Comment: Multivariate Functional Principal Component Analysis for Data Observed on Different (Dimensional) Domains

Steven Golovkine* Edward Gunning† Andrew J. Simpkin‡
Norma Bargary§

July 31, 2023

1 Introduction

We extend our congratulations to Happp and Greven [2018] for their outstanding article. Despite the substantial interest on dimension reduction in the (multivariate) functional data literature, we concur with the authors that existing methods are often limited to either univariate functional data or multivariate functional data defined on a common one-dimensional domain. However, research has introduced a growing presence of data defined on different dimensional domains, such as curves and surfaces, observed in diverse fields, including neuroscience, biomechanics, and other domains. In this discussion, we aim to provide commentary on the estimation of the number of principal components utilising the methodology devised by the authors. To achieve this, we conducted an extensive simulation study and subsequently propose practical guidelines for practitioners to adeptly chose the appropriate number of components for multivariate functional datasets. For ease of presentation, we use the same notation as theirs. Code to reproduce the simulation study in this discussion is available at `https://github.com/FAST-ULxNUIG/variance_mfpca`.

2 Model

Happ and Greven [2018] proposed an extension of functional principal components analysis (Ramsay and Silverman [2005]) to multivariate functional data, named multivariate functional principal components analysis (MFPCA).

We briefly present the estimation of MFPCA. For a detailled description of the estimation process, see [Happ and Greven, 2018, Section 3]. In Happ and Greven [2018], the authors first estimate the principal components for each individual feature and combine them to derive the multivariate components. So, they chose a number of components for each individual feature and then use only these ones to compute the multivariate components. Let K_p be the number of components retained for the pth feature. As the univariate components are concatenated to estimated the multivariate components, the number of multivariate components that can be estimated is $\sum_p K_p$. We however claim that only $\min_p K_p$ can only be accurately estimated. The estimation of the number of components can also be done using the percentage of variance explained. We are interested by the estimation of the eigenvalues of functional datasets. Let $\{\nu_k\}_{1 \leq k \leq K}$ be the set of true eigenvalues and $\{\widehat{\nu}_k\}_{1 \leq k \leq K}$ be the set of estimated eigenfunctions. Simulations are the same as the first setting in Happ and Greven [2018]. The accuracy of the resulting estimates $\widehat{\nu}_j$ is measured by the relative errors $\text{Err}(\widehat{\nu}_j) = (\nu_j - \widehat{\nu}_j)^2 / \nu_j^2$. The percentage of variance explained by the kth component is defined as

$$\text{PVE}_k = \frac{\nu_k}{\sum_{k=1}^{K} \nu_k}.$$

*MACSI, Department of Mathematics and Statistics, University of Limerick, Ireland steven.golovkine@ul.ie

†MACSI, Department of Mathematics and Statistics, University of Limerick, Ireland edward.gunning@ul.ie

‡School of Mathematical and Statistical Sciences, University of Galway, Ireland andrew.simpkin@nuigalway.ie

§MACSI, Department of Mathematics and Statistics, University of Limerick, Ireland norma.bargary@ul.ie

The cumulative percentage of variance explained by the first k components is given by

$$\mathrm{PVE}_{1:k} = \frac{\sum_{l=1}^{k} \nu_l}{\sum_{l'=1}^{K} \nu_{l'}}.$$

Given a certain percentage of variance explained α, the number of components needed to explain at least α% of the variance of the data is

$$\mathrm{NPC}_\alpha = \sum_{k=1}^{K} \mathbf{1}\left\{\mathrm{PVE}_{1:k} < \alpha\right\} + 1 = \min_{k=1,\dots,K} \mathrm{PVE}_{1:k} > \alpha. \tag{1}$$

3 Simulation

We perform a simulation study based on the first setting in the simulation in Happ and Greven [2018]. The data-generating process is based on a truncated version of the Karhunen-Loève decomposition. First, we generate a large orthonormal basis $\{\psi_m\}_{1 \leq k \leq M}$ of $\mathcal{L}^2(\mathcal{T})$ on an interval $\mathcal{T} = [0, T] \subset \mathbb{R}$. We fix $T_1 = 0$ and $T_{p+1} = T$ and we generate $p - 1$ cutting points $T_2, \dots, T_p$ uniformly in $\mathcal{T}$ such that $0 = T_1 < \cdots < T_p < T_{p+1} = T$. Let $s_1, \dots, s_p \in \{-1, 1\}$ be coefficients that randomly flip the eigenfunctions with probability 0.5. The univariate components of the eigenfunctions are then defined as

$$\psi_m^{(j)}(t_j) = s_j \psi_m\big|_{[T_j, T_{j+1}]}\left(\frac{t_j - T_j}{T_{j+1} - T_j}\right), \quad m = 1, \dots, M, \quad j = 1, \dots, p.$$

The notation $\psi_m\big|_{[T_j, T_{j+1}]}$ is the restriction of the function ψ_m to the set $[T_j, T_{j+1}]$. The set of multivariate functions $\{\psi_m\}_{1 \leq m \leq M}$ is an orthonormal system in $\mathcal{H} := \mathcal{L}^2(\mathcal{T}_1) \times \cdots \times \mathcal{L}^2(\mathcal{T}_p)$ with $\mathcal{T}_j = [0, 1]$. Each curve is then simulated using the truncated multivariate Karhunen-Loève expansion:

$$x_i(\mathbf{t}) = \sum_{m=1}^{M} \rho_{i,m} \psi_m(\mathbf{t}), \quad \mathbf{t} \in \mathcal{T}, \quad i = 1, \dots, N,$$

where the scores $\rho_{i,m}$ are sampled as random normal variables with mean 0 and variance ν_m. The eigenvalues ν_m are defined with an exponential decrease, $\nu_m = \exp(-(m+1)/2)$. We simulate, for each replication of the simulation, $N = 25, 50$ and 100 observations. Similarly, each component is sampled on a regular grid of $S = 25, 50$ and 100 sampling points. We use $p = 5$ features and we set $M = 50$. The estimation is done using the R package `MFPCA` [Happ-Kurz, 2020]. For each univariate feature j, we estimate $M_j = 5$ principal components. Then, following the multivariate components estimation procedure, we can estimate $\sum_{j=1}^{p} M_j = 25$ multivariate components.

The outcomes concerning the errors in estimating the eigenvalues are displayed in Figure 1. Notably, the accuracy of the estimation diminishes with an increasing number of components in all scenarios. While there is a distinct jump in accuracy observed for the last five estimated eigenvalues, establishing a general rule proves challenging. It is worth emphasizing that the number of multivariate eigencomponents to be estimated is $\min_p K_p$; otherwise, the univariate components lack sufficient information to accurately recover their corresponding multivariate counterparts. Figure 2 showcases the estimation of the number of components retained across 500 simulation scenarios for a fixed percentage of variance explained. The red dots represent the number of components that should be retained for α% of variance explained, considering an exponential decay of the eigenvalues as defined in equation (1). Additionally, the size of the black dots indicates the frequency of selection for each component over the 500 simulations. Notably, the number of components appears to be consistently underestimated for various combinations of the number of observations N, number of sampling points S, and desired percentage of variance explained α%. These findings may hold considerable significance for practitioners.

4 Conclusion

Happ and Greven [2018] presents a general methodology to estimate principal components for a set of multivariate functional data defined on, possibly, different dimensional domains. Their approach, based on the decomposition of the covariance of each univariate feature, allows easy inference of the components.

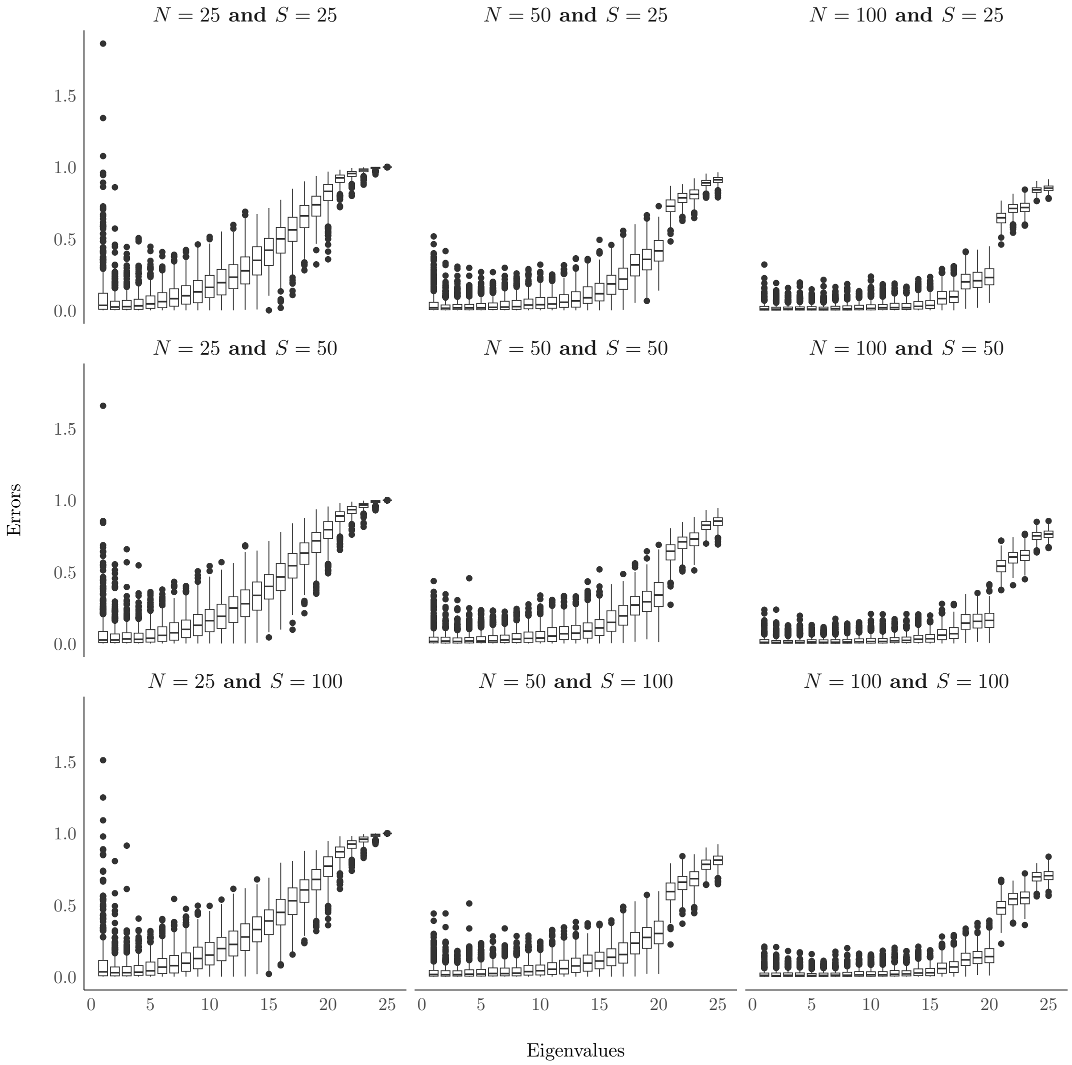

Figure 1: Boxplots of the estimation errors of the eigenvalues. We estimated 5 components for each univariate feature. The number of multivariate eigencomponents that can be estimated is thus 25. N is the number of observations, S is the number of sampling points per curve. We run 500 simulations.

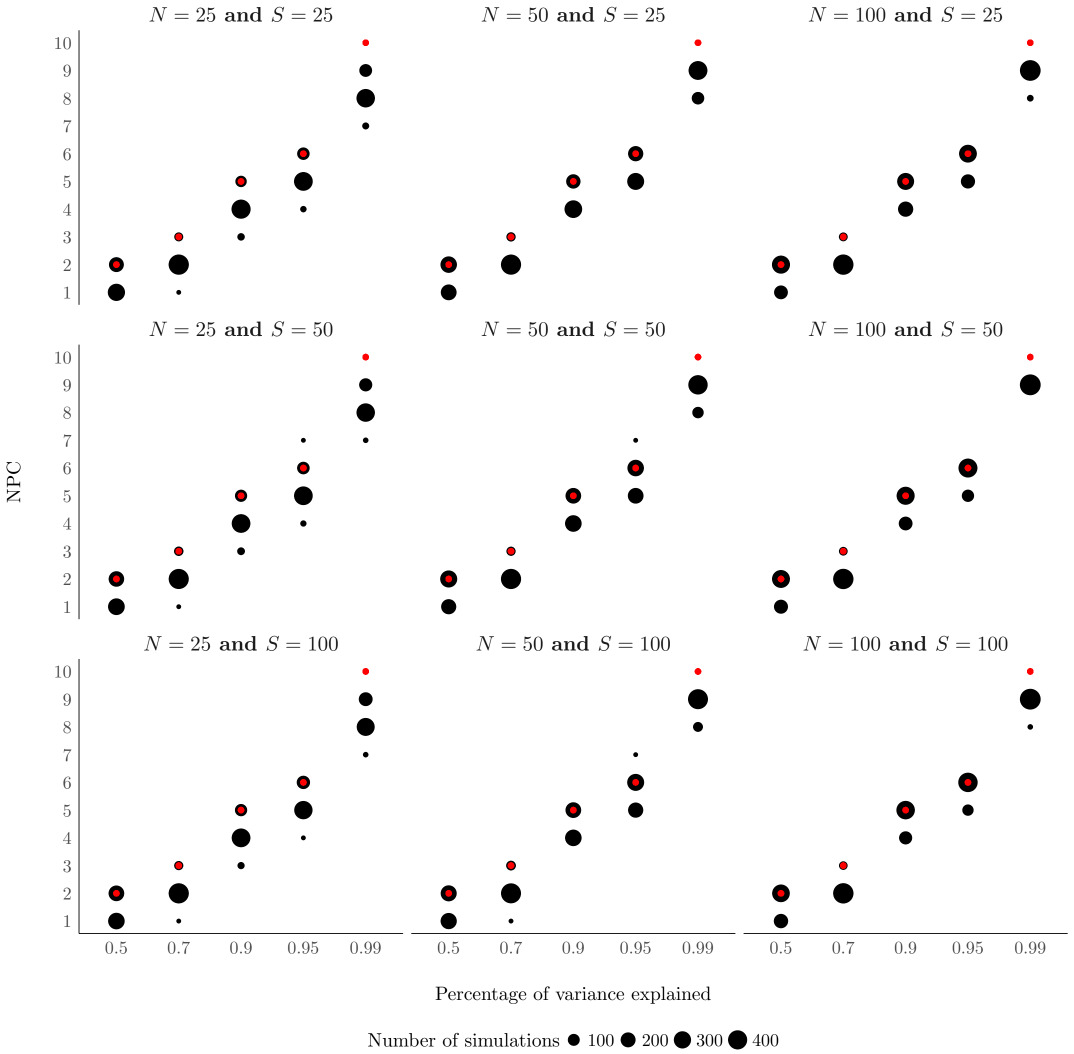

Figure 2: The size of the black dots represents the number of times the number of components has been selected over 500 simulations. The red dots are the true number of components given the percentage of variance explained. N is the number of observations, S is the number of sampling points per curve.

We conducted a simulation study, and the obtained results highlight two important findings. Firstly, although utilizing only a few univariate components may yield a substantial number of multivariate components, their accuracy is notably limited. Secondly, relying on the percentage of variance explained as a criterion for selecting the number of components may result in an underestimation of this number. We, therefore, advise practitioners to exercise caution when determining the number of estimated components. It is prudent to refrain from utilizing more than $\min_j M_j$ estimated multivariate components. Additionally, we strongly recommend conducting simulations that closely resemble the characteristics of the actual data to select the appropriate number of components based on the percentage of variance explained criterion.

Acknowledgment

S. Golovkine, A. J. Simpkin and N. Bargary are partially supported by Science Foundation Ireland under Grant No. 19/FFP/7002 and co-funded under the European Regional Development Fund. E. Gunning is supported in part Science Foundation Ireland (Grant No. 18/CRT/6049) and co-funded under the European Regional Development Fund. The authors also wish to acknowledge the Irish Centre for High-End Computing (ICHEC) for the provision of computational facilities and support.

References

Clara Happ and Sonja Greven. Multivariate Functional Principal Component Analysis for Data Observed on Different (Dimensional) Domains. *Journal of the American Statistical Association*, 113(522):649–659, April 2018. ISSN 0162-1459. doi: 10.1080/01621459.2016.1273115.

Clara Happ-Kurz. Object-Oriented Software for Functional Data. *Journal of Statistical Software*, 93:1–38, April 2020. ISSN 1548-7660. doi: 10.18637/jss.v093.i05.

J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer Series in Statistics. Springer, New York, NY, 2005. ISBN 978-0-387-40080-8 978-0-387-22751-1. doi: 10.1007/b98888.