Functional PCA for Dimensionality Reduction (Machine Learning 2025 Course)

Anita Toleutaeva [1] **Denis Suchkov** [1] **Ildar Saiapov** [1] **Oleg Kobzarev** [1]

Abstract

Dimensionality reduction is a fundamental tool in data analysis, with PCA being its simplest and most widely used linear method. However, many applications involve data, that can not be analyzed in terms of a set of distinct features (like usual tabular data), but rather as a functions, evaluated at some grid (time series, images, video). Functional Principal Component Analysis (FPCA) extends PCA to this domain by extracting smooth basis functions that capture the inherent structure of functional data.

In this work, we review the theoretical foundations and practical implementations of FPCA (with a particular focus on issues observed in the scikit-fda package), discussing its applications in feature extraction, and analyzing its shortcomings and limitations along with proposed strategies to overcome them.

Github repo: github.com/rainbowbrained/ML_FPCA

1. Introduction

Functional data, which are often represented as continuous curves or functions, arise in many modern applications ranging from biomedical signals and spectroscopic data to financial time series. Classical Principal Component Analysis (PCA) is widely used for dimensionality reduction in multivariate data; however, its direct application to functional data is inadequate due to the infinite-dimensional nature and inherent smoothness of such data. Functional Principal Component Analysis (FPCA) extends PCA by decomposing functions into a series of orthogonal basis functions (eigenfunctions) that capture the major modes of variation. This project aims to conduct a comprehensive literature review on FPCA, critically analyze current algorithms and implementations (with a special focus on potential issues in tools like scikit-fda), evaluate its applications in feature extraction, and discuss its limitations along with strategies to overcome them.

2. Preliminaries

In this section, we briefly outline the notation and key concepts underlying PCA and its extension to functional data.

Let $x_i \in \mathbf{R}^p$ for $i = \overline{1,n}$ be a dataset of n observations and p number of features in the classical (multivariate) setting. The data are then centered by subtracting μ and their covariance matrix is given by

$$\Sigma = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^\top.$$

PCA is defined as an orthogonal linear transformation on a real inner product space that transforms the data $X = (x_1, ..., x_n)^T \in \mathbf{R}^{n\times p}$ with column-wise zero empirical mean $\mu^{(j)} = \frac{1}{n}\sum_{i=1}^{n} x_i^j = 0$, $j = \overline{1,p}$, to a new coordinate system with basis being a set of eigenvectors $\{v_k\}_{k=1}^{p}$, such as

$$\Sigma\, v_k = \lambda_k\, v_k, \quad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0.$$

The greatest variance by some scalar projection of the data comes to lie on the first coordinate (called the first principal component PC), the second greatest variance on the second coordinate, and so on (Jolliffe, 2002):

$$z_{ik} = (x_i - \mu)^\top v_k.$$

FPCA extends classical PCA to data observed as functions. For functional data, we model each observation as a function $X_i(t)$ defined on a continuous domain $\mathcal{T}$, then decompose it via the Karhunen–Loeve expansion:

$$X_i(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_{ik}\, \varphi_k(t),$$

[1]Skolkovo Institute of Science and Technology, Moscow, Russia. Correspondence to: Anita Toleutaeva <Anita.Toleutaeva@skoltech.ru>, Denis Suchkov <Denis.Suchkov@skoltech.ru>, Ildar Saiapov <Ildar.Saiapov@skoltech.ru>, Oleg Kobzarev <Oleg.Kobzarev@skoltech.ru>.

Final Projects of the Machine Learning 2025 Course, Skoltech, Moscow, Russian Federation, 2025.

where $\mu(t)$ is the mean function, $\{\varphi_k(t)\}$ are orthonormal eigenfunctions obtained from the covariance operator $G(s,t)$ instead of a finite-dimensional covariance matrix Σ:

$$G(s,t) = \text{Cov}(X(s), X(t)) =$$
$$= \mathbf{E}[(X(s) - \mu(s))(X(t) - \mu(t))].$$

Eigenfunctions and eigenvalues satisfy the integral equation:

$$\int_{\mathcal{T}} G(s,t)\,\varphi_k(s)\,ds = \lambda_k\,\varphi_k(t), \quad t \in \mathcal{T}.$$

ξ_{ik} are the functional principal component scores:

$$\xi_{ik} = \int_{\mathcal{T}} (X_i(t) - \mu(t))\,\varphi_k(t)\,dt \approx$$
$$\approx \sum_{j=1}^{M} (X_i(t_j) - \mu(t_j))\,\varphi_k(t_j)\,\Delta t_j,$$

This representation reduces the infinite-dimensional problem to a finite number of components while retaining the essential structure of the data.

3. Literature Review

Our literature review on topic highlights several key contributions. Together, these works address the theory, computation, implementation, and application of FPCA. They compare discretization, basis expansion, and numerical methods, propose solutions for sparse data (PACE), extend FPCA to multivariate settings, and implement FPCA in software. Despite their strengths, challenges remain (e.g., numerical stability and handling of irregular sampling), motivating ongoing methodological improvements.

Classical PCA has been widely studied for its role in dimensionality reduction in multivariate data (Jolliffe, 2002). FPCA, as introduced in (Karhunen, 1946) and further developed in (Yao et al., 2005), adapts these ideas to settings where observations are curves or functions, providing a more natural representation for time series and other continuous data.

The Annual Review article (Wang et al., 2016) provides a theoretical foundation for FPCA and examine alternative approaches (to mitigate sparse, noisy data and other issues): ranging from mixed-effects modeling and kernel-based smoothing to the aforementioned PACE method. Authors establish convergence rates, for example,

$$\sup_{t \in T} |\hat{\mu}(t) - \mu(t)| = O_p\left(\frac{1}{\sqrt{nh}}\right),$$

and similar rates for smoothed $\hat{G}(s,t)$ and the eigenfunctions. They also emphasize that FPCA is central for dimension reduction and feature extraction in functional data, and inform current research efforts aimed at refining FPCA methodologies.

A Survey of FPCA (Shang, 2013) offers a comprehensive overview of techniques, categorizing the computational strategies into three main approaches. First, a discretization approach mimics classical PCA by renormalizing and interpolating eigenvectors computed on a dense grid. Second, the basis function expansion method represents the underlying stochastic process $X_i(t)$ as a linear combination of fixed basis functions $\{b_j(t)\}$:

$$X_i(t) \approx \sum_{M}^{j=1} c_{ij} b_j(t)$$

Third, numerical approximation methods (mainly quadrature rules) are used to compute integrals for irregularly spaced data. Shang also reviews several extensions of FPCA, including smoothed, robust, sparse, common, and multilevel FPCA.

In a similar vein (Hall et al., 2006) authors address the critical issue of sparsity in longitudinal data through their PACE (Principal Components Analysis through Conditional Expectation) method. Their pipeline includes:

1. Estimating the mean $\hat{\mu}(t)$ and covariance $\hat{G}(s,t)$ via local smoothing;
2. Computing eigenfunctions $\hat{\phi}_k(t)$ and eigenvalues $\hat{\lambda}_k$ from$\hat{G}(s,t)$;
3. Obtaining FPC scores using conditional expectation rather than standard numerical integration:

   $$\hat{\xi}_{ik} = \hat{\lambda}_k \hat{\phi}_k^T \hat{\Sigma}_{Y_i}^{-1} (Y_i - \hat{\mu}), \hat{\Sigma}_{Y_i} = \hat{G} + \hat{\sigma}^2 I.$$

   This method yields consistent estimates and enables confidence band construction.

The article by Happ and Greven (2018) (Happ & Greven, 2018) extend FPCA to multivariate functional data observed on heterogeneous domains (e.g., curves versus images). Their method applies the Karhunen–Loève theorem to jointly decompose data from different domains, incorporating specific weights to balance variation. The approach yields multivariate eigenfunctions $\{\phi_k^{(m)}(t)\}$ and scores that capture joint variation, serving as an alternative to tensor PCA.

More recent studies have focused on algorithmic improvements and practical challenges. The software implementation of FPCA is also critical for its practical adoption. The scikit-fda package (Berrendero et al., 2020) implements FPCA via basis expansion and other methods. It provides tools for mean and covariance estimation and eigenanalysis.

However, issues such as numerical instability and sensitivity to parameter tuning have been noted, especially when handling irregular or sparse data. Our review will examine implementations issues in detail.

4. Project Plan

1. **Planning and literature review**. Survey FPCA theory, algorithms, implementations (with focus on issues in scikit-fda), applications to feature extraction, shortcomings, limitations, and proposed remedies.

2. **Drafting main sections and initial analysis.** Read and annotate selected papers. Develop the section with definitions, notation, and key formulas for PCA and FPCA.

   Summarize motivations, recent advances, and gaps. Detail scikit-fda issues and compare with theoretical FPCA. Finalize mathematical preliminaries and link them coherently with application insights.

3. **FPCA on a toy problem and MNIST Classification Dataset**. Evaluate a noisy polynomial and trigonometrical function on a finite grid and use FPCA to predict its order.

   Apply FPCA to the preprocessed MNIST dataset, determine the optimal number of basis functions, and visualize the basis functions.

4. **PCA on a Multi-Resolution Image Dataset**. Collect or simulate a multi-resolution image dataset. Apply FPCA separately at different resolutions and compare the extracted features. Analyze how resolution impacts the effectiveness of FPCA for feature extraction.

5. **The curse of dimensionality and covariance estimation in FPCA**. Conduct a focused literature review on covariance estimation challenges and high-dimensional generalization. Reproduce selected approaches from the literature to compare performance. Analyze and report on the stability, bias, and variance of different covariance estimation methods.

6. **Revision, integration, and finalization**. Integrate findings into a cohesive final report. Conduct internal reviews, ensure consistency in formatting and style, and finalize all citations. Check references and formatting, prepare final version for submission.

References

Berrendero, J. et al. scikit-fda: A python package for functional data analysis. *Journal of Open Source Software*, 5(52):2979, 2020. URL https://www.researchgate.net/publication/380415756_scikit-fda_A_Python_Package_for_Functional_Data_Analysis.

Hall, P., Müller, H. G., and Wang, J. L. Properties of principal component methods for functional and longitudinal data analysis. *Annals of Statistics*, 34(1):1493–1517, 2006. doi: 10.1214/009053606000001090.

Happ, C. and Greven, S. Multivariate functional principal component analysiswang2016functional for data observed on different (dimensional) domains. *Journal of the American Statistical Association*, 2018. Volume, number, and page details to be added if available.

Jolliffe, I. T. Principal component analysis. springer series in statistics. *Annual Review of Statistics and Its Application*, 2002. doi: doi:10.1007/b98835. URL https://link.springer.com/book/10.1007/b98835.

Karhunen, K. Zur spektraltheorie stochastischer prozesse. *Annales Academiae Scientiarum Fennicae, Series A. I. Mathematica-Physica*, 37:1–79, 1946.

Shang, H. L. A survey of functional principal component analysis. *AStA Advances in Statistical Analysis*, 98(2): 121–142, 2013. doi: 10.1007/s10182-013-0213-1.

Wang, J.-L., Chiou, J.-M., and Müller, H.-G. Functional data analysis. *Annual Review of Statistics and Its Application*, 3:257–295, 2016. doi: 10.1146/annurev-statistics-041715-033624.

Yao, F., Müller, H.-G., and Wang, J.-L. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 2005. URL https://anson.ucdavis.edu/~mueller/jasa03-190final.pdf.

A. Team member's contributions

Explicitly stated contributions of each team member to the final project.

Anita Toleutaeva (25% of work)

- Collecting and organizing relevant literature
- Reviewing literate on the topic (1 papers)
- Preparing the GitHub Repo

Ildar Saiapov (25% of work)

- Reviewing literate on the topic (2 papers)
- Preparing a synthesis matrix of key papers

Denis Suchkov (25% of work)

- Reviewing and summarizing key mathematical foundations of FPCA (1 paper)
- Drafting the background and methods sections

Oleg Kobzarev (25% of work)

- Reviewing literate on FPCA implementation (1 paper)
- Drafting the implementation section