Estimation of the number of components in multivariate functional data analysis: A comment on Happ and Greven

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

July 25, 2023

Abstract

A methodology for multivariate functional data analysis for data observed on different dimensional domains has been recently published Happ and Greven [2018]. It relies on an estimation of principal components for each univariate feature. The authors claim that the number of components can be estimated using ... We proposed to extend the sensitivity analysis in their Supplementary Material. The estimated number of components may not be reliable, and thus we advise practitioners to be careful when chosing the number of components.

In recent years, the analysis of multivariate functional data has become a popular method with applications in several fields. Functional principal component analysis (FPCA) is an extension of principal components analysis to functional data. FPCA has become a prevalent tool in FDA due to its abililty to convert infinite-dimensional functional data into finite-dimensional vectors of random scores. Multivariate functional principal components analysis (MFPCA) is the extension of FPCA to multivariate functional data. It allows to identify and visualize the main sources of variation in the data.

We discuss the estimation of the number of components method in the recently published paper titled "Multivariate Functional Principal Component Analysis for Data Observed on Different (Dimensional) Domains" by Happ and Greven [2018]. For ease of presentation, we use the same notation as theirs.

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 $\mathrm{Err}(\widehat{\nu}_j) = (\nu_j - \widehat{\nu}_j)^2/\nu_j^2$. The percentage of variance explained by the kth component is defined as

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

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

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

*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

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

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

The simulation setting is based on 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 $\{\psi_k\}_{1\leq k\leq K}$ 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

$$\phi_k^{(p)}(t_p) = s_p \psi_k\big|_{[T_p, T_{p+1}]}\left(\frac{t_p - T_p}{T_{p+1} - T_p}\right), \quad k = 1, \dots, K, \quad p = 1, \dots, P.$$

The notation $\phi_k\big|_{[T_p, T_{p+1}]}$ is the restriction of the function ϕ_k to the set $[T_p, T_{p+1}]$. The set of multivariate functions $\{\psi_k\}_{1\leq k\leq K}$ is an orthonormal system in $\mathcal{H} := \mathcal{L}^2(\mathcal{T}_1) \times \cdots \times \mathcal{L}^2(\mathcal{T}_P)$ with $\mathcal{T}_p = [0, 1]$. Each curve is then simulated using the truncated multivariate Karhunen-Loève expansion:

$$X(\mathbf{t}) = \sum_{k=1}^{K} \rho_k \phi_k(\mathbf{t}), \quad \mathbf{t} \in \mathcal{T},$$

where the scores ρ_k are sampled as random normal variables with mean 0 and variance ν_k. The eigenvalues ν_k are defined with an exponential decrease, $\nu_k = \exp(-(k+1)/2)$. We simulate, for each replication of the simulation, $N = 25, 50$ and 100 observations. Similarly, each component is sampled on a regular grid of $M = 25, 50$ and 100 sampling points. We compare the methods for $P = 5$ features and we set $K = 50$. The estimation are done using the R package `MFPCA` [Happ-Kurz].

For each univariate feature p, we estimate $K_p = 5$ principal components. Then, following the multivariate components estimation procedure, we can estimate $\sum_{p=1}^{P} K_p = 25$ multivariate components. The results of the errors of the estimation of the eigenvalues are presented in Figure 1. We remark that the accuracy of the estimation decreases with the number of components in every scenario. While there is a clear jump in the accuracy for the last five estimated eigenvalues, it seems difficult to set a general rule. ... The number of multivariate eigencomponents that should be estimated is $\min_p K_p$, otherwise the univariate components do not contain enough information to recover accurately their multivariate counterpart.

Codes to reproduce the simulations are available at https://github.com/FAST-ULxNUIG/variance_mfpca.

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.

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, 2018. ISSN 0162-1459. doi: 10.1080/01621459.2016.1273115. URL https://doi.org/10.1080/01621459.2016.1273115.

Clara Happ-Kurz. Object-Oriented Software for Functional Data. 93(1):1–38. ISSN 1548-7660. doi: 10.18637/jss.v093.i05. URL https://www.jstatsoft.org/index.php/jss/article/view/v093i05.

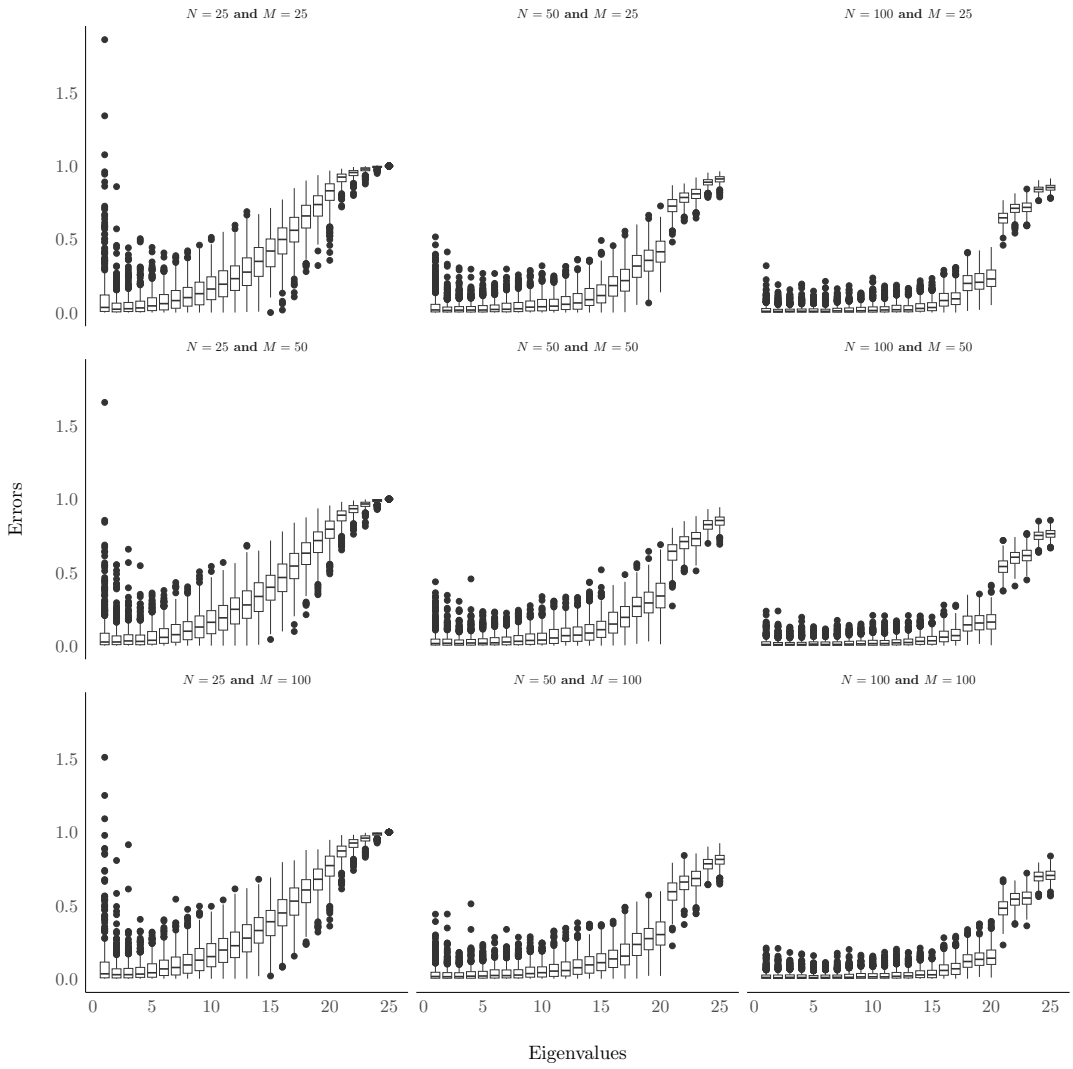

Figure 1: Boxplots of the estimation errors of the eigenvalues. We estimated 5 components for each univariate feature. The number of multivariate eigencomponents that can be estimated is thus 25. N is the number of observations, M is the number of sampling points per curve. We run 500 simulations.

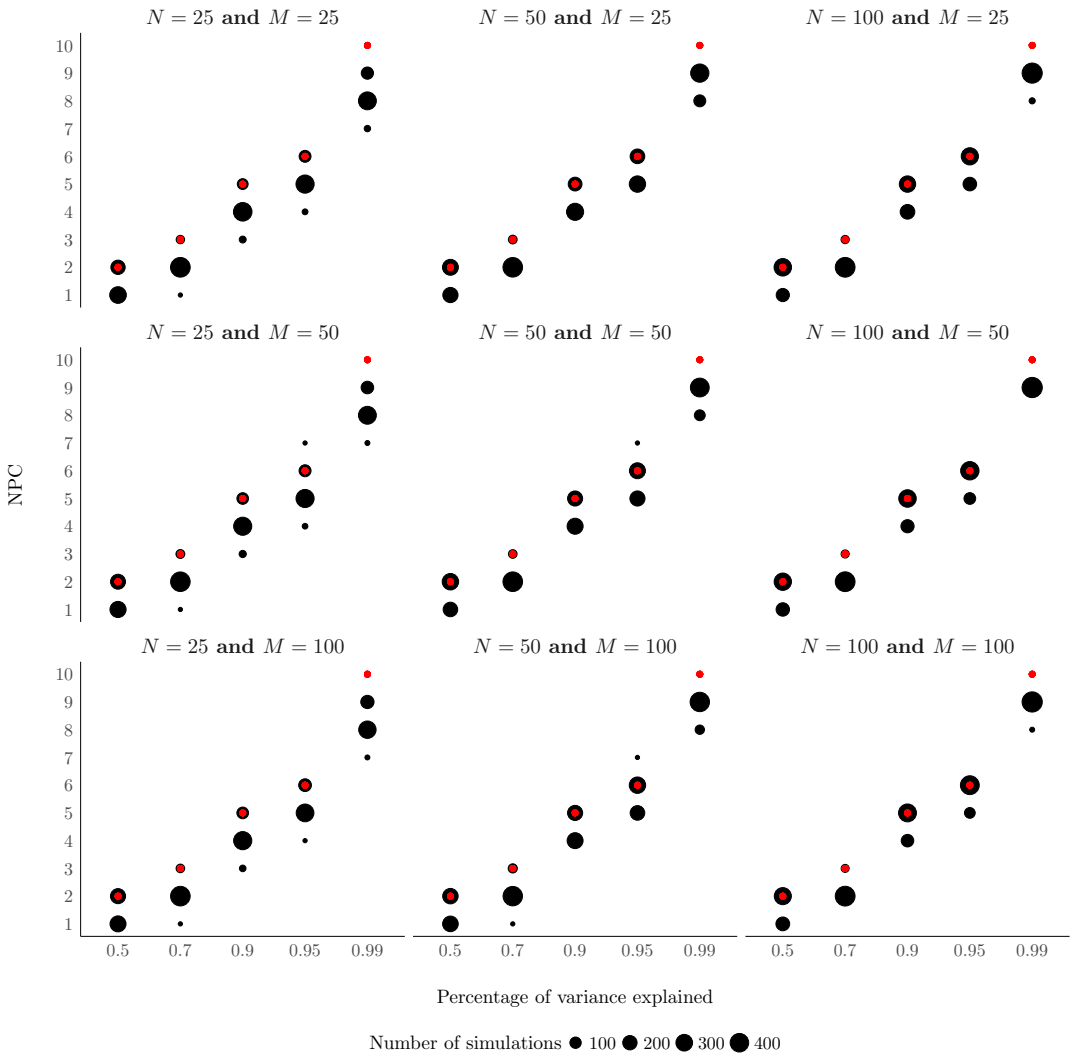

Figure 2: The size of the black dots represents the number of times the number of components has been selected over 500 simulations. The red dots are the true number of components given the percentage of variance explained. N is the number of observations, M is the number of sampling points per curve.

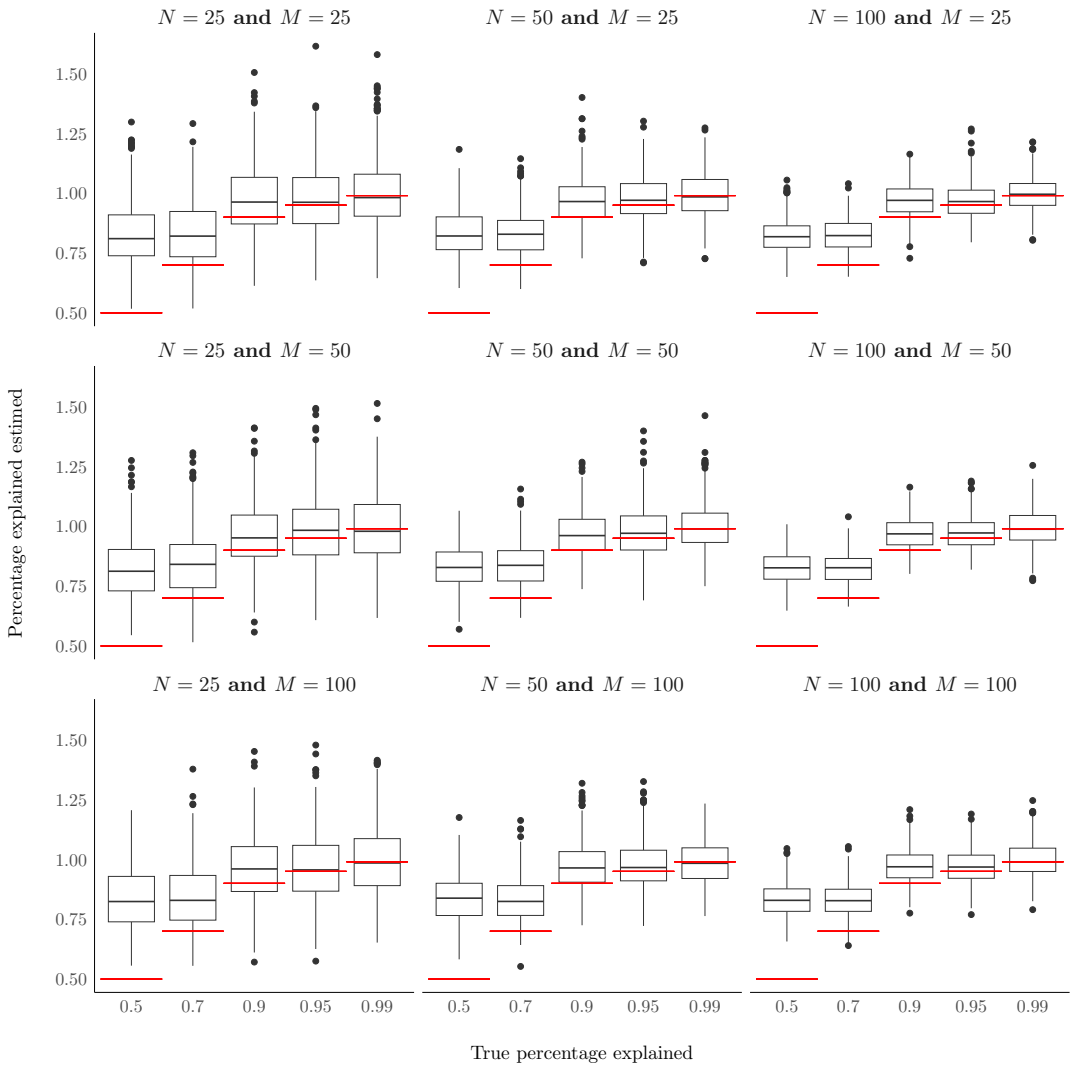

Figure 3: Estimation of the percentage of variance explained over 500 simulations. The red lines are the true percentage of variance explained. N is the number of observations, M is the number of sampling points per curve. The estimation can be larger than 1 has the estimated cumulative eigenvalues are compare to the true cumulative eigenvalues. We run 500 simulations.