Report on

On the use of the Gram matrix for multivariate functional principal components analysis

The aim of this article is to present a new method for estimating the functional principal components for multivariate functional data. The idea is to extend the duality between the rows and the columns for PCA to the context of multivariate functional data.

The context is general : observations are vectors of functions $X = (X^{(1)}, \ldots, X^{(p)})^T$ where, for all $p = 1, \ldots, P$, $X^{(p)}$ is a random function from $\mathcal{T}_p \to \mathbb{R}$ with $\mathcal{T}_p$ a rectangle in $\mathbb{R}^{d_p}$ (the setting then includes both 2D and 3D images).

Multivariate functional PCA consists in estimating the eigenvalues and eigenfunctions of the covariance operator of the data

$$\Gamma : f = (f^{(1)}, \ldots, f^{(p)}) \in \mathcal{L}^1(\mathcal{T}_1) \times \ldots \times \mathcal{L}^P(\mathcal{T}_P) \mapsto \int_{\mathcal{T}_1 \times \mathcal{T}_p} \mathbb{E}\left[(X(\mathbf{s}) - \mathbb{E}[X(\mathbf{s})]) (X(\cdot) - \mathbb{E}[X(\cdot)])^T \right] f(\mathbf{s}) ds,$$

which is based on the estimation of Γ from data that are observed (most of the time) on a discrete grid of $\mathcal{T}_1 \times \ldots \times \mathcal{T}_p$.

The diagonalisation of Γ may be time-consuming (or even impossible) when the number of features P and/or the number of observation points is large.

To overcome this difficulty, the authors prove that the eigenfunctions/eigenvalues (ϕ_k/λ_k) and associated individual scores $\mathfrak{c}_k = \langle\langle X, \phi_k \rangle\rangle$ can be obtained by diagonalising the inner product matrix (or Gram matrix)

$$\mathbf{M} = \left(\frac{1}{N} \langle\langle X_n - \mathbb{E}[X_1], X_{n'} - \mathbb{E}[X_1] \rangle\rangle \right)_{n,n'=1,\ldots,N},$$

where $\langle\langle \cdot, \cdot \rangle\rangle$ is the usual scalar product on the product space $\mathcal{L}^1(\mathcal{T}_1) \times \ldots \mathcal{L}^P(\mathcal{T}_P)$ and $X_1, \ldots, X_N$ are the multivariate functional observations (i.e. random elements of $\mathcal{L}^1(\mathcal{T}_1) \times \ldots \times \mathcal{L}^P(\mathcal{T}_P)$).

More precisely, in Section 4.2 they write a relation between the eigenfunctions of the operator Γ and the eigenvectors of the matrix $\mathbf{M}$. They also establish that the eigenvalues of $\mathbf{M}$ are the same as the eigenvalues of Γ and that the scores can be easily obtained from the diagonalisation of the matrix $\mathbf{M}$. The results are proved in the appendix and a generalisation with weights on the features and individuals are also provided in the supplementary material. The main interest is that, when the number of observations is smaller than the number of grid points, which often happens, the diagonalization of usual estimators of the matrix $\mathbf{M}$ take less computation time than the diagonalization of usual estimators of the covariance operator Γ.

A simulation study is carried out as well as an application to an original dataset.

General comment

My opinion on the manuscript is positive. I find the result presented in the article very interesting for both specialists in the field of functional data analysis and practitioners. It seems to me that it could be a very good contribution to the *Journal of Multivariate Analysis*.

I only have some remarks.

Date: August 30, 2024.

- **p. 1 l. 9–10 of section 1:** It seems to me that the sentence "*FPCA was introduced by Karhunen (1947) and Loève (1945) and developped by Dauxois et al. (1982)*" is not entirely true. In fact, Karhunen and Loève independently proved the so-called Karhunen-Loève decomposition (written on p. 9, eq. (8)) but they were not aware of the possible applications of FPCA at the time. FPCA was later developed by Kleffe (1973); Deville (1974). Note also that at the same time as Kari Karhunen and Michel Loève, an orthogonal series decomposition was also proved by Kosambi (1943).
- **p. 2, l. 26:** The sentence "*the principal components obtained from a PCA run on the rows data matrix are the same as the ones obtained from a PCA run on the columns of the matrix*" is not precise: the eigenvalues are the same but the eigenvectors are different (even if there is a relationship between them).
- **p. 4, l. -4:** The estimation of $\widehat{C}_{p,p}(s_p, s_p)$ depends on the noise variance σ_p^2 which is unknown in practice. How do you estimate it?
- **p. 10, l. -7:** "*the time complexity [...] of the univariate score is* $\mathcal{O}(NM_pK_p)$". It is not written very clearly but I assume that it is the time complexity of computing the scores of **all** individuals?
- **p. 15, l. 7 and l. 13:** It seems that the two sentences "*we focus solely on players who have made more than 1000 shots*" and "*We remove [...] players that have made fewer than 100 shots*" contradict each other. Or maybe I do not understand the overall meaning. Which players exactly are you removing?
- **p. 16, Figure 10:** I see no difference between the eigenfunctions of shots attempted and shots made, the eigenvalues are exactly the same. This seems very surprising to me. Do you have an explanation? Also, plotting the scores can be informative for the interpretation.

References

Jean-Claude Deville. Méthodes statistiques et numériques de l'analyse harmonique. *Annales de l'INSEE*, (15):3–101, 1974.

Jürgen Kleffe. Principal Components of Random Variables with Values in a Seperable Hilbert Space. *Mathematische Operationsforschung und Statistik*, 4(5):391–406, 1973.

DD Kosambi. Statistics in function space. *Journal of the Indian Mathematical Society*, 7:76–88, 1943.