On the estimation of the number of components in multivariate functional principal component analysis

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

July 9, 2024

Abstract

Happ and Greven [2018] developed a methodology for principal components analysis of multivariate functional data for data observed on different dimensional domains. Their approach relies on an estimation of univariate functional principal components for each univariate functional feature. In this paper, we present extensive simulations to investigate choosing the number of principal components to retain. We show empirically that the conventional approach of using a percentage of variance explained threshold for each univariate functional feature may be unreliable when aiming to explain an overall percentage of variance in the multivariate functional data, and thus we advise practitioners to be careful when using it.

1 Introduction

Happ and Greven [2018] develop innovative theory and methodology for the dimension reduction of multivariate functional data on possibly different dimensional domains (e.g., curves and images), which extends existing methods that were limited to either univariate functional data or multivariate functional data on a common one-dimensional domain. Recent research has shown a growing presence of data defined on different dimensional domains in diverse fields such as biomechanics, e.g., Warmenhoven et al. [2019] and neuroscience, e.g., Song and Kim [2022], so we expect the work to have significant practical impact. We aim to provide commentary on the estimation of the number of principal components utilising the methodology devised in Happ and Greven [2018]. To achieve this, we conduct an extensive simulation study and subsequently propose practical guidelines for practitioners to adeptly choose the appropriate number of components for multivariate functional datasets. For ease of presentation, we use the same notation as in Happ and Greven [2018]. 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, each defined on a domain with possibly different dimensions. We denote by $x_n^{(j)}$ the jth entry of the vector x_n, referred to as the jth feature. 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 univariate functional principal components for each feature j. The total number of components that have been estimated over all p features is thus $M_+ = \sum_{j=1}^{p} M_j$. We also define $M_- = \min_{j=1,\dots,p} M_j$ to be the minimum number of univariate components estimated across all univariate features j. The univariate FPCA scores are estimated by projecting the (mean-centered) univariate functional observations onto the estimated eigenfunctions. The

*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

univariate scores from the p features 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 eigenvalues of 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 maximum 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 propose to estimate the number of multivariate components using the percentage of variance explained. For that, they first select M_j univariate components that explain $\alpha\%$ of the variance for each univariate features [Ramsay and Silverman, 2005, Chapter 8.2] and they claim that this number of components is enough to estimate the number of multivariate components that explain $\alpha\%$ of the variance in the multivariate functional data [Happ and Greven, 2018, Section 3.2]. The percentage of variance explained by the mth component and the cumulative percentage of variance explained by the first m components are defined as

$$\text{PVE}_m = 100 \times \nu_m \times \left(\sum_{l=1}^{M_+} \nu_l\right)^{-1} \quad \text{and} \quad \text{PVE}_{1:m} = \sum_{l=1}^{m} \text{PVE}_l, \quad m = 1, \ldots, M_+.$$

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

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

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 $p-1$ cut 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 $\mathcal{H} \coloneqq \mathcal{L}^2(\mathcal{T}_1) \times \cdots \times \mathcal{L}^2(\mathcal{T}_p)$ 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 Gaussian random 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 displays a comparison of the error in the estimated eigenvalues when using $M_j = 5$ and $M_j = 10$, illustrating the effect of M_j on eigenvalue estimation. 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 when $M_j = 5$, while this is not clear when $M_j = 10$. The dependence of the eigenvalue estimates on the number of univariate

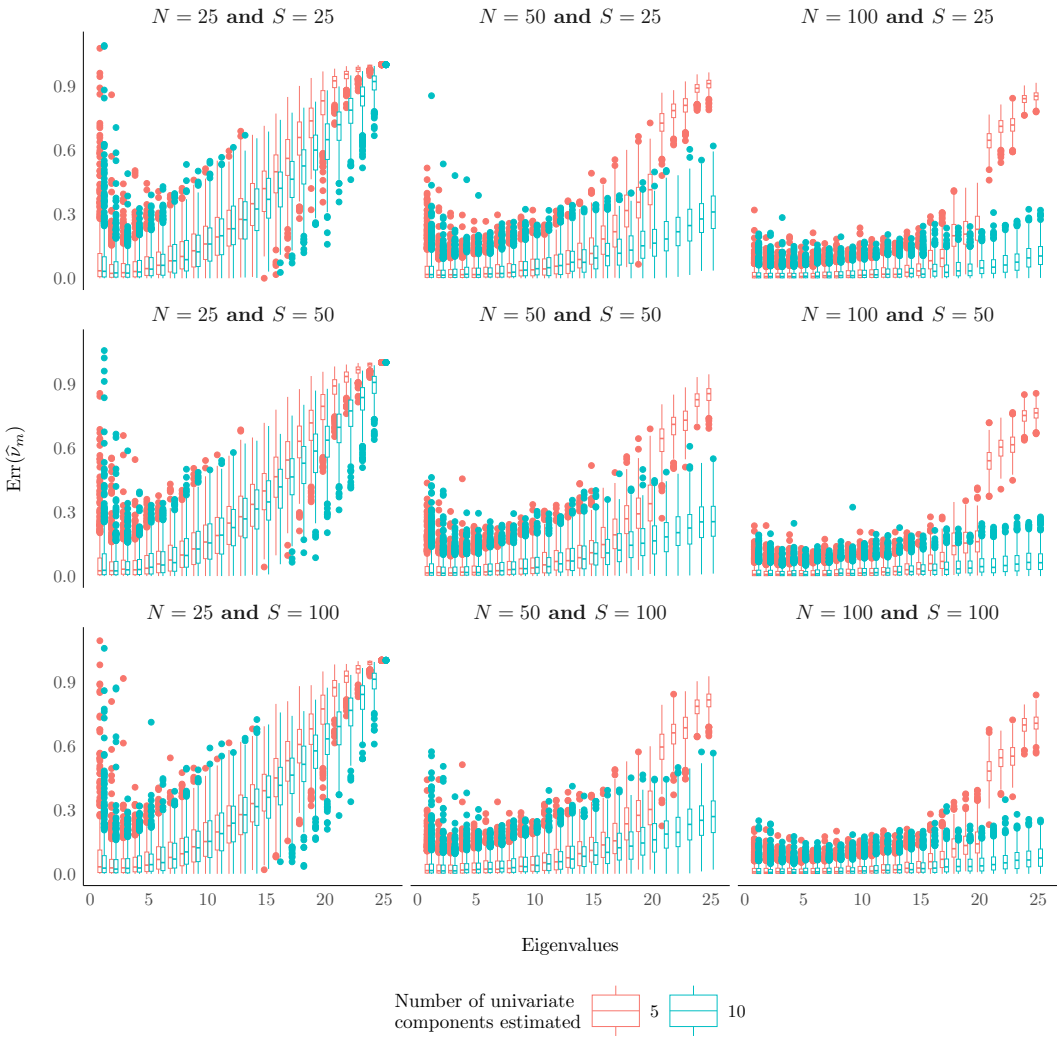


Figure 1: Boxplots of the estimation errors of the eigenvalues. We estimated 5 (red boxplots) and 10 (blue boxplots) univariate functional components for each of $p = 5$ univariate features. The number of multivariate eigencomponents that are estimated is 25. N is the number of observations, S is the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

functional principal components retained M_j is clear, but establishing a general rule based on these observations is challenging. We suggest 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 multivariate components retained across 500 simulation scenarios for a fixed percentage of variance explained. The red dots represent the number of multivariate components that would be needed to explain at least $\alpha\%$ of the variance ($50\%, 70\%, 90\%, 95\%$ and 99%), considering an exponential decay of the eigenvalues as defined in equation (1). Note that, we can compute the number of multivariate components exactly as we know the true eigenvalues. Additionally, the size of the black dots indicates the frequency of selection for each number of multivariate 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 of the 500, two multivariate components were selected and for around 300 simulations of the 500 only one multivariate component was selected to explain 50% of the variance, while the true number of multivariate components is 2 (red dot). 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 $\alpha\%$. Therefore, this simulation scenario shows that using a percentage of variance explained of level α to choose the number of univariate components M_j is not sufficient to estimate the number of multivariate functional principal components that explain $\alpha\%$ of the variance in the multivariate functional data. These findings may hold considerable significance for practitioners.

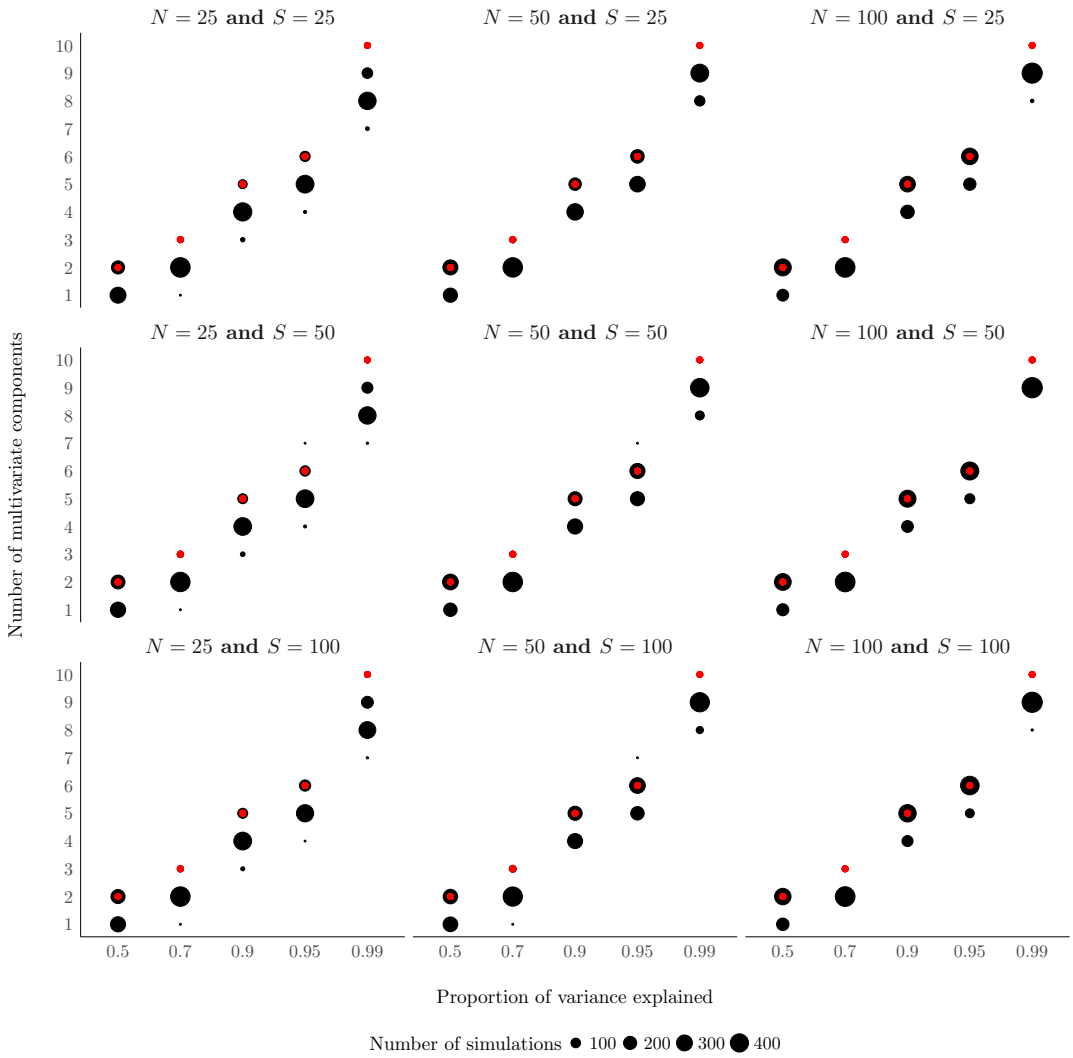
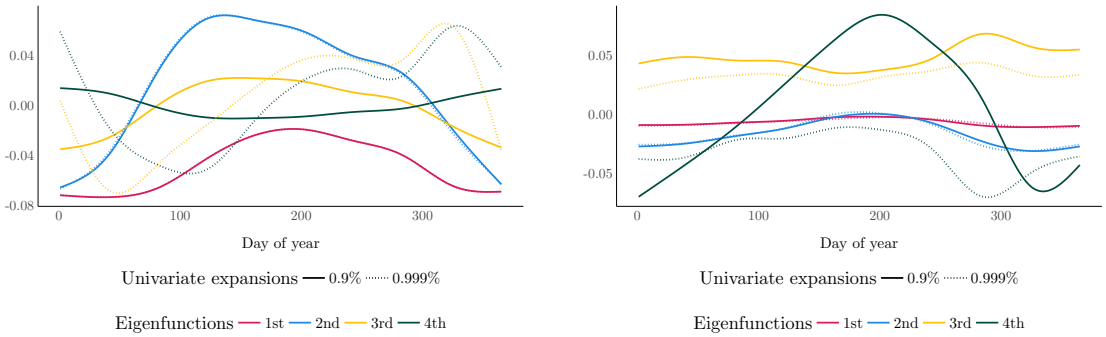


Figure 2: The size of the black dots represents the number of times the number of components has been selected over 500 simulations. The red dots are the true number of components given the percentage of variance explained. N is the number of observations, S is the number of sampling points per curve.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

Eigenvalues	1st	2nd	3rd	4th
0.9%	15844	1675	307	45
0.999%	15850	1679	438	213

Table 1: Eigenvalues



(a) Temperature

(b) Precipitation

Figure 3: The first four eigenfunctions of the Canadian weather dataset.

4 Application: Canadian weather dataset

We show how this fact may influenced on a real dataset. We consider the Canadian weather dataset, available in the `fda` R package. It contains daily measurements of temperature (in Celsius) and precipitation (in millimeters) for 35 Canadian weather stations, averaged over the years 1960 to 1994. This is an example of multivariate functional data with $p = 2$ defined on one dimensional domains.

The data are first expanded in a B-splines basis of 10 functions. We consider two cases: in the first, we estimate the univariate eigencomponents that represent 90% of the variance and in the second, we estimate the univariate eigencomponents that represent 99.9% of the variance.

Estimation of the eigenvalues 1. Figure 3 presents the estimation of the eigenfunctions for the two cases.

5 Conclusion

Happ and Greven [2018] present 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 estimation of the components.

We have 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 univariate components may result in an underestimation of the number of multivariate components. We, therefore, advise practitioners to exercise caution when determining the number of estimated components required in their analysis. We suggest to estimate at most M_- 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.

Jun Song and Kyongwon Kim. Sparse multivariate functional principal component analysis. *Stat*, 11(1):e435, 2022. ISSN 2049-1573. doi: 10.1002/sta4.435.

John Warmenhoven, Stephen Cobley, Conny Draper, Andrew Harrison, Norma Bargary, and Richard Smith. Bivariate functional principal components analysis: Considerations for use with multivariate movement signatures in sports biomechanics. *Sports Biomechanics*, 18(1):10–27, January 2019. ISSN 1476-3141. doi: 10.1080/14763141.2017.1384050.