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 24, 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 visalize 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].

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. Simulations are the same as the first setting in Happ and Greven [2018]. The accuracy of the resulting estimates $\widehat{\lambda}_j$ is measured by the relative errors $\mathrm{Err}(\widehat{\lambda}_j) = (\lambda_j - \widehat{\lambda}_j)^2/\lambda_j^2$.

1 Introduction

We aim to show that the procedure proposed by Happ and Greven [2018] may leaad to inconsistency in the retained number of components, based on extensive simulation.

The simulation may varies has follow:

- Number of curves $N = 25, 50, 100, 200$
- Number of sampling points $M = 25, 50, 100, 200$
- Number of components $P = 2, 5, 10, 20, 50$

*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

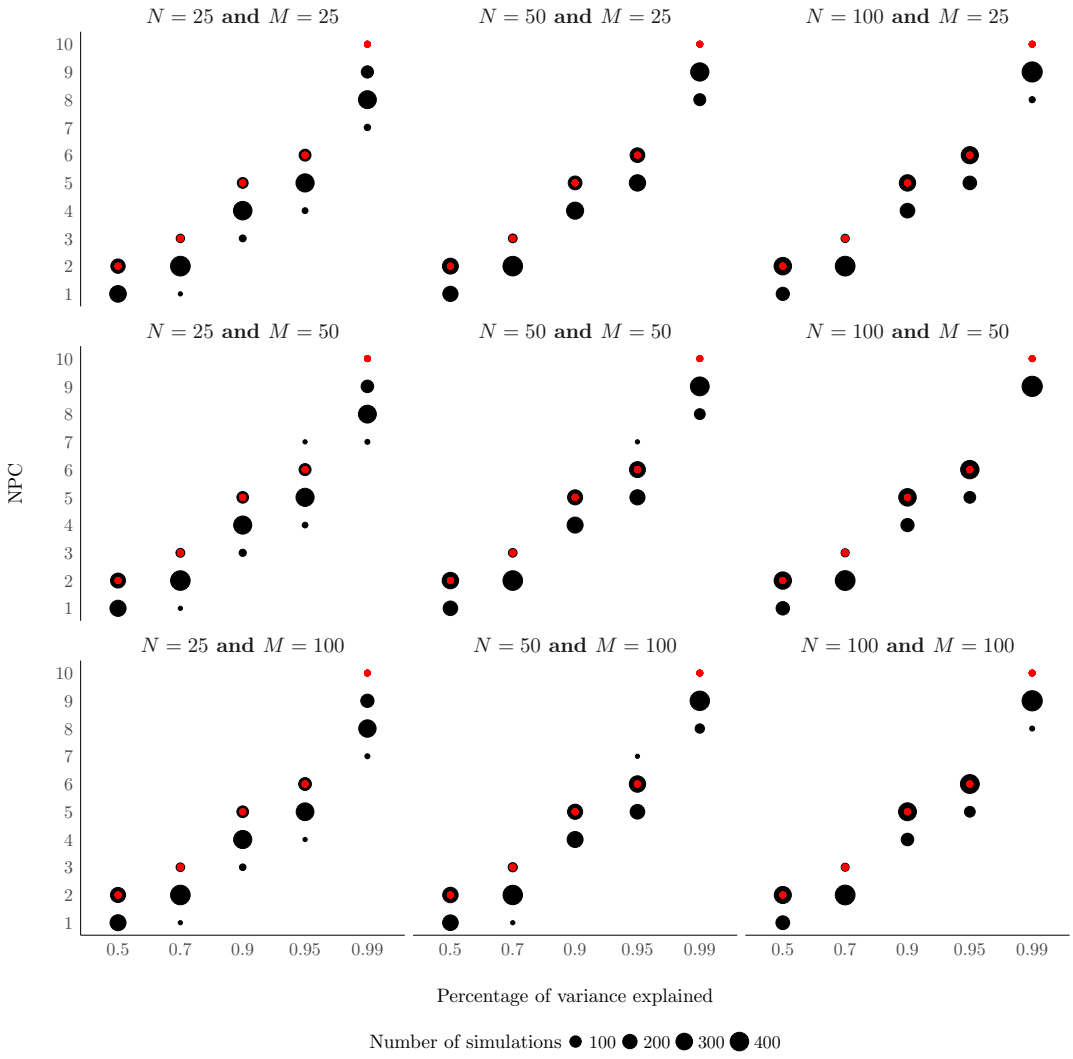

Figure 1: N is the number of observations, M is the number of sampling points per curve.

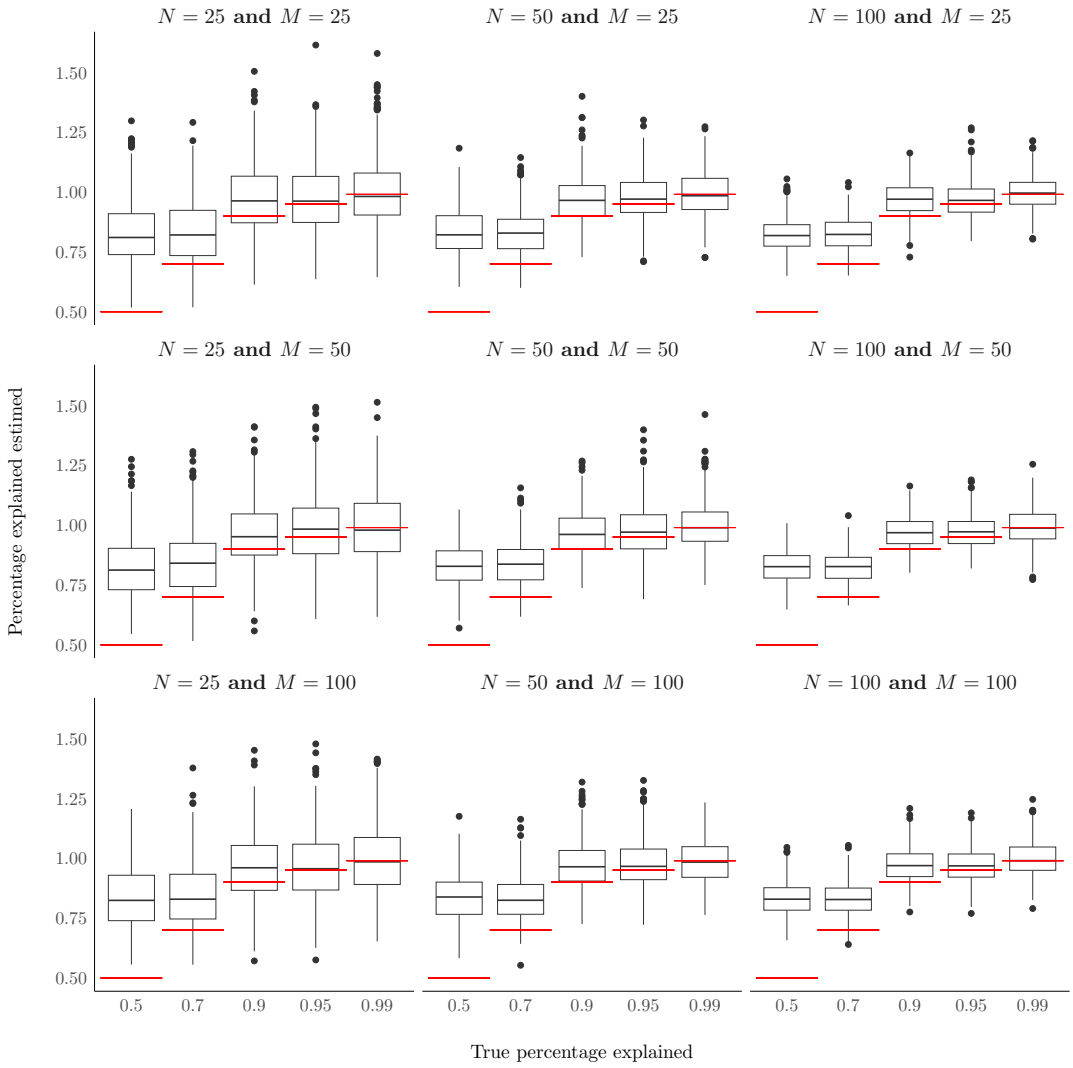

Figure 2: N is the number of observations, M is the number of sampling points per curve.

- No noise and we assume the curve are sampled on a common grid.
- Based on the Karhunen-Loève decomposition, make sure that the decreasing of the eigenvalues is coherent with KL assumptions. The data are defined with a large number of components and different decreasing of the eigenvalues scenarios.
- We do the same for the percentage of variance explained. We set the percentage of variance explained for the multivariate components to be $\alpha\%$ ($\alpha = 50, 75, 90, 95, 99$) and we change the percentage of variance explained by the univariate components.
- The quality of the estimation is based on different measures: the number of retained components (only if we set the percentage of variance explained), the estimation of the eigenvalues ($\log -\text{AE}$), the estimation of the multivariate eigenfunctions (ISE) and the reconstruction of the curves (MISE).
- Data are simulated with Happ and Greven [2018] setting and we can use ICHEC to run them.

2 Ideas

- This should be a quick paper on the selection of the number of components for MFPCA.
- Based on how the number of components is selected in MFPCA Happ and Greven [2018].
- Just considering the number of components, based on an univariate expansion, we speculate that we need for that say K univariate components to effectively estimate K multivariate components. Let K_p be the number of estimated components for the pth feature and K the number of multivariate components we want to estimate. Computationally speaking, we can estimate up to $\sum_{p=1}^{P} K_p$ multivariate components. We however claim that the number of accurately estimated components is only $\min_{p=1,\dots,P} K_p$.
- The same phenomenon appears with the percentage of variance explained, we can not retrieve $\alpha\%$ of the variance with the multivariate curves, if the univariate components also explained $\alpha\%$ of the univariate curves. This might be related to the "multivariate testing" phenomena. Mentioned in the Supplementary material of Happ and Greven [2018]. We rerun the sensitivity analysis and aim to show that the percentage of variance explained might be different of one expects.
- This is going to be a practical paper, no proof, except pratical proofs will be presented here. Only work on extensive simulations.

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.