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

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

August 4, 2023

1 Introduction

We extend our congratulations to Happ and Greven [2018] for their outstanding article. Despite the substantial interest on the topic of 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, recent research has shown 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 (FPCA, Ramsay and Silverman [2005]) to multivariate functional data defined on different dimensional domains, named multivariate functional principal components analysis (MFPCA). We briefly present the estimation procedure of the principal components given a sample $x_1, \dots, x_N$ of multivariate functional data. The detailed estimation procedure is given in [Happ and Greven, 2018, Section 3] For all $n = 1, \dots, N$, the observation x_n is a vector of p functions. The first step is to perform a univariate FPCA for each individual feature j using $x_1^{(j)}, \dots, x_N^{(j)}$. We estimate M_j components for each feature j. The total number of components that have been estimated is thus $M_+ = \sum_{j=1}^{p} M_j$. We also define $M_- = \min_{j=1,\dots,p} M_j$, being the minimum number of univariate components estimated across univariate feature j. Using the univariate eigenfunctions, an estimation of the univariate scores is performed, defined as the projection of the curves onto the eigenfunctions. The univariate scores are then concatenated in a matrix of size $N \times M_+$. An eigenanalysis of this matrix is performed resulting in eigenvalues ν_m and eigenvectors $\mathbf{c}_m$. Finally, the multivariate eigenfunctions and scores are estimated as a linear combination of the univariate eigenfunctions and scores weighted by the eigenvectors $\mathbf{c}_m$. The multivariate eigenvalues are the same as the matrix of the concatenated scores ν_m. In this context, our focus lies in investigating how the selection of the parameter M_j impacts the estimation of the eigenvalues ν_m.

Using this methodology, the number of multivariate eigenvalues that can be estimated is M_+. Let $\{\nu_m\}_{1 \leq m \leq M_+}$ be the set of true eigenvalues and $\{\widehat{\nu}_m\}_{1 \leq m \leq M_+}$ be the set of estimated eigenvalues. We use the relative errors $\text{Err}(\widehat{\nu}_m) = (\nu_m - \widehat{\nu}_m)^2 / \nu_m^2$ to assess the accuracy of the estimates. The authors also claim that the percentage of variance explained is suitable to find appropriate

*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

M_j [Ramsay and Silverman, 2005, Chapter 8.2]. The percentage of variance explained by the mth component and the cumulative percentage of variance explained by the first m components are defined as

$$\text{PVE}_m = \nu_m \times \left(\sum_{l=1}^{M_+} \nu_l \right)^{-1}, \quad \text{and} \quad \text{PVE}_{1:m} = \sum_{l=1}^{m} \text{PVE}_l, \quad m = 1, \ldots, M_+.$$

If we fix the percentage of variance explained to be $\alpha\%$, the number of components needed to explain $\alpha\%$ of the variance is given by

$$\text{NPC}_\alpha = \sum_{m=1}^{M_+} \mathbf{1}\left\{\text{PVE}_{1:m} < \alpha\right\} + 1 = \min_{m=1,\ldots,M_+} \text{PVE}_{1:m} > \alpha. \tag{1}$$

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 $T_2, \ldots, T_p$ uniformly in $\mathcal{T}$ such that $0 = T_1 < \cdots < T_p < T_{p+1} = T$. Let $s_1, \ldots, s_p \in \{-1, 1\}$ be coefficients that randomly flip the eigenfunctions with probability 0.5. The univariate components of the eigenfunctions are then defined as

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

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:

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

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 $N = 25, 50$ and 100 observations for each replication of the simulation. 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 principal components. Then, following the multivariate components estimation procedure, we can estimate M_+ multivariate components. The simulations are replicated 500 times.

Figure 1 illustrates the obtained results regarding the errors in estimating the eigenvalues. Remarkably, the accuracy of the estimation declines with an increasing number of components in all scenarios. In particular, there is a notable jump in accuracy observed for the last five estimated eigenvalues. Nonetheless, establishing a general rule based on these observations is challenging. We recommend to estimate at most M_- multivariate components; otherwise, the univariate components may not contain enough information to effectively recover their corresponding multivariate counterparts. Figure 2 presents 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 $\alpha\%$ of variance explained ($50\%, 70\%, 90\%, 95\%$ and 99%), 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. For example, for the panel where $N = 25$ and $S = 25$, and for a proportion of variance explained $\alpha = 0.5$, for approximately 200 simulations over 500, two multivariate components were required while for around 300 simulations over 500 only one multivariate component was required to explain 50% of the variance. 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 $\alpha\%$. These findings may hold considerable significance for practitioners.

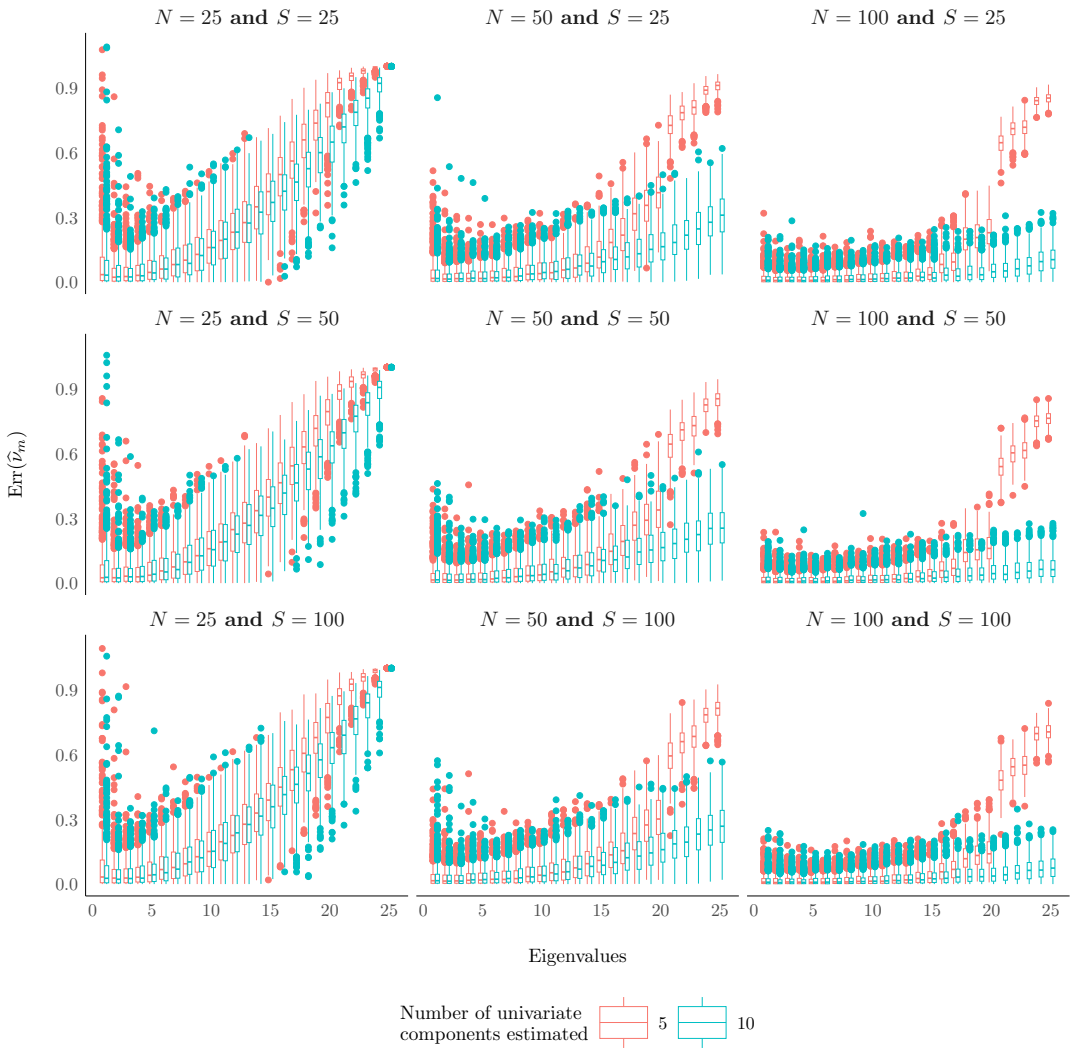

Figure 1: Boxplots of the estimation errors of the eigenvalues. We estimated 5 (red boxplots) and 10 (blue boxplots) components for each of $p = 5$ univariate feature. The number of multivariate eigencomponents that are estimated is 25. N is the number of observations, S is the number of sampling points per curve. We run 500 simulations.

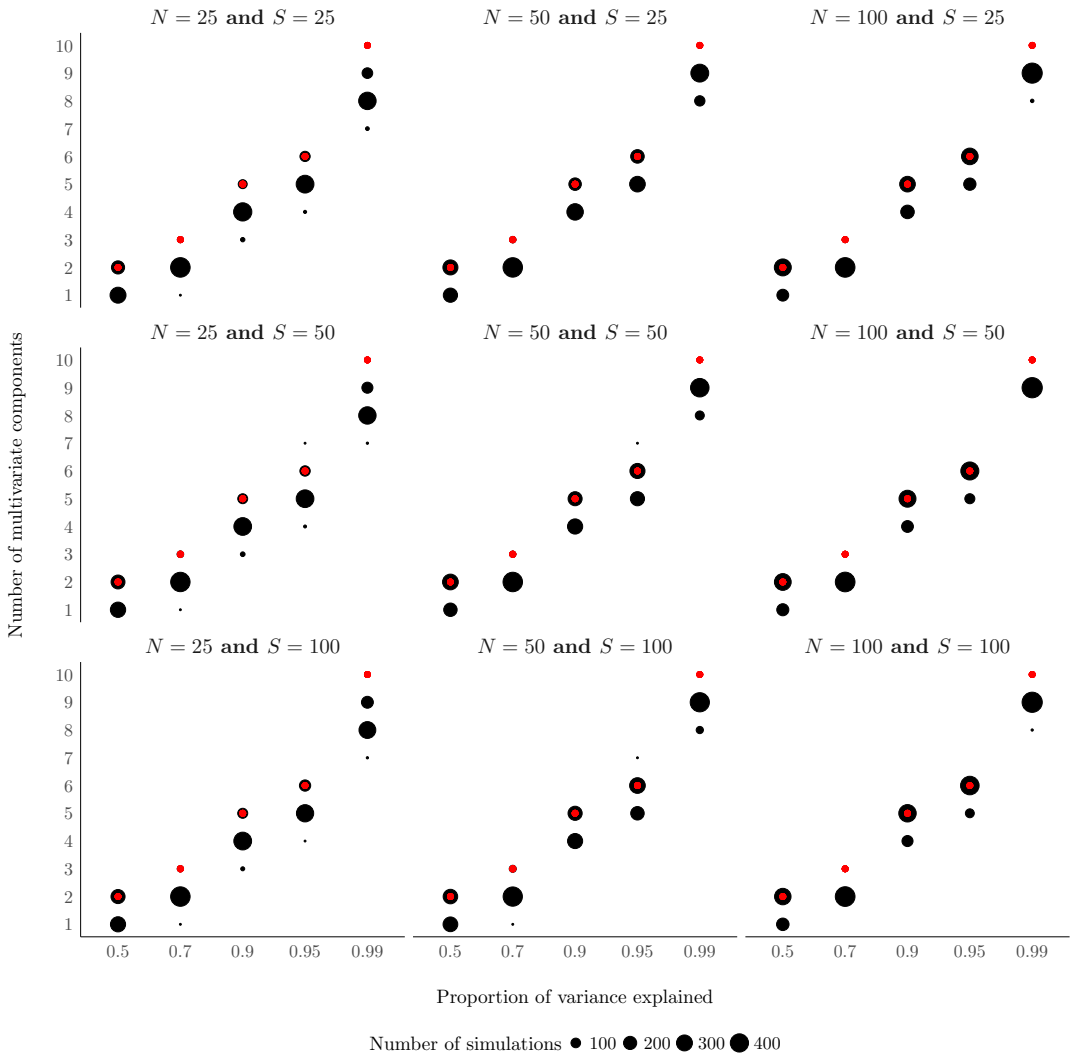

Figure 2: The size of the black dots represents the number of times the number of components has been selected over 500 simulations. The red dots are the true number of components given the percentage of variance explained. N is the number of observations, S is the number of sampling points per curve.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

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.

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 require in their analysis. It is prudent to refrain from utilizing more than M_- 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.