



Multivariate functional principal component analysis for endogenously stratified data

Christelle Judith Agonkoui^(1,*), Freedath Djibril Moussa⁽²⁾ and Sophie Dabo-Niang⁽³⁾

¹Institute of Mathematics and Physics Sciences, University of Abomey Calavi, Benin

²Faculté des Sciences et Techniques, University of Abomey Calavi, Benin

³Laboratoire Paul Painlevé UMR CNRS 8524, INRIA-MODAL, University of Lille, France

Received on May 12, 2023; Accepted on August 1, 2023; Published on September 30, 2023

Copyright © 2022, Afrika Statistika and The Statistics and Probability African Society (SPAS). All rights reserved

Abstract. . We address the problem of performing dimension reduction on multivariate functional data observed on different domains in an endogenously stratified sampling context. The aim is to propose a new multivariate functional principal component analysis (MFPCA) approach for data sampled by a stratification of a population according to a binary variable of interest. This estimation strategy is derived from a direct relationship between univariate and multivariate FPCA for finite Karhunen-Loève decompositions. The proposed methodology yields encouraging results and can be applied to data with measurement errors. Computational results on simulated data highlight the good performance of the proposed methodology compared to the classical *MFPCA*, which ignores the type of data sampling. A real-life application considering breast cancer cells data is also presented.

Key words: choice-based sampling; case-control; multivariate functional data; multivariate functional principal components.

AMS 2010 Mathematics Subject Classification Objects : 62D05; 62H99; 62P10.

*Corresponding author Christelle Judith Agonkoui : christelle.agonkoui@imsp-uac.org

Freedath Djibril Moussa : freedath.djibril@imsp-uac.org

Sophie Dabo-Niang : sophie.dabo@univ-lille.fr

Résumé (Abstract in French) Nous abordons le problème de la réduction de dimension des données fonctionnelles multivariées observées sur différents domaines dans un contexte d'échantillonnage endogène stratifié. Nous proposons une nouvelle approche d'analyse en composantes principales fonctionnelles multivariées (ACPFM) pour des données échantillonnées par une stratification de la population suivant une variable d'intérêt binaire. Cette stratégie d'estimation est obtenue à partir d'une relation entre les ACP fonctionnelles univariée et multivariée pour les décompositions finies de Karhunen-Loève. Notre méthodologie donne des résultats encourageants et peut être appliquée à des données comportant des erreurs de mesure. Des simulations mettent en évidence les bonnes performances de la méthodologie proposée par rapport à l'ACPFM classique, qui ignore la nature d'échantillonnage des données. Nous présentons également une application en considérant des données réelles portant sur des cellules cancéreuses du sein.

The authors.

Christelle Judith Agonkoui, M.Sc, is preparing a PhD Thesis at the Institute of Mathematics and Physics Sciences, University of Abomey Calavi, Benin, under the supervision third author.

Freedath Djibril Moussa, Ph.D., is an Associate professor at :Faculté des Sciences et Techniques, University of Abomey Calavi, Benin.

Sophie Dabo-Niang, Ph.D., is a full professor of Statistics at: Laboratoire Paul Painlevé UMR CNRS 8524, INRIA-MODAL, University of Lille, France.

1. Introduction

Nowadays, due to the proliferation of information systems and technologies, a significant amount of large-scale data is collected in several domains like epidemiology, health and economics. In these areas, case-control or choice-based studies (also known as response-based studies) are very frequent.

Typically, the considered sampling scheme is the result of an endogenous stratification where each stratum is defined according to the individual responses, described by the discrete values of a response variable. Such design is usual in several situations where it often happens that one or more outcomes occur infrequently in the population but are important for determining some key parameters. By stratifying the population with respect to the responses, one can gather information on those infrequent outcomes at a much lower cost than would be incurred by simply increasing the size of a random sample. A main motivation for this sampling scheme is the possibility to improve the accuracy of the analysis at reduced survey costs.

Case-control approach is then a powerful method for investigating factors that may explain a particular event. It is extensively used in biomedical studies when one of the treatments being compared is uncommon and in epidemiology to study the incidence of a disease, one of the best-known examples being the investigation by Doll and Hill (1950) of the possible connection between cigarette consumption and lung cancer.

A case-control study is a retrospective study, an alternative to a prospective study. Cases (diseased individuals) are identified in an underlying population and a comparable control (individuals without the disease) group is sampled.

Case-control approach is also used in the econometrics literature where it is usually named choice-based or endogenous sampling. In econometric analysis, the variable of interest comes from an individual decision, while in epidemiology or a health research having a disease is not a choice, but the consequence of exposure to one or more risk factors. The extensive literature on choice-based sampling includes the papers of Manski and Lerman (1977), Manski and McFadden (1981), Cosslett (1981), Cosslett (1981), Imbens (1992), and Ramalho (2002). Endogenous stratification has been studied, reviewed and discussed in Didelez *et al.*(2010), Langholz (2010), Hynes and Greene (2013), and Abadie *et al.*(2018), to name a few. Interest in choice-based or case-control studies continues to grow (Martínez *et al.*(2019), Yari *et al.*(2021)) with technological advances and increasing access to large patient databases of various types and sizes to support health research. This is for example the case in the joint study of several risk factors for certain diseases where the response is often binary (having the disease or not).

When these factors evolve in time or space and can be seen as curves, images, one may analyze them in a joint and exhaustive way using the theory of multivariate functional data analysis. Functional data analysis (FDA) is a very active research field (see the monographs of Ramsay and Silverman (2005), Ferraty and Vieu (2006), Wang *et al.*(2016)) that requires a step of transformation of raw data into functional objects and projection to reduce the infinite dimension of the space where the functional data evolved. A fundamental tool for this dimension reduction step is Functional principal component analysis (FPCA). In particular, multivariate functional data (finite dimension vector of functions) analysis has been widely studied through several approaches including multivariate principal component analysis (MFPCA) for vector of functions observed on the same finite one-dimensional interval (Ramsay and Silverman (2005), Ferraty and Vieu (2006), Jacques and Preda (2014), Chiou *et al.*(2014), Wang *et al.*(2016), Schmutz *et al.*(2020)).

Recently, Happ and Greven (2018) developed a theory for MFPCA, that focuses on multivariate functional data on different domains (e.g. time functions, images or more complex structures). They showed that their method is competitive with respect to existing approaches for data observed on a common one-dimensional

domain. The theoretical foundations of their MFPCA are provided in terms of a Karhunen-Loève theorem. For the practically relevant case of a finite or truncated Karhunen-Loève representation, they established a direct theoretical relationship between univariate and multivariate FPCA. This suggests an estimation strategy for multivariate functional principal components functions and scores based on the univariate counterparts.

This approach has been extended (Golovkine *et al.*(2022)) to a multivariate functional principal component framework based on the notion of local regularity of the functional variables, to cluster multivariate functional data. Although the multivariate principal component analysis methods of Happ and Greven (2018) and Golovkine *et al.*(2022) are more flexible than other methods, they are not suitable for the stratified sampling of interest in this contribution. The very recent review of [Koner and Staicu (2023) gives more details on multivariate functional data analysis.

For stratified sampling, univariate functional principal component analysis (FPCA) has been investigated in few papers. Cardot *et al.*(2010) proposed a univariate FPCA in stratified sampling context, based on Horvitz-Thompson estimators of the covariance operator which can be seen as nonlinear functionals. The reader may also refer to Cardot and Josserand (2011) who proposed consistent estimators of mean and variance functions based on the Horvitz-Thompson estimator. More recently, Ahmed *et al.*(2018) adapted univariate FPCA for choice-based sampling in a context of binary functional linear regression and focused on inference.

In this paper, we propose a MFPCA approach that extends Happ and Greven (2018)'s work to choice-based sampling context, according to a binary response variable. The remainder of the document is organized as follows. Section 2 presents the multivariate functional principal component analysis (MFPCA) methodology under the sampling context of interest. The asymptotic properties of the MFPCA resulting estimators are given in Section 3 while the computational results to assess the performance of the proposed method are stated in Section 4. The proofs of the theoretical results are postponed to the appendix.

2. Multivariate Functional Principal Component Analysis under Choice-Based Sampling: MFPCA-CBS

Let $X = (X_1, \dots, X_P)^\top$ be a vector-valued stochastic process with $P \geq 1$ an integer. For $1 \leq p \leq P$, let I_p be a compact set in $\mathbb{R}$, with finite (Lebesgue-) measure and such that $X_p : I_p \rightarrow \mathbb{R}$ is assumed to belong to $\mathcal{L}^2(I_p)$, equipped with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{L}^2(I_p)}$ and norm $\|\cdot\|_{\mathcal{L}^2(I_p)}$. We denote by $\mathbf{I} := I_1 \times \dots \times I_P$, the P -Fold Cartesian product of I_p . So, X is a stochastic process indexed by $\mathbf{t} = (t_1, \dots, t_P) \in \mathbf{I}$ and taking values in the P -Fold Cartesian product space $\mathcal{H} := \mathcal{L}^2(I_1) \times \dots \times \mathcal{L}^2(I_P)$.

Let the scalar product $\langle\langle \cdot, \cdot \rangle\rangle : \mathcal{H} \times \mathcal{H} \rightarrow \mathbb{R}$, be defined as:

Journal home page: <http://www.jafristat.net>, www.projecteuclid.org/euclid.as,
www.ajol.info/afst

This issue [Vol 17 (4)] was open on October 1, 2022 and closed on September 30, 2023

$$\langle\langle f, g \rangle\rangle := \sum_{p=1}^P \langle f_p, g_p \rangle_{\mathcal{L}^2(I_p)} = \sum_{p=1}^P \int_{I_p} f_p(t_p) g_p(t_p) dt_p, \quad f, g \in \mathcal{H}.$$

Then, $\mathcal{H}$ is a Hilbert space with respect to $\langle\langle \cdot, \cdot \rangle\rangle$ (Prop.1 of Happ and Greven (2018)). We denote by $\|\cdot\|$ the norm induced by $\langle\langle \cdot, \cdot \rangle\rangle$.

2.1. Sampling context

Let us describe the Choice-Based Sampling (CBS) process of interest. Consider that in a given population we observe a discrete random variable Y that takes values in $\{0, 1\}$. Let $Q(i) = \mathbb{P}(Y = i)$ be the share of individuals such that $Y = i$ in the considered population with $i \in \{0, 1\}$. We assume that the population is divided, according to the values of the responses variables Y , into two strata $\mathcal{S}(0) = \{(0, X), X \in \mathcal{H}\}$ and $\mathcal{S}(1) = \{(1, X), X \in \mathcal{H}\}$. Let $0 < H(i) < 1$ be the probability with which one will draw an individual from stratum $\mathcal{S}(i)$. Set $Q^* = Q(1)$ and $H^* = H(1)$. The sample of interest is composed of i.i.d observations $X^{(1)}, \dots, X^{(N)}$ of X sampled at the rates $H^*, 1 - H^*$ from the two different strata. Then it is composed of H^*N and $(1 - H^*)N$ individuals from strata $\mathcal{S}(1)$ and $\mathcal{S}(0)$ respectively.

The remaining if the section gives a principal component analysis method for the stratified sample $X^{(1)}, \dots, X^{(N)}$.

The results obtained in this work may be extending to the case of multiple strata (more than two discrete values of Y). Note that other stratified sampling methods (Manski and Lerman (1977), Manski and McFadden (1981), Cosslett (1981), Imbens (1992), Cardot *et al.*(2010)) may also be used depending on the situation or domain of interest. The results obtained in this paper could then be extended to more complex sampling.

2.2. Multivariate Karhunen-Loève representation

We assume that $\mathbb{E}[X(\mathbf{t})] := (\mathbb{E}[X_1(t_1)], \dots, \mathbb{E}[X_P(t_P)]) = \mathbf{0}$, $\forall \mathbf{t} \in \mathbf{I}$. Let C denote the $P \times P$ matrix-valued covariance function which, for $\mathbf{s}, \mathbf{t} \in \mathbf{I}$, is defined as

$$C(\mathbf{s}, \mathbf{t}) := \mathbb{E}[X(\mathbf{s}) \otimes X(\mathbf{t})],$$

where the (p, q) th element of the matrix $C(\mathbf{s}, \mathbf{t})$, for $1 \leq p, q \leq P$, is the covariance function between the p -th and the q -th components of X :

$$C_{p,q}(s_p, t_q) := \mathbb{E}[X_p(s_p)X_q(t_q)] = \text{Cov}(X_p(s_p), X_q(t_q)), \quad s_p \in I_p, t_q \in I_q.$$

In particular, $C_{p,q}(\cdot, \cdot)$ belongs to $\mathcal{L}^2(I_p \times I_q)$.

The covariance C defines a linear self-adjoint and positive operator, $\Gamma : \mathcal{H} \rightarrow \mathcal{H}$, the covariance operator of X on the Hilbert space $\mathcal{H}$, where for $f \in \mathcal{H}$ and $\mathbf{t} \in \mathbf{I}$, the q th component of $\Gamma f(\mathbf{t})$ is given by

$$(\Gamma f)^{(q)}(t_q) := \langle\langle C_{\cdot, q}(\cdot, t_q), f(\cdot) \rangle\rangle = \sum_{p=1}^P \int_{I_p} C_{p, q}(s_p, t_q) f_p(s_p) ds_p, \quad t_q \in I_q, f \in \mathcal{H}.$$

Assume that there exist constants M_{pq} such that

$$\int_{I_q} (C_{p, q}(s_p, t_q))^2 dt_q < M_{pq} < \infty, \quad \forall s_p \in I_p, \quad (1)$$

where $C_{p, q}(s_p, \cdot) : I_p \times I_q \longrightarrow \mathbb{R}$ is uniformly continuous (see Happ and Greven (2018), Prop. 2, for more details), then Γ is compact, self-adjoint and positive operator and it can be diagonalized.

By the theory of Hilbert-Schmidt operators (e.g. Reed and Simon (1980), Theorem. VI.16), there exists a complete orthonormal basis $\{\phi_j, j = 1, 2, \dots\} \subset \mathcal{H}$ and a sequence of real numbers $\lambda_1 \geq \lambda_2 \geq \dots \geq 0$ such that

$$\Gamma \phi_j = \lambda_j \phi_j \quad \text{and} \quad \lambda_j \rightarrow 0 \quad \text{as} \quad j \rightarrow \infty.$$

The λ_j 's are the eigenvalues of the covariance operator Γ and the ϕ_j 's are the associated eigenfunctions. The multivariate version of the Karhunen-Loève's representation (Prop. 4, p. 6 in Happ and Greven (2018)) is:

$$X(\mathbf{t}) = \sum_{j=1}^{\infty} \xi_j \phi_j(\mathbf{t}), \quad \mathbf{t} \in \mathbf{I},$$

with zero mean random variables $\xi_j = \langle\langle X, \phi_j \rangle\rangle$ and $\text{Cov}(\xi_j, \xi_l) = \lambda_l \mathbf{1}_{\{j=l\}}$. Let $J \geq 1$ and assume that the first J eigenvalues are nonzero, i.e. $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_J \geq \lambda_{J+1}$. By an easy extension of Theorem 3.2 in Horváth and Kokoszka (2012), one can deduce that, up to a sign, the elements of the MFPCA basis are characterized by the following property :

$$\begin{aligned} \phi_1 &= \arg \max_{\phi} \langle\langle \Gamma \phi, \phi \rangle\rangle \quad \text{such that} \quad \|\phi\| = 1, \\ \phi_2 &= \arg \max_{\phi} \langle\langle \Gamma \phi, \phi \rangle\rangle \quad \text{such that} \quad \|\phi\| = 1 \quad \text{and} \quad \langle\langle \phi, \phi_1 \rangle\rangle = 0, \\ &\vdots \\ \phi_{J+1} &= \arg \max_{\phi} \langle\langle \Gamma \phi, \phi \rangle\rangle \quad \text{such that} \quad \|\phi\| = 1 \quad \text{and} \quad \langle\langle \phi, \phi_l \rangle\rangle = 0, \quad \forall l \leq J. \end{aligned}$$

In practice, the MFPCA basis is privileged because it induces the most accurate truncation for a given J . Then, the truncated Karhunen-Loève expansion of the process X is

$$X_{[J]}(\mathbf{t}) = \sum_{j=1}^J \xi_j \phi_j(\mathbf{t}), \quad \mathbf{t} \in \mathbf{I}, \quad J \geq 1; \quad (2)$$

and the truncated Karhunen-Loève expansion of the components of the process X is

$$X_{p,[J_p]}(t_p) = \sum_{j=1}^{J_p} \xi_{p,j} \phi_{p,j}(t_p), \quad t_p \in I_p, \quad J_p \geq 1, \quad 1 \leq p \leq P, \quad (3)$$

where $\{\phi_{p,j}, j = 1, 2, \dots\}$ is the univariate FPCA basis associated to the covariance operator Γ_p of X_p and the scores are $\xi_{p,j} = \langle X_p, \phi_{p,j} \rangle$. Happ and Greven (2018) (Prop.5, p.7) derived a direct relationship between the truncated representations (3) of the single elements X_p and the truncated representation (2) of the multivariate functional data X .

2.3. Empirical MFPCA under CBS

In practice, the eigenvalues $\{\lambda_j\}_{j \geq 1}$, eigenfunctions $\{\phi_j\}_{j \geq 1}$ and scores $\{\xi_j\}_{j \geq 1}$ have to be estimated using a CBS sample since they are not known. The second part of Prop. 5 p.7 in Happ and Greven (2018) drives a convenient estimation approach.

Given a sample of N i.i.d observations $X^{(1)}, \dots, X^{(N)}$ of X drawn through the above CBS sampling, the new proposed estimation procedure for MFPCA consists of four steps:

Step 1. For each element X_p , estimate a univariate FPCA based on the observations $X_p^{(1)}, \dots, X_p^{(N)}$. The eigenfunctions constructed are those of the integral operator associated with the integral kernel defined by the variance-covariance function of X_p , which is estimated for each $s, t \in I_p$ as follows (Ahmed *et al.*(2018), p. 142):

$$\hat{K}_p(s, t) = \frac{1}{N-1} \sum_{n=1}^N X_p^{(n)}(s) X_p^{(n)}(t) \left(\frac{Q^*}{H^*} \mathbb{I}(Y_n = 1) + \frac{1-Q^*}{1-H^*} \mathbb{I}(Y_n = 0) \right), \quad (4)$$

where $Y_n, n = 1, \dots, N$ are the i.i.d. observations of the binary random variable Y related to the sampling.

This results in estimated eigenfunctions $\hat{\phi}_{p,j}$ and scores $\hat{\xi}_{p,j}^{(n)}, n = 1, \dots, N, j = 1, \dots, J_p$ for a given truncation integer J_p .

This FPCA helps to construct a eigenbasis adapted to the sampling of interest compared with the classical FPCA (Cardot *et al.*(2010) p. 84).

Step 2. Define the matrix $\Xi \in \mathbb{R}^{N \times J_+}$ with $J_+ = \sum_{p=1}^P J_p$, where each row $(\hat{\xi}_{1,1}^{(n)}, \dots, \hat{\xi}_{1,J_1}^{(n)}, \dots, \hat{\xi}_{P,1}^{(n)}, \dots, \hat{\xi}_{P,J_P}^{(n)})$ contains the estimated scores for the P components of the n -th observation. Let's consider the matrix $\mathbf{Z} \in \mathbb{R}^{J_+ \times J_+}$ (Prop.5., p.7 Happ and Greven (2018)) consisting of blocks $\mathbf{Z}^{(pq)} \in \mathbb{R}^{J_p \times J_q}$ with entries

$$Z_{jk}^{(pq)} = \text{Cov}(\xi_{p,j}, \xi_{q,k}), \quad j = 1, \dots, J_p, \quad k = 1, \dots, J_q, \quad p, q = 1, \dots, P.$$

An estimate $\hat{\mathbf{Z}} \in \mathbb{R}^{J_+ \times J_+}$ of the matrix $\mathbf{Z}$ in the context of stratified sampling is given by

$$\hat{\mathbf{Z}} = \frac{1}{N-1} \Xi^T \Xi \left(\frac{Q^*}{H^*} \mathbb{I}(Y_n = 1) + \frac{1-Q^*}{1-H^*} \mathbb{I}(Y_n = 0) \right). \quad (5)$$

Note that when the sampling is random, it suffices to use (4) and (5) with $H^* = Q^*$.

Step 3. Perform a matrix eigen-analysis for $\hat{\mathbf{Z}}$ resulting in eigenvalues $\hat{\lambda}_j$ and orthonormal eigenvectors $\hat{\mathbf{v}}_j$.

Step 4. Elements of the estimated multivariate eigenfunctions are given by

$$\hat{\phi}_{p,j}(t_p) = \sum_{k=1}^{J_p} [\hat{\mathbf{v}}_j]_{p,k} \hat{\phi}_{p,k}(t_p), \quad t_p \in I_p, \quad j = 1, \dots, J_+$$

and the corresponding multivariate scores can be calculated via

$$\hat{\xi}_j^{(n)} = \sum_{p=1}^P \sum_{k=1}^{J_p} [\hat{\mathbf{v}}_j]_{p,k} \hat{\xi}_{p,k}^{(n)} = \Xi_{n,\cdot} \hat{\mathbf{v}}_j.$$

The estimated eigenvalues and functions are derived under the assumption of a finite sample size N and a finite Karhunen-Loève representation for each univariate function X_p . They are relevant in practice with an appropriate choice of the truncation orders. Indeed, these orders $J_{p,p} = 1, \dots, P$ may increase along with the sample size. Hence, it is worth to investigate the infinite sample properties of the proposed methodology as $J_{p,p} = 1, \dots, P$ and N tend to infinity.

Note that the univariate FPCA estimation procedure of Ahmed *et al.*(2018) is a particular case ($P = 1$) of the above proposed method.

Remark: In the finite sample properties, for seek of comparison with Happ and Greven (2018), we give results based on the MFPCA estimation algorithm under CBS using a smoothed covariance operator associated with the estimated covariance function $\hat{K}_p(\cdot, \cdot)$.

3. Asymptotic Properties

We give in this section infinite sample properties of the eigen-elements estimators and functional variables approximations.

Let us consider the superpopulation asymptotic framework introduced by Isaki and Fuller (1982) which supposes that the population and the sample size tends to infinity. In the remainder C is a constant that may varies from line to line. Let M be the size of the population, not necessarily known, in which the CBS was performed. The next assumptions will be used to show the convergence results of the estimated eigenelements.

- (H1) $\sup_{1 \leq n \leq N} \|X^{(n)}\| \leq C < \infty$,
 (H2) $\lim_{M \rightarrow \infty} (N/M) = \pi \in]0, 1[$,
 (H3) $\min_{1 \leq n \leq N} \left(\frac{Q^*}{H^*} \mathbb{I}(Y_n = 1) + \frac{1 - Q^*}{1 - H^*} \mathbb{I}(Y_n = 0) \right) \geq \varepsilon > 0$, with ε a constant,
 (H4) $\Delta_{p,J_p} = \sup_{j=1, \dots, J_p} (\lambda_{p,j} - \lambda_{p,j+1})^{-1} < \infty$ for every finite J_p ,
 (H5) $\int_{I_p} \int_{I_q} \mathbb{E} [X_p(s_p)^2 X_q(t_q)^2] ds_p dt_q < \infty$, $\forall q = 1, \dots, P$,
 (H6) $\langle \phi_{p,j}, \hat{\phi}_{p,j} \rangle \geq 0$, for all $j = 1, \dots, J_p$,
 (H7) $\hat{\xi}_{p,j}^{(n)} = \langle X_p^{(n)}, \hat{\phi}_{p,j} \rangle$, for all $j = 1, \dots, J_p$ and $n = 1, \dots, N$.

Assumption (H1) is classical in functional data analysis, it does not implies that the variables $X^{(n)}(\mathbf{t})$ are uniformly bounded in n and $\mathbf{t} \in \mathbf{I}$. It is for instance fulfills when X_p is a standard brownian motion (see e.g, Cardot *et al.*(2010)). Assumptions (H2) and (H3) are verified by sampling plans such as simple random sampling without replacement or stratified sampling (see Cardot *et al.*(2010) for more details) including the CBS of interest.

Assumptions (H4) and (H5) are about theoretical properties of X_p and Γ_p , while assumptions (H6) and (H7) depend on the univariate decompositions used; they are similar to assumptions (A1), (A2), (A4) and (A5) in Happ and Greven (2018). Assumption (H4) guarantees that the first J_p univariate eigenvalues of each element have multiplicity 1.

Proposition 1. (Estimation Error). Assume (H4) – (H7) hold. Then for $u_{\max} = \max_{p=1, \dots, P} J_p$ and $\Delta_J := \max_{p=1, \dots, P} \Delta_{p,J_p}$, the maximal eigenvalue of $\mathbf{Z} - \hat{\mathbf{Z}}$ can be characterized by

$$\nu_{\max}(\mathbf{Z} - \hat{\mathbf{Z}}) = O_p \left(u_{\max} N^{-1/2} \max(1, \Delta_J) \right).$$

Let $\lambda_j^{[J]}, \phi_j^{[J]}, j \in \mathbb{N}$, be respectively the eigenvalues and eigenfunctions of the co-variance operator $\Gamma^{[J]}$ associated with $X^{[J]}$ of truncated univariate elements $X_p^{[J]} = \sum_{j=1}^{J_p} \xi_{p,j} \phi_{p,j}$ and scores $\xi_j^{[J]} := \langle X^{[J]}, \phi_j^{[J]} \rangle$. It holds for all $j = 1, \dots, J$, that

$$\begin{aligned} \left| \lambda_j^{[J]} - \hat{\lambda}_j \right| &= O_p \left(u_{\max} N^{-1/2} \max(1, \Delta_J) \right), \\ \left\| \phi_j^{[J]} - \hat{\phi}_j \right\| &= O_p \left(u_{\max}^{3/2} N^{-1/2} \max(1, \Delta_J) \right), \\ \left| \xi_j^{[J(n)]} - \hat{\xi}_j^{(n)} \right| &= O_p \left(u_{\max}^{3/2} N^{-1/2} \max(1, \Delta_J) \right), \\ \left\| X^{[J(n)]} - \hat{X}^{[J(n)]} \right\| &= O_p \left(u_{\max} \Delta_J N^{-1/2} \right), \end{aligned}$$

with $\hat{X}_p^{[J(n)]} = \sum_{j=1}^{J_p} \hat{\xi}_{p,j}^{(n)} \hat{\phi}_{p,j}$ the univariate components of $\hat{X}^{[J(n)]}$ ($\mathbf{Z}$ and $\hat{\mathbf{Z}}$ defined in Section 2.3).

4. Numerical experiments

Based on the MFPCA procedure presented in the previous sections, the properties of the functional analysis estimation are computationally studied in the first two parts of the following section, whereas an application using real data is presented in the last part. We compare the results of the proposed PCA methodology with that of Happ and Greven (2018) under different strata sizes of the population and the sample.

In all the numerical experiments, we considered a B-spline smoothing of the covariance operator estimate with 15 basis functions. This smoothing step implemented in the R package MFPCA of Happ (2017b) is adapted to our case study. It uses one of the univariate FPCA methods implemented in the R package refund (Crainiceanu *et al.*(2014)) with global smoothing compare to the local one of 'PACE' algorithm (Yao *et al.*(2005)). Several other packages have been used to simulate the data, namely the R packages *fda* (Ramsay *et al.*(2022)) and *fundata* (Happ (2017a)). Note that, numerical analyses with the unsmoothed version of the covariance estimate, not presented here, have also been done. They give less performances than those based on the smoothed estimate.

4.1. Simulation study

We study here the performance of the proposed procedure based on synthetic stratified data with three different models, highlighting the importance of considering the choice-based sampling design. For this settings, the multivariate random function is of the form

$$X(\mathbf{t}) = \sum_{j=1}^J \xi_j \phi_j(\mathbf{t}) + \epsilon(\mathbf{t}), \quad \epsilon(\mathbf{t}) \stackrel{\text{iid}}{\sim} \mathcal{N}_p(\mathbf{0}, \sigma_\epsilon \mathbf{I}), \quad \mathbf{t} \in \mathbf{I}, \quad (6)$$

where J is the number of the used basis functions $\{\phi_j\}_j$.

Model 1: Densely observed bivariate functional data on the same interval with different numbers of observation points.

We consider the sampling method defined in Section 2. We simulate the population functional data using the first $J_1 = J_2 = 25$, functions of the Fourier basis $\{\phi_{1,j}(t) = \phi_{2,j}(t) = \sqrt{2} \sin(j\pi t), t \in [0, 1]\}$ and consider the bivariate functional variable

$$X = (X_1, X_2)^\top + \epsilon, \quad (7)$$

with

$$\begin{cases} X_1(t_1) = \sum_{j=1}^{J_1} \xi_{1,j} \phi_{1,j}(t_1) & t_1 \in I_1 = [0, 1] \\ X_2(t_2) = \sum_{j=1}^{J_2} \xi_{2,j} \phi_{2,j}(t_2) & t_2 \in I_2 = [0, 1] \end{cases}$$

where the random coefficients $\xi_{1,j} \sim \mathcal{N}(0, (J_1 + 1 - j)/J_1 + 1)$ and $\xi_{2,j} \sim \mathcal{N}(0, 1/j + 1)$. The data are considered without ($\sigma_\epsilon = 0$) and with measurement error ($\sigma_\epsilon = 1.6$). The data generation process follows.

Step 1. We generate a population data $\{X^{(n)}(\mathbf{t}), \mathbf{t} \in \mathbf{I}\}_{n=1, \dots, M}$ of size $M = 2000$ where $X^{(n)}$ is generated as described in (7). The functions $X_1^{(n)}$ and $X_2^{(n)}$ are respectively on an equispaced grids of $E_1 = 100$ and $E_2 = 50$ sampling points in $[0, 1]$.

Step 2 Let $Y_*^{(n)} = \beta_0 + \langle \beta, X^{(n)} \rangle + e^{(n)}$, a latent variable with $\beta = (\beta_1, \beta_2)^\top$ a multivariate functional parameter defined as $\beta(\mathbf{t}) = \left(\sum_{j=1}^{25} \theta_{1,j} \phi_{1,j}(t_1), \sum_{j=1}^{25} \theta_{2,j} \phi_{2,j}(t_2) \right)^\top$, with $\theta_{1,j} = \theta_{2,j} = 0$ for $j \geq 4$. The intercept $\beta_0 = 0.8$ and the first three multivariate coefficients are: $(\theta_{1,1} = 1.0, \theta_{2,1} = 1.3)$, $(\theta_{1,2} = -0.7, \theta_{2,2} = 0.7)$ and $(\theta_{1,3} = 0.9, \theta_{2,3} = 0.4)$. We consider that the errors $e^{(n)} \sim \mathcal{N}(0, 0.01)$, with a signal-to-noise ratio of the order of 4.1, $n = 1, \dots, M$.

Step 3 We set a binary variable $Y^{(n)}$, $n = 1, \dots, M$ such that

$$Y^{(n)} = \begin{cases} 1 & \text{if } Y_*^{(n)} > q_\alpha, \\ 0 & \text{else} \end{cases}$$

where q_α is the empirical quantile of order α , $0 < \alpha < 1$, associated with $Y_* = (Y_*^{(n)})_{n=1, \dots, M}$. Thus $Y^{(n)}$ is a Bernoulli random variable, and α allows to fix different values of the proportion of $Y^{(n)} = 1$ (cases), that is Q^* . For values of $\alpha = 0.90, 0.70, 0.50$, and 0.15 we have respectively $Q^* = 0.10, 0.30, 0.50$, and 0.85 , low, moderate and high cases in the populations.

Step 4. We draw different stratified samples $(X^{(1)}, \dots, X^{(N)})$ of size $N = 250$ with different proportions of cases ($Y^{(n)} = 1$) in the sample: $H^* = 0.1, 0.5$, and 0.7 . If $H^* = 0.5$, this means that in the obtained stratified sample, the number of

individuals with response $Y^{(n)} = 1$ (called cases) is equal to $N/2$, as the number of individuals with response $Y^{(n)} = 0$ (controls). Note that varying Q^* with a fixed value of H^* allows to observe the influence of the ratio Q^*/H^* on the proposed MFPCA_{CBS} method. $\square$

Model 2: Densely observed bivariate functional data on different intervals

Here, we consider a similar scheme as above ($\sigma_\epsilon = 0, 1.6$ and $Q^* = 0.10, 0.30, 0.50, 0.85$) as in Model 1 but on different intervals $I_1 = [-1, 1]$ and $I_2 = [0, 1.5]$, using the first $J_1 = J_2 = 25$ functions of the Fourier basis on these intervals. The observations are sampled on equidistant grids with $E_1 = 100$ for $X_1^{(n)}$ and $E_2 = 60$ for $X_2^{(n)}$.

Model 3: Trivariate functional data on different intervals.

We extend the simulation setting 2 of Happ and Greven (2018) to the sampling design of interest. Model (6) is used with a multivariate functional process

$$X = (X_1, X_2, X_3)^\top + \epsilon,$$

where $\sigma_\epsilon = 0$ and 0.5 , random coefficients $\xi_j \sim \mathcal{N}(0, \lambda_j)$, eigenvalues with exponential ($\lambda_j^{\text{exp}} = \exp(-(j+1)/2)$) or linear ($\lambda_j^{\text{lin}} = (J+1-j)/J$) decrease, and eigenfunctions ϕ_j simulated using the first J Fourier basis functions on different intervals $I_1 = [-1, 0.5]$, $I_2 = [0, 1]$ and $I_3 = [1.5, 2]$ (for more details, see Happ and Greven (2018)). In the following, we take $J = 8$ and $J = 20$.

We generate a population dataset $\{X^{(n)}(\mathbf{t}), \mathbf{t} \in \mathbf{I}\}_{n=1, \dots, M}$ of size $M = 10000$ and use the same stratification method as in Cardot *et al.* (2010) to draw a sample. The variables $X_p^{(n)}$, $p = 1, 2, 3$, are generated on an equi-spaced grids of $E_1 = E_3 = 50$ and $E_2 = 100$ sampling points.

First, we construct two strata $\mathcal{S}(0)$ and $\mathcal{S}(1)$ as follows. The first strata is composed of $\tilde{X}^{(n)} = \alpha_1 X^{(n)}$ with corresponding response $\tilde{Y}^{(n)} = 0$, $\alpha_1 = 2$, $n = 1, \dots, 7000$. For the second strata let $\tilde{X}^{(n)} = \alpha_2 X^{(n)}$ and $\tilde{Y}^{(n)} = 1$, $n = 7001, \dots, M$, $\alpha_2 = 4$. This means that the proportion of cases in the population is $Q^* = 0.30$. Note that the two strata have similar functional shape, the second strata functions are translations of the first ones.

Next, we draw different stratified samples according to three different sizes $N = 100, 250$, and 1000 with different proportions of cases in the samples: $H^* = 0.1, 0.5$,

and 0.7. Each stratified sample is built by drawing independently two simple random samples of sizes N_{cases} in the strata $\mathcal{S}(1)$ and $N_{controls} = N - N_{cases}$ in the strata $\mathcal{S}(0)$. The sample sizes are chosen such as in the strata:

$$\frac{N_{cases}}{N} = H^* \quad \text{and} \quad \frac{N_{controls}}{N} = 1 - H^*.$$

To evaluate our estimation methodology, the size of the maximal eigenvalue gap Δ_J is computed for each of the kind of eigenfunctions. Namely, we have $\Delta_j^{\exp} = C \exp(-J)$, $C > 0$ and $\Delta_j^{\ln} = J$. To vary this gap, we just need to vary the maximum rank J (8 and 20) of the multivariate components. The results for Model 3 are given in Tables 7-8 and Figure 3.

4.1.1. Discussion

The samples generated with the three models are used to investigate the performance of the MFPCA_{CBS} estimation method based on the univariate functional principal components that are calculated by the FPCA under CBS algorithm implemented in Ahmed *et al.*(2018). Optimal choice of $\hat{\xi}$ and $\hat{\phi}$ depends on the PCA method and the choice of the number J of eigenfunctions used in the truncation strategy. The choice of the number J_p of univariate eigenfunctions used in the truncation strategy is a key step in functional data analysis and many approaches are used. Ahmed *et al.*(2018) focused on the selection of J_p by Akaike Information Criterion (AIC) since they were mainly interested to a regression problem. Among the various choices of the number of principal components, those associated with cumulative inertia values lower than some threshold (for instance 80%), is of interest here. This approach is considered as the most stable in terms of "optimal" choice. Other choices related to the number of eigenvalues larger the mean ("Kaiser") or the "elbow" method that selects all components just before the eigenvalues curve flattens out is also usual.

In order to study the performance of the proposed method, we compare our approach to that of Happ and Greven (2018) using the same local smoothing of the covariance operator estimate of univariate FPCA, implemented in the R package MFPCA (Happ (2017b)). After applying MFPCA_{CBS} and its random sampling counterpart MFPCA on the three simulated models, we reconstruct the data using the first J estimated eigenfunctions. The goodness of the reconstructed observations $\hat{X}^{(n)} = \sum_{j=1}^J \hat{\xi}_j \hat{\phi}_j(t)$ is evaluated by the mean relative squared error (see Happ and Greven (2018))

$$MRSE = \frac{1}{N} \sum_{n=1}^N \left(\frac{\|X^{(n)} - \hat{X}^{(n)}\|^2}{\|X^{(n)}\|^2} \right),$$

based on the $J = 14; 20$ first multivariate principal components for Models 1 and 2, and on the $J = 8; 20$ multivariate principal components for Model 3. The accuracy

of the resulting eigenfunctions estimates $\hat{\phi}_j$ (resp. eigenvalues estimates $\hat{\lambda}_j$) is measured by the relative errors $\text{Er}(\hat{\phi}_j) = \frac{\|\phi - \hat{\phi}_j\|^2}{\|\phi\|^2}$ for Models 1 and 2, using the $J = 9$ first estimated eigenfunctions (resp. $\text{Er}(\hat{\lambda}_j) = \frac{(\lambda_j - \hat{\lambda}_j)^2}{\lambda_j^2}$ for Model 3, using the $J = 5$ first estimated eigenvalues for the sample sizes $N = 250$ and 1000). In order to show the relevance of the noise levels chosen, we give in each of the tables of the three models a with and without signal-to-noise ratio (SNR_X) on the functional variable X . A signal to noise ratio $SNR_Y \approx 4$ on the response variable Y has been considered.

The obtained results (based on 100 Monte Carlo replications) of Model 1 are presented in Tables 1, 2, 3 and Figure 1. Tables 1 and 2 give the average *MRSE* for all the proportions Q^* . It shows that *MFPCA_{CBS}* yields significantly smaller *MRSE* than *MFPCA* without ($\sigma_\epsilon = 0$) or with measurement error ($\sigma_\epsilon = 1.6$). The presence of measurement error has an impact on the *MRSE* in both methods. The comparison of the eigenfunctions estimation (Figure 1) shows that *MFPCA_{CBS}* gives more accurate estimates with lower errors.

In Model 1, increasing the proportion Q^* does not affect the quality of the results obtained by our method. *MFPCA_{CBS}* leads to very good results for the goodness of reconstruction of the data and for the accuracy of eigenfunctions obtained, it outperforms *MFPCA*. To investigate the choice of estimated number of components, we compute for Model 1, the number of components to get different percentage of explained variance (80%, 95%, 99%). The results are given in Table 3 which shows that the number of components estimated by *MFPCA_{CBS}* to achieve a given percentage of explained variance is higher but close to that of the *MFPCA* up to two components, keeping in mind that the true number of component is 25.

The results for Model 2 are given in Tables 4, 5 and Figure 2. In this model, *MFPCA_{CBS}* gives good estimates of the eigenfunctions with small *MRSE* in the reconstruction of the data compared to *MFPCA* (Tables 4, 5 and Figure 2). Both methods are sensitive to the presence of measurement error and the proportion Q^* of cases in the population does not influence *MFPCA_{CBS}*'s performance. The reconstruction of $X_1^{(n)}$ by *MFPCA_{CBS}* is as good as in Model 1 compare to the case of $X_2^{(n)}$, although it's much better than *MFPCA*. This can be explained by the smaller number of sampling points ($E_2 = 60$) considered for $X_2^{(n)}$. This highlights the well known influence of sampling points when transforming raw data to a functional object.

Note that for *MFPCA_{CBS}* when fixing the proportions of cases in the population Q^* , a sample with similar size of cases (Q^*) yields better results lower or higher Q^* while the choice of Q^* does not actually influence the performance of *MFPCA*. When the structure of the sample is close to that of the population then the reconstruction of the data by *MFPCA_{CBS}* is improved, which is not surprising.

The results presented in the first tables and figures for Model 1 and Model 2, were obtained with $\hat{\mathbf{Z}} = \frac{1}{N-1} \mathbf{\Xi}^T \mathbf{\Xi}$ as in Happ and Greven (2018). We also run the same numerical studies with $\hat{\mathbf{Z}} = \frac{1}{N-1} \mathbf{\Xi}^T \mathbf{\Xi} \left(\frac{Q^*}{H^*} \mathbb{I}(Y_n = 1) + \frac{1-Q^*}{1-H^*} \mathbb{I}(Y_n = 0) \right)$ (see Section 2.3) for MFPCA_{CBS}. We observe a slight but not significant difference in the *MRSE* associated to MFPCA_{CBS} (Table 6). Hence including Q^* a second time through $\hat{\mathbf{Z}}$ (see steps 1 and 2 in Section 2.3) to take into account the sampling design in the MFPCA procedure is not necessary.

In Model 3 setting, MFPCA_{CBS} and MFPCA yield similar results for almost all the values of H^* considered when the number of components is $J = 8$ (small eigenvalue gap compare to $J = 20$ in the linear case) with generally a decrease of *MRSE* for both methods when the sample size N increases (Tables 7-8 and Figure 3). Moreover, the variation of the maximal eigenvalue gap shows the performance of MFPCA_{CBS} when the number of multivariate components increases ($J = 20$). With a larger number of components ($J = 20$) the gap is large for the linear eigenvalue setting and very small for the exponential setting. The first eigenfunctions are better estimated compare to the last ones in particular for the MFPCA method. Note that in the setting of Model 3, the shapes of the functions in the two strata are not very different in shape, then classical PCA works well.

Overall, the simulation results support the fact that the sampling method must be taken into account in the principal component empirical analysis in particular when the two strata have different behaviors. This finite sample study shows the effectiveness of the new method compared to MFPCA when the data are drawn under choice-based sampling.

4.2. Real data

Most cancer patient die due to metastasis, and the early onset of this multi-step process is usually missed by current staging tumor modalities. Advanced techniques exist to enrich disseminated tumor cells from patient blood and bone marrow as cancer progression marker. However, these cells present high heterogeneity, only some of them can exhibit stem cell phenotype and tumor development potential, others can have plasticity potential to reprogram into cancer stem cells. So, detection and characterization are challenging due to lack of clear phenotypic markers. Therefore, there is a critical need to find new ways to anticipate and predict metastasis development at an early stage of patient care. Cancer progression involves many cellular morphological effects, which have been revealed by biophysical studies. The relevance of the characterization of cancer cells by their bio-mechanical phenotype is attested by reports pointing out their physical alteration as reduced cell stiffness with invasiveness for lung, breast and colon cancers, while the deformability of circulating lymphocytes is reduced in the case of acute lymphoblastic leukemia. The physical properties even allow identifying different malignant breast epithelial cell lines by their deform-ability and their

viscoelastic behavior. Even though the analysis capability of cancer cells by their physical characteristics has been demonstrated, reports mostly compare different known states of cancer cells. So far to our knowledge, no prediction capability has been reported to detect cancer cells and evaluate their invasiveness only by bio-mechanical characterization. Nowadays, there is a growing interest in the use of cell physical phenotyping to detect and categorize disseminating cells population by several types of physical parameters (electrical and mechanical) using MEMS (Micro Electro Mechanical Systems) technology performing. The MEMS devices Takayama *et al.*(2018), Yamahata *et al.*(2008), Ahmadian *et al.*(2022) capture a cell for stimulation and provides the information on the mechanical/electrical properties of the captured cell. They performed a compression assay protocol on each cell and measured changes in the resonance frequency and amplitude values for single-cell biophysical properties. They monitored resonance frequency and amplitude related to the mechanical property of the cell (stiffness and viscosity) and physical property as a function of time in addition to the initial measurement on the cell dimension.

This application aims to apply the developed functional multivariate principal component analysis on a case-control sample based on a population of cancer cells of two different categories (one most metastatic than the other). Namely, we extract several samples from a population of cells. The population data of size almost 3000 is derived from the MEM devices, they are composed of $p = 5$ characteristics including the amplitude (size) and the 4 electrical properties (function of time during the compression period, at 1 and 5 Mega hertz, real and imaginary signal) of single cells from three different breast cancer cell lines in a controlled environment. The compression assays on two sets of breast cancer cell lines, one with less metastatic potential (named case) while the other is of higher metastatic potential is divided into two different cell line categories (named control1 and control2). We then consider two populations named P1 and P2. The first is composed with higher metastatic (control1) and less metastatic potential (case) cell lines while the P2 is composed of the same sub-population of less metastatic potential cell line (case) and another different higher metastatic cell line (control2). The percentage of cases in the two populations is respectively $Q_1^* = 67.06\%$ and $Q_2^* = 48.07\%$. For each cell we have a curve on the amplitude (size) and four electrical signal curves (real and imaginary time functions at 1 and 5 Mega hertz) of same time domain, scaled in $[0, 3]$ but different number of observed time points.

The raw data are first smoothed using B-splines. Note that the cell properties are of different length and different observed time points. Some of the raw data are sparse, then smoothing allows for a more realistic consideration of the whole dataset as functional data. The number of spline basis functions used for this case study is 10. The real data curves for each of the features are given in Figure 4 to Figure 8 showing different variations with respect to cell types. The less metastatic group (case) has higher amplitude (size).

Then we apply the proposed methodology on different constructed choice-based samples ($H^* = 0.15; 0.25; 0.5$) in the population. We compare the reconstruction of the original data with that obtained by the *MFPCA* Happ and Greven (2018) method ignoring the non-random nature of the sample using.

The results show that the first four multivariate functional principal components, achieve over 95% of the percentages of explained variance, as shown in Figure 9 and Figure 10 for both methods. Note the non correlation of the FPC scores. For J selection, we use the scree plot criterion looking for the “elbow” in the curve and selects all components just before the line flattens out (here $J = 4$ for both methods and different proportions $H^* = 0.15, 0.25, 0.50$). This choice combined with that of the explained variance rate allows to reach more than 95% of variance. The reconstruction results given in Tables 9 and 10 are based on 50 replications and $J = 5$. They show that the proposed method outperforms compare to *MFPCA*. Note that the mean squared error of *MFPCA* is larger than that of *MFPCA_{CBS}*, with the same explained variance achieving a similar number of principal components J . The heatmaps of the first four estimated eigenfunctions by both methods are shown in Figure 11. For seek of simplicity, we only present the map of the second population (with $Q_2^* = 48.07\%$) to compare *MFPCA_{CBS}*. The five physical parameters features reconstructions, for a given cell, by *MFPCA_{CBS}* and *MFPCA* are given in Figure 12 to Figure 16, and compared to the true functions, for the case $H^* = 0.5$. Overall, the reconstructed curves obtained by both methods have a very similar shape to the real functions, but the proposed method outperforms. These results confirm the simulations finding regarding the usefulness of the proposed method when the sample of interest is far from random.

5. Conclusion

This work developed a multivariate functional principal component analysis approach, in which the sample is stratified under a choice-based design. Specifically, we extended the multivariate functional PCA method proposed by Happ and Greven (2018), which requires a random sample and the univariate stratified PCA method of Ahmed *et al.*(2018). To take into account the non-random nature of the sample we develop a new covariance operator, functional principal component functions and scores estimation strategy using a truncated series expansion. We establish that this strategy is asymptotically a correct approximation, when the truncation dimension grows large with the sample size.

The numerical results on synthetic data show that the proposed methodology outperform compare to that ignores the non-random nature of the data. Most importantly, as the proportions of cases of the population Q^* grows large, Happ and Greven (2018)’s method appears to be less performing while when the sample is distributed as the population in terms of cases ($Q^* \approx H^*$) then the proposed methodology gives better result. In practice, we adapt *MFPCA* estimation procedure. We use the raw functional data to estimate the smoothed covariance operator as in *MFPCA*, the only difference is that our covariance estimate takes

into account the endogeneous sampling. We also numerically investigate the behavior of the PCA when the data are first smoothed (as in the theoretical setting where the functional data are smoothed) and an un-smoothed covariance operator is estimated, these results (not shown here) give a good performance, but they cannot be compared with *MFPCA* of Happ and Greven (2018) who considered raw data and a smoothed covariance operator in all their numerical experiments. The results obtained suggest that the RMSE is sensitive to the smoothing of the covariance operator. Further investigations will be done in the future to investigate deeply the effect of the smoothing of raw and covariance functions on *MFPCA*_{CBS}.

An empirical application using cancer cell lines population dataset where different stratified samples are derived, provides good results and shows the performance of the propose methodology. Finally, the proposed multivariate functional principal component analysis constitutes a fundamental method to reduce dimension of multivariate functional data when the sample is stratified under a choice-based or case-control sampling from a given population. This open an array of potential extensions to other kind of stratified sampling and applications in clustering or supervised learning models where one is interested to estimate a binary regression model described linking Y to the multivariate functional variable X .

Table 1. Average *MRSE* over 100 replications for Model 1 (with the corresponding standard deviation in brackets) depending on small, moderate and large proportions of Q^* and H^* of events ($Y = 1$), and different measurement error variance (σ_ϵ with signal to noise ratios for X and Y denoted SNR_X and SNR_Y) with the $J = 20$ first multivariate principal components.

Model 1.					
Parameters	Q^*	$\sigma_\epsilon = 0$ ($SNR_X=0$, $SNR_Y=4.1$)		$\sigma_\epsilon = 1.6$ ($SNR_X=5$, $SNR_Y=4.1$)	
		<i>MFPCA</i> _{CBS}	<i>MFPCA</i>	<i>MFPCA</i> _{CBS}	<i>MFPCA</i>
$H^* = 0.1$	0.10	0.113 (4.0 10^{-3})	0.261 (6.9 10^{-3})	0.419 (6.2 10^{-3})	0.520 (7.3 10^{-3})
	0.30	0.114 (4.0 10^{-3})	0.262 (6.8 10^{-3})	0.418 (6.6 10^{-3})	0.515 (7.3 10^{-3})
	0.50	0.113 (3.4 10^{-3})	0.260 (6.9 10^{-3})	0.420 (5.9 10^{-3})	0.515 (6.7 10^{-3})
	0.85	0.117 (4.9 10^{-3})	0.264 (4.1 10^{-3})	0.440 (9.7 10^{-3})	0.518 (4.6 10^{-3})
$H^* = 0.5$	0.10	0.111 (3.0 10^{-3})	0.265 (6.0 10^{-3})	0.417 (4.8 10^{-3})	0.519 (5.4 10^{-3})
	0.30	0.113 (3.8 10^{-3})	0.262 (8.0 10^{-3})	0.417 (5.7 10^{-3})	0.517 (6.3 10^{-3})
	0.50	0.114 (4.0 10^{-3})	0.262 (7.6 10^{-3})	0.419 (6.2 10^{-3})	0.519 (6.9 10^{-3})
	0.85	0.113 (3.4 10^{-3})	0.263 (6.7 10^{-3})	0.416 (5.6 10^{-3})	0.516 (6.9 10^{-3})
$H^* = 0.7$	0.10	0.111 (2.2 10^{-3})	0.267 (4.5 10^{-3})	0.421 (4.3 10^{-3})	0.523 (4.8 10^{-3})
	0.30	0.114 (3.3 10^{-3})	0.264 (5.9 10^{-3})	0.420 (6.1 10^{-3})	0.519 (6.9 10^{-3})
	0.50	0.114 (3.9 10^{-3})	0.264 (7.6 10^{-3})	0.418 (6.3 10^{-3})	0.518 (6.5 10^{-3})
	0.85	0.113 (3.7 10^{-3})	0.262 (6.5 10^{-3})	0.417 (6.2 10^{-3})	0.516 (7.4 10^{-3})

Table 2. Average *MRSE* over 100 replications for Model 1 (with the corresponding standard deviation in brackets) depending on small, moderate and large proportions of Q^* and H^* of events ($Y = 1$), and different measurement error variance (σ_ϵ with signal to noise ratios for X and Y denoted SNR_X and SNR_Y) with the $J = 14$ first multivariate principal components.

Model 1.					
Parameters	Q^*	$\sigma_\epsilon = 0$ ($SNR_X=0$, $SNR_Y=4.1$)		$\sigma_\epsilon = 1.6$ ($SNR_X=5$, $SNR_Y=4.1$)	
		<i>MFPCA</i> _{CBS}	MFPCA	<i>MFPCA</i> _{CBS}	MFPCA
$H^* = 0.1$	0.10	0.163 (4.9 10^{-3})	0.271 (7.0 10^{-3})	0.455 (6.0 10^{-3})	0.523 (7.1 10^{-3})
	0.30	0.164 (4.7 10^{-3})	0.272 (6.8 10^{-3})	0.455 (6.2 10^{-3})	0.522 (6.4 10^{-3})
	0.50	0.163 (4.7 10^{-3})	0.270 (7.0 10^{-3})	0.457 (6.3 10^{-3})	0.521 (6.8 10^{-3})
	0.85	0.168 (3.4 10^{-3})	0.274 (4.2 10^{-3})	0.473 (5.3 10^{-3})	0.528 (4.3 10^{-3})
$H^* = 0.5$	0.10	0.161 (3.9 10^{-3})	0.276 (6.1 10^{-3})	0.456 (5.1 10^{-3})	0.525 (5.3 10^{-3})
	0.30	0.162 (5.0 10^{-3})	0.272 (8.2 10^{-3})	0.455 (5.5 10^{-3})	0.523 (6.1 10^{-3})
	0.50	0.163 (4.8 10^{-3})	0.272 (7.7 10^{-3})	0.457 (6.6 10^{-3})	0.525 (6.6 10^{-3})
	0.85	0.164 (4.8 10^{-3})	0.274 (6.7 10^{-3})	0.455 (5.8 10^{-3})	0.522 (6.6 10^{-3})
$H^* = 0.7$	0.10	0.161 (2.7 10^{-3})	0.278 (4.6 10^{-3})	0.458 (4.5 10^{-3})	0.530 (4.5 10^{-3})
	0.30	0.163 (4.3 10^{-3})	0.274 (6.8 10^{-3})	0.456 (6.1 10^{-3})	0.525 (6.2 10^{-3})
	0.50	0.163 (4.4 10^{-3})	0.272 (6.6 10^{-3})	0.458 (5.6 10^{-3})	0.526 (6.5 10^{-3})
	0.85	0.164 (4.8 10^{-3})	0.272 (7.1 10^{-3})	0.456 (6.0 10^{-3})	0.524 (6.7 10^{-3})

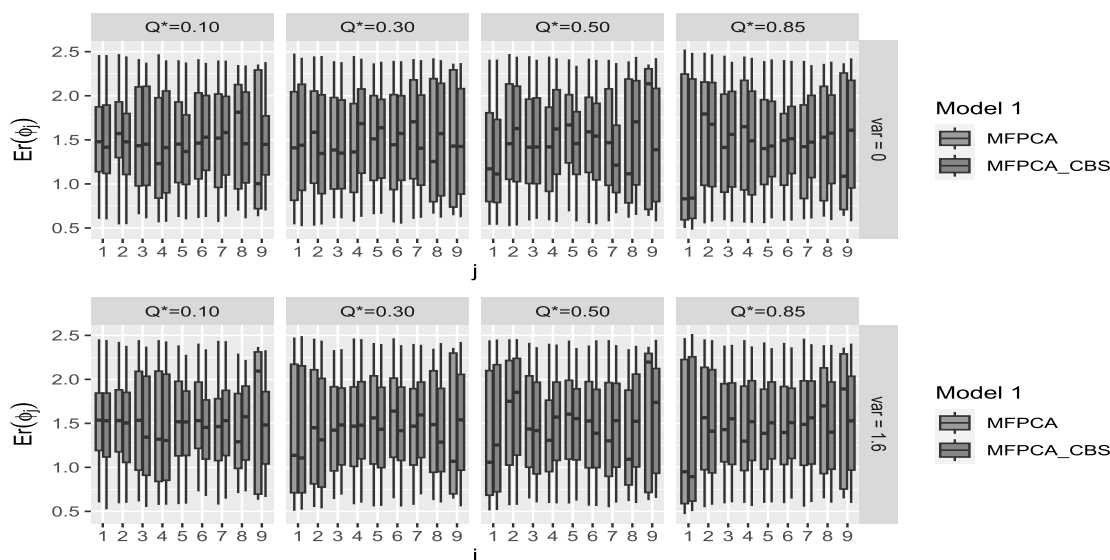


Fig. 1. Comparison of relative errors for eigenfunctions of the two methods for Model 1, depending on proportion Q^* of events ($Y = 1$) and measurement error ($\sigma_\epsilon = 0$ ($\text{SNR}_X=0$) and $\sigma_\epsilon = 1.6$ ($\text{SNR}_X=5$)) with $H^* = 0.5$ fixed.

Table 3. Average over 100 replications of the number of multivariate functional principal components estimated for Model 1, with different proportion Q^* of events ($Y = 1$), measurement error, with J chosen to count 80%, 95%, 99% of the variance (pve).

Model 1.	Percentage of variance explained (pve)							True J
		0.80		0.95		0.99		
	Q*	MFPCA _{CBS}	MFPCA	MFPCA _{CBS}	MFPCA	MFPCA _{CBS}	MFPCA	
σ _ε = 0	0.10	9.00	6.75	13.16	10.00	16.66	14.00	25
(SNR _X =0,	0.30	8.99	6.85	13.22	9.98	16.63	13.92	
SNR _Y =4.1)	0.85	9.00	6.90	13.39	10.00	16.51	13.96	
σ _ε = 1.6	0.10	8.23	6.00	12.96	9.00	15.15	11.47	25
(SNR _X =5,	0.30	8.18	6.01	12.96	9.00	15.16	11.44	
SNR _Y =4.1)	0.85	8.22	6.02	12.93	9.00	15.18	11.48	

Table 4. Average *MRSE* over 100 replications for Model 2 (with the corresponding standard deviation in brackets) depending on small, moderate and large proportions of Q^* and H^* of events ($Y = 1$), and different signal to noise ratio, with the $J = 20$ first multivariate principal components.

Model 2.					
Parameters	Q^*	$\sigma_\epsilon = 0$ ($\text{SNR}_X=0, \text{SNR}_Y=4.1$)		$\sigma_\epsilon = 1.6$ ($\text{SNR}_X=3.5, \text{SNR}_Y=4.1$)	
		<i>MFPCA</i> _{CBS}	MFPCA	<i>MFPCA</i> _{CBS}	MFPCA
$H^* = 0.1$	0.10	0.114 (3.9 10^{-3})	0.263 (7.2 10^{-3})	0.527 (6.0 10^{-3})	0.609 (6.5 10^{-3})
	0.30	0.113 (3.3 10^{-3})	0.260 (7.5 10^{-3})	0.527 (6.1 10^{-3})	0.609 (7.1 10^{-3})
	0.50	0.113 (3.3 10^{-3})	0.260 (6.6 10^{-3})	0.532 (5.9 10^{-3})	0.609 (6.1 10^{-3})
	0.85	0.116 (3.3 10^{-3})	0.264 (4.5 10^{-3})	0.561 (9.0 10^{-3})	0.610 (4.2 10^{-3})
$H^* = 0.5$	0.10	0.111 (2.9 10^{-3})	0.267 (6.3 10^{-3})	0.527 (5.5 10^{-3})	0.609 (5.7 10^{-3})
	0.30	0.113 (3.8 10^{-3})	0.262 (8.0 10^{-3})	0.527 (5.7 10^{-3})	0.608 (5.7 10^{-3})
	0.50	0.114 (4.0 10^{-3})	0.262 (6.8 10^{-3})	0.527 (6.6 10^{-3})	0.609 (6.7 10^{-3})
	0.85	0.113 (3.4 10^{-3})	0.264 (6.2 10^{-3})	0.528 (5.9 10^{-3})	0.609 (5.7 10^{-3})
$H^* = 0.7$	0.10	0.112 (2.7 10^{-3})	0.270 (5.0 10^{-3})	0.539 (5.0 10^{-3})	0.624 (4.2 10^{-3})
	0.30	0.114 (3.3 10^{-3})	0.263 (6.9 10^{-3})	0.530 (5.9 10^{-3})	0.613 (6.2 10^{-3})
	0.50	0.113 (3.8 10^{-3})	0.263 (7.7 10^{-3})	0.529 (6.7 10^{-3})	0.613 (6.9 10^{-3})
	0.85	0.113 (3.7 10^{-3})	0.263 (7.2 10^{-3})	0.528 (5.3 10^{-3})	0.612 (6.2 10^{-3})

Table 5. Average *MRSE* over 100 replications for Model 2 (with the corresponding standard deviation in brackets) depending on small, moderate and large proportions of Q^* and H^* of events ($Y = 1$), and different signal to noise ratio, with the $J = 14$ first multivariate principal components.

Model 2.					
Parameters	Q^*	$\sigma_\epsilon = 0$ ($\text{SNR}_X=0, \text{SNR}_Y=4.1$)		$\sigma_\epsilon = 1.6$ ($\text{SNR}_X=3.5, \text{SNR}_Y=4.1$)	
		<i>MFPCA</i> _{CBS}	MFPCA	<i>MFPCA</i> _{CBS}	MFPCA
$H^* = 0.1$	0.10	0.164 (4.6 10^{-3})	0.273 (7.3 10^{-3})	0.558 (5.6 10^{-3})	0.615 (5.7 10^{-3})
	0.30	0.163 (4.2 10^{-3})	0.271 (7.6 10^{-3})	0.561 (5.7 10^{-3})	0.617 (6.0 10^{-3})
	0.50	0.163 (4.4 10^{-3})	0.270 (6.6 10^{-3})	0.561 (6.5 10^{-3})	0.615 (6.4 10^{-3})
	0.85	0.168 (3.3 10^{-3})	0.274 (4.5 10^{-3})	0.584 (5.9 10^{-3})	0.621 (4.5 10^{-3})
$H^* = 0.5$	0.10	0.162 (3.9 10^{-3})	0.277 (6.4 10^{-3})	0.559 (5.4 10^{-3})	0.614 (5.5 10^{-3})
	0.30	0.162 (4.9 10^{-3})	0.272 (8.1 10^{-3})	0.558 (5.9 10^{-3})	0.613 (5.8 10^{-3})
	0.50	0.164 (4.8 10^{-3})	0.273 (6.9 10^{-3})	0.559 (6.7 10^{-3})	0.614 (6.6 10^{-3})
	0.85	0.164 (4.7 10^{-3})	0.274 (6.2 10^{-3})	0.559 (5.5 10^{-3})	0.614 (5.6 10^{-3})
$H^* = 0.7$	0.10	0.162 (3.2 10^{-3})	0.280 (4.7 10^{-3})	0.566 (4.8 10^{-3})	0.624 (4.5 10^{-3})
	0.30	0.163 (4.4 10^{-3})	0.273 (6.7 10^{-3})	0.561 (6.1 10^{-3})	0.618 (6.6 10^{-3})
	0.50	0.163 (4.2 10^{-3})	0.273 (6.3 10^{-3})	0.561 (6.8 10^{-3})	0.618 (6.5 10^{-3})
	0.85	0.162 (4.6 10^{-3})	0.273 (6.4 10^{-3})	0.560 (5.6 10^{-3})	0.616 (5.5 10^{-3})

Table 6. Average *MRSE* over 100 replications for Model 1 and Model 2 depending on small, moderate and large proportions Q^* of events ($Y = 1$) and different signal to noise ratio with $\hat{\mathbf{Z}} = \frac{1}{N-1} \mathbf{\Xi}^T \mathbf{\Xi} \left(\frac{Q^*}{H^*} \mathbb{I}(Y_n = 1) + \frac{1-Q^*}{1-H^*} \mathbb{I}(Y_n = 0) \right)$ and $H^* = 0.5$.

Model 1		
Q^*	$MFPCA_{CBS}$	
	$\sigma_\epsilon = 0$ ($SNR_X=0, SNR_Y=4.1$)	$\sigma_\epsilon = 1.6$ ($SNR_X=5, SNR_Y=4.1$)
0.10	0.112	0.417
0.30	0.113	0.417
0.50	0.114	0.419
0.85	0.113	0.417

Model 2		
Q^*	$MFPCA_{CBS}$	
	$\sigma_\epsilon = 0$ ($SNR_X=0, SNR_Y=4.1$)	$\sigma_\epsilon = 1.6$ ($SNR_X=3.5, SNR_Y=4.1$)
0.10	0.112	0.527
0.30	0.113	0.527
0.50	0.114	0.527
0.85	0.114	0.528

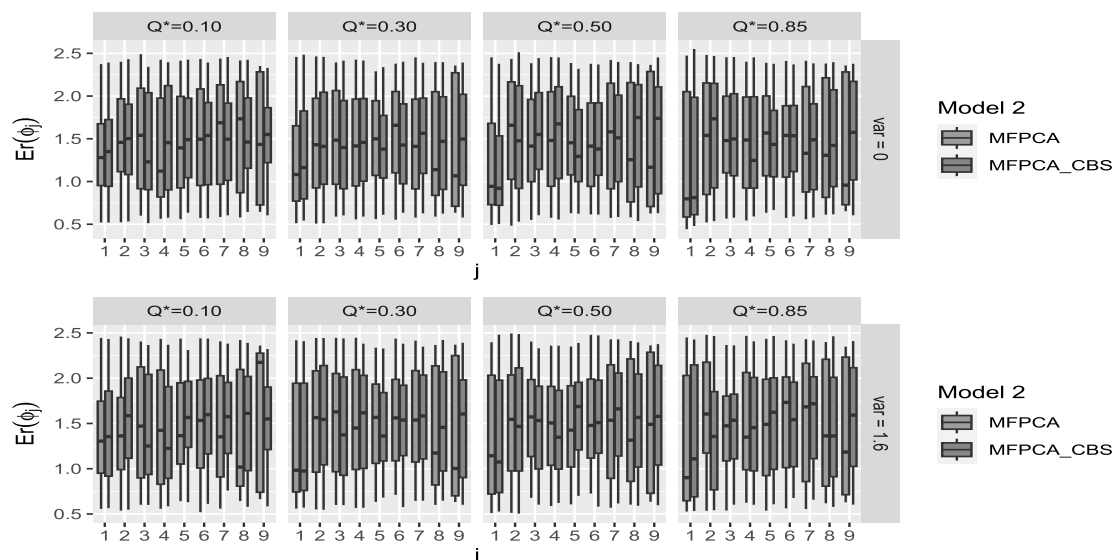


Fig. 2. Comparison of relative errors for eigenfunctions of the two methods for Model 2, depending on proportion Q^* of events ($Y_n = 1$) and measurement error ($\sigma_\epsilon = 0$ ($\text{SNR}_X=0$) and $\sigma_\epsilon = 1.6$ ($\text{SNR}_X=3.5$)) with $H^* = 0.5$ fixed.

Table 7. Average *MRSE* (in %) over 100 replications for Model 3 (with the corresponding standard deviation in brackets) based on the 8 functional principal components used for each method with three different sample sizes $N = 100, 250$ and 1000, $Q^* = 0.30$ and H^* varying.

Model 3							
Parameters	N	J		$\sigma_\epsilon = 0$		$\sigma_\epsilon = 0.5$	
				λ_j^{lin} (SNR _X =0)	λ_j^{exp} (SNR _X =0)	λ_j^{lin} (SNR _X =2.6)	λ_j^{exp} (SNR _X =2)
$H^* = 0.15$	100	8	<i>MFPCA_{CBS}</i>	0.222 (0.051)	0.298 (0.090)	16.955 (0.876)	27..901 (1.396)
		8	<i>MFPCA</i>	0.223 (0.049)	0.287 (0.057)	16.988 (0.879)	27.987 (1.414)
		8	<i>MFPCA_{CBS}</i>	0.206 (0.008)	0.295 (0.017)	16.993 (0.601)	27.876 (0.842)
		8	<i>MFPCA</i>	0.208 (0.007)	0.296 (0.016)	17.010 (0.602)	27.939 (0.839)
	250	8	<i>MFPCA_{CBS}</i>	0.204 (0.003)	0.291 (0.010)	17.062 (0.249)	27.998 (0.428)
		8	<i>MFPCA</i>	0.207 (0.003)	0.293 (0.010)	17.068 (0.249)	28.017 (0.429)
	1000	8	<i>MFPCA_{CBS}</i>	0.221 (0.048)	0.302 (0.086)	16.989 (0.840)	27.833 (1.367)
		8	<i>MFPCA</i>	0.214 (0.029)	0.301 (0.082)	17.023 (0.843)	27.927 (1.367)
	250	8	<i>MFPCA_{CBS}</i>	0.213 (0.035)	0.293 (0.019)	16.988 (0.516)	28.163 (0.941)
		8	<i>MFPCA</i>	0.213 (0.028)	0.294 (0.019)	17.003 (0.517)	28.228 (0.943)
	1000	8	<i>MFPCA_{CBS}</i>	0.204 (0.003)	0.291 (0.010)	17.075 (0.218)	28.132 (0.455)
		8	<i>MFPCA</i>	0.207 (0.003)	0.293 (0.010)	17.080 (0.217)	28.153 (0.459)
$H^* = 0.50$	100	8	<i>MFPCA_{CBS}</i>	0.235 (0.053)	0.328 (0.133)	17.150 (0.964)	28.264 (1.348)
		8	<i>MFPCA</i>	0.213 (0.024)	0.300 (0.084)	17.127 (0.965)	28.278 (1.336)
	250	8	<i>MFPCA_{CBS}</i>	0.214 (0.023)	0.297 (0.019)	17.093 (0.544)	28.341 (0.794)
		8	<i>MFPCA</i>	0.209 (0.008)	0.294 (0.018)	17.083 (0.543)	28.364 (0.794)
	1000	8	<i>MFPCA_{CBS}</i>	0.206 (0.003)	0.293 (0.008)	17.084 (0.248)	28.190 (0.433)
		8	<i>MFPCA</i>	0.208 (0.003)	0.294 (0.008)	17.084 (0.247)	28.223 (0.433)

Journal home page: <http://www.jafristat.net>, www.projecteuclid.org/euclid.as,
www.ajol.info/afst

This issue [Vol 17 (4)] was open on October 1, 2022 and closed on September 30, 2023

Table 8. Average *MRSE* (in %) over 100 replications for Model 3 (with the corresponding standard deviation in brackets) based on the 20 functional principal components used for each method with three different sample sizes $N = 100, 250$ and 1000, $Q^* = 0.30$ and H^* varying.

Model 3							
Parameters	N	J		$\sigma_\epsilon = 0$		$\sigma_\epsilon = 0.5$	
				$\lambda_j^{\text{lin}} (\text{SNR}_X=0)$	$\lambda_j^{\text{exp}} (\text{SNR}_X=0)$	$\lambda_j^{\text{lin}} (\text{SNR}_X=3.8)$	$\lambda_j^{\text{exp}} (\text{SNR}_X=2)$
$H^* = 0.15$	100	20	<i>MFPCA_{CBS}</i>	0.572 (0.207)	0.839 (0.129)	7.051 (0.340)	26.700 (1.262)
		20	MFPCA	1.706 (0.176)	0.833 (0.115)	8.082 (0.305)	26.997 (1.304)
		250	20 <i>MFPCA_{CBS}</i>	0.453 (0.149)	0.857 (0.055)	6.896 (0.211)	26.946 (0.712)
		20	MFPCA	1.636 (0.160)	0.860 (0.055)	8.073 (0.231)	27.047 (0.718)
	1000	20	<i>MFPCA_{CBS}</i>	0.428 (0.136)	0.872 (0.030)	6.902 (0.142)	27.242 (0.352)
		20	MFPCA	1.606 (0.115)	0.875 (0.030)	8.072 (0.134)	27.257 (0.353)
	250	20	<i>MFPCA_{CBS}</i>	0.567 (0.195)	0.854 (0.097)	7.085 (0.324)	26.389 (1.275)
		20	MFPCA	1.698 (0.181)	0.853 (0.085)	8.187 (0.324)	26.691 (1.266)
		250	20 <i>MFPCA_{CBS}</i>	0.459 (0.155)	0.859 (0.062)	6.893 (0.218)	26.970 (0.752)
		20	MFPCA	1.670 (0.150)	0.861 (0.062)	8.076 (0.222)	27.086 (0.759)
	1000	20	<i>MFPCA_{CBS}</i>	0.439 (0.142)	0.867 (0.028)	6.841 (0.122)	27.067 (0.370)
		20	MFPCA	1.632 (0.139)	0.870 (0.028)	8.043 (0.148)	27.078 (0.369)
$H^* = 0.50$	100	20	<i>MFPCA_{CBS}</i>	0.666 (0.254)	0.903 (0.194)	7.254 (0.332)	26.806 (1.123)
		20	MFPCA	1.700 (0.179)	0.836 (0.112)	8.109 (0.310)	26.986 (1.166)
		250	20 <i>MFPCA_{CBS}</i>	0.520 (0.164)	0.863 (0.063)	7.008 (0.259)	26.839 (0.827)
		20	MFPCA	1.689 (0.166)	0.859 (0.052)	8.037 (0.230)	26.935 (0.832)
	1000	20	<i>MFPCA_{CBS}</i>	0.467 (0.148)	0.867 (0.039)	6.883 (0.163)	26.988 (0.363)
		20	MFPCA	1.660 (0.140)	0.870 (0.028)	8.092 (0.151)	26.985 (0.365)

Journal home page: <http://www.jafristat.net>, www.projecteuclid.org/euclid.as,
www.ajol.info/afst

This issue [Vol 17 (4)] was open on October 1, 2022 and closed on September 30, 2023

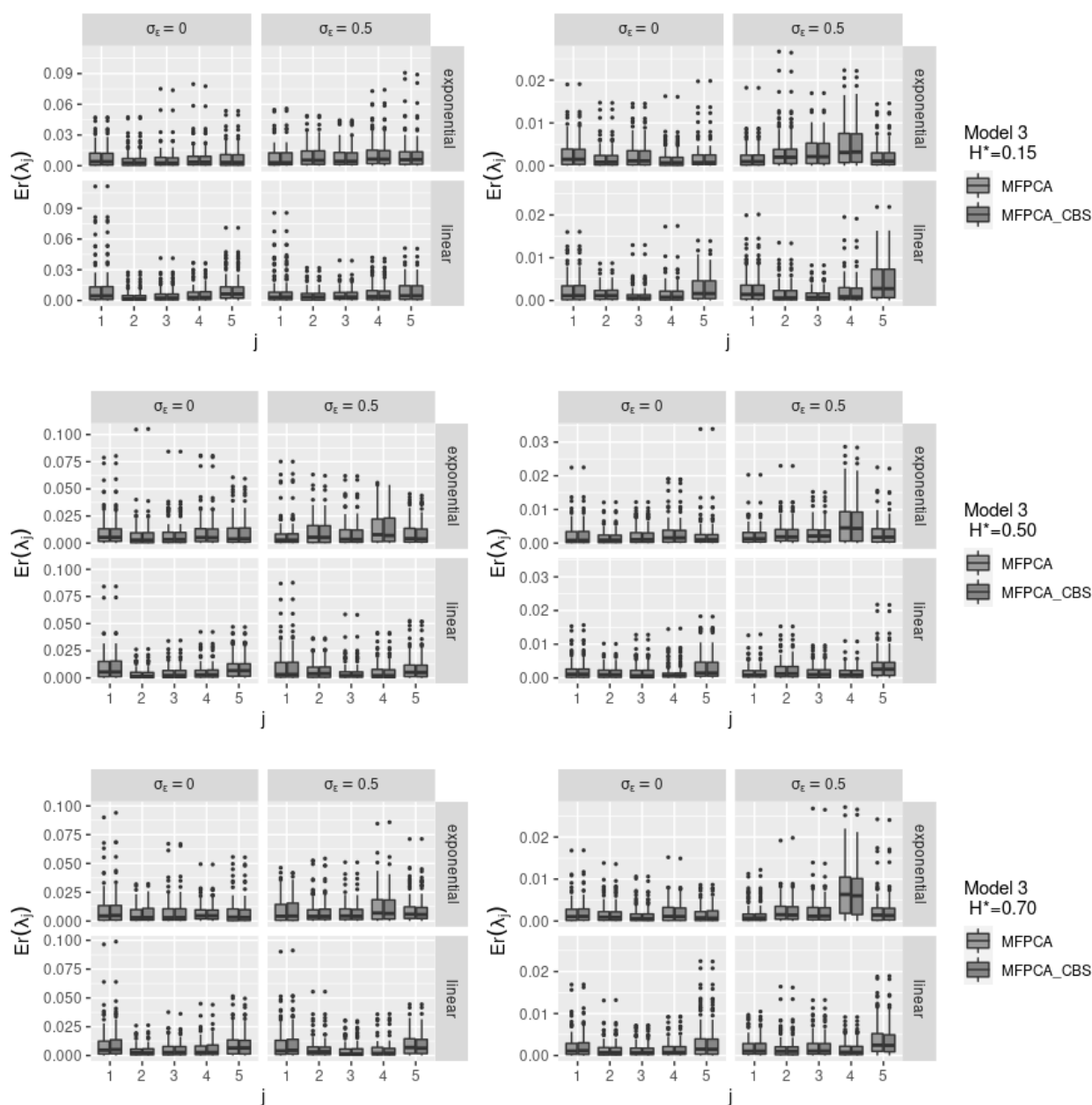


Fig. 3. Comparison of relative errors for eigenvalues of the two methods for Model 3, depending on proportion H^* and the sample size N (in each line, the left hand size is for $N = 250$, right hand side is for $N = 1000$).

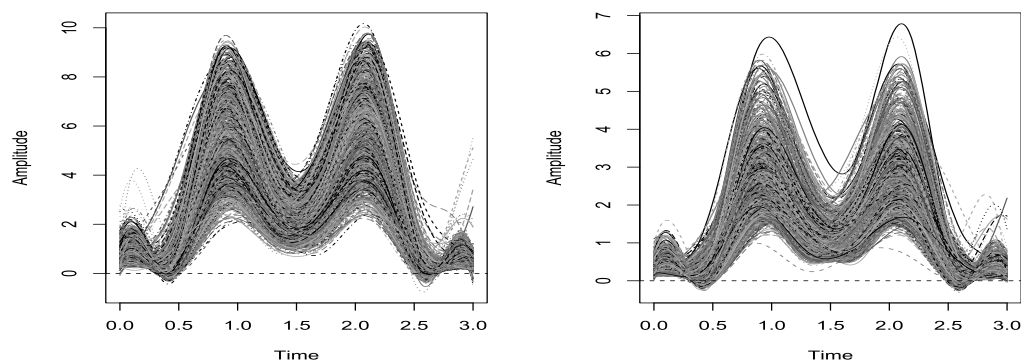


Fig. 4. Smoothed amplitude (size) curves of the population cancer cells per group. The right curves are the most softer (metastatic) group

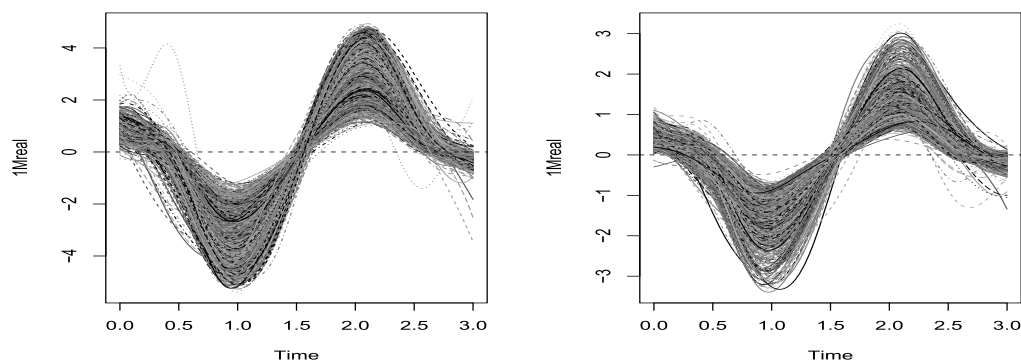


Fig. 5. Smoothed electrical (real signal at 1 Mega hertz) curves of the population cancer cells per group. The right curves are the most softer (metastatic) group

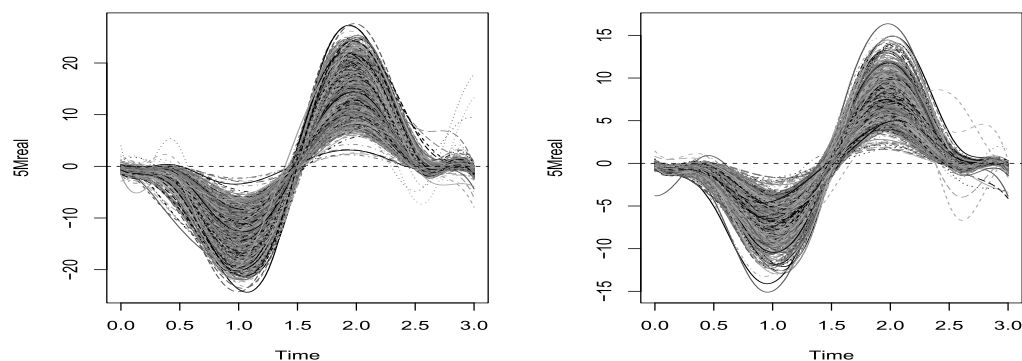


Fig. 6. Smoothed electrical (real signal at 5 Mega hertz) curves of the population cancer cells per group. The right curves are the most softer (metastatic) group

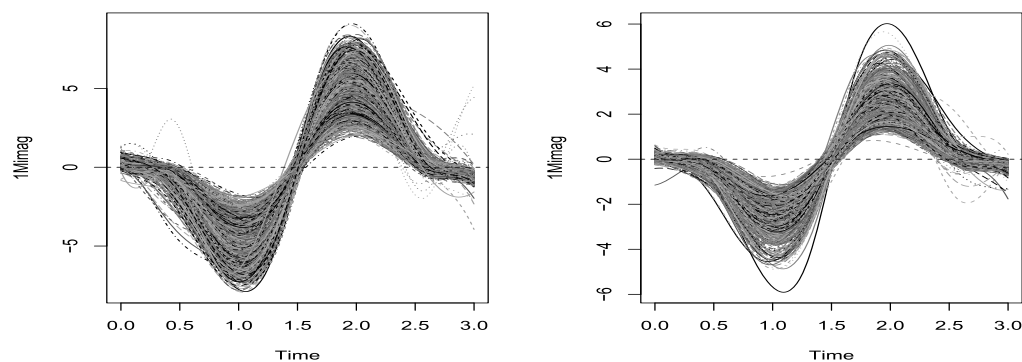


Fig. 7. Smoothed electrical (imaginary signal at 1 Mega hertz) curves of the population cancer cells per group. The right curves are the most softer (metastatic) group

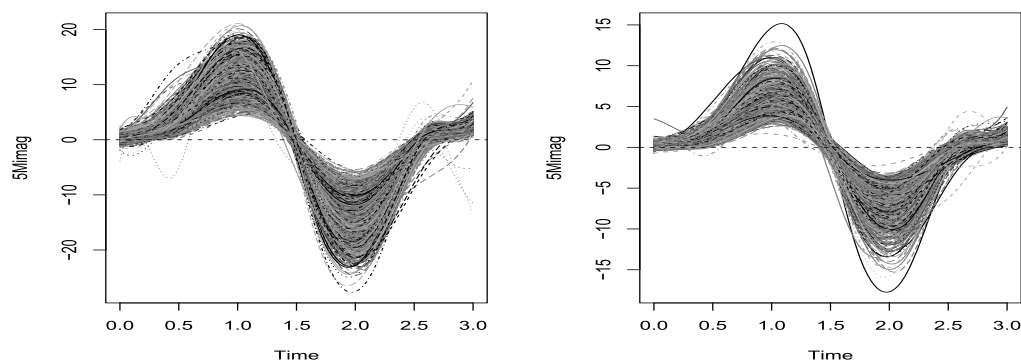


Fig. 8. Smoothed electrical (imaginary signal at 5 Mega hertz) curves of the population cancer cells per group. The right curves are the most softer (metastatic) group

Table 9. Results (mean relative squared errors of reconstruction; *MRSE* and corresponding standard deviation (in parentheses) using the 5 first principal components of each method over 50 replications) with a CBS sample of size $N = 400$ drawn in the population P_1 , $Q_1^* = 67.06\%$ and H^* varying.

	<i>MRSE</i> (in %)	
	<i>MFPCA</i> _{CBS}	<i>MFPCA</i>
$H^* = 0.15$	0.209 (0.020)	0.329 (0.022)
$H^* = 0.25$	0.199 (0.020)	0.336 (0.020)
$H^* = 0.50$	0.200 (0.019)	0.344 (0.022)

Table 10. Results (mean relative squared errors of reconstruction; $MRSE$ and corresponding standard deviation (in parentheses) using the 5 first principal components of each method over 50 replications) with a CBS sample of size $N = 400$ drawn in the population $P2$, $Q_2^* = 48.07\%$ and H^* varying.

	$MRSE$ (in %)	
	$MFPCA_{CBS}$	MFPCA
$H^* = 0.15$	0.224 (0.027)	0.350 (0.023)
$H^* = 0.25$	0.211 (0.020)	0.351 (0.020)
$H^* = 0.50$	0.202 (0.020)	0.350 (0.025)

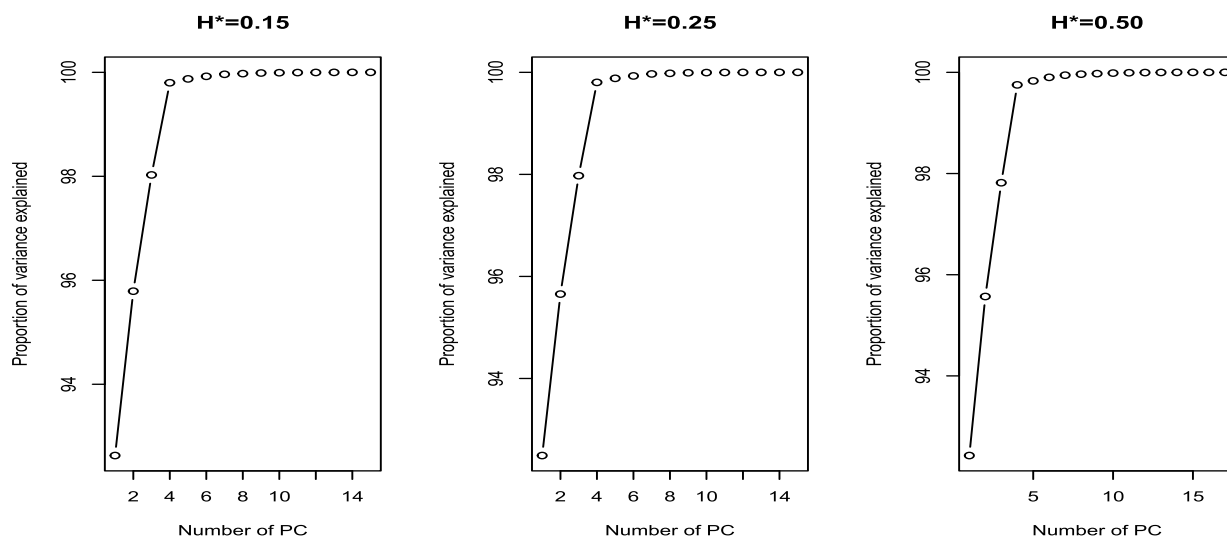


Fig. 9. Percentages of explained variance of the real data for the $MFPCA_{CBS}$ method for $P2$.

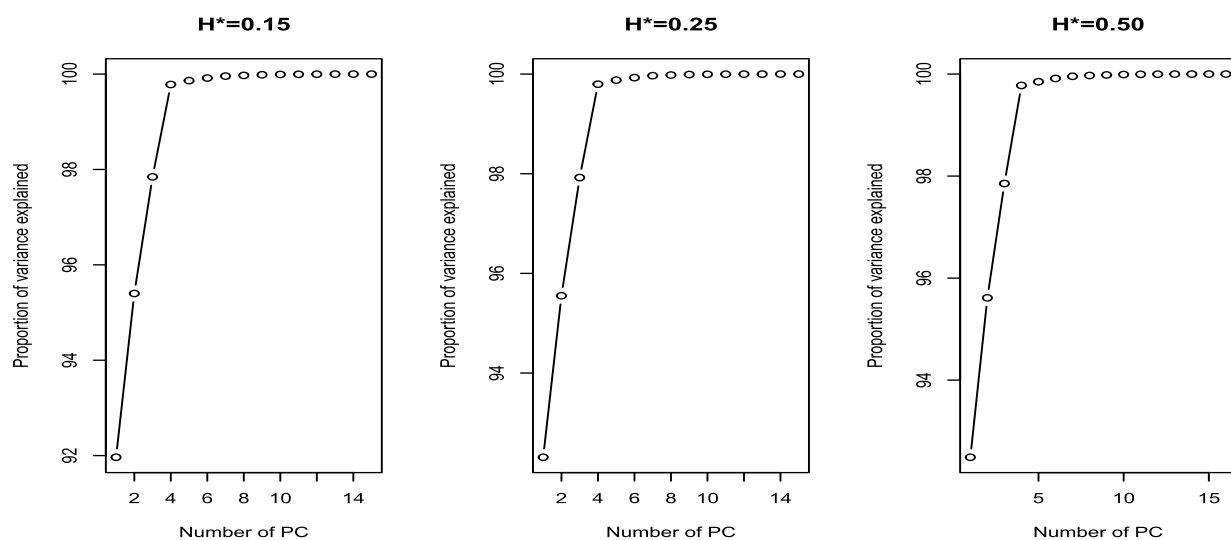


Fig. 10. Percentages of explained variance of the real data for the MFPCA method for P2.

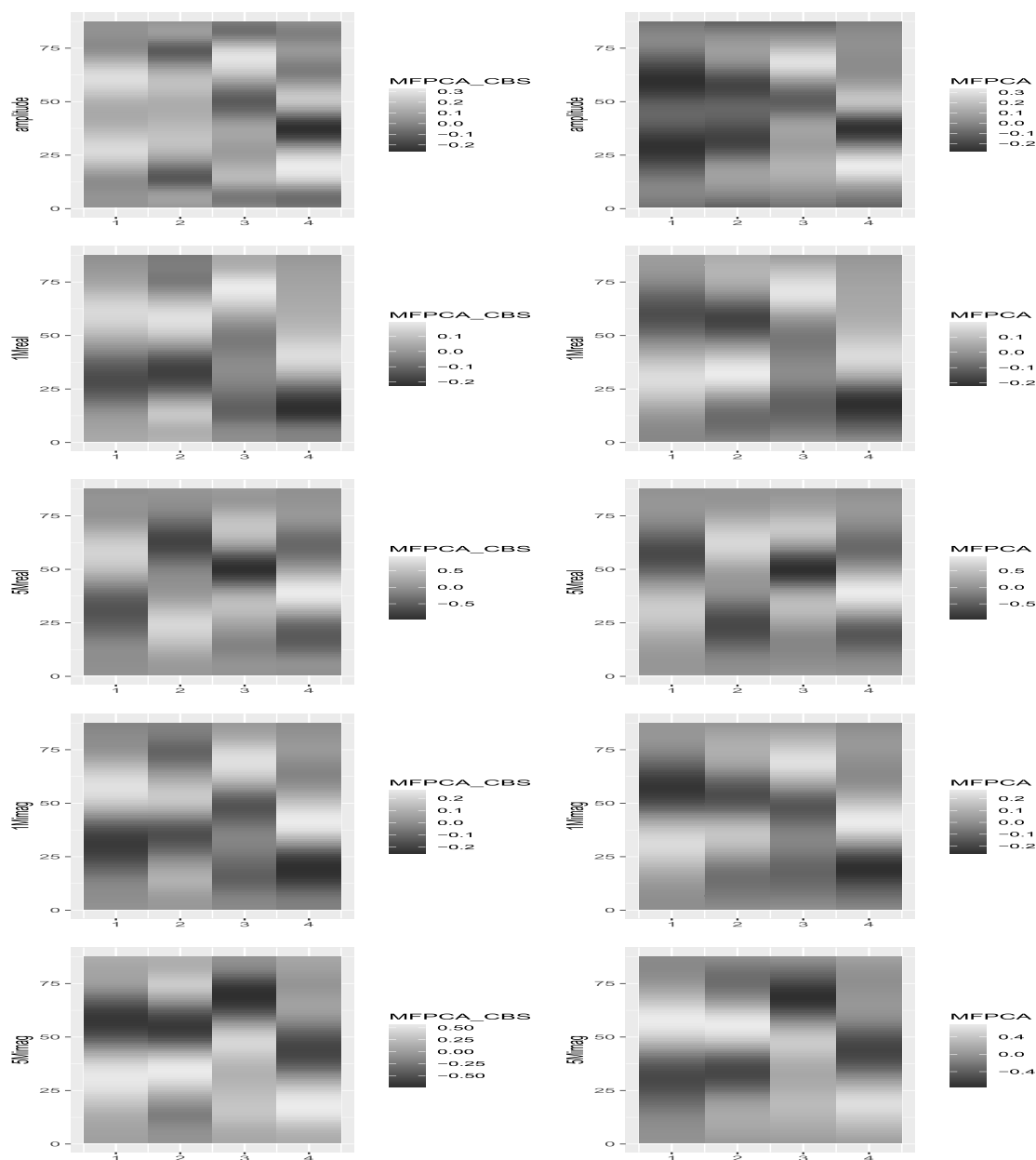


Fig. 11. Heatmaps of the first four estimated multivariate eigenfunctions for the two methods and each of the 5 components of the functional variable (population P2).

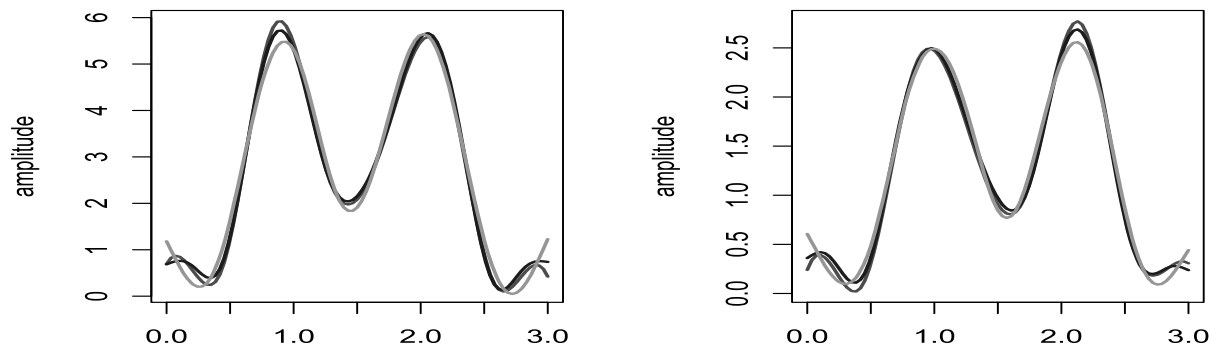


Fig. 12. Real amplitude data (curves in red) and reconstructed curves by $MFPCA_{CBS}$ method (blue curves) and the MFPCA method (green curve). The right curves are the most softer (metastatic) group, $H^* = 0.5$ (population P2).

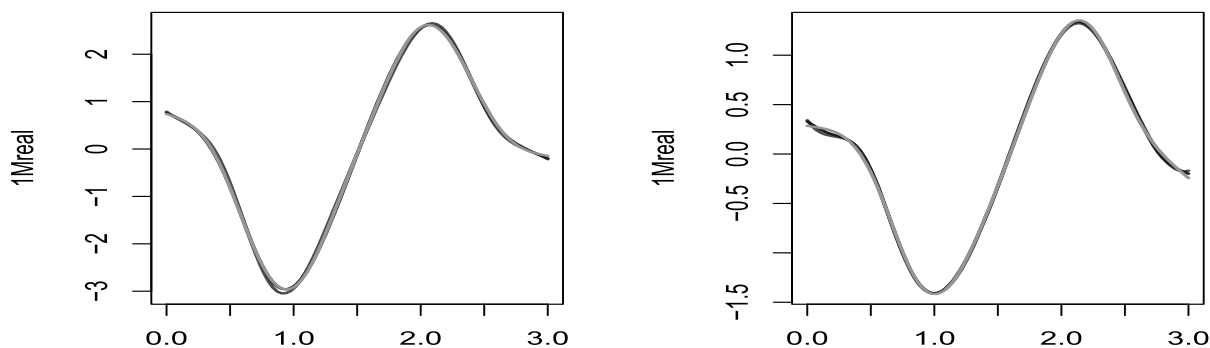


Fig. 13. Real data for signal (real) at 1 Mega hertz (curves in red) and reconstructed curves by $MFPCA_{CBS}$ method (blue curves) and the MFPCA method (green curve). The right curves are the most softer (metastatic) group, $H^* = 0.5$ (population P2).

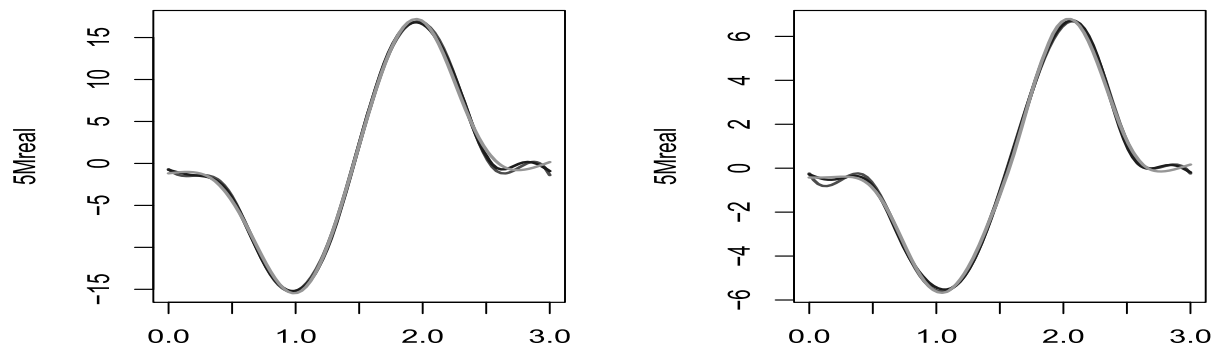


Fig. 14. Real data for signal (real) at 5 Mega hertz (curves in red) and reconstructed curves by $MFPCA_{CBS}$ method (blue curves) and the MFPCA method (green curve). The right curves are the most softer (metastatic) group, $H^* = 0.5$ (population P2).

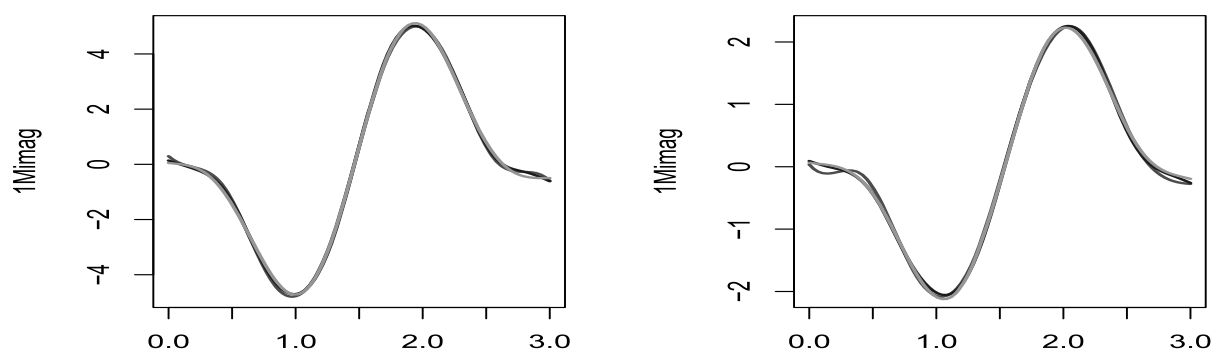


Fig. 15. Real data for signal (imaginary) at 1 Mega hertz (curves in red) and reconstructed curves by $MFPCA_{CBS}$ method (blue curves) and the MFPCA method (green curve). The right curves are the most softer (metastatic) group, $H^* = 0.5$ (population P2).

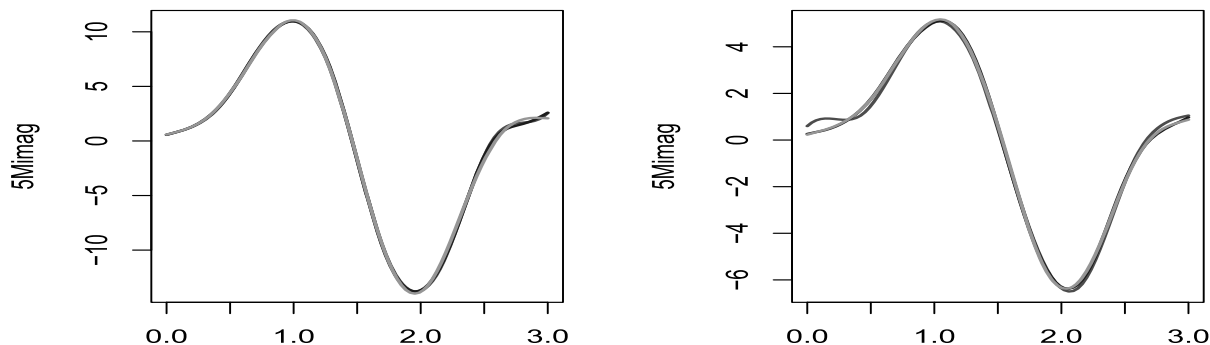


Fig. 16. Real data for signal (imaginary) at 5 Mega hertz (curves in red) and reconstructed curves by $MFPCA_{CBS}$ method (blue curves) and the $MFPCA$ method (green curve). The right curves are the most softer (metastatic) group, $H^* = 0.5$ (population P2).

Appendix.

This section contains the proofs details of the infinite sample properties. Let us first study the univariate elements X_p of X with covariance operator Γ_p and associated eigenvalues $\lambda_{p,j}$ and eigenfunctions $\phi_{p,j}$, $j = 1, \dots, J_p$. Let $\|\cdot\|_{op}$ the Hilbert–Schmidt norm for operators mapping $\mathcal{L}^2(I_p)$ to $\mathcal{L}^2(I_p)$. Let us present a preliminary lemma.

Lemma 1. (Cardot et al.(2010), p. 80). Under hypotheses (H1)-(H3), we have

$$\|\Gamma_p - \hat{\Gamma}_p\|_{op} = O_p(N^{-1/2})$$

This result shows the consistency of the covariance operator estimators of the functions X_p , $p = 1, \dots, P$. The following result states the convergence of the truncated eigen functions and scores to the true ones.

Proposition 2. (Happ and Greven (2018), Prop. 6 (Approximation Error))

The approximation error $\|X^{[J]} - X\|$ converges to 0 in probability for $J_1, \dots, J_p \rightarrow \infty$. For each $j \in \mathbb{N}$, $\lambda_j^{[J]}$ converges to λ_j including multiplicity and the total projection $\mathcal{P}_j^{[J]}$ of $\mathcal{H}$ onto the eigenspace of $\Gamma^{[J]}$ associated with $\lambda_j^{[J]}$ converges in norm to the total projection $\mathcal{P}_j$ of $\mathcal{H}$ onto the eigenspace of Γ associated with λ_j .

In particular, if λ_j and $\lambda_j^{[J]}$ both have multiplicity 1 with associated eigenfunctions ϕ_j and $\phi_j^{[J]}$, such that $\langle\langle\phi_j, \phi_j^{[J]}\rangle\rangle \geq 0$, then

$$\|\phi_j^{[J]} - \phi_j\| \rightarrow 0 \quad \text{for } J_1, \dots, J_p \rightarrow \infty.$$

The scores $\xi_j^{[J]} := \langle\langle X^{[J]}, \phi_j^{[J]} \rangle\rangle$ converge in probability to ξ_j for all $j \in \mathbb{N}$.

All nonzero eigenvalues λ_j are assumed to have multiplicity 1, then (see the above proposition) the eigenfunctions $\phi_j^{[J]}$ converge to ϕ_j , if $\langle\langle\phi_j, \phi_j^{[J]}\rangle\rangle \geq 0$.

Proof of Prop. 2. For $f \in \mathcal{H}$, $t \in \mathbf{I}$ and $p = 1, \dots, P$, the covariance operator $\Gamma^{[J]}$ associated with $X^{[J]}$ is given by

$$(\Gamma^{[J]}f)^{(p)}(t_p) = \sum_{q=1}^P \int_{I_q} \text{Cov}\left(X_q^{[J]}(s_q), X_p^{[J]}(t_p)\right) f_q(s_q) ds_q, \quad t_p \in I_p.$$

Let $C_{q,p}^{[J]}(s_q, t_p) := \text{Cov}\left(X_q^{[J]}(s_q), X_p^{[J]}(t_p)\right)$. Similarly to Happ and Greven (2018), using Mercer's Theorem [Mercer (1909)], the univariate Karhunen-Loève Theorem (e.g. Bosq (2000), Thm 1.5.) and (1), we have:

$$C_{p,p}^{[J]}(t_p, t_p) = \sum_{j=1}^{J_p} \lambda_{p,j} \phi_{p,j}(t_p)^2 \rightarrow \sum_{j=1}^{\infty} \lambda_{p,j} \phi_{p,j}(t_p)^2 = C_{p,p}(t_p, t_p) \quad \text{for } J_p \rightarrow \infty, t_p \in I_p,$$

$$\text{Var}\left(X_p(t_p) - X_p^{[J]}(t_p)\right) \rightarrow 0 \quad \text{for } J_p \rightarrow \infty,$$

$$\text{Var}(X_p(t_p)) = \text{Cov}(X_p(t_p), X_p(t_p)) = C_{p,p}(t_p, t_p) \leq \|C_{p,p}\|_{\infty} < \infty \quad \forall t_p \in I_p,$$

and

$$C_{q,p}^{[J]}(s_q, t_p) \rightarrow C_{q,p}(s_q, t_p) \quad \text{for } J_q, J_p \rightarrow \infty.$$

Then, it follows that $\Gamma^{[J]}$ is bounded and converges in norm to Γ for $J_1, \dots, J_p \rightarrow \infty$. Moreover, from the symmetry $C_{p,q}^{[J]}(t_p, s_q) = C_{q,p}^{[J]}(s_q, t_p)$, $\Gamma^{[J]}$ is self-adjoint. Now, we shall prove the convergence results for $\lambda_j^{[J]}$, $\phi_j^{[J]}$ and $\xi_j^{[J]}$. Γ being a compact operator, it is also bounded [Reed and Simon (1980), Chapter VI.5.]. As Γ and $\Gamma^{[J]}$ are both bounded, self-adjoint and such that $\Gamma^{[J]}$ converges in norm to Γ , the eigenvalues $\lambda_j^{[J]}$ of $\Gamma^{[J]}$ converge to the eigenvalues λ_j of Γ including multiplicity and the associated total projections converge in norm [Weidmann (1997)]. Since the difference of two projections is also a projection and the orthogonal projection is 1-lipschitzian, we have

$$\left\| \mathcal{P}_j f - \mathcal{P}_j^{[J]} f \right\| = \left\| \left(\mathcal{P}_j - \mathcal{P}_j^{[J]} \right) f \right\| \leq \|f\|.$$

Then,

$$\left\| \mathcal{P}_j - \mathcal{P}_j^{[J]} \right\|_{op} = \sup_{\|f\|=1} \left\| \mathcal{P}_j f - \mathcal{P}_j^{[J]} f \right\| < \infty.$$

Thus, $\mathcal{P}_j^{[J]}$ converges in norm to $\mathcal{P}_j$.

If the j -th eigenvalue has multiplicity 1, then the projections on the eigenspaces spanned by ϕ_j and $\phi_j^{[J]}$, respectively, are given by

$$\mathcal{P}_j f = \langle\langle f, \phi_j \rangle\rangle \phi_j, \quad \mathcal{P}_j^{[J]} f = \langle\langle f, \phi_j^{[J]} \rangle\rangle \phi_j^{[J]}, \quad f \in \mathcal{H}.$$

Without loss of generality one may choose the orientation of ϕ_j and $\phi_j^{[J]}$ such that $\langle\langle \phi_j, \phi_j^{[J]} \rangle\rangle \geq 0$. In this case, as ϕ_j and $\phi_j^{[J]}$ both have norm 1,

$$\begin{aligned} \left\| \mathcal{P}_j - \mathcal{P}_j^{[J]} \right\|_{op}^2 &\geq \left\| \mathcal{P}_j \phi_j - \mathcal{P}_j^{[J]} \phi_j^{[J]} \right\|^2 \\ &= \left\| \phi_j - \langle\langle \phi_j, \phi_j^{[J]} \rangle\rangle \phi_j^{[J]} \right\|^2, \\ &= \left(1 - \langle\langle \phi_j, \phi_j^{[J]} \rangle\rangle \right) \left(1 + \langle\langle \phi_j, \phi_j^{[J]} \rangle\rangle \right), \\ &\geq \frac{1}{2} \left\| \phi_j - \phi_j^{[J]} \right\|^2. \end{aligned}$$

Norm convergence of the total projections hence implies $\left\| \phi_j - \phi_j^{[J]} \right\| \rightarrow 0$ for $J_1, \dots, J_p \rightarrow \infty$. For fixed $j = 1, \dots, J$, it holds that,

$$\begin{aligned} \left| \lambda_j - \lambda_j^{[J]} \right| &= \left| \phi_j^\top \Gamma \phi_j - \phi_j^{[J]\top} \Gamma^{[J]} \phi_j^{[J]} \right|, \\ &\leq \left| (\phi_j - \phi_j^{[J]})^\top \Gamma \phi_j \right| + \left| \phi_j^{[J]\top} \Gamma (\phi_j - \phi_j^{[J]}) \right| + \left| \phi_j^{[J]\top} (\Gamma - \Gamma^{[J]}) \phi_j^{[J]} \right|, \\ &\leq \left\| \phi_j - \phi_j^{[J]} \right\| \cdot \lambda_j \left\| \phi_j \right\| + \left\| \phi_j^{[J]} \right\| \cdot \left\| \Gamma \right\|_{op} \cdot \left\| \phi_j - \phi_j^{[J]} \right\| + \left\| \Gamma - \Gamma^{[J]} \right\|_{op} \cdot \left\| \phi_j^{[J]} \right\|^2, \\ &= o(1). \end{aligned}$$

Thus, the eigenvalues $\lambda_j^{[J]}$ converge to the eigenvalues λ_j . To derive convergence of the scores $\xi_j^{[J]}$, note that for $\delta > 0$ and

$$k := \left(\frac{2}{\delta} \sum_{p=1}^P \|C_{p,p}\|_{\infty} \sigma(I_p) \right)^{1/2},$$

$$\begin{aligned} \mathbb{P}(\|X\| > k) &\stackrel{Chebychev}{\leq} \frac{\mathbb{E}(\|X\|^2)}{k^2}, \\ &\stackrel{Fubini}{\leq} \frac{1}{k^2} \sum_{p=1}^P \int_{I_p} \mathbb{E}(X_p^2(t_p)) dt_p \leq \frac{1}{k^2} \sum_{p=1}^P \int_{I_p} C_{p,p}(t_p, t_p) dt_p, \\ &\leq \frac{1}{k^2} \sum_{p=1}^P \int_{I_p} \|C_{p,p}\|_{\infty} dt_p \leq \frac{1}{k^2} \sum_{p=1}^P \|C_{p,p}\|_{\infty} \sigma(I_p) = \frac{\delta}{2} < \delta, \end{aligned}$$

i.e. the norm of X is bounded in probability ($\|X\| = O_p(1)$). Moreover,

$$\mathbb{E} \left[\left\| X - X^{[J]} \right\|^2 \right] \stackrel{Fubini}{=} \sum_{p=1}^P \int_{I_p} \mathbb{E} \left[|X_p(t_p) - X_p^{[J]}(t_p)|^2 \right] dt_p \rightarrow 0,$$

for $J_1, \dots, J_p$, as the expectation in the integral converges uniformly to 0 and is thus bounded (by univariate Karhunen-Loève Theorem). As I_p has finite measure, the overall integral converges to 0. Hence $\|X - X^{[J]}\|$ converges in the second mean to 0, thus $\|X - X^{[J]}\| = o_p(1)$. Finally, this leads to

$$\begin{aligned} |\xi_j - \xi_j^{[J]}| &= |\langle X, \phi_j \rangle - \langle X^{[J]}, \phi_j^{[J]} \rangle| = |\langle X, \phi_j - \phi_j^{[J]} \rangle + \langle X - X^{[J]}, \phi_j^{[J]} \rangle|, \\ &\leq |\langle X, \phi_j - \phi_j^{[J]} \rangle| + |\langle X - X^{[J]}, \phi_j^{[J]} \rangle|, \\ &\stackrel{Cauchy-Schwarz}{\leq} \|X\| \|\phi_j - \phi_j^{[J]}\| + \|X - X^{[J]}\| \|\phi_j^{[J]}\|, \\ &= o_p(1), \end{aligned}$$

i.e. $\xi_j^{[J]}$ converges in probability to ξ_j .

Lemma 2. Under the assumptions of Prop. 1 it holds that

$$\nu_{\max}(\mathbf{Z} - \hat{\mathbf{Z}}) \leq O_p \left(u_{\max} N^{-1/2} \max(1, \Delta_J) \right),$$

$u_{\max} = \max_{p=1, \dots, P} J_p$ and $\Delta_J := \max_{p=1, \dots, P} \Delta_{p, J_p}$.

Proof of Lemma 2. For $p, q = 1, \dots, P$ and $f \in \mathcal{L}^2(I_p)$ define the bounded operator $\Gamma_{p,q} : \mathcal{L}^2(I_p) \rightarrow \mathcal{L}^2(I_q)$ via

Journal home page: <http://www.jafristat.net>, www.projecteuclid.org/euclid.as,
www.ajol.info/afst

This issue [Vol 17 (4)] was open on October 1, 2022 and closed on September 30, 2023

$$(\Gamma_{p,q}f)(t) := \int_{I_p} \text{Cov}(X_p(s), X_q(t))f(s)ds = \int_{I_p} C_{p,q}(s,t)f(s)ds.$$

Analogously, define $\hat{\Gamma}_{p,q} : \mathcal{L}^2(I_p) \rightarrow \mathcal{L}^2(I_q)$ by

$$(\hat{\Gamma}_{p,q}f)(t) := \int_{I_p} \widehat{\text{Cov}}(X_p(s), X_q(t))f(s)ds = \int_{I_p} \hat{C}_{p,q}(s,t)f(s)ds,$$

with

$$\hat{C}_{p,q}(s,t) := \widehat{\text{Cov}}(X_p(s), X_q(t)) = \frac{1}{N-1} \sum_{n=1}^N X_p^{(n)}(s)X_q^{(n)}(t) \left(\frac{Q^*}{H^*} \mathbb{I}(Y_n = 1) + \frac{1-Q^*}{1-H^*} \mathbb{I}(Y_n = 0) \right).$$

Since the $X^{(n)}$ are independent copies of the process X , it holds that

$$\begin{aligned} & \mathbb{E} \left[\int_{I_p} \int_{I_q} \left(C_{p,q}(s,t) - \hat{C}_{p,q}(s,t) \right)^2 ds dt \right], \\ &= \mathbb{E} \left[\int_{I_p} \int_{I_q} \left(C_{p,q}(s,t) - \frac{1}{N-1} \sum_{n=1}^N X_p^{(n)}(s)X_q^{(n)}(t) \left(\frac{Q^*}{H^*} \mathbb{I}(Y_n = 1) + \frac{1-Q^*}{1-H^*} \mathbb{I}(Y_n = 0) \right) \right)^2 ds dt \right], \\ &= \mathbb{E} \left[\frac{1}{(N-1)^2} \int_{I_p} \int_{I_q} \sum_{n_1=1}^N \sum_{n_2=1}^N \left(\frac{N-1}{N} C_{p,q}(s,t) - X_p^{(n_1)}(s)X_q^{(n_1)}(t) \left(\frac{Q^*}{H^*} \mathbb{I}(Y_{n_1} = 1) + \frac{1-Q^*}{1-H^*} \mathbb{I}(Y_{n_1} = 0) \right) \right) \right. \\ &\quad \left. \left(\frac{N-1}{N} C_{p,q}(s,t) - X_p^{(n_2)}(s)X_q^{(n_2)}(t) \left(\frac{Q^*}{H^*} \mathbb{I}(Y_{n_2} = 1) + \frac{1-Q^*}{1-H^*} \mathbb{I}(Y_{n_2} = 0) \right) \right) ds dt \right], \\ &= \frac{1}{(N-1)^2} \int_{I_p} \int_{I_q} \sum_{n_1=1}^N \mathbb{E} \left[\left(\frac{N-1}{N} C_{p,q}(s,t) - X_p^{(n_1)}(s)X_q^{(n_1)}(t) \left(\frac{Q^*}{H^*} \mathbb{I}(Y_{n_1} = 1) + \frac{1-Q^*}{1-H^*} \mathbb{I}(Y_{n_1} = 0) \right) \right)^2 \right] ds dt, \\ &\stackrel{\text{with } \varepsilon \text{ defined in (H3)}}{\leq} \frac{N}{(N-1)^2} \int_{I_p} \int_{I_q} \mathbb{E} \left[\left(\frac{N-1}{N} C_{p,q}(s,t) - \varepsilon X_p(s)X_q(t) \right)^2 \right] ds dt, \\ &= \frac{N}{(N-1)^2} \int_{I_p} \int_{I_q} \left[\frac{(N-1)^2}{N^2} C_{p,q}^2(s,t) - 2\varepsilon \frac{N-1}{N} C_{p,q}^2(s,t) + \varepsilon^2 \mathbb{E}(X_p^2(s)X_q^2(t)) \right] ds dt, \\ &\leq \frac{N}{(N-1)^2} \int_{I_p} \int_{I_q} \left[C_{p,q}^2(s,t) - 2\varepsilon \frac{N-1}{N} C_{p,q}^2(s,t) + \varepsilon^2 \mathbb{E}(X_p^2(s)X_q^2(t)) \right] ds dt, \\ &= \frac{N}{(N-1)^2} \int_{I_p} \int_{I_q} \left[\varepsilon^2 \mathbb{E}(X_p^2(s)X_q^2(t)) + (1-2\varepsilon \frac{N-1}{N}) C_{p,q}^2(s,t) \right] ds dt, \\ &= \frac{N}{(N-1)^2} \int_{I_p} \int_{I_q} \left[\varepsilon^2 \mathbb{E}(X_p^2(s)X_q^2(t)) + (1-2\varepsilon) C_{p,q}^2(s,t) + \frac{2\varepsilon}{N} C_{p,q}^2(s,t) \right] ds dt, \\ &= O\left(\frac{N}{(N-1)^2}\right) + O\left(\frac{1}{(N-1)^2}\right) = O\left(\frac{1}{N}\right). \end{aligned}$$

The last step follows from the fact that the integral term does not depend on N and is finite by assumption (H5) and condition (1) for $C_{p,q}$ in Section 2.2. This implies

$$\begin{aligned}\|\Gamma_{p,q} - \hat{\Gamma}_{p,q}\|_{\text{op}} &= \sup_{\|f\|=1} \left(\int_{I_q} \left(\Gamma_{p,q} f - \hat{\Gamma}_{p,q} f \right)^2(t) dt \right)^{1/2}, \\ &= \sup_{\|f\|=1} \left(\int_{I_q} \left(\int_{I_p} \left(C_{p,q}(s,t) - \hat{C}_{p,q}(s,t) \right) f(s) ds \right)^2 dt \right)^{1/2}, \\ &\stackrel{\text{Cauchy Schwarz}}{\leq} \sup_{\|f\|=1} \left(\int_{I_q} \left(\int_{I_p} \left(C_{p,q}(s,t) - \hat{C}_{p,q}(s,t) \right)^2 ds \right) \|f\|^2 dt \right)^{1/2}, \\ &\leq \left(\int_{I_q} \int_{I_p} \left(C_{p,q}(s,t) - \hat{C}_{p,q}(s,t) \right)^2 ds dt \right)^{1/2} = O_p(N^{-1/2}).\end{aligned}$$

Recall that

$$Z_{lk}^{(pq)} = \text{Cov}(\xi_{p,l}, \xi_{q,k})$$

and

$$\hat{Z}_{lk}^{(pq)} = \frac{1}{N-1} \sum_{n=1}^N \hat{\xi}_{p,l}^{(n)} \hat{\xi}_{q,k}^{(n)} \left(\frac{Q^*}{H^*} \mathbb{I}(Y_n = 1) + \frac{1-Q^*}{1-H^*} \mathbb{I}(Y_n = 0) \right)$$

for $l = 1, \dots, J_p$, $k = 1, \dots, J_q$, $p, q = 1, \dots, P$. As $\mathbf{Z}$ and $\hat{\mathbf{Z}}$ are both symmetric matrices in $\mathbb{R}^{J \times J}$ it holds (cf. Chapter 3.7 in Horn and Johnson (1991))

$$\nu_{\max}(\mathbf{Z} - \hat{\mathbf{Z}}) \leq \max_{p=1, \dots, P} \max_{l=1, \dots, J_p} \sum_{q=1}^P \sum_{k=1}^{J_q} \left| Z_{lk}^{(pq)} - \hat{Z}_{lk}^{(pq)} \right|. \quad (8)$$

Let now $l = 1, \dots, J_p$, $k = 1, \dots, J_q$, $p, q = 1, \dots, P$ be fixed. By assumption (H7), we have

$$\begin{aligned}
 \left| Z_{lk}^{(pq)} - \hat{Z}_{lk}^{(pq)} \right| &= \left| \text{Cov}(\xi_{p,l}, \xi_{q,k}) - \frac{1}{N-1} \sum_{n=1}^N \xi_{p,l}^{(n)} \xi_{q,k}^{(n)} \left(\frac{Q^*}{H^*} \mathbb{I}(Y_n = 1) + \frac{1-Q^*}{1-H^*} \mathbb{I}(Y_n = 0) \right) \right|, \\
 &\stackrel{(H7)}{=} \left| \text{Cov}(\langle X_p, \phi_{p,l} \rangle, \langle X_q, \phi_{q,k} \rangle) \right. \\
 &\quad \left. - \frac{1}{N-1} \sum_{n=1}^N \langle X_p^{(n)}, \hat{\phi}_{p,l} \rangle \cdot \langle X_q^{(n)}, \hat{\phi}_{q,k} \rangle \left(\frac{Q^*}{H^*} \mathbb{I}(Y_n = 1) + \frac{1-Q^*}{1-H^*} \mathbb{I}(Y_n = 0) \right) \right|, \\
 &= \left| \int_{I_p} \int_{I_q} \mathbb{E} \left(X_p(s) X_q(t) \right) \phi_{p,l}(s) \phi_{q,k}(t) \, ds \, dt \right. \\
 &\quad \left. - \int_{I_p} \int_{I_q} \frac{1}{N-1} \sum_{n=1}^N \left(X_p^{(n)}(s) X_q^{(n)}(t) \right) \hat{\phi}_{p,l}(s) \hat{\phi}_{q,k}(t) \left(\frac{Q^*}{H^*} \mathbb{I}(Y_n = 1) \right. \right. \\
 &\quad \left. \left. + \frac{1-Q^*}{1-H^*} \mathbb{I}(Y_n = 0) \right) \, ds \, dt \right|, \tag{9} \\
 &= \left| \int_{I_p} \int_{I_q} \text{Cov} \left(X_p(s) X_q(t) \right) \phi_{p,l}(s) \phi_{q,k}(t) - \widehat{\text{Cov}} \left(X_p(s), X_q(t) \right) \hat{\phi}_{p,l}(s) \hat{\phi}_{q,k}(t) \, ds \, dt \right| \\
 &\leq \left| \int_{I_p} \int_{I_q} C_{p,q}(s, t) \left[\phi_{p,l}(s) \phi_{q,k}(t) - \hat{\phi}_{p,l}(s) \hat{\phi}_{q,k}(t) \right] \, ds \, dt \right| \\
 &\quad + \left| \int_{I_p} \int_{I_q} \left[C_{p,q}(s, t) - \hat{C}_{p,q}(s, t) \right] \hat{\phi}_{p,l}(s) \hat{\phi}_{q,k}(t) \, ds \, dt \right|, \\
 &\leq \int_{I_p} \int_{I_q} \left| C_{p,q}(s, t) \left[\phi_{p,l}(s) \phi_{q,k}(t) - \hat{\phi}_{p,l}(s) \hat{\phi}_{q,k}(t) \right] \right| \, ds \, dt \\
 &\quad + \left| \int_{I_q} \left((\Gamma_{p,q} - \hat{\Gamma}_{p,q}) \hat{\phi}_{p,l} \right) (t) \hat{\phi}_{q,k}(t) \, dt \right|, \\
 &\leq \int_{I_p} \int_{I_q} C_{p,p}(s, s)^{1/2} C_{q,q}(t, t)^{1/2} \left| \phi_{p,l}(s) \phi_{q,k}(t) - \hat{\phi}_{p,l}(s) \hat{\phi}_{q,k}(t) \right| \, ds \, dt \\
 &\quad + \int_{I_q} \left| \left((\Gamma_{p,q} - \hat{\Gamma}_{p,q}) \hat{\phi}_{p,l} \right) (t) \hat{\phi}_{q,k}(t) \right| \, ds \, dt, \\
 &\leq \|C_{p,p}\|_{\mathcal{L}}^{1/2} \|C_{q,q}\|_{\mathcal{L}}^{1/2} \left(\int_{I_p} \int_{I_q} |\phi_{p,l}(s)| |\phi_{q,k}(t) - \hat{\phi}_{q,k}(t)| \, ds \, dt + \int_{I_p} \int_{I_q} |\phi_{p,l}(s) - \hat{\phi}_{p,l}(s)| |\hat{\phi}_{q,k}(t)| \, ds \, dt \right) \\
 &\quad + \left\| (\Gamma_{p,q} - \hat{\Gamma}_{p,q}) \hat{\phi}_{p,l} \right\| \cdot \left\| \hat{\phi}_{q,k} \right\|, \\
 &\leq \|C_{p,p}\|_{\mathcal{L}}^{1/2} \|C_{q,q}\|_{\mathcal{L}}^{1/2} \left(\left[\int_{I_p} \int_{I_q} |\phi_{p,l}(s)|^2 \, ds \, dt \times \int_{I_p} \int_{I_q} |\phi_{q,k}(t) - \hat{\phi}_{q,k}(t)|^2 \, ds \, dt \right]^{1/2} \right. \\
 &\quad \left. + \left[\int_{I_p} \int_{I_q} |\phi_{p,l}(s) - \hat{\phi}_{p,l}(s)|^2 \, ds \, dt \times \int_{I_p} \int_{I_q} |\hat{\phi}_{q,k}(t)|^2 \, ds \, dt \right]^{1/2} \right) + \left\| \Gamma_{p,q} - \hat{\Gamma}_{p,q} \right\|_{\text{op}} \cdot \left\| \hat{\phi}_{q,k} \right\|,
 \end{aligned}$$

$$\begin{aligned}
 &= (\|C_{p,p}\|_{\infty} \|C_{q,q}\|_{\infty} \sigma(I_p) \sigma(I_q))^{1/2} \left(\|\phi_{p,l}\| \|\phi_{q,k} - \hat{\phi}_{q,k}\| + \|\phi_{p,l} - \hat{\phi}_{p,l}\| \|\hat{\phi}_{q,k}(t)\| \right) \\
 &\quad + \|\Gamma_{p,q} - \hat{\Gamma}_{p,q}\|_{\text{op}}, \\
 &= O_p \left(\Delta_{q,J_q} N^{-1/2} \right) + O_p \left(\Delta_{p,J_p} N^{-1/2} \right) + O_p \left(N^{-1/2} \right), \\
 &= O_p \left(\max \left\{ \Delta_{q,J_q} N^{-1/2}, \Delta_{p,J_p} N^{-1/2}, N^{-1/2} \right\} \right).
 \end{aligned}$$

The rate for $\phi_{q,k}$ and $\phi_{p,l}$ in the last steps is shown at the beginning of the proof of Proposition 1. Finally, equation (8) gives

$$\begin{aligned}
 \nu_{\max}(\mathbf{Z} - \hat{\mathbf{Z}}) &\leq \max_{p=1,\dots,P} \max_{l=1,\dots,J_p} \sum_{q=1}^P \sum_{k=1}^{J_q} O_p \left(\max \left\{ \Delta_{q,J_q} N^{-1/2}, \Delta_{p,J_p} N^{-1/2}, N^{-1/2} \right\} \right), \\
 &\leq \max_{p=1,\dots,P} \max_{l=1,\dots,J_p} \sum_{q=1}^P \sum_{k=1}^{J_q} O_p \left(\max \left\{ \Delta_J N^{-1/2}, \Delta_J N^{-1/2}, N^{-1/2} \right\} \right), \\
 &\leq O(u_{\max}) O_p \left(\max \left\{ \Delta_J N^{-1/2}, N^{-1/2} \right\} \right) \leq O_p \left(u_{\max} N^{-1/2} \max(1, \Delta_J) \right),
 \end{aligned}$$

with $u_{\max} = \max_{p=1,\dots,P} J_p$ and $\Delta_J := \max_{p=1,\dots,P} \Delta_{p,J_p}$.

Proof of Prop. 1. Under assumption (H6), using Lemma 4.3. in Bosq (2000) and the convention $\lambda_0 := \infty$, we get: for $j = 1, \dots, J_p$,

$$\begin{aligned}
 \|\phi_{p,j} - \hat{\phi}_{p,j}\| &\leq 2^{3/2} \max [(\lambda_{p,j-1} - \lambda_{p,j})^{-1}, (\lambda_{p,j} - \lambda_{p,j+1})^{-1}] \|\Gamma_p - \hat{\Gamma}_p\|_{\text{op}} \\
 &= 2^{3/2} \Delta_{p,J_p} \|\Gamma_p - \hat{\Gamma}_p\|_{\text{op}} = O_p \left(\Delta_{p,J_p} N^{-1/2} \right).
 \end{aligned}$$

Based on this result, Lemma 2 states that $\nu_{\max}(\mathbf{Z} - \hat{\mathbf{Z}}) \leq O_p(u_{\max} N^{-1/2} \max(1, \Delta_J))$ with $u_{\max} = \max_{p=1,\dots,P} J_p$ and $\Delta_J := \max_{p=1,\dots,P} \Delta_{p,J_p}$. Then, similarly to Happ and Greven (2018), we deduce the announced convergence results using Corollary 1 in Yu *et al.*(2015).

Acknowledgment. The authors would like to thank the editorial team and the referees for their valuable comments. The first author thanks the International Mathematical Union (IMU) for grant number "IMU-BGF-2020-22" awarded for research purposes, as well as CEA-SMIA Benin.

We would like to thank the editorial board for their handling of our paper. We are grateful for the opportunity to share our research with the academic community, thanks to their support.

References

- Abadie, A.; Chingos, M.M. and West, M.R., 2018. Endogenous stratification in randomized experiments. *Review of Economics and Statistics*, 100(4), 567-580.
- Ahmadian, B.; Mjujamba, D.; Gerbedoen, J.-C.; Kumemura, M.; Fujita, H.; Collard, D.; Dabo-Niang, S.; Lagadec, C. and Tarhan, M.C., 2022. Single cell classification using statistical learning on mechanical properties measured by MEMS tweezer. *MEMS 2022*, 9-13.
- Ahmed, M.-S.; Attouch, M.K. and Dabo-Niang, S., 2018. Binary functional linear models under choice-based sampling. *Econometrics and statistics*, 7, 134-152.
- Berrendero, J.; Justel, A. and Svarc, M., 2011. Principal components for multivariate functional data. *Computational Statistics & Data Analysis*, 55(9), 2619-2634.
- Bosq, D.(2000) *Linear processes in function spaces: theory and applications*. Publisher: Springer Science & Business Media. 1rst ed.
- Cardot, H. and Josserand, E.(2011) Horvitz-Thompson estimators for functional data: Asymptotic confidence bands and optimal allocation for stratified sampling. *Biometrika*. Vol 98(1), pp.107-118.
- Cardot, H.; Chaouch, M.; Goga, C. and Labruère, C., 2010. Properties of design-based functional principal components analysis. *Journal of statistical planning and inference*, 140(1), 75-91.
- Chiou, J.-M.; Chen, Y.-T. and Yang, Y.-F., 2014. Multivariate functional principal component analysis: A normalization approach. *Statistica Sinica*, 1571-1596.
- Cosslett, S.R.(1981) *Efficient estimation of discrete-choice models*. Structural analysis of discrete data with econometric applications. 3, 51-111.
- Cosslett, S.R.(1981) Maximum likelihood estimator for choice-based samples. **Econometrica: Journal of the Econometric Society**. 1289-1316.
- Crainiceanu, C.; Reiss, P.; Goldsmith, J.; Huang, L.; Huo, L.; and Scheipl, F., 2014. Regression with functional data. *R package version 3.0. 1*.
- Didelez, V.; Kreiner, S. and Keiding, N., 2010. Graphical models for inference under outcome-dependent sampling. *Statistical Science*, 25(3), 368-387.
- Doll, R. and Hill, A.B.(1950) Smoking and carcinoma of the lung. *British medical journal*. Vol 2(4682), pp.739.
- Ferraty, F. and Vieu, P.(2006) Nonparametric functional data analysis: theory and practice. Publisher: Springer Science & Business Media.
- Golovkine, S.; Klutchnikoff, N. and Patilea, V., 2022. Clustering multivariate functional data using unsupervised binary trees. *Computational Statistics and Data Analysis*, 168, 107376.
- Happ, C. and Greven, S.(2018) Multivariate functional principal component analysis for data observed on different (dimensional) domains. *Journal of the American Statistical Association*. Vol 113(522), pp.649-659.
- Happ, C.(2017) *funData: An S4 class for functional data*. R package version.
- Happ, C.(2017) *MFPCA: multivariate functional principal component analysis for data observed on different dimensional domains*., R package version.
- Horn, R.A. and Johnson, C.R.(1991) Topics in Matrix Analysis. Publisher: Cambridge University Press. 3rd ed.
- Horváth, L. and Kokoszka, P.(2012) Inference for Functional Data with Applications. Publisher: Springer-Verlag.
- Hynes, S. and Greene, W.(2013) A panel travel cost model accounting for endogenous stratification and truncation: a latent class approach. *Land Economics*. Vol 89(1), pp.177-192.
- Imbens, G.W.(1992) *An efficient method of moments estimator for discrete choice models with choice-based sampling*. Econometrica: Journal of the Econometric Society. 1187-1214.

- Isaki, C.T. and Fuller, W.A.(1982) Survey design under the regression superpopulation model. *Journal of the American Statistical Association*. Vol 77(377), pp.89-96.
- Jacques, J. and Preda, C.(2014) Model-based clustering for multivariate functional data. *Computational Statistics and Data Analysis*. Vol 71, pp.92-106.
- Koner, S. and Staicu, A.M.(2023) Second-Generation Functional Data. *Annual Review of Statistics and Its Application*. Vol 10, pp.547-572.
- Langholz, B.(2010) Case-control studies= odds ratios: blame the retrospective model. *Epidemiology*. 21(1), 10-12 .
- Manski, C.F. and Lerman, S.R.(1977) The estimation of choice probabilities from choice based samples. *Econometrica: Journal of the Econometric Society*. pp.1977-1988.
- Manski, C.F. and McFadden, D.(1981) Alternative estimators and sample designs for discrete choice analysis. *Structural analysis of discrete data with econometric applications*. Vol 2, pp.2-50.
- Martinez, D.; Papuzinski, C.; Stojanova, J. and Arancibia, M., 2019. General concepts in biostatistics and clinical epidemiology: observational studies with case-control design. *Medwave*, 19(10), e7716.
- Mercer, J.(1909) *Functions of positive and negative type, and their connection the theory of integral equations..* Philosophical transactions of the royal society of London. Series A. 209(441-458), 415-446.
- Müller, H.-G. and Stadtmüller, U.(2005) Generalized functional linear models. *The Annals of Statistics*. Vol 33(2), pp.774-805.
- Ramvalho, E.A.(2002) *Regression models for choice-based samples with misclassification in the response variable*. *Journal of Econometrics*, 106(1), 171-201.
- Ramsay, J.O.; Wickham, H.; Ramsay, M.J. and deSolve, S., 2022. Package 'fda'.
- Ramsay, J.O. and Silverman, B.W.(2005) Principal components analysis for functional data. *Functional data analysis*. pp.147-172.
- Reed, M. and Simon, B.(1980) *Methods of Modern Mathematical Physics I: Functional Analysis*. Publisher: Acad Press, San Diego, USA. 3rd ed.
- Saporta, G.(1981) *Méthodes exploratoires d'analyse de données temporelles*. Cahiers du Bureau universitaire de recherche opérationnelle Série Recherche. 37, 7-194.
- Schmutz, A.; Jacques, J.; Bouveyron, C.; Cheze, Laurence and Martin, P., 2020. Clustering multivariate functional data in group-specific functional subspaces. *Computational Statistics*, 1-31.
- Takayama, Y.; Perret, G.; Kumemura, M.; Ataka, M.; Meignan, S.; Karsten, S.L.; Fujita, H.; Collard, D.; Lagadec, C. and Tarhan, M.C., 2018. Developing a MEMS device with built-in microfluidics for biophysical single cell characterization. *Multidisciplinary Digital Publishing Institute*, 9(6), 275.
- Wang, J.-L.; Chiou, J.-M. and Müller, H.-G., 2016. Functional data analysis. *Annual Review of Statistics and Its Applications*, 3, 257-295.
- Weidmann, J.(1997) *Strong operator convergence and spectral theory of ordinary differential operators..* Univ. Iagel. Acta. Math-. 1208, 153-163.
- Yamahata, C.; Collard, D.; Legrand, B.; Takekawa, T.; Kumemura, M.; Hashiguchi, G. and Fujita, H., 2008. Silicon nanotweezers with subnanometer resolution for the micromanipulation of biomolecules. *Journal of microelectromechanical systems*, 17(3), 623-631.
- Yao, F.; Müller, H.-g. and Wang, J.-L., 2005. Functional data analysis for sparse longitudinal data. *Journal of the American statistical association*, 100(470), 577-590.
- Yari, A.; Yousefi K.H.; Zarezadeh Y.; Ardalan A.; Soufi B.M.; Ostadtaghizadeh A. and Motlagh E.M., 2021. Behavioral, health-related and demographic risk factors of death in floods: A case-control study. *PLoS one*, 16(12), e0262005.
- Yu, Y.; Wang, T. and Samworth, R.J., 2015. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2), 315-323.