoted by $C : L^2(\mathbb{R}) \to L^2(\mathbb{R})$, is defined as

$$C(v)(t) = \int c(t,s)v(s)ds,\ v \in L^2(\mathbb{R}). \tag{1}$$

The goal is to find $v \in L^2(\mathbb{R})$ with norm 1 that maximizes the variance in the foreground and minimizes the variance in the background, which is in line with the key idea in its finite dimensional counterpart (Abid et al. 2017). Let $C_x(t), C_y(t)$ be the covariance operators for the foreground curve $x(t)$ and background curve $y(t)$, respectively, and let $c_x(t,s), c_y(t,s)$ be the corresponding covariance functions. This optimization problem is:

$$\underset{v\in L^2(\mathbb{R}),\|v\|=1}{\text{argmax}} \int\int (c_x(t,s) - \alpha c_y(t,s))v(t)v(s)dsdt, \tag{2}$$

where $\alpha \in \mathbb{R}$ is a hyperparameter. The bigger α, the more important the background group is. One extreme case is $\alpha = 0$, so that CFPCA degenerates to FPCA. In another direction, when $\alpha = \infty$, CFPCA finds the curve orthogonal to the FPCs of the background group. The practical choice of α is deferred to the Application section.

Similar to FPCA, which is an eigen-problem, the above optimization problem is also equivalent to a eigen-problem:

$$\int c(t,s)v(s)ds = \lambda v(t) \tag{3}$$

where $c(t,s) = c_x(t,s) - \alpha c_y(t,s)$. The resulting v is the low dimensional curve unique to foreground group.

In practice, with finite observations often discretized at specific time points, we represent the functional data by a matrix, with the caveat that the order in which measurements appear elicits a natural meaning. Consider the observed curves $\{x_i(t_k)\}_{i=1}^N$ and $\{y_j(t_k)\}_{j=1}^M$, where t_k are discrete time points, $k = 1, \cdots, n$. We refer to $\{x_i\}$ the foreground dataset and $\{y_j\}$ the background dataset.

We further assume that the n time points $t_1, \cdots, t_n$ are equally spaced apart on a finite time interval with length T. In scenarios where this uniformity is absent and the data is sufficiently dense, we can still apply CFPCA by interpolating the data first (Zhang and Wang 2016; Cardot 2000;

Ramsay and Silverman 2005; Zhang and Chen 2007), so that each observation has consistent temporal measurements. We denote $c_x^{N,n}(t_i, t_j)$ and $c_y^{M,n}(t_i, t_j)$ as the empirical centered covariance functions for the foreground and background at the discretized time points. Similarly, define the centered sample covariance matrices $C_x^{N,n}$, $C_y^{M,n} \in \mathbb{R}^{n\times n}$. $(C_x^{N,n})_{ij} = c_x^{N,n}(t_i, t_j)$ and $(C_y^{M,n})_{ij} = c_y^{M,n}(t_i, t_j)$. Define $C^{N,M,n} = C_x^{N,n} - \alpha C_y^{M,n} \in \mathbb{R}^{n\times n}$. In addition, for any function v with unit norm, let v be the n-vector of values $v(t_j)$. Let $w = T/n$, $c^{N,M,n}(t_i, t_j) = c_x^{N,n}(t_i, t_j) - \alpha c_y^{M,n}(t_i, t_j)$. Then for each t_i, we can approximate Equation (3) as

$$\int c(t_i, s)v(s)ds \approx w \sum_k c^{N,M,n}(t_i, t_k)v(t_k)$$

Therefore, we can approximate the solutions in Equation (3) by solving the following eigen-problem:

$$wC^{N,M,n}v = \lambda v. \tag{4}$$

To obtain an approximate eigenfunction, we can use any interpolation method from the discrete eigenvectors v, such as splines (Micula and Micula 2012). However, if the points are dense enough, this will not have a significant effect. CFPCA algorithm is summarized in Algorithm 1.

As in PCA and cPCA, we call the leading eigenfunctions the contrastive functional principal components (CFPCs). Moreover, we provide error bounds on the estimated eigenfunctions $\widehat{v}_k$ and eigenvalues $\widehat{\lambda}_k$ in the Theory section.

Algorithm 1: CFPCA

1: **Input** target and background data: $\{x_i(t_k)\}_{i=1}^N$, $\{y_j(t_k)\}_{j=1}^M$, $k = 1, \cdots, n$; contrast parameter α, number of components L, number of time points n, length of interval T. Define $w = \frac{T}{n}$.

2: Center the data $\{x_i(t_k)\}_{i=1}^N$, $\{y_j(t_k)\}_{j=1}^M$.

3: Calculate empirical covariance functions:

$$c_x^{N,n}(t_k, t_l) = \frac{1}{N}\sum_{i=1}^{N} x_i(t_k)x_i(t_l)$$

$$c_y^{M,n}(t_k, t_l) = \frac{1}{M}\sum_{j=1}^{M} y_j(t_k)y_j(t_l)$$

4: Perform eigendecomposition on the discretized covariance matrices:

$$w \cdot C^{N,M,n} = w \cdot (C_x^{N,n} - \alpha C_y^{M,n})$$

5: For each eigenvector $(u_1, \ldots, u_L)$ and eigenvalue $(\rho_1, \ldots, \rho_L)$,

$$\widehat{v}_j = w^{-1/2}u_j, \ \widehat{\lambda}_j = w^{-1}\rho_j$$

6: **Return** $(\widehat{v}_1, \ldots, \widehat{v}_L)$ and $(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_L)$.

In FPDA, the implementation in Python uses a different technique (Ramos-Carreo et al. 2019) for better scalability and numerical robustness. For the same purpose, we have adapted CFPCA to incorporate this approach (see Appendix). For each application, we still use Algorithm 1 outlined here. Empirically however, we found that the results of both algorithms produced similar results.

We now consider a special parametric model. Motivated by the observation on shared components and foreground-specific components, we propose the following model assumption, which shares the same philosophy as CLVM (Severson, Ghosh, and Ng 2019):

$$x_i(t) = \sum_{l=1}^{L} a_{il}w_l(t) + \sum_{k=1}^{K} \eta_{ik}s_k(t) + \epsilon_i(t), \quad i = 1 \ldots N,$$

$$y_j(t) = \sum_{k=1}^{K} \gamma_{jk}s_k(t) + \epsilon_j(t), \quad j = 1 \ldots M. \tag{5}$$

where $w_l(t)$ and $s_k(t)$ are latent, orthogonal curves, and a_{il}, η_{ik}, $\gamma_{jk} \in \mathbb{R}$ are the corresponding coefficients and are mutually independent. The noise components ϵ_i and ϵ_j are independent and distributed as $\epsilon_i, \epsilon_j \sim \mathcal{N}(0, f(t))$, where $f(t)$ is a noise level function. The primary interest lies in $\{w_l(t)\}_{l=1}^L$, the lower-dimensional representation unique to the foreground dataset. In contrast, $\{s_k(t)\}_{k=1}^K$ are shared by two groups and are to be removed.

Observe that when Equations (5) hold, then w_l's are the top eigenfunctions of $C_x - C_y$. As a result, we suggest to set $\alpha = 1$ as the default in practice, or as a reference for parameter tuning. In this situation, we expect the estimated eigenfunctions $\widehat{v}_j$ to be consistent estimators of the foreground-specific components w_j, as detailed in the next section.

Theory

This section provides error bounds for the estimated eigenfunctions and eigenvalues derived from CFPCA. We begin by defining the functional spaces and introducing key notations. Let $L^2([a, b])$ be space of measurable real-valued functions defined on interval $[a, b] \subset \mathbb{R}$ with finite squared integrals. This forms a separable Hilbert space, with inner product $\langle x, y\rangle = \int_a^b x(t)y(t)dt, \ x, y \in L^2([a, b])$. The space of bounded linear operators on $L^2([a, b])$ is denoted by $\mathcal{L}$, with the norm defined as

$$\|\Psi\|_{\mathcal{L}} = \sup\{\|\Psi(x)\| : \|x\| \le 1\}, \ \Psi \in \mathcal{L}.$$

Integral operators are within this space:

$$\Psi(x)(t) = \int_a^b \psi(t, s)x(s)ds, \quad x \in L^2([a, b]),$$

with the Hilbert-Schmidt norm is defined as $\|\Psi\|_{\mathcal{S}}^2 = \int\int \psi^2(t, s)\, dt\, ds$.

Assume the foreground $\{x_i\}_{i=1}^N$ and the background $\{y_j\}_{j=1}^M$ be iid functions with zero mean. Define C_x and C_y as the covariance operators of the foreground and the background, respectively, and C_x^N and C_y^M as their empirical counterparts. Consider v_j and $\widehat{v}_j^{N,M}$ as the j-th eigenfunction of $C = C_x - \alpha C_y$ and $C^{N,M} = C_x^N - \alpha C_y^M$, respectively, with $\alpha \ge 0$. The corresponding eigenvalues are

denoted as λ_j and $\widehat{\lambda}_j^{N,M}$. Additionally, let $\widehat{v}_j^{N,M,n}$ be the discretized vector of $\widehat{v}_j^{N,M}$. We use $\widehat{p}_j^{N,M} = \text{sign}(\langle \widehat{v}_j^{N,M}, v_j \rangle)$ to account for potential sign indeterminacy in eigenfunctions.

To quantify the convergence rate of these estimates, we present the following theorem:

Theorem 1. *Assume the $L+1$ largest eigenvalues of C satisfy $\lambda_1 > \lambda_2 > \cdots > \lambda_L > \lambda_{L+1} \geq 0$ and $x_i, y_k \in L^4(\mathbb{R})$ for $i = 1, \ldots, N$ and $k = 1, \ldots, M$, then the following hold for $j = 1, \cdots, L$:*

$$\mathbb{E}\left[\left\|\widehat{p}_j^{N,M}\widehat{v}_j^{N,M} - v_j\right\|\right] = O(N^{-\frac{1}{2}}) + O(M^{-\frac{1}{2}}) \quad (6)$$

$$\mathbb{E}\left[\sup_{1\leq j\leq L}\left|\widehat{\lambda}_j^{N,M} - \lambda_j\right|\right] = O(N^{-\frac{1}{2}}) + O(M^{-\frac{1}{2}}). \quad (7)$$

This theorem demonstrates the effectiveness of CFPCA in providing robust estimations for the eigenfunctions and eigenvalues with convergence rates that match traditional parametric rates. Building on this theoretical foundation, the following theorem specifically addresses the recovery of foreground-specific components under the model assumption in Equations (5).

Theorem 2. *When x_i, y_k follow the model assumption in Equations* (5)*, let $\widehat{v}_j^{N,M}$ be the solution of CFPCA with $\alpha = 1$, then for $j = 1, \cdots, L$,*

$$\mathbb{E}\left[\left\|\widehat{p}_j^{N,M}\widehat{v}_j^{N,M} - w_j\right\|\right] = O(N^{-\frac{1}{2}}) + O(M^{-\frac{1}{2}})$$

that is, CFPCA can recover the foreground-specific components w_j at the parametric rate.

The theorem presented below provides an extension to account for discretization.

Theorem 3. *Under the same assumptions as Theorem 1 , let n be the number of observed time points and $\widehat{v}_j^{N,M,n}$ be the discretized solution of CFPCA, then*

$$\mathbb{E}\left[\left\|\widehat{p}_j^{N,M,n}\hat{v}_j^{N,M,n} - v_j\right\|\right] = O(N^{-\frac{1}{2}} + M^{-\frac{1}{2}} + n^{-\frac{1}{2}}).$$

Under the same assumptions as Theorem 2, let $\widehat{v}_j^{N,M,n}$ be the discretized solution of CFPCA with $\alpha = 1$, then

$$\mathbb{E}\left[\left\|\widehat{p}_j^{N,M,n}\hat{v}_j^{N,M,n} - w_j\right\|\right] = O(N^{-\frac{1}{2}} + M^{-\frac{1}{2}} + n^{-\frac{1}{2}}).$$

The above theorems pave the way for applying CFPCA in varied contexts, which we explore through simulations and applications in the next two sections. Proofs can be found in the Supplement.

Simulation

This section presents a series of four simulations designed to demonstrate the effectiveness of CFPCA over FPCA. Table 1 details the configuration of how the foreground and background datasets were generated, following the model assumptions in Equations (5). Table 2 compares performance of FPCA applied solely to the foreground, FPCA applied to the union of the foreground and background, and CFPCA, measured by the bias, computed by finding the L^2 norm of the difference of the true and estimated eigenfunction, of the first and the second PCs.

The results consistently show that CFPCA outperforms both versions of FPCA by achieving significantly lower bias values for both principal components. This highlights CFPCA's capability to more effectively distinguish and visualize the unique characteristics of the foreground data.

Sim	Simulation Settings	
	Shared	Foreground-specific
1	$s_1(t) = \cos(2\pi t + 1)$	$w_1(t) = \frac{3}{4}(3t^2 - 1)$, $w_2(t) = \frac{9}{10}(5t^3 - 3t)$
2	$s_1(t) = t^2 e^t$, $s_2(t) = te^{-t}$	$w_1(t) = \cos(2\pi t)$, $w_2(t) = \sin(2\pi t)$
3	$s_1(t) = \cos(\pi t + 1)$, $s_2(t) = \sin(2\pi t + 1)$, $s_3(t) = \cos(4\pi t + 1)$	$w_1(t) = t^2 - 1$, $w_2(t) = \frac{1}{2}(t^3 - 3t)$
4	$s_1(t) = t^3 - 2t + 2$, $s_2(t) = -5.75t^2 + t + 1$	$w_1(t) = t^2 - 1$, $w_2(t) = \frac{1}{2}(t^3 - 3t)$

Table 1: The foreground was generated from a linear combination of the shared functions and the foreground-specific (FG) functions, while the background (BG) generated from just the shared functions. For both groups, $\mathcal{N}(0, 1)$ coefficients were used.

Sim	FPCA on FG		FPCA on FG + BG		CFPCA	
	PC1	PC2	PC1	PC2	PC1	PC2
1	0.72	0.43	1.01	0.50	**0.08**	**0.08**
	(0.01)	(0.01)	(0.002)	(0.01)	(0.06)	(0.05)
2	1.32	1.13	1.33	1.15	**0.14**	**0.14**
	(0.01)	(0.01)	(0.003)	(0.003)	(0.14)	(0.14)
3	1.24	1.24	1.26	1.35	**0.05**	**0.04**
	(0.01)	(0.01)	(0.003)	(0.006)	(0.02)	(0.03)
4	1.39	1.25	1.39	1.25	**0.14**	**0.13**
	(0.004)	(0.002)	(0.003)	(0.001)	(0.13)	(0.13)

Table 2: Bias of the first and second FPCs. 40 replicates were conducted with the mean (standard deviation) reported in this table. The samples size are $N = M = 100000$ over the time interval $[-1, 1]$.

Fig. 2 illustrates the first and second FPCs from Simulation 2, as a representative example. While FPCA captures some aspects of the true function, CFPCA provides a much clearer and more accurate recovery of the foreground-specific patterns.

Further empirical support for CFPCA's theoretical advantages is provided in Fig. 3, which correlates with the convergence rate described in Theorem 2. The boxplots and log plots display the bias reduction as the sample size increases, confirming the convergence rates of approximately $-1/2$ for

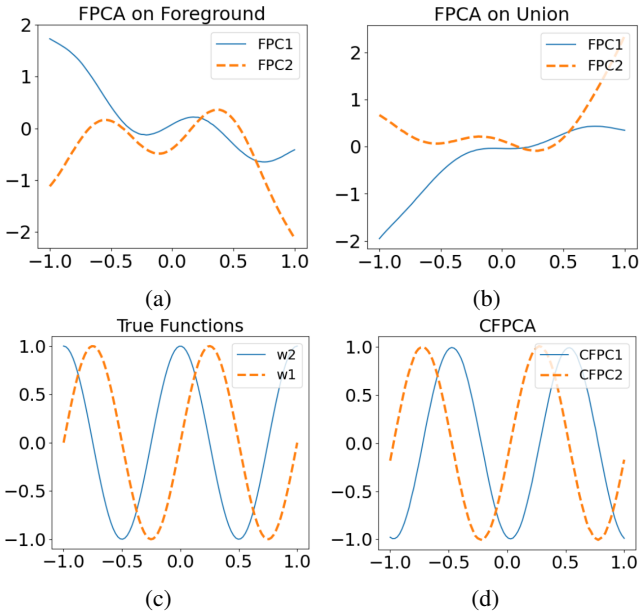


Figure 2: FPCs for simulation 2. FPCA is unable to recover the unique functions (c) in (a) and (b), while CFPCA is able to recover them almost exactly in (d).

the estimated eigenfunctions. For additional experimental detail and similar plots for the other three simulations, see Supplement.

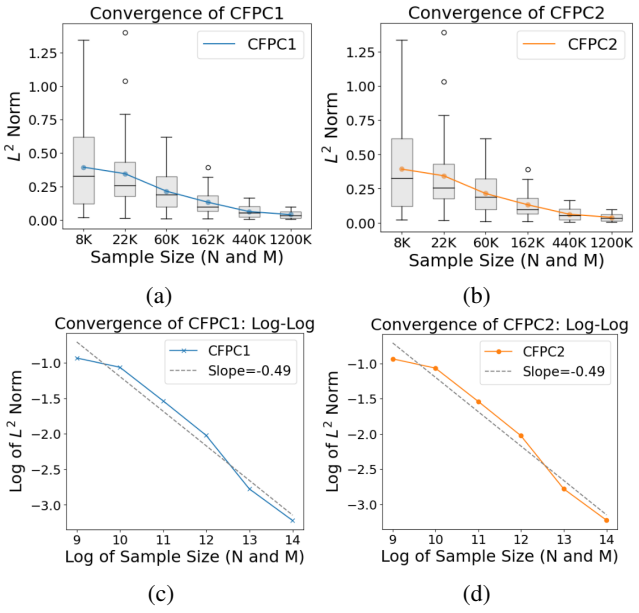


Figure 3: Bias for simulation 2. (a) and (b): box plot (with 40 replicates) of the bias of the first two CFPCs with increasing sample sizes . (c) and (d): log-log plot to illustrate the convergence rate. Best fit line is plotted, with slope close to $-1/2$.

Application

In this section, we demonstrate the application of CFPCA on two datasets: the gait cycle and stock market. When implementing CFPCA, two critical hyperparameters must be considered: $\alpha > 0$, which controls the influence of the background dataset, and L, the number of dimensions for reduction. We typically set $\alpha = 1$ as suggested by Equations (5); however, the optimal α can vary depending on the level of shared information between the foreground and background datasets. In scenarios with discernible subgroups within the foreground group, cross-validation can be used to optimize α. For the applications discussed in this section, we chose an α that not only ensures effective separation but also facilitates interpretation, with details on the selection process and comprehensive results available in the Supplement. In the situation without such a subgroup structure, tuning α remains a challenging problem (Li, Jones, and Engelhardt 2020), see also Discussion. The choice of the dimensionality, L, arguably an open question (Wang, Chiou, and Müller 2016), is guided by the specific needs of the analysis, with $L = 1$ or 2 typically sufficient for visualization purposes.

In selecting which datasets serve as the foreground or background groups, the distinction is straightforward in some scenarios. For instance, in biomedical case-control studies, cases would naturally constitute the foreground dataset, as they are the primary focus of the study, while healthy patients serve as the background dataset, helping to isolate and remove non-target variations. However, in less clear-cut scenarios, the decision rests with the researcher based on the specific focus on the analysis.

In each application, we present the Silhouette score (SS) that evaluates each cluster's tightness and separation (Rousseeuw 1987), and the Davies-Bouldin (DB) index that measures intracluster similarity and inter-cluster differences (Singh et al. 2020), as metrics. The higher the SS or the lower the DB, the better the clustering. Table 3 summarizes the datasets we considered.

Experiment (Fig #)	N	M	n
Berkeley Growth Dataset (Fig. 1)	39	54	31
Gait Cycle (Fig. 4)	200	200	100
Stock Market (Fig. 5)	63	49	252

Table 3: Summary of characteristics of datasets on each application. N, M is the size of foreground and background respectively, n is the number of time points.

Gait Cycle

The first application investigates the gait cycle, defined as the period from when one foot makes contact with the ground to its next contact. The purpose of the study was to examine how knee and ankle joint angular displacement are affected by bracing. In particular, the investigators wanted to see how asymmetric gait, which was simulated using a brace on the knee or ankle, affected joint angular displacement. The data consists of 10 subjects, each with 3 possible conditions: unbraced, knee brace, or ankle brace, and 2 possible

placements: the left or right leg. Angular displacements were consistently measured at both the knee and ankle, irrespective of the brace's location. Ten consecutive gait cycles were repeated for each possible combination. The data were collected by (Shorter et al. 2008), processed by (Helwig et al. 2011), and published by (Helwig et al. 2016).

Our specific interest lies in the knee's angular displacement when an ankle brace is worn. To establish a baseline, we run FPCA on just braced subjects (Fig. 4a), and also on the union of braced and unbraced subjects (Fig. 4c). FPCA cannot reveal any significant clustering within the foreground in either case. In contrast, CFPCA, using unbraced subjects as a background dataset, successfully differentiated between left and right knee behaviors, unveiling subtle differences not captured by FPCA (Fig. 4e).

This result indicates that ankle brace may affect knee mobility differently, which to our knowledge, has not been previously studied. By using a background dataset, CFPCA was able to isolate and highlight unique, low-dimensional structures in the foreground data, effectively focusing on the variation specific to the braced condition.

To further understand why CFPCA successfully identifies this unique structure, we examine the basis functions unique to the foreground, as shown in Fig. 4f.

Fig. 4b indicates that the left and right knee in the foreground generally share the same shape, which explains why FPCA's difficulty in distinguishing the two. These commonalities are also present in the background (Fig. 4d), enabling CFPCA to effectively isolate and highlight the distinctive information pertinent to the braced condition. In Fig. 4f, the intervals 0-20, 30-40, 70-90 show increased variation, corresponding to noticeable differences in the right knee when braced at the ankle. This suggests that bracing the ankle introduces changes in knee mobility, particularly accentuating variance during these phases of the gait cycle. The minimal variance in these phases when unbraced implies a significant impact of the brace on knee movement. Further downstream analysis is necessary to fully understand the biomechanical impact of the ankle brace on knee movement and to explore potential implication. See Supplement for more discussion.

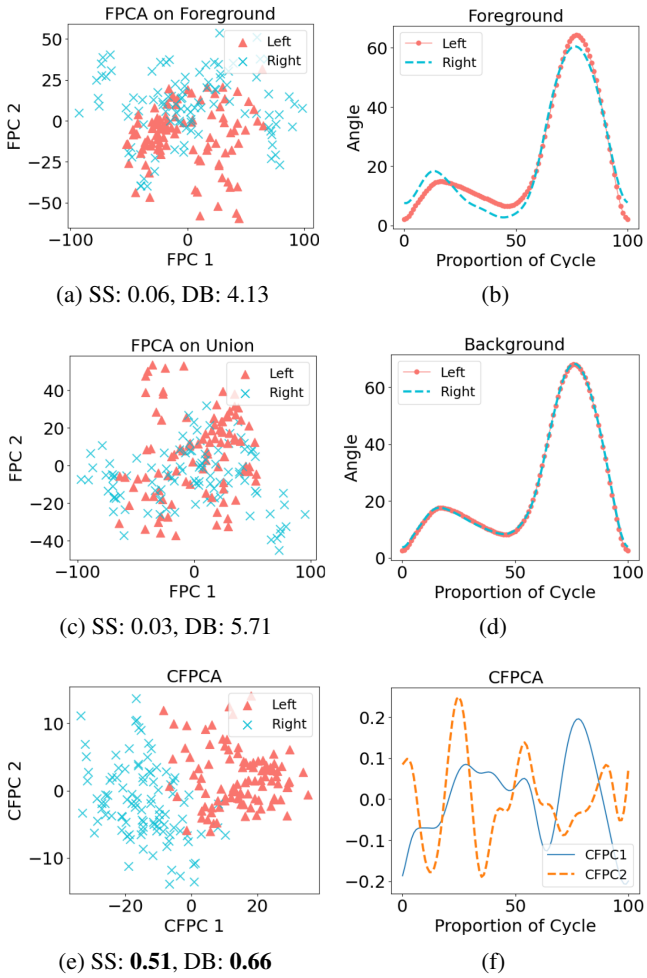


Figure 4: (First column) The first two FPCs from (a) FPCA on just the braced subjects; (c) FPCA on the union of braced and unbraced; (e) CFPCA with braced as the foreground and unbraced as the background. (Second column) (b) Mean angle displacements of the braced subjects separated by knee; (d) Mean angle displacements on the unbraced subjects separated by knee; (f) The top two CFPCs.

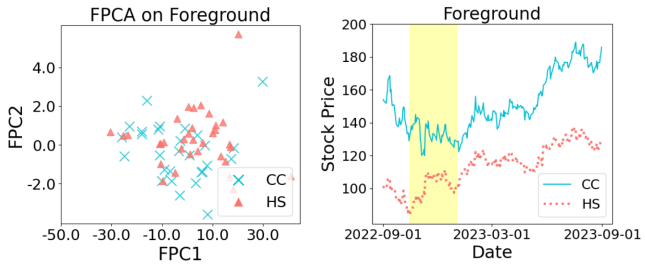
Stock Market

In our second case study, we explore the stock market dynamics, focusing on technology and non-technology companies. We sourced daily closing prices from Yahoo Finance using the yfinance package (Aroussi 2024). Fig. 5a illustrates the outcome of applying FPCA solely to the technology sector, revealing no significant patterns. Similarly, Fig. 5c shows the results for both sectors combined, which also fails to demonstrate any notable clustering or distinct patterns.

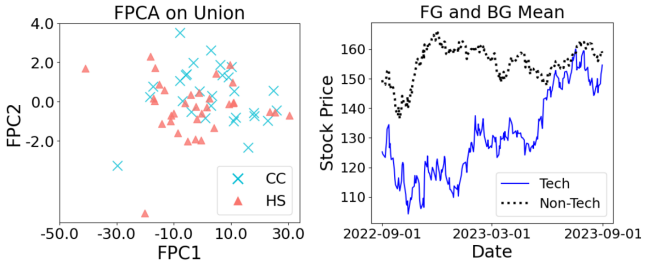
Using non-technology stocks as the background dataset, CFPCA aims to highlight technology-specific trends. As shown in Fig. 5e, CFPCA successfully differentiates between companies specializing in cloud computing (CC) services and those in hardware and semiconductors (HS).

Further analysis of mean behaviors reveals that both cloud computing and hardware and semiconductors generally show upward trends despite fluctuations. Fig. 5b shows subtle differences between groups highlighted in yellow. CFPCA effectively isolates unique trends within the foreground, as shown in Fig. 5f, where variances from October 2022 to January 2023 aligns precisely with the periods with divergent trends between sectors. During this period, the hardware and semiconductor sectors saw high demand, especially for chips (SIA 2022). Additionally, there was also notable innovation and product development across multiple companies. For instance, Nvidia unveiled its next-generation GPU architecture in September (Holt 2022), while Intel introduced the 13th Gen Intel Core, touted as the world's fastest desktop processor (Intel 2022). To address the surging demand for semiconductors, Applied Materials announced plans to expand its operations in Singapore, increasing its global manufacturing and research and development capacities (Applied Materials 2022).

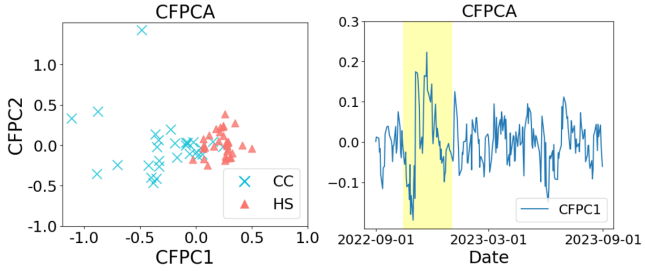
With the rise of artificial intelligence and expansion in cloud services, competition reached its peak (Seth 2023). However, several companies experienced a moderate decrease in performance briefly. This showed in technology giant Google, who although continued to allocate large amounts of money in cloud computing, was still operating at losses during this period (HG Insights 2023). Salesforce experienced a rough end to 2022, partly due to the competition in cloud services and lagging behind its top technology rivals (Monica 2022). The financial news corresponding to this time period seems to be reflective of what we observe in the performance in Fig. 5b and the heightened variation in Fig. 5f.



(a) SS: -0.008, DB: 8.33 (b)



(c) SS: -0.008, DB: 8.37 (d)



(e) SS: **0.33**, DB: **1.09** (f)

Figure 5: (Left) The first two FPCs from (a) FPCA on technology; (c) FPCA on the union of technology and non-technology; (e) CFPCA with technology as the foreground and non-technology as the background. (Right) (b) Mean stock price of technology by sectors; (d) Mean of technology and non-technology; (f) The first CFPC.

Additional insights into the underlying causes of these trends were sought through an examination of relevant financial news, with more detailed findings in the Supplement.

Building on our discussion, it is crucial to consider the potential societal impacts of these findings. On one hand, this analysis could help investors and policymakers make better-informed decisions. However, there is also a risk that these insights could be misused for market manipulation or to create unfair competitive advantages, emphasizing the need for ethical guidelines and transparency in financial data usage.

Discussion

In this paper, we introduced CFPCA as a method to discover unique variation within a foreground group while minimizing the variation in a background group.

The strategic use of a background dataset effectively eliminates more dominant, non-distinct variations from the foreground analysis, enhancing the clarity of our analyses. However, like FPCA, the dimension reduction performed by CFPCA does not yield statistically significant metrics by itself, necessitating further hypothesis testing to substantiate the findings. Although our current implementation is limited to time series data, its methodology is flexible and can be adapted for other types of functional data, including spatial data and images.

To enhance the utility and applicability of CFPCA, several promising directions can be pursued.

Kernel method: Incorporating kernel methods to create a kernel version of CFPCA could allow the method to capture nonlinear relationships. This would be particularly beneficial in fields like genomics and image processing, where data often exhibit nonlinear patterns.

Multiple foreground groups: Extending CFPCA to manage scenarios with multiple foreground groups, particularly in biomedical studies where multiple treatment groups are compared against a single control. This extension requires a framework to distinguish and analyze each group's unique features relative to the control, especially important in clinical trials and personalized medicine, where understanding the differential effects of treatments is crucial.

Real-time analysis: Adapting CFPCA for real-time data analysis, such as in financial market or continuous health monitoring systems would expand its usability in dynamic settings where data are continuously updated.

Sparsity and irregularity: Extending CFPCA to better handle datasets that are sparse and irregularly sampled would broaden its applicability in areas like environmental monitoring and internet of things (IoT) where sensors might record data at non-uniform intervals. Such extensions could involve integrating statistical and machine learning techniques that specialize in dealing with missing data and interpolation.

Automated hyperparameter tuning: Developing methods for automated tuning of the contrastive parameter α and the number of unique components L, would simplify the use of CFPCA, making it more accessible to users who may not be experts in statistical methodology.

Acknowledgements

EZ was supported by NIH T32ES007018; DL was supported by R01 AG079291, R56 LM013784, R01 HL149683, UM1 TR004406, R01 LM014407, and P30 ES010126.

References

Abid, A.; Zhang, M. J.; Bagaria, V. K.; and Zou, J. 2017. Contrastive principal component analysis. *arXiv preprint arXiv:1709.06716*.

Applied Materials. 2022. Applied Materials Launches Singapore 2030 Plan to Expand its Operations and Innovation Capabilities. Applied Materials.

Aroussi, R. 2024. yfinance. Python package index. Python library for accessing historical market data from Yahoo Finance.

Boente, G.; Rodriguez, D.; and Sued, M. 2011. Testing the equality of covariance operators. In *Recent advances in functional data analysis and related topics*, 49–53. Springer.

Cai, T. T.; and Yuan, M. 2011. Optimal estimation of the mean function based on discretely sampled functional data: Phase transition. *Annals of Statistics*.

Cardot, H. 2000. Nonparametric estimation of smoothed principal components analysis of sampled noisy functions. *Journal of Nonparametric Statistics*, 12(4): 503–538.

Fan, J.; and Lin, S.-d. 1998. Test of significance when data are curves. *Journal of the American Statistical Association*, 93(443): 1007–1021.

Hall, P. 2010. Principal component analysis for functional data: Methodology, theory, and discussion. *Journal of the American Statistical Association*.

Hall, P.; and Hosseini-Nasab, M. 2006. On properties of functional principal components analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(1): 109–126.

Hall, P.; and Van Keilegom, I. 2007. Two-sample tests in functional data analysis starting from discrete data. *Statistica Sinica*, 1511–1531.

Helwig, N. E.; Hong, S.; Hsiao-Wecksler, E. T.; and Polk, J. D. 2011. Methods to temporally align gait cycle data. *Journal of biomechanics*, 44(3): 561–566.

Helwig, N. E.; Shorter, K. A.; Ma, P.; and Hsiao-Wecksler, E. T. 2016. Smoothing spline analysis of variance models: A new tool for the analysis of cyclic biomechanical data. *Journal of biomechanics*, 49(14): 3216–3222.

HG Insights. 2023. Market Report on Google Cloud Platform. HG Insights.

Holt, K. 2022. NVIDIA looks set to reveal its next-gen GeForce RTX GPUs on September 20th. engadget.

Hörmann, S.; and Kokoszka, P. 2012. Functional time series. In *Handbook of statistics*, volume 30, 157–186. Elsevier.

Hotelling, H. 1933. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6): 417.

Intel. 2022. Intel Launches 13th Gen Intel Core Processor Family Alongside New Intel Unison Solution. Intel.

Joliffe, I. T.; and Morgan, B. 1992. Principal component analysis and exploratory factor analysis. *Statistical methods in medical research*, 1(1): 69–95.

Jolliffe, I. T.; and Cadima, J. 2016. Principal component analysis: a review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, 374(2065): 20150202.

Kokoszka, P.; and Reimherr, M. 2017. *Introduction to functional data analysis*. Chapman and Hall/CRC.

Li, D.; Jones, A.; and Engelhardt, B. 2020. Probabilistic contrastive principal component analysis. *arXiv preprint arXiv:2012.07977*.

Micula, G.; and Micula, S. 2012. *Handbook of splines*, volume 462. Springer Science & Business Media.

Monica, P. 2022. Salesforce stock and earnings update. CNN.

Panaretos, V. M.; Kraus, D.; and Maddocks, J. H. 2010. Second-order comparison of Gaussian random functions and the geometry of DNA minicircles. *Journal of the American Statistical Association*, 105(490): 670–682.

Ramos-Carreo, C.; Surez, A.; Torrecilla, J. L.; Carbajo Berrocal, M.; Marcos Manchn, P.; Prez Manso, P.; Hernando Bernab, A.; Garca Fernndez, D.; Hong, Y.; Rodrguez-Ponga Eyris, P. M.; Snchez Romero, .; Petrunina, E.; Castillo, .; Serna, D.; and Hidalgo, R. 2019. GAA-UAM/scikit-fda: Functional Data Analysis in Python.

Ramsay, J.; and Silverman, B. 2005. Principal components analysis for functional data. *Functional data analysis*, 147–172.

Ramsay, J. O. 1997. BW Silverman Functional data analysis.

Rogol, A. D.; Roemmich, J. N.; and Clark, P. A. 2002. Growth at puberty. *Journal of adolescent health*, 31(6): 192–200.

Rousseeuw, P. J. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20: 53–65.

Seth, A. 2023. Cloud war heats up. Twimbit.

Severson, K. A.; Ghosh, S.; and Ng, K. 2019. Unsupervised learning with contrastive latent variable models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 4862–4869.

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

Shorter, K. A.; Polk, J. D.; Rosengren, K. S.; and Hsiao-Wecksler, E. T. 2008. A new approach to detecting asymmetries in gait. *Clinical Biomechanics*, 23(4): 459–467.

SIA. 2022. State of U.S. Semiconductor Industry. Semiconductor Industry Association.

Singh, A. K.; Mittal, S.; Malhotra, P.; and Srivastava, Y. V. 2020. Clustering evaluation by davies-bouldin index (dbi) in cereal data using k-means. In *2020 Fourth international conference on computing methodologies and communication (ICCMC)*, 306–310. IEEE.

Tuddenham, R.; and Snyder, M. 1954. Physical growth of California boys and girlsfrom birth to age 18. *Calif. Publ. Child Develop*, 1: 183–364.

Wang, J.-L.; Chiou, J.-M.; and Müller, H.-G. 2016. Functional data analysis. *Annual Review of Statistics and its application*, 3: 257–295.

Yao, F.; Müller, H.-G.; and Wang, J.-L. 2005. Functional data analysis for sparse longitudinal data. *Journal of the American statistical association*, 100(470): 577–590.

Zhang, J.-T.; and Chen, J. 2007. Statistical inferences for functional data. *The Annals of Statistics*.

Zhang, X.; and Wang, J.-L. 2016. From sparse to dense functional data and beyond. *Electronic Journal of Statistics*.