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

Steven Golovkine*
MACSI, Department of Mathematics and Statistics,
University of Limerick, Ireland
Edward Gunning
Department of Biostatistics and Epidemiology,
University of Pennsylvania, USA
Andrew J. Simpkin
School of Mathematical and Statistical Sciences,
University of Galway, Ireland
and
Norma Bargary
MACSI, Department of Mathematics and Statistics,
University of Limerick, Ireland

July 12, 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.

Keywords: Dimension reduction; Functional data analysis; Variance

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 and data analysis 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 $x_1, \ldots, x_N$ of multivariate functional data. The detailed estimation procedure is given in Happ and Greven (2018), Section 3. For all $n = 1, \ldots, N$, 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 $x_1^{(j)}, \ldots, x_N^{(j)}$. We esti-
mate 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 $M_- = \min_{j=1,\ldots,p} M_j$ 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 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 combi-
nation 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 $\mathrm{Err}(\widehat{\nu}_m) = (\nu_m - \widehat{\nu}_m)^2/\nu_m^2$ 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

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

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 $\mathcal{H} := \mathcal{L}^2(\mathcal{T}_1) \times \cdots \times \mathcal{L}^2(\mathcal{T}_p)$ 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 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
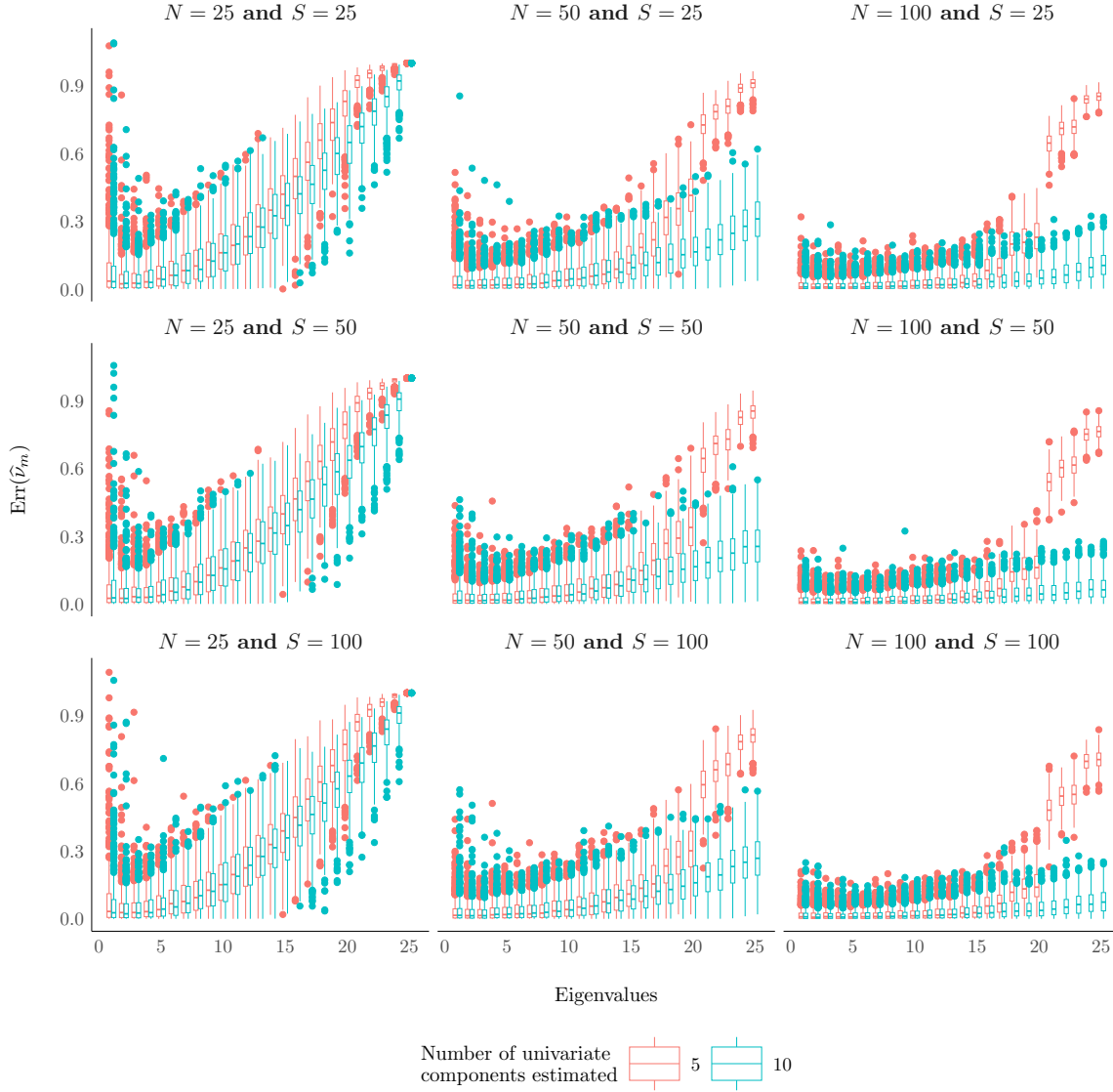
Figure 1: Boxplots of the estimation errors of the eigenvalues. We estimated 5 (red boxplots) and 10 (blue boxplots) univariate functional components for each of $p = 5$ univariate features. The number of multivariate eigencomponents that are estimated is 25. N is the number of observations, S is the number of sampling points per curve. We run 500 simulations.

in the multivariate functional data. These findings may hold considerable significance for practitioners.

4   Application: Canadian weather dataset

To illustrate our simulation results, we apply the same idea on a real dataset, the Canadian weather dataset, available in the R package `fda` (Ramsay et al., 2023). The dataset 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 domain, the temperature being the first feature and the precipitation being the second feature. We aim to estimate M multivariate eigencomponents of the data using different number of univariate eigencomponents and compare the results. For that, we define two scenarios, one where $M = M_+$ and one where $M = M_-$.

We first expand the data in a B-splines basis with 10 functions. In the first scenario, for each feature, we estimate two univariate eigencomponents ($M_1 = 2$ and $M_2 = 2$). In the second scenario, for each feature, we estimate four univariate eigencomponents ($M_1 = 4$ and $M_2 = 4$). In both scenarios, we then estimate $M = 4$ multivariate eigencomponents. So, for the first scenario, $M = M_+ = 4$ and for the second scenario, $M = M_- = 4$. Based on the simulation and Figure 1, we expect the first two multivariate eigencomponents to be similar and the other two to be unequal.

Table 1 presents the estimation of the eigenvalues for the Canadian weather dataset for both scenarios. We notice that the values are similar for the first two eigenvalues, but quite different for the other two.

Figure 3 presents the estimation of the eigenfunctions for the Canadian weather dataset for both scenarios. As with the eigenvalues, we notice that the first two (multivariate) eigenfunctions are approximately the same, but the other two are not equal. While we do
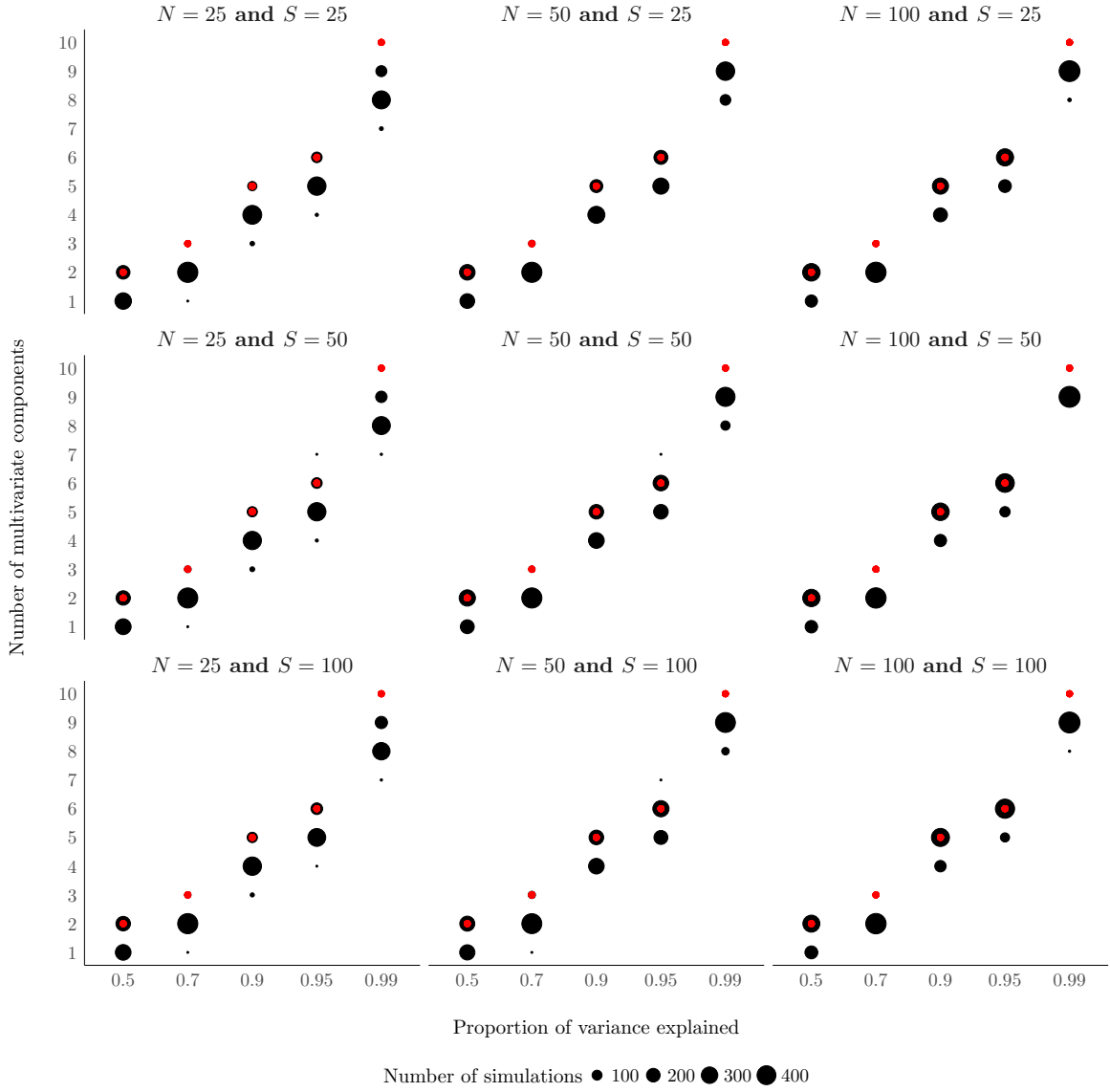
Figure 2: The size of the black dots represents the number of times the number of components has been selected over 500 simulations. The red dots are the true number of components given the percentage of variance explained. N is the number of observations, S is the number of sampling points per curve. The size of the black dots is continuous.

| | | Eigenvalues | | | |
Scenario	Univariate expansions	1st	2nd	3rd	4th
1	2 components	15845	1675	308	45
2	4 components	15850	1679	438	213

Table 1: Estimation of the first four eigenvalues of the Canadian weather dataset using two and four univariate components for the univariate expansions.

not know the true eigenfunctions, based on simulation results, the estimated multivariate eigenfunctions from the second scenario should be closer to the truth than the estimation from the first scenario. We may explain this result as we have not estimated enough information in the univariate decomposition in the first scenario to effectively estimate the third and fourth multivariate eigencomponents. We reiterate the suggestion to estimate at most M_- multivariate eigencomponents (which will be $M_- = 2$ for the first scenario).
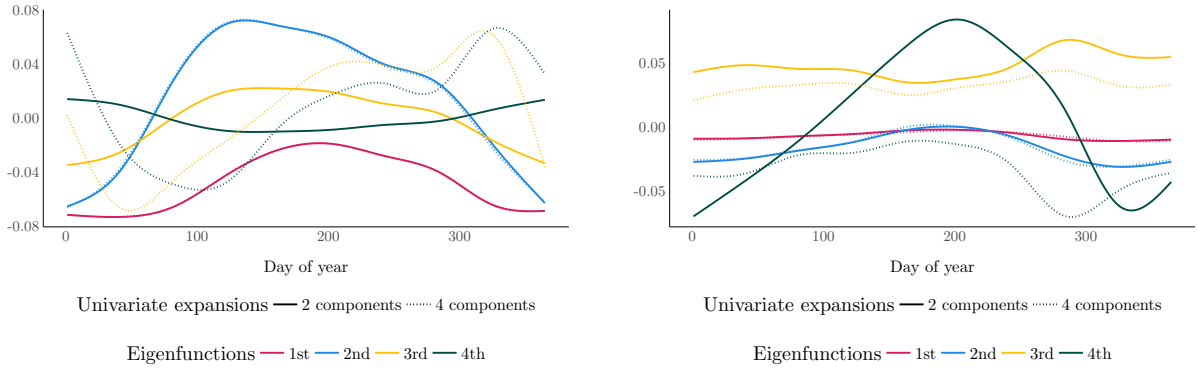
(a) Temperature (first component)   (b) Precipitation (second component)

Figure 3: Estimation of the first four eigenfunctions of the Canadian weather dataset using using two and four univariate components for the univariate expansions.

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 an example on a real dataset, 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.

References

Happ, C. and S. Greven (2018, April). Multivariate Functional Principal Component Analysis for Data Observed on Different (Dimensional) Domains. *Journal of the American Statistical Association 113*(522), 649–659.

Happ-Kurz, C. (2020, April). Object-Oriented Software for Functional Data. *Journal of Statistical Software 93*, 1–38.

Ramsay, J., G. Hooker, and S. Graves (2023, May). Fda: Functional Data Analysis.

Ramsay, J. O. and B. W. Silverman (2005). *Functional Data Analysis*. Springer Series in Statistics. New York, NY: Springer.

Song, J. and K. Kim (2022). Sparse multivariate functional principal component analysis. *Stat 11*(1), e435.

Warmenhoven, J., S. Cobley, C. Draper, A. Harrison, N. Bargary, and R. Smith (2019, January). Bivariate functional principal components analysis: Considerations for use with multivariate movement signatures in sports biomechanics. *Sports Biomechanics 18*(1), 10–27.