Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Robust estimators under a functional common principal components model



^a CONICET, Argentina

^b CITEDEF, Argentina

^c Universidad de Buenos Aires, Argentina

^d Universidad Nacional de San Martín, Argentina

ARTICLE INFO

Article history: Received 23 February 2016 Received in revised form 27 August 2016 Accepted 27 August 2016 Available online 4 September 2016

Keywords: Common principal component model Fisher-consistency Functional data analysis Outliers Projection-pursuit Robust estimation

ABSTRACT

When dealing with several populations of functional data, equality of the covariance operators is often assumed even when seeking for a lower-dimensional approximation to the data. Usually, if this assumption does not hold, one estimates the covariance operator of each group separately, which leads to a large number of parameters. As in the multivariate setting, this is not satisfactory since the covariance operators may exhibit some common structure, as is, for instance, the assumption of common principal directions. The existing procedures to estimate the common directions are sensitive to atypical observations. For that reason, robust projection-pursuit estimators for the common directions under a common principal component model are considered. A numerical method to compute the first directions is also provided. Under mild conditions, consistency results are obtained. A Monte Carlo study is performed to compare the finite sample behaviour of the estimators based on robust scales and on the standard deviation. The usefulness of the proposed approach is illustrated on a real data set.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

The common principal components model introduced by Flury (1984) for *p*th dimensional data, generalizes proportionality of the covariance matrices by allowing the matrices to have different eigenvalues but identical eigenvectors, that is, $\Sigma_i = \beta \Lambda_i \beta^{\text{T}}$, i = 1, ..., k, where Λ_i are diagonal matrices and β is the orthogonal matrix of the common eigenvectors. This model can be viewed as a generalization of principal components to *k* groups, since the principal transformation is identical in all populations considered while the variances associated with them vary among groups. In biometric applications, principal components are frequently interpreted as independent factors determining the growth, size or shape of an organism. It seems therefore reasonable to consider a model in which the same factors arise in different, but related species. The common principal components model clearly serves this purpose.

In this paper, we go further and consider several populations of functional data instead of finite-dimensional ones. As in the *p*-dimensional case, in many situations, one collects functional data $X_{i,1}, \ldots, X_{i,n_i}$ from *k* independent samples with mean μ_i and different covariance operators Γ_i which may exhibit some common structure to be taken into account in the

E-mail addresses: lbali@dm.uba.ar (J.L. Bali), gboente@dm.uba.ar (G. Boente).

http://dx.doi.org/10.1016/j.csda.2016.08.017 0167-9473/© 2016 Elsevier B.V. All rights reserved.







^{*} Correspondence to: Departamento de Matemáticas, FCEyN, UBA, Ciudad Universitaria, Pabellón 1, Buenos Aires, C1428EHA, Argentina. Fax: +54 11 45763335.

estimation procedure. The simplest generalization of equal covariance operators consists of assuming their proportionality, i.e., $\Gamma_i = \rho_i \Gamma_1$, for i = 1, ..., k and $\rho_1 = 1$. On the other hand, a natural extension of functional principal components to several populations, which also corresponds to a generalization to the functional setting of the common principal components model introduced by Flury (1984), is to assume that the covariance operators Γ_i have common eigenfunctions ϕ_i but different eigenvalues $\lambda_{i,i}$, i.e.,

$$\Gamma_i = \sum_{j=1}^{\infty} \lambda_{i,j} \phi_j \otimes \phi_j, \tag{1}$$

where to identify the directions, we assume that the eigenvalues of the first population are ordered in decreasing order, that is, $\lambda_{1,1} \ge \lambda_{1,2} \ge \cdots \ge \lambda_{1,j} \ge \lambda_{1,j+1} \cdots$. This model is usually denoted the functional common principal component model (FCPC) and provides a framework for analysing different population data that share their main modes of variation ϕ_1, ϕ_2, \ldots using a parsimonious approach. When the eigenvalues preserve the order across populations, i.e., if

$$\lambda_{i,1} \ge \lambda_{i,2} \ge \dots \ge \lambda_{i,j} \ge \lambda_{i,j+1} \dots, \quad \text{for } i = 1, \dots, k$$
⁽²⁾

as assumed, for instance, in Benko et al. (2009) and Boente et al. (2010), the common directions will represent, as in the one-population setting, the main modes of variation for each population. A more general setting than (2), is to assume that the largest d eigenvalues may not preserve the order among populations, that is,

$$\lambda_{i,j} \ge \lambda_{i,d+1} \ge \lambda_{i,d+2} \ge \dots \ge 0 \quad \text{for } 1 \le i \le k \text{ and } 1 \le j \le d.$$
(3)

In this case, ϕ_1, \ldots, ϕ_d represent the modes of variation that are common to each group, even when the ordering across groups changes. As mentioned in Coffey et al. (2011), the eigenvalues $\lambda_{i,j}$, $1 \le j \le d$, determine the order of the common directions in each group and may allow to study the differences in the distribution of the variation across groups. As in principal component analysis, the functional common principal component model could be used to reduce the dimensionality of the data, retaining as much as possible of the variability present in each of the populations.

When dealing with several populations, one possibility to identify and examine the main sources of variability of the data, is to perform a functional principal component analysis (FPCA) separately on each population, using either classical or robust estimators. However, as mentioned for instance in Coffey et al. (2011) and Fengler et al. (2003), if the principal directions show a similar structure across population as in (1), it may be more sensible from a practical point of view to obtain common eigenfunctions estimators for all the groups. Besides, a separate analysis for each population leads to difficulties in the interpretation of the obtained principal directions specially if the first *d* components can change their order from group to group as in (3). In this sense, the functional common principal component model leads to a more parsimonious model reducing the dimensionality of the functions to be estimated. A related problem was studied by Benko et al. (2009) who considered the case of k = 2 populations and provided tests for equality of means and equality of a fixed number of eigenfunctions. Another possibility considered in the literature, see for instance Donoghue et al. (2008), is to aggregate into a single data set the observations from all populations and then perform a functional principal component analysis based on the combined data set. As is well known, this approach is not satisfactory unless the data from the different groups come from a unique population. To allow for location differences and avoid mixing between group and within population variability, the practitioner should centre each group data with respect to an estimator of its mean μ_i before applying a functional principal component analysis. It is worth noting that the directions obtained using classical FPCA applied to the data after group-mean centering correspond to the eigenfunctions of the pooled sample covariance operator. Boente et al. (2010) studied the asymptotic distribution of these family of estimators under (1) and (2) and proposed estimators that correspond, in the finite-dimensional setting, to the multivariate normal maximum likelihood ones. As shown in Boente et al. (2010), for Gaussian processes with proportional covariance operators, the eigenfunctions of the pooled covariance operator are asymptotically less efficient than those obtained using a functional version of the estimating maximum likelihood equations obtained in Flury (1984) for multivariate normally distributed observations.

Besides the above mentioned asymptotic property, the advantage of an analysis based on the FCPC model over a functional principal component analysis of the pooled group-mean centred data set was clearly illustrated by several authors, including Benko and Härdle (2005) and Benko et al. (2009) who analyse the implied volatility of German stock markets and Coffey et al. (2011) who studied human movement data. More precisely, Benko and Härdle (2005) and Benko et al. (2009) used a functional common principal component analysis to analyse log-returns of the implied volatility of options with different maturities and describe the advantages of the methodology to construct lower-dimensional approximations for each population and to provide a more parsimonious model for the implied volatility surface. On the other hand, Coffey et al. (2011) consider the Achilles tendon, ankle dorsiflexion and leg abduction angles between injured and control subjects and show that the analysis based on a FCPC model reveals differences in the variation of movement patterns of injured versus control subjects that were not detected by considering the functional principal component analysis of the combined data. They also mention that, due to these differences, the scores from a functional common principal model can be used to discriminate groups.

The estimators defined in Boente et al. (2010), as well as the procedures used in Benko et al. (2009) and Coffey et al. (2011), are based on the sample covariance operators of each population being, therefore, sensitive to atypical trajectories. Up to our knowledge, robust proposals for functional principal components consider only the one-population case. For

instance, Gervini (2008) studies a fully functional approach to robust estimation of the principal components by considering a functional version of the spherical principal components defined in Locantore et al. (1999). Sawant et al. (2012) provide a robust approach of principal components based on a robust eigen-analysis of the coefficients of the observed data on some known basis, while Lee et al. (2013) propose a procedure that combines *M*-estimation with a smoothness penalty leading to *M*-type smoothing spline estimators. Besides, Hyndman and Ullah (2007) give an application of a robust projection-pursuit approach, applied to smoothed trajectories, but do not study the properties of their method in detail. On the other hand, Bali et al. (2011) introduce robust estimators of the principal directions based on robust projection-pursuit combined with different smoothing methods through a penalization in the scale or in the norm and establish their strong consistency. For sparsely and irregularly observed functional data, Gervini (2009) develops robust functional principal component estimators and uses them for outlier detection. In the functional setting, when second moment exists, Kraus and Panaretos (2012) define a ρ -dispersion operator less sensitive to outliers, that has the same eigenfunctions as the covariance operator. The ρ -dispersion operator is used to construct a test for comparing the second-order characteristics of two functional samples. However, the order among eigenvalues of the ρ -dispersion operator and the covariance operator may be not be preserved.

On the other hand, when dealing with several populations of multivariate observations, robust estimators under a common principal components model are considered in Boente and Orellana (2001). Further developments are given by Boente et al. (2006) who define a general class of projection-pursuit estimators in order to improve the efficiency of the robust estimators for a given scale and also, to recover the maximum likelihood estimators when the scale is the standard deviation.

Taking into account that the common principal component model leads to a more parsimonious model, reducing the dimensionality of the functions to be estimated and the sensitivity of the classical estimation proposals to atypical curves, the problem of robust estimation of the common direction becomes relevant. For that reason, the main purpose of this paper is to introduce a general class of robust estimators of the common directions under a FCPC model providing consistency results from a fully functional approach to the problem, as done in the one-population setting by Gervini (2008), Bali et al. (2011) and Kraus and Panaretos (2012), among others. The estimators to be considered are defined through a projection-pursuit procedure and can be viewed as an extension of those proposed by Boente et al. (2006) in the multivariate setting as well as a generalization to several populations of those given in Bali et al. (2011).

In the finite-dimensional setting, several authors, such as Critchley (1985), Jaupi and Saporta (1993), Shi (1997), Croux and Haesbroeck (1999) and Croux and Ruiz-Gazen (2005) have suggested statistical diagnostics and graphical displays for detecting outliers in principal component analysis for one population, such as side-by-side boxplots of the scores obtained from a robust principal component analysis. Under a FCPC model, the scores obtained from a robust functional common principal analysis may also be used to detect influential observations in the samples. Furthermore, as in the one-population setting, the estimation of the common directions provides a dimension reduction tool where the use of robust methods becomes important to obtain reliable directions. As noted above, the functional common principal model assumes that the same directions contribute to the variation in each population, while the distribution of their variation may differ between populations. Hence, as in Coffey et al. (2011), the robust estimation of the common directions and their size provide a meaningful and more resistant way of comparing the dispersion structure across populations.

This paper is organized as follows. In Section 2, we introduce the notation to be used and we recall the definition of the projection-pursuit estimators defined in Bali et al. (2011). In Section 3, the robust projection-pursuit estimators for the common directions and their related functional are introduced. An algorithm to compute the estimators is also described in Section 3.1. Fisher-consistency of the proposals is discussed in Section 4.1, while strong consistency results for the given proposals and for the numerical approximation described in Section 3.1 are stated in Sections 4.2 and 4.3, respectively. Section 5 summarizes the results of a Monte Carlo study conducted to compare the performance of the robust proposals between them and also with that of the classical estimators based on the sample variance. The proposed estimators are applied to a real data set in Section 6 where they are helpful to detect influential observations and to discriminate between groups. Some final comments are given in Section 7. Proofs are relegated to the online supplement where also some additional results may be found.

2. Basic definitions and notation

The observations to be considered in this paper are elements of a separable Hilbert space \mathcal{H} with inner product $\langle \cdot, \cdot \rangle$ and related norm $\|\alpha\|^2 = \langle \alpha, \alpha \rangle$. Let \otimes stand for the tensor product on \mathcal{H} , e.g., for $u, v \in \mathcal{H}$, the operator $u \otimes v : \mathcal{H} \to \mathcal{H}$ is defined as $(u \otimes v)w = \langle v, w \rangle u$. From now on, \mathcal{V}_1 stands for $\{\alpha : \|\alpha\| \le 1\}$ while $\delta_1 = \{\alpha : \|\alpha\| = 1\}$.

If $X \in \mathcal{H}$ is a random element with finite second moment, i.e., $\mathbb{E}(||X||^2) < \infty$, we denote by $\Gamma_X = \mathbb{E}\{(X - \mu) \otimes (X - \mu)\}$ its covariance operator, where $\mu = \mathbb{E}(X)$. The operator $\Gamma_X : \mathcal{H} \to \mathcal{H}$ is a linear, self-adjoint, positive semi-definite and continuous operator. Moreover, Γ_X is a Hilbert–Schmidt operator so, it has a countable number of eigenvalues, all of which are real and non-negative. As is well known, one can choose the eigenfunctions of a Hilbert–Schmidt operator so that they form an orthonormal basis for \mathcal{H} . Let $\{\phi_j : j \ge 1\}$ and $\{\lambda_j : j \ge 1\}$ be respectively an orthonormal basis of eigenfunctions of Γ_X and their corresponding eigenvalues with $\lambda_j \ge \lambda_{j+1}$. With this notation, the spectral value decomposition for Γ_X can be expressed as $\Gamma_X = \sum_{j=1}^{\infty} \lambda_j \phi_j \otimes \phi_j$. In this setting, principal components analysis has been successfully extended from the multivariate setting to accommodate functional data. The *j*th principal component variable is defined as $Z_j = \langle \phi_j, X - \mu \rangle$, leading to the Karhunen–Loève expansion $X = \mu + \sum_{j=1}^{\infty} Z_j \phi_j$ with the Z_j 's being uncorrelated and having variances λ_j in descending order.

When dealing with several populations, i.e., when $X_{i,1}, \ldots, X_{i,n_i}$ are observations from k independent samples with mean μ_i and covariance operators $\Gamma_i = \Gamma_{X_{i,1}}$ satisfying (1) and (3) for some $d \ge 1$, the processes $X_{i,1}$, $i = 1, \ldots, k$, can be written as $X_{i,1} = \mu_i + \sum_{j=1}^{\infty} \lambda_{i,j}^{1/2} \xi_{ij} \phi_j$, where ξ_{ij} are zero mean random variables such that $\mathbb{E}(\xi_{ij}^2) = 1$, $\mathbb{E}(\xi_{ij} \xi_{is}) = 0$ for $j \ne s$. In particular if $\lambda_{i,1} \ge \lambda_{i,2} \ge \cdots \ge 0$ for all $i = 1, \ldots, k$, the common eigenfunctions exhibit the same major modes of variation across populations. It is worth noticing that when considering a functional proportional model, $X_{i,1}$, $i = 1, \ldots, k$, can be written as $X_{i,1} = \mu_i + \rho_i^{1/2} \sum_{j=1}^{\infty} \lambda_j^{1/2} \xi_{ij} \phi_j$, with $\rho_1 = 1, \lambda_1 \ge \lambda_2 \ge \cdots \ge 0$ and ξ_{ij} random variables as described above.

2.1. Projection-pursuit principal component functional in the one-population setting

We recall some definitions given in Bali et al. (2011) which will help to generalize the functional common principal component model to the situation in which second moments do not exist.

Denote by \mathcal{G} the set of all univariate distributions. From now on, $\sigma_{\mathbb{R}} : \mathcal{G} \to [0, +\infty)$ stands for a scale functional, that is, a functional over the set of univariate distributions which is location invariant and scale equivariant, i.e., if $G_{a,b}$ stands for the distribution of aY + b when $Y \sim G$, then, $\sigma_{\mathbb{R}}(G_{a,b}) = |a|\sigma_{\mathbb{R}}(G)$, for all real numbers a and b.

Two well known examples of scale functionals are the standard deviation which is not resistant to outliers, $sD(G) = [\mathbb{E}\{Y - \mathbb{E}(Y)\}^2]^{1/2}$, where $Y \sim G$, and the median absolute deviation about the median, $MAD(G) = c \text{ median}\{|Y - \text{median}(Y)|\}$, where c is a normalization constant typically chosen as $c = 1/\Phi^{-1}(0.75)$ that leads to a simple resistant scale estimator. A broader class of robust scale functionals, which includes as special cases the two previous examples, is the *M*-scale functionals (see Huber (1981)). To be more precise, as in Maronna et al. (2006), let $\rho : \mathbb{R} \rightarrow [0, \infty)$ be a ρ -function, that is, an even function, non-decreasing on |x|, increasing for x > 0 when $\rho(x) < \lim_{t \to +\infty} \rho(t)$ and such that $\rho(0) = 0$. When ρ is bounded, it is assumed that $\sup_{u \in \mathbb{R}} \rho_c(u) = \|\rho\|_{\infty} = 1$. Given a univariate distribution *G* and $Y \sim G$, a location parameter μ and a continuous ρ -function, the *M*-functional $\sigma(G)$ satisfies $\mathbb{E}[\rho_c\{(Y - \mu)/\sigma(G)\}] = \delta$, where $\rho_c(u) = \rho(u/c)$, and c > 0 is a user-chosen tuning constant. Their empirical versions are known as *M*-scale estimators. To ensure consistency of the *M*-scale estimator when the data are normally distributed the tuning parameter c is chosen to satisfy $\delta = \mathbb{E}\{\rho_c(Z)\}$ where *Z* has a standard normal distribution. If, in addition, $\delta = \|\rho\|_{\infty}/2$ then, the *M*-estimate of scale has maximal breakdown point 50%.

From now on, when $X \sim P$ and $\alpha \in \mathcal{H}$, $P[\alpha]$ stands for the distribution of the real random variable $\langle \alpha, X \rangle$. Given a probability measure *P* and a scale functional σ_{R} , Bali et al. (2011) define the robust functional principal components direction functionals as

$$\begin{split} \phi_{\mathsf{R},1}(P) &= \operatorname*{argmax}_{\|\alpha\|=1} \sigma_{\mathsf{R}}\left(P[\alpha]\right), \\ \phi_{\mathsf{R},m}(P) &= \operatorname*{argmax}_{\|\alpha\|=1, \alpha \in \mathcal{R}_{m}} \sigma_{\mathsf{R}}\left(P[\alpha]\right) \quad \text{for } m \geq 2, \end{split}$$

where $\mathcal{B}_m = \{\alpha \in \mathcal{H} : \langle \alpha, \phi_{\mathsf{R},j}(P) \rangle = 0, j = 1, ..., m-1\} = \langle \phi_{\mathsf{R},1}(P), ..., \phi_{\mathsf{R},m-1}(P) \rangle^{\perp}$ with $\langle \alpha_1, ..., \alpha_\ell \rangle$ the linear space spanned by $\alpha_1, ..., \alpha_\ell$ and \mathcal{L}^{\perp} the orthogonal of the closed linear space \mathcal{L} . These authors also define the robust eigenvalue functionals as $\lambda_{\mathsf{R},j}(P) = \sigma_{\mathsf{R}}^2 \{P[\phi_{\mathsf{R},j}(P)]\}$. If σ_{R} is the standard deviation, the usual definition of principal components is obtained. As mentioned in Bali et al. (2011), if the scale functional σ_{R} is (weakly) continuous, the maximum above is attained.

The functionals $\phi_{R,m}(P)$ are Fisher-consistent for the functional elliptical family, that is, in this case the functionals $\phi_{R,m}(P)$ have a simple interpretation. Elliptical distributions were defined in Bali and Boente (2009) and characterized in Boente et al. (2014). For the sake of completeness, we recall their definition.

Let *X* be a random element in a separable Hilbert space \mathcal{H} and $\mu \in \mathcal{H}$. Moreover, let $\Gamma : \mathcal{H} \to \mathcal{H}$ be a self-adjoint, positive semi-definite and compact operator. We say that *X* has an elliptical distribution with parameters (μ , Γ), denoted as $X \sim \mathcal{E}(\mu, \Gamma)$, if for any $d \geq 1$ and for any linear and bounded operator $A : \mathcal{H} \to \mathbb{R}^d$, AX has a multivariate elliptical distribution with parameters $A\mu$ and $A\Gamma A^*$, i.e., $AX \sim \mathcal{E}_d(A\mu, A\Gamma A^*)$, where $A^* : \mathbb{R}^p \to \mathcal{H}$ stands for the adjoint operator of *A*. As in the finite-dimensional setting, if the covariance operator, Γ_X , of *X* exists then, $\Gamma_X = a \Gamma$, for some a > 0. The operator Γ is called the dispersion operator while μ is the location parameter. Elliptical distributions in \mathcal{H} include the Gaussian distributions, while other elliptical distributions can be obtained as mixtures of Gaussian processes.

Recall that if $X \sim P = \mathcal{E}(\mu, \Gamma)$, then $\sigma_{R}^{2}(P[\alpha]) = c \langle \alpha, \Gamma \alpha \rangle$ for any scale functional and for some c > 0 depending on the scale, which allows to derive the Fisher-consistency of $\phi_{R,m}(P)$, i.e., that $\phi_{R,m}(P)$ correspond to the eigenfunctions of Γ .

2.2. The several population setting

In the context of several independent populations with probability measures P_1, \ldots, P_k , that is, when $X_{i,1}, \ldots, X_{i,n_i}$ are independent and such that $X_{i,j} \sim X_{i,1} \sim P_i$, we will denote by \widehat{P}_{i,n_i} the empirical measure under P_i , i.e., $\widehat{P}_{i,n_i}(A) = (1/n_i) \sum_{j=1}^{n_i} I_A(X_{i,j})$, while $\widehat{P}_{i,n_i}[\alpha]$ stands for the empirical distribution of the real random variables { $\langle X_{i,1}, \alpha \rangle, \ldots, \langle X_{i,n_i}, \alpha \rangle$ }. Moreover, P will denote the product measure $P = P_1 \times \cdots \times P_k$.

On the other hand, given a scale functional $\sigma_{\mathbb{R}}$, we denote $\sigma_i : \mathcal{H} \to [0, +\infty)$ the function $\sigma_i(\alpha) = \sigma_{\mathbb{R}} (P_i[\alpha])$. The empirical version of σ_i^2 , denoted $s_{i,n_i}^2 : \mathcal{H} \to \mathbb{R}$ is defined as $s_{i,n_i}^2(\alpha) = \sigma_{\mathbb{R}}^2(\widehat{P}_{i,n_i}[\alpha])$.

Note that if $c \in \mathbb{R}$, then using that σ_{R} is a scale functional we get that $\sigma_{i}^{2}(c\alpha) = c^{2}\sigma_{i}^{2}(\alpha)$.

As in Boente et al. (2006), given an increasing function $f : (0, +\infty) \to \mathbb{R}$ (eventually defined on 0) and $0 < q_i < 1$ fixed numbers such that $\sum_{i=1}^{k} q_i = 1$, define $\zeta_f(\alpha) = \sum_{i=1}^{k} q_i f\{\sigma_i^2(\alpha)\}$. It is worth noting, that the numbers q_i play the role of the asymptotic proportions of the samples. To define its empirical version, let $N = \sum_{i=1}^{k} n_i$ and $\hat{q}_i = n_i/N$. Then, the estimator $\hat{\varsigma} : \mathcal{H} \to \mathbb{R}$ of ζ_f is defined as $\hat{\varsigma}(\alpha) = \sum_{i=1}^{k} \hat{q}_i f\{s_{i,n_i}^2(\alpha)\}$ with $\hat{q}_i = n_i/N$. In particular, when the scale is the standard deviation and $\mathcal{H} = \mathbb{R}^p$, the choice $f = \log$ leads to the maximum likelihood estimators for multivariate normally distributed observations which, as it is well known, are more efficient than those related to the identity function for multivariate normal samples.

3. The common direction functional and their estimators

The principal directions can be estimated applying a sample version of the functional $\phi_{\mathbf{x},j}(\cdot)$ to each population. However, in most cases, even if a functional common principal component model holds, the estimators obtained in such a way will not be equal over all the populations. Hence, as mentioned in the Introduction, a unified approach is preferred.

If we assume that $\mathbb{E}||X_{i,1}||^2 < \infty$, and that the scale functional is the standard deviation, we have that $\sigma_R^2(P_i[\alpha]) = \langle \alpha, \Gamma_i \alpha \rangle$ with Γ_i the covariance operator of the *i*th population. Then, under a functional common principal component model, if Γ_i satisfy (2), $\phi_{R,j}(P_i) = \phi_j$ for all $j \ge 1$ and i = 1, ..., k. On the other hand, when the weaker assumption (3) holds, we have that $\phi_{R,j}(P_i) = \phi_j$ for all $j \ge d + 1$ and i = 1, ..., k while, for $j \le d$, $\phi_{R,j}(P_i)$ equals the eigenfunction among $\phi_1, ..., \phi_d$ related to the *j*th largest value among $\lambda_{i,1}, ..., \lambda_{i,d}$.

Similarly, when $X_{i,1} \sim \mathcal{E}(\mu_i, \Gamma_i)$ with dispersion operators Γ_i satisfying (1) and (2), we have that $\phi_{R,j}(P_i) = \phi_j$ for all $j \ge 1$ and i = 1, ..., k for any scale functional σ_R . Hence, for any i = 1, ..., k, $\phi_{R,1}(P_i)$ will maximize $\sum_{i=1}^k q_i \sigma_R^2 (P_i[\alpha])$ over $\vartheta_1 = \{\alpha \in \mathcal{H} : \|\alpha\| = 1\}$. More generally, it will maximize $\varsigma_f(\alpha)$ over ϑ_1 for any strictly increasing function f. This motivates to define the common directions projection-pursuit functional as

$$\begin{cases} \phi_{f,1}(P) = \operatorname*{argmax}_{\|\alpha\|=1} \varsigma_f(\alpha) = \operatorname*{argmax}_{\|\alpha\|=1} \sum_{i=1}^k q_i f\{\sigma_R^2(P_i[\alpha])\}, \\ \phi_{f,m}(P) = \operatorname*{argmax}_{\|\alpha\|=1,\alpha\in\mathscr{B}_{f,m}} \varsigma_f(\alpha) = \operatorname*{argmax}_{\|\alpha\|=1,\alpha\in\mathscr{B}_{f,m}} \sum_{i=1}^k q_i f\{\sigma_R^2(P_i[\alpha])\} \quad 2 \le m, \end{cases}$$

$$\tag{4}$$

where $\mathcal{B}_{f,m} = \langle \phi_{f,1}(P), \dots, \phi_{f,m-1}(P) \rangle^{\perp}$. We also define the robust principal values functionals as

$$\lambda_{f,i,m}(P) = \sigma_{R}^{2} \{ P_{i}[\phi_{f,m}(P)] \} = \sigma_{i}^{2} \{ \phi_{f,m}(P) \}.$$
(5)

Lemmas S.1 and S.2 in the online supplement show that the maximum in (4) is attained, so the functionals $\phi_{f,m}(\cdot)$ are well defined.

When $X_{i,1} \sim \mathcal{E}(\mu_i, \Gamma_i)$ with dispersion operators Γ_i satisfying (1) and the order among eigenvalues is not preserved among populations, we can only ensure that $\phi_{\mathbb{R},1}(P_i)$ corresponds to the eigenfunction ϕ_{ℓ_i} related to the largest value among $\{\lambda_{i,j}\}_{j\geq 1}$. In this sense, the functional $\phi_{f,1}(P)$ defined in (4) provides a unified approach representing the main mode of variation of the whole population with respect to the considered function f and the scale $\sigma_{\mathbb{R}}$.

The practitioner can select as function f, the identity function, labelled id, or the logarithm function, among others. As mentioned above, when f = id and σ_R^2 is the variance, the functionals defined in (4) correspond to the eigenfunctions of the pooled covariance operator whose sample version was studied in Boente et al. (2010). On the other hand, the function $f = \log \log s$, in the multivariate setting, to the maximum likelihood estimators when considering the sample variance. When considering robust scale estimators, the choice $f = \log was$ recommended in Boente et al. (2006) for multivariate observations, based on their simulation results and on the fact that, under a proportional model, the related estimators maximize the asymptotic variance of the common principal directions over the class of strictly increasing twice continuously differentiable functions f, for a given choice of σ_R .

Note also when f is defined at 0, we can assume without loss of generality that f(0) = 0.

Let $X_{i,1}, \ldots, X_{i,n_i}$ in \mathcal{H} be independent observations from k independent populations, that is, $X_{i,j}$ are independent and such that $X_{i,j} \sim P_i$ and denote $N = \sum_{i=1}^k n_i$ and $\hat{q}_i = n_i/N$. Throughout this paper, we will assume that $\hat{q}_i \rightarrow q_i$ with $0 < q_i < 1$, for $i = 1, \ldots, k$, and $\sum_{i=1}^k q_i = 1$. As in Boente et al. (2006), the general projection-pursuit functional common direction estimators are now naturally defined by considering the empirical version of (4), that is, as

$$\begin{cases} \widehat{\phi}_{f,1} = \operatorname*{argmax}_{\|\alpha\|=1} \sum_{i=1}^{k} \widehat{q}_{i} f\{s_{i,n_{i}}^{2}(\alpha)\} = \operatorname*{argmax}_{\|\alpha\|=1} \widehat{\varsigma}(\alpha), \\ \widehat{\phi}_{f,m} = \operatorname*{argmax}_{\|\alpha\|=1,\alpha\in\widehat{\mathscr{B}}_{f,m}} \sum_{i=1}^{k} \widehat{q}_{i} f\{s_{i,n_{i}}^{2}(\alpha)\} = \operatorname*{argmax}_{\|\alpha\|=1,\alpha\in\widehat{\mathscr{B}}_{f,m}} \widehat{\varsigma}(\alpha) \quad 2 \le m, \end{cases}$$
(6)

where $\widehat{\mathcal{B}}_m = \langle \widehat{\phi}_{f,1}, \dots, \widehat{\phi}_{f,m-1} \rangle^{\perp}$ while the estimators of their size in the *i*th population are defined as $\widehat{\lambda}_{i,m} = s_{i,n}^2 (\widehat{\phi}_{f,m})$.

3.1. An algorithm to compute the estimators

As it is well known, in infinite-dimensional spaces, the unit ball is not compact making difficult to effectively compute the estimators defined through (6). An approximation to the true estimators can be obtained generalizing the algorithm given in Bali and Boente (2014) to the functional *k*-populations setting. The algorithm described in Bali and Boente (2014) is an extension of that introduced by Croux and Ruiz-Gazen (1996) to deal with functional data. The algorithm is defined as follows

(a) Compute robust location estimators $\hat{\mu}_i$ for each population and centre the observations $\widetilde{X}_{ij} = X_{ij} - \hat{\mu}_i$. A choice for $\hat{\mu}_i$ is the functional spatial median defined in Gervini (2008), that is,

$$\widehat{\mu}_{i} = \underset{\theta \in \mathcal{H}}{\operatorname{argmin}} \sum_{j=1}^{n_{i}} \left(\|X_{ij} - \theta\| - \|X_{ij}\| \right).$$

$$\tag{7}$$

The spatial median is sometimes referred to as the multivariate L^1 median, but this is a misnomer since the norm in (7) is the L^2 norm. Note that when the norm is replaced by the square of the norm in (7), the resulting parameter is the mean.

- (b) Normalize the observations $\alpha_{i,j} = \widetilde{X}_{i,j}/\|\widetilde{X}_{i,j}\|$ and consider the set of possible directions $\mathcal{A}_N = \{\alpha_{i,j} = \widetilde{X}_{i,j}/\|\widetilde{X}_{i,j}\|, 1 \le i \le k, 1 \le j \le n_i\} \subset \mathcal{S}_1$, where $N = \sum_{i=1}^k n_i$.
- (c) Estimate $\phi_{f,1}(P)$ by

$$\widehat{\phi}_{f,CR,1} = \operatorname*{argmax}_{1 \le s \le k, 1 \le j \le n_s} \sum_{i=1}^k \widehat{q}_i f\{\sigma_R^2(P_{i,n_i}[\alpha_{s,j}])\} = \operatorname*{argmax}_{\alpha \in \mathcal{A}_N} \widehat{\varsigma}(\alpha).$$

The subsequent directions are obtained as follows.

For $2 \le \ell \le q$, define recursively $Y_{ij}^{(\ell)} = \widetilde{X}_{ij} - \pi_{v_{\ell-1}}\widetilde{X}_{ij}$, where $\pi_{v_{\ell-1}}\alpha$ stands for the orthogonal projection of α over the linear space $v_{\ell-1}$ spanned by $\widehat{\phi}_{f,CR,1}, \ldots, \widehat{\phi}_{f,CR,\ell-1}$. Let the set of candidate directions for the ℓ th common principal direction be $\mathcal{A}_{N,\ell} = \{Y_{ij}^{(k)}/\|Y_{ij}^{(k)}\|, 1 \le i \le k, 1 \le j \le n_i\}$ and define $\widehat{\phi}_{f,CR,\ell} = \operatorname{argmax}_{\alpha \in \mathcal{A}_{N,\ell}}\widehat{\varsigma}(\alpha)$. It is worth noting that the above algorithm also corresponds to the extension to the functional setting of the procedure

It is worth noting that the above algorithm also corresponds to the extension to the functional setting of the procedure considered in Boente and Orellana (2001) and Boente et al. (2006). Theorem 2 establishes the strong consistency of $\hat{\phi}_{f,CR,1}$ for elliptical families. Furthermore, it allows to derive consistency results for the first common principal component direction estimator computed using the algorithm described in Boente and Orellana (2001) and Boente et al. (2006).

Usually, in practice even when $X_{i,1} \in L^2([0, 1])$, one rarely observes the entire trajectories. The functional datum for the *j*th replication usually corresponds to a finite set of discrete values $x_{i,j,1}, \ldots, x_{i,j,m_{i,j}}$ with $x_{ijs} = X_{i,j}(t_{ijs})$, $1 \le s \le m_{i,j}$. In what follows, we will assume that the grid points are dense on [0, 1]. The sparse situation is an interesting topic which deserves a careful treatment as in the one-population setting, but is beyond the scope of the paper. Depending on the characteristics of the grid where observations are obtained, one can employ different strategies to analyse the observations. When the data are observed at the same grid points $t_{ijs} = t_s$, $1 \le s \le m$, for all $1 \le j \le n_i$, $1 \le i \le k$, i.e., if $x_{ijs} = X_{i,j}(t_s)$ one may apply the algorithm to the discretized observations $\mathbf{x}_{ij} = (x_{ij1}, \ldots, x_{ijm})^T$. As mentioned in Coffey et al. (2011), even if the observations since in functional data analysis time-ordering becomes important. We refer also to Gervini (2008) for a discussion on the advantages of a fully functional approach to the problem of robust estimation in FPCA. In a one-sample setting, the basis expansion approach to obtain the principal component estimators is discussed in Ramsay and Silverman (2005) where they argue that the number *p* of basis functions should depend on the sample size *n*, on the number *p* of sampling points and on the level of smoothing imposed by using $p_n < n$ and on how efficient the basis reproduces the behaviour of the data, among others. Moreover, they recommend to use a basis expansion of order *p* only to calculate more than a fairly small proportion of eigenfunctions.

When k = 1, robust estimators based on basis expansion were considered by Locantore et al. (1999) and Sawant et al. (2012), among others. Let $\{\zeta_s\}_{s\geq 1}$ be any orthonormal basis of \mathcal{H} , $p = p_N$ an increasing sequence of integers such that $p_N < N = \sum_{i=1}^k n_i$ and define $y_{i,j,s} = \langle \zeta_s, X_{i,j} \rangle$, for $1 \leq s \leq p$. Note that if $\zeta_s = \phi_s$ and if the covariance operators or the dispersion operators of $X_{i,j}$ satisfy (1), then the covariance/dispersion matrices, Σ_i , of $\mathbf{y}_{i,1} = (y_{i,1,1}, \ldots, y_{i,1,p})^T$ satisfy a cpc model since they are diagonal. However, since the eigenfunctions are our target, we have to consider a known given orthonormal basis of \mathcal{H} . In order to obtain a solution $\hat{\phi}_j$ of (6), we will considered the multivariate vectors $\mathbf{y}_{i,j}$, for $1 \leq j \leq n_i$ and $1 \leq i \leq k$, and we compute the general projection-pursuit estimators $\hat{\boldsymbol{\beta}}_j = (\hat{\beta}_{j1}, \ldots, \hat{\beta}_{jp})^T$ defined in Boente et al. (2006). The common principal projection-pursuit estimator may then be obtained as $\hat{\phi}_j = \sum_{s=1}^p \hat{\beta}_{js} \zeta_s$. Consistency results for this *sieve* approach are given in Bali (2012), where also consistency results analogous to those given in Theorem 2 are obtained for the finite-dimensional basis expansion approximate estimators.

Table 1			
Average computing time of the algorithm ((in CPU	seconds	.).

m		n _i	Wiener process				Smooth process							
			f = id			$f = \log$			f = id			$f = \log$		
			d = 1	d = 2	<i>d</i> = 3	d = 1	d = 2	<i>d</i> = 3	d = 1	d = 2	<i>d</i> = 3	d = 1	d = 2	<i>d</i> = 3
50	MAD	100	0.194	0.379	0.568	0.192	0.383	0.577	0.208	0.416	0.612	0.217	0.411	0.609
50	MAD	200	0.718	1.433	2.123	0.713	1.422	2.161	0.771	1.536	2.279	0.763	1.518	2.280
50	MAD	300	1.639	3.304	5.056	1.678	3.361	5.028	1.700	3.376	5.055	1.668	3.360	5.033
50	M-scale	100	0.439	0.871	1.310	0.432	0.863	1.313	0.440	0.875	1.300	0.444	0.856	1.304
50	M-scale	200	1.357	2.734	4.075	1.368	2.719	4.067	1.371	2.725	4.074	1.373	2.693	4.040
50	M-scale	300	2.819	5.574	8.332	2.765	5.523	8.265	2.816	5.638	8.369	2.737	5.545	8.172
75	MAD	100	0.213	0.419	0.621	0.213	0.426	0.625	0.223	0.423	0.638	0.220	0.433	0.633
75	MAD	200	0.791	1.594	2.400	0.810	1.624	2.401	0.802	1.615	2.417	0.804	1.614	2.409
75	MAD	300	1.808	3.551	5.335	1.793	3.538	5.269	1.802	3.532	5.268	1.786	3.519	5.260
75	M-scale	100	0.455	0.898	1.343	0.451	0.895	1.341	0.450	0.893	1.326	0.446	0.891	1.313
75	M-scale	200	1.405	2.793	4.180	1.394	2.782	4.184	1.393	2.766	4.159	1.401	2.779	4.138
75	M-scale	300	2.900	5.698	8.628	2.885	5.710	8.598	2.869	5.717	8.510	2.904	5.748	8.505
100	MAD	100	0.218	0.422	0.621	0.216	0.417	0.620	0.215	0.419	0.622	0.216	0.418	0.621
100	MAD	200	0.796	1.562	2.336	0.795	1.561	2.326	0.798	1.560	2.333	0.798	1.565	2.325
100	MAD	300	1.761	3.478	5.160	1.757	3.460	5.167	1.768	3.471	5.180	1.757	3.470	5.234
100	M-scale	100	0.429	0.847	1.260	0.429	0.848	1.263	0.432	0.848	1.262	0.430	0.847	1.259
100	M-scale	200	1.344	2.658	3.965	1.347	2.660	3.963	1.339	2.651	3.962	1.339	2.659	3.963
100	M-scale	300	2.759	5.476	8.194	2.760	5.480	8.189	2.761	5.470	8.200	2.764	5.477	8.225

3.1.1. Computing time

Although a formal computational complexity analysis of the algorithm described in Steps (a)-(c) is beyond the scope of this paper, the numerical experiments reported in Section 5 show that the algorithm works very well. We tested the speed of our R code, using the same settings considered in Section 5.1 of our simulation study, on an Intel i7-2600K CPU (3.4 GHz) machine. More precisely, we consider two different situations for a proportional model generating either a Wiener or a smooth Gaussian process labelled Wiener process and Smooth process in Table 1, respectively. We generated $N = n_1 + n_2$ observations in $L^2([0, 1])$ from k = 2 populations, with $n_1 = n_2$ and we use a discretization of the domain of the observed function $X \in L^2([0, 1])$, over *m* equally spaced points in [0, 1]. Under the first model, the observations $X_{i,i}$ correspond to Brownian motion processes with covariance kernels $\gamma_1(s, t) = 10 \min(s, t)$ and $\gamma_2(s, t) = 10\gamma_1(s, t)$, while for the smooth process considered, the observations are Gaussian with covariance kernels $\gamma_1(s, t) = (1/2)(1/2)^{0.9(s-t)^2}$ and $\gamma_2(s,t) = 10\gamma_1(s,t)$. To obtain the set of centred candidates $\hat{\mu}_i$ is taken as the spatial median of each population defined in (7) and computed using the algorithm of Hössier and Croux (1995). Table 1 reports the average time in CPU minutes over 20 random samples for different combinations of the grid size (m), the number of components to be estimated (d) and the function f. We first generated all the samples and then we computed the estimators for each of the given samples to evaluate the speed of our code. The results in Table 1 show that the algorithm is very fast. As expected, the computing time increases with the number of components to be estimated and with the sample size. The results in Table 1 suggest that the computing time of the algorithm increases linearly on the number of principal direction to be estimated and quadratically on the sample size. Besides, the grid size *m* does not seem to have a significant effect on the computing time. On the other hand, the procedure based on the M-scale is much slower since it involves a re-weighting algorithm to compute the scale.

4. Main results

4.1. Fisher-consistency

Assume now that we are dealing with several populations with finite second moment, i.e., $\mathbb{E}||X_{i,1}||^2 < \infty$, and that the scale functional is the standard deviation. Then, $\sigma_R^2(P_i[\alpha]) = \langle \alpha, \Gamma_i \alpha \rangle$ with Γ_i the covariance operator of the *i*th population. Then, under a functional common principal component model (1) such that (2) holds, we have that $\phi_{R,j}(P_i) = \phi_j$ for all $j \ge 1$ and i = 1, ..., k. This ensures that $\phi_{f,j}(P) = \phi_j$ for any strictly increasing function, when σ_R is the standard deviation.

However, as we mentioned in the Introduction, our goal is to consider situations in which second moments may not exist. In those situations, an important point to highlight is what the functions $\phi_{f,j}(P)$ defined in Section 3 represent, at least in some particular situations. The aim of this Section is to extend the definition of a functional common principal component model to the situation in which the covariance operator does not exist or when the underlying distribution is not elliptical, in order to ensure Fisher-consistency of the functionals $\phi_{f,i}(P)$ defined in Section 3.

As mentioned above, when the different samples $X_{i,1}$ are elliptically distributed, i.e., $X_{i,1} \sim \mathcal{E}(\mu_i, \Gamma_i)$ with dispersion operators Γ_i satisfying (1) and (2), for any scale functional σ_R we have that $\phi_{R,j}(P_i) = \phi_j$ for all $j \ge 1$ and i = 1, ..., k, so that $\phi_{f,j}(P) = \phi_j$ for any strictly increasing function and for any scale functional. This suggests that a possible way to characterize the functional $\phi_{f,j}(P)$ in a more general context is through the behaviour of the functionals $\phi_{R,j}(P_i)$. In this way, we avoid

requiring second moment conditions or an elliptical distribution to the random elements. We will also give a notion of a partially functional common principal component model in which the populations share only the first *s* principal directions.

Definition 1. We say that P_1, \ldots, P_k are partially weakly-FCPC of order $s \ge 1$ for the scale functional σ_R if $\phi_{R,j}(P_i) = \phi_{R,j}(P_1)$ (except for a sign change) for all $j = 1, \ldots, s$ and $i = 1, \ldots, k$. Furthermore, we say that P_1, \ldots, P_k are weakly-FCPC for the scale functional σ_R if P_1, \ldots, P_k are partially weakly-FCPC of any order, that is if $\phi_{R,j}(P_i) = \phi_{R,j}(P_1)$ (except for a sign change) for all $j \ge 1$ and $i = 1, \ldots, k$.

Definition 2. We say that P_1, \ldots, P_k are partially strongly-FCPC of order $s \ge 1$ if there exist constants $c_i > 0$ and self-adjoint, positive semi-definite and compact operators Γ_i such that for any $\alpha \in \mathcal{H}, \sigma_R^2(P_i[\alpha]) = c_i \langle \alpha, \Gamma_i \alpha \rangle$, where $\Gamma_i = \sum_{j=1}^s \lambda_{i,j} \phi_j \otimes \phi_j + \sum_{j=s+1}^\infty \lambda_{i,j} \phi_{i,j} \otimes \phi_{i,j}$ with $\lambda_{i,1} \ge \cdots \ge \lambda_{i,j} \ge \cdots$, for $i = 1, \ldots, k$. When P_1, \ldots, P_k are partially strongly-FCPC for any order $s \ge 1$, we say that P_1, \ldots, P_k are strongly-FCPC, in which case, the operators Γ_i satisfy (1) and (2).

Clearly, partially strong-FCPC implies partially weak-FCPC. It is worth noting that partially strong-FCPC still assumes that all the populations have the largest eigenvalues in the same order and that the shared directions correspond to these eigenvalues. This is a difference with the finite-dimensional setting. If second moment exists and σ_R^2 is the variance, P_1, \ldots, P_k are strongly-FCPC when the covariance operators satisfy (1) and (2). Besides, if P_1, \ldots, P_k are strongly-FCPC, then the operators Γ_i have the same eigenfunctions and the order among the eigenvalues is preserved along populations. The constant c_i can be absorbed by the operator Γ_i . However, we prefer to distinguish them, since in some situations, the operator Γ_i has a simple interpretation and the constant c_i corresponds to a normalizing constant related to the scale functional.

Note that, when $X_{i,1} \sim \mathcal{E}(\mu_i, \Gamma_i)$ with dispersion operators Γ_i satisfying (1) and (2), we have that P_1, \ldots, P_k are strongly-FCPC.

The following lemma shows that, if P_1, \ldots, P_k are weakly-FCPC, the weights q_i and the score function f do not play a major role when defining the functional $\phi_{f,j}(P)$.

Lemma 1. Assume that $q_i \ge 0$, $\sum_{i=1}^k q_i = 1$, $f : \mathbb{R} \to \mathbb{R}$ is a strictly increasing function and that P_1, \ldots, P_k are partially weakly-FCPC of order $s \ge 1$ under $\sigma_{\mathbb{R}}$. Then, $\phi_{f,j}(P) = \phi_{\mathbb{R},j}(P_1)$ for all $j = 1, \ldots, s$. In particular, if P_1, \ldots, P_k are weakly-FCPC under $\sigma_{\mathbb{R}}$, we have that $\phi_{f,j}(P) = \phi_{\mathbb{R},j}(P_1)$ for all $j \ge 1$.

It is worth noting that the above result does not ensure uniqueness of the solution of (4) which will be a condition needed to ensure consistency of the estimators. As in the one-population setting, one important issue is what the functions $\phi_{f,m}(P)$ represent, at least in some particular situations. Lemma 2 shows that, as mentioned above, for functional elliptical families, the functionals $\phi_{f,m}(P)$ and $\lambda_{f,i,m}(P)$ are well defined, that is, the solution of (4) is unique, and have a simple interpretation. In particular, our result holds if all the populations have an elliptical distribution, but is not restricted to them.

Lemma 2. Let $\phi_{f,m}$ and $\lambda_{f,i,m}$ be the functionals defined in (4) and (5), respectively. Let $X_{i,1} \sim P_i$ be random elements such that P_1, \ldots, P_k are partially strongly-FCPC of order s. Assume that for some $i_0 \in \{1, \ldots, k\}$ there exists $d \ge 2$, $\lambda_{i_0,1} > \lambda_{i_0,2} > \cdots > \lambda_{i_0,d} > \lambda_{i_0,d+1}$. Then, if f is a strictly increasing function and $q_{i_0} > 0$, we have that, for all $j = 1, \ldots, \min(d, s)$, $\phi_{f,j}(P) = \phi_j$ and $\lambda_{f,i}(P) = c_i \lambda_{i,i}$.

It is also easy to see, that if P_1, \ldots, P_k are strongly-FCPC and for each $m \ge 1$ there exists $i_m = 1, \ldots, k$ such that $\lambda_{i_m,m+1}$, then $\phi_{f,j}(P) = \phi_j$, for all j.

If the operator Γ_i given in the definition of strongly-FCPC is the covariance operator of P_i , then the eigenfunctions functionals $\phi_{f,m}$ are the common principal components. Besides, we also have that $\lambda_{f,j} = \sigma_i^2(\phi_{f,j}) = c_i\lambda_{i,j}$ where $\lambda_{i,j}$ is the *j*-eigenvalue of the covariance operator of *i*th population, that is, the traditional principal value in the classical approach. Therefore, the robust eigenvalue functional will be Fisher-consistent except by multiplying factor c_i that can be chosen to be equal to 1 for all populations under a common central Gaussian model to ensure Fisher-consistency of the robust eigenvalue functionals.

When $X_{i,1} \sim \mathcal{E}(\mu_i, \Gamma_i)$ with dispersion operators Γ_i satisfying (1) and (2), Lemma 2 entails that $\phi_{f,j}(P) = \phi_j$ for any strictly increasing function and for any scale functional. However, in some applications the order among eigenvalues may vary across populations and in this case, it is important to identify if $\phi_{f,j}(P)$ corresponds to the common eigenfunctions ϕ_{ℓ} . Lemma 3 gives an answer to this problem when the function *f* is convex.

Lemma 3. Let $X_{i,1} \sim P_i$ be random elements such that for any $\alpha \in \mathcal{H}$, $\sigma_R^2(P_i[\alpha]) = c_i \langle \alpha, \Gamma_i \alpha \rangle$, where $\Gamma_i = \sum_{j \ge 1} \lambda_{i,j} \phi_j \otimes \phi_j$ and $c_i > 0$. Let f be a strictly increasing and convex function, continuous in $[0, +\infty)$. Define $v_j = \sum_{i=1}^k q_i f(c_i \lambda_{i,j})$ and assume (without loss of generality in the selection of the indexes) that $v_1 \ge v_2 \ge \cdots \ge v_s$ and $v_j \le v_s$, for all $j \ge s + 1$. Let $\phi_{f,m}$ and $\lambda_{f,i,m}$ be the functionals defined in (4) and (5), respectively. If for some $2 \le d \le s$, $v_1 > v_2 > \cdots > v_d$, we have that, for all $1 \le j \le d$, $\phi_{f,j}(P) = \phi_j$ and $\lambda_{f,j}(P) = c_i \lambda_{i,j}$ for $i = 1, \dots, k$. Note that if Γ_i satisfy (1) and (3) and f is strictly increasing, then $v_{d+1} \ge v_{d+2} \ge \ldots$, while $v_j \le v_{d+1}$, for $1 \le j \le d$. Hence, Lemma 3 provides a way of characterizing $\phi_{f,j}(P)$ as the *j*th main mode of variation of the whole population (according to the robust scale σ_R and the score function f), since it corresponds to the common eigenfunction related to the *j*th largest value among v_1, v_2, \ldots, v_d .

As mentioned above, the assumption $\sigma_{R}^{2}(P_{i}[\alpha]) = c_{i}\langle \alpha, \Gamma_{i}\alpha \rangle$ is satisfied when $P_{i} = \mathcal{E}(\mu_{i}, \Gamma_{i})$ and the constant c_{i} may be absorbed by Γ_{i} . Assume for simplicity that $c_{i} = 1$ for i = 1, ..., k and note that when f = id, $v_{j} = \sum_{i=1}^{k} q_{i}\lambda_{i,j}$ is an eigenvalue of $\sum_{i=1}^{k} q_{i}\Gamma_{i}$. Then, when f = id, Lemma 3 entails that $\phi_{f,1}$ is the eigenfunction related to the largest eigenvalue of the pooled dispersion operator $\sum_{i=1}^{k} q_{i}\Gamma_{i}$, as when considering the standard deviation.

4.2. Consistency of the estimators $\widehat{\phi}_{f,m}$ defined through (6)

As mentioned above, in the finite-dimensional case, if dispersion operators are proportional, that is, under the second level of hierarchy defined by Flury (1984), the function $f = \log \min i$ and $g_f(\alpha)$ and $g(\alpha)$ are not defined when $\alpha = 0$. We refer to Boente et al. (2006) for details. The main disadvantage of log is that $g_f(\alpha)$ and $\widehat{g}(\alpha)$ are not defined when $\alpha = 0$. Moreover, if \mathcal{H} is infinite dimensional, $g_f(\alpha)$ and $\widehat{g}(\alpha)$ will not be weakly continuous due to the singularity at $\alpha = 0$. For that reason, the statements of our results consider on one side, the case of a continuous function $f : [0, \infty) \to \mathbb{R}$ and on the other one, the case of a continuous function $f : (0, \infty) \to \mathbb{R}$ with $\lim_{t\to 0^+} f(t) = -\infty$. In the second case some extra hypothesis over the distribution will be necessary. From now on and for the sake of simplicity, if f is not defined at 0, which is the case of $f = \log_f f(0)$ stands for $\lim_{t\to 0^+} f(t)$.

To derive consistency results for the estimators defined in Section 3, we will consider the following set of assumptions.

- **C0.** For some $d \ge 2$ and j = 1, ..., d, $\phi_{f,j}(P)$ are unique up to a sign change where $P = P_1 \times \cdots \times P_k$.
- **C1.** $\sigma_i : \mathcal{H} \to \mathbb{R}$ is a weakly continuous function, i.e., continuous with respect to the weak topology in \mathcal{H} .
- **C2.** $f : (0, +\infty) \rightarrow \mathbb{R}$ is a strictly increasing and continuous function. Moreover, if f is defined at 0 and $|f(0)| < \infty$, f is continuous at 0.
- **C3.** $\sup_{\|\alpha\|=1} \left| s_{i,n_i}^2(\alpha) \sigma_i^2(\alpha) \right| \xrightarrow{\text{a.s.}} 0$ almost surely, for any $i = 1, \ldots, k$.

C4.
$$\widehat{q}_i \longrightarrow q_i, 0 < q_i < 1.$$

It is clear that **CO** holds if for some $1 \le i \le k$, $\lambda_{f,i,1} > \cdots > \lambda_{f,i,d} > \lambda_{f,i,d+1}$. On the other hand, if **CO** holds then, for any $\ell = 1, \ldots, d$, there exists $i = i_{\ell} \in \{1, \ldots, k\}$ such that $\lambda_{f,i,\ell} > \lambda_{f,i,\ell+1}$.

Remark 1. It is worth noticing that **C1** and **C2** imply that $\varsigma_f : \mathcal{H} \to \mathbb{R}$ is a weakly continuous function. Note also that **C1** holds if the univariate scale functional σ_R is qualitatively robust, that is, continuous with respect to the weak topology on the space of probability measures, which is induced by the Prohorov distance. Nevertheless, this is not strictly necessary. For instance, if the scale functional is such that $\sigma_i^2(\alpha) = c_i \langle \alpha, \Gamma_i \alpha \rangle$, for some positive constants c_i and self-adjoint, positive semi-definite, compact operators Γ_i , as is the case when σ_R is the standard deviation or when the observations have an elliptical distribution, we also obtain weak continuity of σ_i . Assumption **C1** also implies that the functional σ_i^2 is weakly uniformly continuous in the unit sphere ϑ_1 . Besides, assumption **C3** follows from the consistency of the sample covariance operators (see Dauxois et al. (1982)), if σ_R equals the standard deviation, while for any scale functional σ_R continuous with respect to the weak topology, **C3** follows from Corollary 6.1 in Bali et al. (2011), which only requires the weak continuity of the scale functional σ_R . Finally, **C2** to **C4** imply that $\sup_{\|\alpha\|=1} |\widehat{\varsigma}(\alpha) - \varsigma_f(\alpha)| \stackrel{\text{a.s.}}{\longrightarrow} 0$ almost surely (see Lemma S.3 in the online supplement).

The following lemma will be useful for deriving consistency of the general eigenfunction estimators. An extra condition on the principal values $\lambda_{f,i,j}$ is needed when $f(0) = -\infty$ to avoid singularities.

Lemma 4. Let $P = P_1 \times \cdots \times P_k$, $\phi_{f,m} = \phi_{f,m}(P)$ and $\lambda_{f,i,m} = \lambda_{f,i,m}(P)$ be defined as in (4) and (5) and let $\widehat{\phi}_m \in \mathcal{V}_1$ be such that $\widehat{\phi}_m \neq 0$, $\|\widehat{\phi}_m\| \stackrel{\text{a.s.}}{\longrightarrow} 1$ almost surely and $\langle \widehat{\phi}_m, \widehat{\phi}_j \rangle \stackrel{\text{a.s.}}{\longrightarrow} 0$ almost surely. Assume that **C0** to **C2** hold. Besides, if $f(0) = -\infty$ we also assume that for any $i = 1, \ldots, k$, $\lambda_{f,i,1} > \cdots > \lambda_{f,i,d} > \lambda_{f,i,d+1}$.

- (a) If $\varsigma_f(\widehat{\phi}_1) \xrightarrow{a.s.} \varsigma_f(\phi_{f,1})$, then, $\langle \widehat{\phi}_1, \phi_{f,1} \rangle^2 \xrightarrow{a.s.} 1$. Hence, with an appropriate sign choosing, i.e., taking $\widehat{\phi}_1$ such that $\langle \widehat{\phi}_1, \phi_{f,1} \rangle > 0$, we get that $\|\widehat{\phi}_1 \phi_{f,1}\| \xrightarrow{a.s.} 0$ and thus $\widehat{\phi}_1 \xrightarrow{a.s.} \phi_{f,1}$.
- (b) Given m = 2, ..., d, if $\varsigma_f(\widehat{\phi}_m) \xrightarrow{\text{a.s.}} \varsigma_f(\phi_{f,m})$ and $\widehat{\phi}_s \xrightarrow{\text{a.s.}} \phi_{f,s}$, for s = 1, ..., m-1, we have that $\langle \widehat{\phi}_m, \phi_{f,m} \rangle^2 \xrightarrow{\text{a.s.}} 1$. Hence, if $\widehat{\phi}_m$ is chosen such that $\langle \widehat{\phi}_m, \phi_{f,m} \rangle > 0$, we obtain that $\|\widehat{\phi}_m - \phi_{f,m}\| \xrightarrow{\text{a.s.}} 0$, which means that we can choose the sign of $\widehat{\phi}_m$ so that $\widehat{\phi}_m \xrightarrow{\text{a.s.}} \phi_{f,m}$.

Theorem 1 establishes the continuity of the functionals defined in (4) and (5), for general continuous score functions defined at 0 and for $f(0) = -\infty$ and hence the asymptotic robustness of the estimators derived from them, as defined in Hampel (1971). This can be seen just by replacing almost sure convergence by convergence in its statement and by taking P_{i,n_i} , i = 1, ..., k, fixed sequences of probability measures instead of random ones. Moreover, the consistency of the estimators $\hat{\phi}_{f,m}$ and $\hat{\lambda}_{i,m}$ defined in Section 3 is obtained from **C0** to **C4**, taking $P_{i,n_i} = \hat{P}_{i,n_i}$ for i = 1, ..., k.

Theorem 1. Let $P = P_1 \times \cdots \times P_k$ be a probability measure satisfying **CO** and $\phi_{f,m} = \phi_{f,m}(P)$ and $\lambda_{f,i,m} = \lambda_{f,i,m}(P)$ be defined as in (4) and (5), respectively. Furthermore, let P_{i,n_i} , $i = 1, \ldots, k$, be random sequences of probability measures, \widehat{q}_{i,n_i} be random variables such that $\widehat{q}_{i,n_i} \xrightarrow{a.s.} q_i$ almost surely as $N \to \infty$ with $0 < q_i < 1$, $\sum_{i=1}^k q_i = 1$ and $N = \sum_{i=1}^k n_i$. Let $f : (0, +\infty) \to \mathbb{R}$ be an increasing function and denote $\sigma_{i,n_i}^2(\alpha) = \sigma_{\mathbb{R}}^2(P_{i,n_i}[\alpha])$ and $\varsigma_N(\alpha) = \sum_{i=1}^k \widehat{q}_i f\{\sigma_{i,n_i}^2(\alpha)\}$. Define $\widehat{\lambda}_{i,m} = \sigma_{i,n_i}^2(\widehat{\phi}_m)$ with $\widehat{\phi}_1 = \operatorname{argmax}_{\|\alpha\|=1} \varsigma_N(\alpha)$ and $\widehat{\phi}_m = \operatorname{argmax}_{\alpha \in \widehat{\mathscr{B}}_m} \varsigma_N(\alpha)$, for $m \ge 2$, where $\widehat{\mathscr{B}}_m = \langle \widehat{\phi}_1, \ldots, \widehat{\phi}_{m-1} \rangle^{\perp}$. Assume that

(i) C1 and C2 hold,

(ii) $\sup_{\|\alpha\|=1} |\sigma_{i,n_i}^2(\alpha) - \sigma_i^2(\alpha)| \xrightarrow{a.s.} 0$ almost surely.

Moreover, if $f(0) = -\infty$, assume that for any i = 1, ..., k, $\lambda_{f,i,1} > \cdots > \lambda_{f,i,d} > \lambda_{f,i,d+1}$, Then, we have that $\langle \widehat{\phi}_m, \phi_{f,m} \rangle^2 \xrightarrow{\text{a.s.}} 1$ and $\widehat{\lambda}_{i,m} \xrightarrow{\text{a.s.}} \lambda_{f,i,m}$, for m = 1, ..., d. Hence, with an appropriate sign choosing, i.e., if $\widehat{\phi}_m$ is chosen such that $\langle \widehat{\phi}_m, \phi_{f,m} \rangle > 0$, we obtain that we have that $\|\widehat{\phi}_m - \phi_{f,m}\| \xrightarrow{\text{a.s.}} 0$.

Remark 2. Note that assumption (ii) in Theorem 1 corresponds to **C3** when $P_{i,n_i} = \widehat{P}_{i,n_i}$, the empirical probability measure of the *i*th population. On the other hand, when $\sigma_{\mathbb{R}}(\cdot)$ is a continuous scale functional, Theorem 6.2 in Bali et al. (2011) implies that (ii) holds whenever P_{i,n_i} converges weakly to P_i . Moreover, if $\sigma_{\mathbb{R}}(\cdot)$ is a continuous scale functional and P_i satisfy **C0**, Theorem 1 entails the continuity of the functionals $\phi_{f,j}(\cdot)$ and $\lambda_{f,i,j}(\cdot)$ at P, for $j = 1, \ldots, d$. Hence, the proposed estimators are qualitatively robust and consistent. In particular, the estimators are robust if the populations are independent each with an elliptical distribution $\mathscr{E}(\mu_1, \Gamma_1) \times \cdots \times \mathscr{E}(\mu_k, \Gamma_k)$ such that, for some $i = 1, \ldots, k$, the *d* largest eigenvalues of the operators Γ_i are all distinct.

4.3. Consistency of the approximate estimator $\widehat{\phi}_{f,CR,1}$

For the sake of simplicity, we shall assume that the location of each population μ_i is known and equal to 0, so that, Step (a) is not performed and $\tilde{X}_i = X_i$. Theorem 2 states the consistency of the first common direction estimator computed using the algorithm described in Section 3.1 for elliptical families. Recall that Lemmas 2 and 3 in Section 4.1 provide conditions ensuring that in this situation $\phi_{f,1}(P) = \phi_1$.

Theorem 2. Let \mathcal{H} is a separable Hilbert space \mathcal{H} and $X_{i,j} \sim P_i$, $1 \leq j \leq n_i$, $1 \leq i \leq k$, independent. Let $q_{i,N} = n_i/N$ where $N = \sum_{i=1}^k n_i$. Assume that

- (i) $P_i \sim \mathcal{E}(0, \Gamma_i)$ where Γ_i is a compact operator such that $\Gamma_i = \sum_{\ell \geq 1} \lambda_{i,\ell} \phi_\ell \otimes \phi_\ell$, where $\lambda_{i,\ell}$ are the eigenvalues, ordered so that $\lambda_{i,1} > \lambda_{i,2} \geq \ldots$, and ϕ_ℓ are the common eigenfunctions of the Γ_i ,
- (ii) there exists $1 \le i \le k$ such that $\mathbb{P}(X_{i,1} = 0) = 0$,
- (iii) $q_{i,N} \rightarrow q_i$ with $0 < q_i < 1$,
- (iv) **CO**, **C2** and **C3** hold.

Then, $\langle \widehat{\phi}_{f,CR,1}, \phi_1 \rangle^2 \xrightarrow{a.s.} 1$, that is, except perhaps for a sign change $\widehat{\phi}_{f,CR,1} \xrightarrow{a.s.} \phi_1$. More precisely, if $\widehat{\phi}_{f,CR,1}$ is chosen such that $\langle \widehat{\phi}_{f,CR,1}, \phi_1 \rangle > 0$ we have that $\| \widehat{\phi}_{f,CR,1} - \phi_1 \| \xrightarrow{a.s.} 0$.

5. Monte Carlo study

In this simulation study, to compute the estimators, we apply the algorithm proposed in Section 3.1, using a discretization of the domain of the observed function $X \in L^2(\mathfrak{L})$, over m = 50 equally spaced points in \mathfrak{L} . To obtain the set of centred candidates $\hat{\mu}_i$ is taken as the point-to-point mean of each population when the scale is the standard deviation, while for the robust procedures, we choose $\hat{\mu}_i$ as the spatial median of each population defined in (7) and computed using the algorithm of Hössjer and Croux (1995). In all cases, we performed 1000 replications.

Corresponding to non-resistant and robust estimators of the principal common directions, three scale functions are considered, the classical standard deviation (SD), the median absolute deviation (MAD) and an *M*-estimator of scale (*M*-scALE) with breakdown point 1/2. For the *M*-estimator, we used the score function introduced by Beaton and Tukey (1974), $\rho_c(y) = \rho(y/c)$ with $\rho(y) = \min(3y^2 - 3y^4 + y^6, 1)$, with tuning constant c = 1.56 and $\delta = 1/2$ that ensure that the scale estimator is Fisher-consistent at the normal distribution and has breakdown point 50%. To compute the *M*-scale, we use a re-weighted algorithm with initial estimator the MAD. In all tables, the estimators corresponding to each scale choice

are labelled as SD, MAD, *M*-SCALE. Even if any scale estimator can be used to compute the projection-pursuit estimators in Step (c) of Section 3.1, we recommend to use a simple and fast robust scale as those considered here.

Two settings were considered in our simulation study. The first one is a two population proportional model and the second one is a three population one. The first one is considered to see if some advantage is observed when using $f = \log_{10}$, while the second one is considered as a case where the proportionality does not hold.

For each situation, we compute the estimators of the first three principal directions and the square distance between the true and the estimated direction, that is, $\|\widehat{\phi}_{f,j} - \phi_j\|^2$. Mean values over replications, denoted $M_j(f)$, which hereafter is referred to as mean square error, are reported in the tables summarizing the results.

5.1. Proportional model

To analyse the effect of having smooth or rough trajectories on the algorithm leading to the estimation procedure, we consider two different situations for a proportional model. For that purpose, the uncontaminated observations correspond to Gaussian processes being either a Wiener or a smooth process. We generated $N = n_1 + n_2$ observations in $L^2([0, 1])$ from k = 2 populations, with $n_1 = n_2 = 100$.

For the uncontaminated observations, labelled C_0 , the proportionality constant was equal to 10. To be more precise, we considered the models

- Model 1: The observations $X_{i,j}$, $j = 1, ..., n_i$, correspond to Brownian motion processes with covariance kernels $\gamma_1(s, t) = 10 \min(s, t)$ and $\gamma_2(s, t) = 10\gamma_1(s, t)$. This model will be labelled Wiener process in the tables.
- Model 2: Corresponds to a smooth process. In this case, the observations $X_{i,j}$, $j = 1, ..., n_i$, are Gaussian with covariance kernels $\gamma_1(s, t) = (1/2)(1/2)^{0.9(s-t)^2}$ and $\gamma_2(s, t) = 10\gamma_1(s, t)$. This model will be labelled Smooth process in the tables.

Let ϕ_j stand for the eigenfunctions of the covariance operator Γ_1 related to the covariance kernel $\gamma_1(s, t)$ and let $Bi(1, \epsilon)$ be the Bernoulli distribution with probability of success ϵ . For each model, a contamination in the fourth eigenfunction was considered and denoted $C_{4,\epsilon}$, where ϵ corresponds to the contamination level. The contaminated observations denoted $X_{i,j}^{(c)}$

are generated as $X_{i,j}^{(c)} = (1 - V_{i,j})X_{i,j} + V_{i,j}W_{i,j}$, where $V_{i,j} \sim Bi(1, \epsilon)$ and $W_{i,j} \sim N(\mu_i, \sigma_c)\phi_4$ independent of $X_{i,j}$ with $\sigma_c = 0.1$, $\mu_1 = 10, \mu_2 = 30$. Two values for the proportion of atypical data are considered $\epsilon = 0.05$ and $\epsilon = 0.1$.

The results for the first three common principal directions estimates are reported in Table 2. For the uncontaminated samples, the advantage of using $f = \log$ can be appreciated. In the multivariate setting, the better performance of the estimators computed with the logarithm function can be explained since, when $\sigma_{\rm R}$ is the standard deviation, they lead to the maximum likelihood estimators. Besides, for any fixed scale $\sigma_{\rm R}$, estimators obtained using $f = \log$ maximize the asymptotic variance of the common principal directions over the class of strictly increasing twice continuously differentiable score functions f (see Boente et al. (2006)). Our simulation results show that the same improvement is obtained in the functional setting. The advantage of using $f = \log$ is more evident for the smooth process, where the ratio $R_i(\log, id) = M_i(\log)/M_i(id)$ between the mean square errors when considering $f = \log \operatorname{and} f = \operatorname{id} \operatorname{takes} \operatorname{values} \operatorname{between} 65\%$ and 75% for all i = 1, 2, 3. In particular, there is also an advantage when combining the logarithm with an M-scale, since in this case the ratios $R_i(\log, id)$ are around 65% for all the components. For the Brownian motion, the benefit of considering the logarithm function is smaller, in particular, for the standard deviation. For this process, the lower values of the ratios $R_i(\log, id)$ are achieved by the MAD (around 79%), followed by the M-scale with values around 88%. On the other hand, as expected, when using robust scales a loss of efficiency is expected with respect to the classical procedure, when no outliers are present. The M-scale leads to efficiencies around 80% with respect to the standard deviation for the Brownian motion, while much lower efficiencies are obtained for the smooth process. On the other hand, the procedure based on the MAD leads to much larger mean square errors. In the finite-dimensional setting, this fact has also been observed by Boente et al. (2006) and was explained by a jump in the MAD influence function that leads to a slower rate of convergence of the projection-pursuit estimators (see Croux and Ruiz-Gazen (2005) and Cui et al. (2003) for a discussion). It is also worth mentioning that the variability of the estimators of the common directions may be dictated by the accuracy of the algorithm in locating the maximum. In the classical case and when considering f = id, the true estimators can be computed as the eigenfunctions of the pooled sample covariance operator. However, for the robust ones, the exact solution cannot be expressed as the solution of an eigen-problem. As mentioned in Cui et al. (2003), the lack of a good optimization algorithm increases the variability of the direction estimates and leads to low efficiencies under Gaussian models. In this sense, the results given in Theorems and 2 only show that the procedure described in Section 3.1 leads to consistent estimators, further research is needed to have a better insight on their efficiency.

For contaminated samples, classical principal component analysis comes out poorly in the comparison. Effectively, the procedures based on the standard deviation breakdown since it often homed the second and third direction estimators on the linear space orthogonal to the true ones. Note that since both the true direction and the estimated one have norm 1, mean square errors close to 2 mean that the estimated direction is almost orthogonal to the true one. This behaviour is even more striking when considering a smooth process, where all the common direction estimators are affected. The robust procedures are more stable, in particular, they lead to reliable results when estimating the first two common principal directions. This performance is much better for Model 2 which has smooth trajectories. On the other hand, when $\epsilon = 0.1$, the amount of outliers affects the robust estimators of the third common principal direction, although much less than when using the

Table 2

Mean values of $\|\widehat{\phi}_{f,j} - \phi_j\|^2$ under a proportional model and different contaminations, with uncontaminated trajectories generated from a Wiener process or from a smooth process.

	Scale	$\widehat{\phi}_{f,1}$		$\widehat{\phi}_{f,2}$		$\widehat{\phi}_{f,3}$		
		f = id	$f = \log$	f = id	$f = \log$	f = id	$f = \log$	
		Wiener proce	SS					
<i>C</i> ₀	SD	0.0143	0.0138	0.0956	0.0914	0.2483	0.2329	
	MAD	0.0514	0.0407	0.2496	0.1929	0.5786	0.4619	
	M-scale	0.0188	0.0167	0.1254	0.1095	0.3207	0.2834	
<i>C</i> _{4,0.05}	SD	0.9135	0.9303	1.9097	1.9045	1.7811	1.8049	
	MAD	0.0676	0.0588	0.3564	0.3011	0.8252	0.7757	
	M-scale	0.0441	0.0372	0.2486	0.2170	0.7857	0.7130	
C _{4,0.1}	SD	1.7338	1.8015	1.9111	1.9172	1.7865	1.8020	
	MAD	0.1037	0.0941	0.5680	0.5092	1.1108	1.1202	
	M-scale	0.0919	0.0895	0.5684	0.5499	1.1866	1.2269	
		Smooth proce	ess					
<i>C</i> ₀	SD	0.0013	0.0009	0.0017	0.0011	0.0012	0.0009	
	MAD	0.0281	0.0221	0.0511	0.0380	0.0528	0.0402	
	M-scale	0.0045	0.0029	0.0060	0.0039	0.0040	0.0026	
<i>C</i> _{4,0.05}	SD	1.9263	1.9249	1.9439	1.9541	1.9655	1.9723	
	MAD	0.0421	0.0373	0.0974	0.0835	0.5355	0.5694	
	M-scale	0.0321	0.0296	0.0584	0.0500	0.7801	0.8243	
C _{4,0.1}	SD	1.9587	1.9666	1.9448	1.9545	1.9646	1.9719	
	MAD	0.0615	0.0568	0.1624	0.1372	1.0835	1.1511	
	M-scale	0.0540	0.0528	0.1212	0.1094	1.3264	1.3317	

classical methods. Even if both the MAD and the *M*-scale estimator have a 50% breakdown point, the estimators based on the MAD are more resistant than those based on an *M*-scale. It is worth noting that, when considering the robust scales, even if the contaminated samples do not follow a proportional model, choosing $f = \log$ leads to smaller mean square errors than those obtained with f = id in most cases. In this sense, using a robust *M*-scale combined with the logarithm function seems to be the better choice giving a good compromise between robustness and efficiency under the central Gaussian model.

5.2. Three population model

We considered $N = \sum_{i=1}^{k} n_i$ observations in $L^2([0, 1])$ from k = 3 populations, with $n_i = 100, i = 1, ..., 3$. Under the central model, labelled C_0 , all the populations are Gaussian with distribution as follows

- For the first population, $X_{1,\ell} \sim P_1$ where P_1 corresponds to a Brownian motion in the interval [0, 1] with covariance kernel $\gamma_1(s, t) = 10 \min(s, t)$. This choice of the covariance operator leads to principal directions $\phi_n(t) = \sqrt{2} \sin \{(2n-1)\pi t/2\}$ with related principal values $\lambda_{1,n} = 10 [2/\{(2n-1)\pi\}]^2$.
- The second population is also a Brownian process but with covariance kernel $\gamma_2(s, t) = 2\gamma_1(s, t)$, that is, proportional to the previous one.
- The third population is a finite-range one, generated as $X_{3,\ell} = Z_{1,\ell}\phi_1 + Z_{2,\ell}\phi_2 + Z_{3,\ell}\phi_3$, where $\phi_n(t) = \sqrt{2} \sin\{(2n-1)\pi t/2\}, Z_{k,\ell} \sim N(0, \sigma_k^2)$, with $\sigma_1 = 3, \sigma_2 = 1$ and $\sigma_3 = 1/2$. Thus, $\lambda_{3,1} = \sigma_1^2 = 9, \lambda_{3,2} = \sigma_2^2 = 1$ and $\lambda_{3,3} = \sigma_3^2 = 1/4$ and $\lambda_{3,j} = 0$ for $j \ge 4$.

Note that the first two populations have continuous but rough trajectories while the third one has smooth trajectories. Hence, among the candidates to be considered in our maximization procedure, we have smooth candidates to approximate the true common principal direction estimators.

Each of the three populations is contaminated with a contaminating distribution highly concentrated on the fourth principal direction denoted $C_{4,\epsilon}$ as in Section 5.1. In this framework, the contaminated observations denoted $X_{i,j}^{(c)}$ are generated as $X_{i,j}^{(c)} = (1 - V_{i,j})X_{i,j} + V_{i,j}W_{i,j}$, where $V_{i,j} \sim Bi(1,\epsilon)$ and $W_{i,j} \sim N(\mu_i, \sigma_c)\phi_4$ independent of $X_{i,j}$ with $\sigma_c = 0.1$, $\mu_1 = 10, \mu_2 = 15$ and $\mu_3 = 20$. Denote $P_{\epsilon}^{(c)}$ the joint distribution of $(X_{1,1}^{(c)}, X_{2,1}^{(c)}, X_{3,1}^{(c)})$. Two values for the proportion of atypical data are considered $\epsilon = 0.1$ and $\epsilon = 0.2$.

Table 3 summarizes the results of the simulation. The fact that the robust estimators, based on the MAD and *M*-scale, are more resistant under the presence of the contamination model than the classical estimator based on the standard deviation is confirmed. It is worth noticing that the robust methods are sensitive to 20% of contamination, especially when considering the third direction. Evidently, for this contamination level, we are getting close to the breakdown point of the estimator due to the closeness of the eigenvalues. An approach to the computation of the breakdown point, in the finite-dimensional case, was given by Boente and Orellana (2001). Nevertheless, in the case of functional data the problem is more complex and is beyond the scope of the paper. However, in the one-population setting, it is well known that the sensitivity of the robust

Table 3
Mean values of $\ \widehat{\phi}_{f,j} - \phi_j\ ^2$ for the three population model under different contamination:

	Scale	$\widehat{\phi}_{f,1}$		$\widehat{\phi}_{f,2}$		$\widehat{\phi}_{f,3}$	
		f = id	$f = \log$	f = id	$f = \log$	f = id	$f = \log$
	SD	0.0037	0.0036	0.0043	0.0043	0.0036	0.0036
C_0	MAD	0.0279	0.0279	0.0585	0.0576	0.0752	0.0672
	M-scale	0.0061	0.0065	0.0130	0.0130	0.0109	0.0101
	SD	1.9274	1.9050	1.9298	1.9291	1.9209	1.9251
$C_{4,0,1}$	MAD	0.0870	0.0845	0.2588	0.2360	0.8667	0.8521
	M-scale	0.0934	0.0897	0.2635	0.2292	1.1009	1.0773
	SD	1.9475	1.9418	1.9308	1.9295	1.9174	1.9222
$C_{4.0.2}$	MAD	0.1946	0.1864	0.7966	0.7280	1.5080	1.4755
	M-scale	0.2108	0.2003	0.9394	0.8511	1.6430	1.6300

estimators is related to the relative size of the eigenvalues. On the other hand, when dealing with several populations and f = id, we argue that the sensitivity of the robust estimators to a given contamination is related to the relative size of the eigenvalues of the pooled covariance operator $\sum_{i=1}^{3} q_i \Gamma_i$. When considering $f = \log$, in the finite-dimensional case, Boente et al. (2006) showed that the performance of the robust estimators is related to the closeness of the eigenvalues of the matrix $\log \left(\prod_{i=1}^{k} \Sigma_i^{q_i}\right)$, where Σ_i stands for the covariance matrix of the *i*th population. These eigenvalues are related to the relative size of $\varsigma_f(\phi_i)$ and for that reason, using $f = \log$ leads to slightly better results. The closeness between the values of $\varsigma_f(\phi_j)$ was also discussed in the finite-range study available in Bali (2012).

The results in Table 3 show that the function f plays a relevant role. As in the proportional two population setting considered in Section 5.1, in most cases, the performance of the estimator is better or equal, when we use $f = \log$ than when f = id is considered. With respect to the behaviour of the estimators based on the two robust scales, based on the obtained results we recommend using a robust *M*-scale since it provides estimators of the common directions more efficient under the central Gaussian model with a similar behaviour than those based on the MAD under the considered contaminations.

6. Example: Notch shape data

To illustrate the proposed procedures, we apply our estimators to the notch shape data set analysed in Ramsay and Silverman (2002) where a principal component analysis over the complete data and a discrimination analysis are considered. Our goal is to use the robust procedure to detect atypical or influential observations in the sample through their scores and to perform a robust discriminant analysis.

The data represent the shape of the knees of different individuals which are classified as healthy or suffering an arthritic condition. For each individual, we have information regarding the shape of the joint. It has been suggested that osteoarthritis can alter this shape. In particular, the intercondylar notch is considered important by medical specialists. We refer to Ramsay and Silverman (2002) for details. The data set consists of N = 96 notch outlines, on each of which we have some concomitant information which provides evidence of arthritic bone damage. For simplicity, the labels 1 and 2 refer to the arthritic and healthy group, respectively. Among the data, $n_1 = 21$ femur belong to arthritic individuals and $n_2 = 75$ to individuals showing no signs of arthritic bone change. We first perform a robust common functional principal components analysis over these two groups to detect possible influential observations in the sample. As in Ramsay and Silverman (2002), the data is parametrized by arc-length, so the functional datum corresponds to the two dimensional function $(X(t), Y(t))^{T}$, where $t \in (0, 1)$. Fig. 1 depicts, in the X-Y plane, the trajectories (in grey) together with the spatial median $\hat{\mu}_i$, i = 1, 2, in a solid black line.

The first four common principal directions are computed, using an *M*-scale estimator. Since choosing f = id or $f = \log_{10}$, leads to similar results we only report the conclusions obtained when f = id. The robust estimators are obtained maximizing (6) over a set of candidates A as described in Section 5. However, to enlarge the set of candidates, we also include the classical directions and the uncentred data in A.

Fig. 2 presents the parallel boxplots of the centred scores $\hat{z}_{i,j,\ell} = \langle X_{i,j} - \hat{\mu}_i, \hat{\phi}_\ell \rangle$ when $\hat{\phi}_\ell$ are the robust common principal direction estimators. The inner product is taken as the standard dot product in the space of functions from [0, 1] to $\mathbb{R} \times \mathbb{R}$. As is well known, the boxplots of the scores over the classical estimators should not be used due to a possible masking effect. On the other hand, when using the robust projection-pursuit estimators the largest values of $|\hat{z}_{i,j,\ell}|$ indicate the presence of atypical observations which may influence the estimation of the common principal directions, the first eigenvalue estimator related to the Arthritic group is almost the double of its robust relative (see Table 4). This fact may be explained by the presence of an individual with a very flat trajectory as shown in Fig. 1.

Fig. 2 shows that, in the Arthritic group, one observation (labelled as 18) appears with an extremely small score in the first direction, while two other ones (labelled 11 and 14) have large scores. Besides, observation 19 has a large score in the fourth one. These observations are highlighted in Fig. 3, where solid and dashed lines correspond to individuals with positive and negative scores, respectively. Furthermore, curves given in shades of black, blue (with circles) and red (with triangles)



Fig. 1. Notch shape of the 96 individuals. The black solid line is the estimated spatial median.



Fig. 2. Boxplot of the scores.

Table 4 Values of $1000 \times \hat{\lambda}_{i,j}$ with $\hat{\lambda}_{i,j} = s_{i,n_i}^2(\hat{\phi}_{f,m})$ and s_{i,n_i} the standard deviation or the *M*-scale estimator.

Scale		$\widehat{\lambda}_{i,1}$	$\widehat{\lambda}_{i,2}$	$\widehat{\lambda}_{i,3}$	$\widehat{\lambda}_{i,4}$	$\widehat{\lambda}_{i,5}$	$\widehat{\lambda}_{i,6}$
M-scale	Arthritic	3.3820	0.7108	0.3413	0.2480	0.2724	0.0245
	Healthy	3.5858	0.8456	0.3296	0.2055	0.1242	0.0631
SD	Arthritic	6.0289	0.9754	0.4690	0.1974	0.0788	0.0525
	Healthy	4.3071	0.9674	0.3819	0.2956	0.0792	0.0602

correspond to outliers detected by the boxplots in the first, third and fourth scores, respectively. Note that observation 11 corresponds to the observation with a plateau. On the other hand, in the group with no signs of arthritic bone change, six scores are flagged as outliers. The observation with the largest score in the first direction, labelled as 22, corresponds to the smallest curve in Fig. 3 among those identified as possible influential. On the other hand, the individual labelled 1 has the smallest value of $\hat{z}_{4,j,2}$ and is the one with the roundest and most flat behaviour among the four data detected as atypical using $\hat{z}_{4,j,2}$. The observations with the larger values of $\hat{z}_{4,j,2}$ are labelled as 23, 49 and 64 and correspond to two observations showing some torsion to the left and one small round curve. Finally, individual 34 that has the largest value among the third scores of the healthy data set, shows a narrow trajectory with a peak.

As in Ramsay and Silverman (2002), we project the data over the linear space spanned by the first six robust common principal directions. Denote $v_{i,j,\ell} = \langle X_{i,j}, \hat{\phi}_{\ell} \rangle$ and $\mathbf{v}_{i,\ell} = (v_{i,1,\ell}, \ldots, v_{i,6,\ell})^{\mathrm{T}}$. As most classical estimators, which are optimal under normality assumptions, linear and quadratic discriminant rules are not robust due to the lack of robustness of the sample covariance matrix. To solve this problem robust alternatives have been considered among others by Campbell (1978), Critchley and Vitiello (1991), Croux and Dehon (2001), Hubert and Van Driessen (2004), Croux and Joossens (2005), Croux et al. (2008) and Pires and Branco (2010). As in the classical situation, if the covariance matrices are quite



Fig. 3. Notch shape of the 96 individuals, with atypical observations plotted in colour. The solid and dashed lines correspond to individuals with positive and negative scores, respectively. Curves plotted in shades of black, blue (with circles) and red (with triangles) correspond to outliers detected by the boxplots in the first, third and fourth scores, respectively. Different triangles allow to identify the trajectories. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

different robust quadratic discrimination is always preferred to robust linear discrimination, see loossens and Croux (2004). Due to the presence of the influential trajectories described above, we performed both a robust linear and a quadratic discriminant analysis. For the first one we considered a projection-pursuit and also plug-in approach using as robust covariance estimators the minimum covariance determinant estimator and an M-estimator. The robust linear and guadratic discriminant rules lead to similar conclusions which may be explained by the robust eigenvalue estimators reported in Table 4. The obtained values suggest that no differences in size exist between the arthritic and the healthy group. Note that when performing a classical common principal component analysis the size of the first common component seems to differ between both groups due to the presence of the atypical trajectories mentioned above. For that reason, we only report here, the results obtained with the robust projection-pursuit approach proposed by Pires and Branco (2010) using M-estimators of location and scale. The computations were done using the function LdaPP from the R library rrcov, see Todorov (2006). Fig. 4 shows the boxplots of the obtained discriminant scores for each population, i.e., the values of $\hat{\mathbf{a}}^{\mathsf{T}}\mathbf{v}_{i,\ell}$ where $\hat{\mathbf{a}}$ is the robust discriminant direction. Note that the boxes in the given plots do not overlap showing the capability to discriminate the given projection. Even if this conclusion is quite similar to that reported in Ramsay and Silverman (2002), using a common principal component model provides a parsimonious framework less restrictive than projecting the data over the linear space spanned by the first principal directions of all the data. In this particular example, when performing a separate robust principal component analysis, the fourth and fifth principal directions seem to be reversed between individuals showing arthritic bone damage and the healthy ones. This is also illustrated by the robust common principal component analysis since $(\lambda_{1,4}, \lambda_{1,5})^{T} = (0.000248, 0.000272)^{T}$ while $(\lambda_{2,4}, \lambda_{2,5})^{T} = (0.000206, 0.000124)^{T}$. Note also that the close values obtained for the Arthritic group suggest that when considering only this population, the fourth and fifth principal directions may not be uniquely identified, while the linear space spanned by them will be identified since the next eigenvalue estimator is much smaller (see Table 4). In this sense, a common functional principal component model provides a better way to reduce the dimensionality of the data. Furthermore, as discussed in the Introduction, the procedure that combines all the observations assumes that the directions preserve the size over groups and may not be appropriate if there are location differences. Besides, as is well known, the use of robust procedures allows to detect atypical trajectories and to provide more reliable results.

7. Concluding remarks

In this paper, we present a simple definition of the functional common principal component model, which provides an extension of the model considered by Flury (1984) in the finite-dimensional setting. The defined functionals $\phi_{f,j}(P)$ represent the main modes of variation of the whole population with respect to the considered function f and the scale σ_{R} . Besides, under mild assumptions they are related to the common eigenfunctions of the dispersion operators of elliptical processes.

We introduce a new family of robust estimators of the common principal directions for functional data using a projectionpursuit approach. The proposed procedure adapts the ideas introduced in Boente et al. (2006) for multivariate samples and in Bali et al. (2011) for the functional one-population setting to provide resistant estimators of the common directions based on robust scales. As in the multivariate setting, the robust procedures introduced may be helpful, through their scores, to detect atypical observations as well as to discriminate between groups. The estimators turn out to be qualitatively robust and consistent under mild conditions on the related robust scale estimators. When considering as scale the standard deviation,



Fig. 4. Boxplot of the discriminant scores based on the first six robust common principal directions.

the proposed estimators generalize those considered in Boente et al. (2010) since the proposed estimators combine a choice for the scale with a score function.

Due to the sparseness of the unit ball in infinite-dimensional spaces, it is important to provide a method to effectively compute the first common directions. In this paper, we introduce estimators which approximate the true ones using an extension of the algorithm given in Croux and Ruiz-Gazen (1996) and Bali and Boente (2014) to the functional *k*-populations setting. For elliptical families, the first common direction estimator obtained through this algorithm is shown to be consistent.

Finally, a simulation study confirms the expected inadequate behaviour of the classical estimators in the presence of outliers, with the robust procedures performing significantly better. In particular, the procedure based on an *M*-scale combined with the logarithm function is recommended. Through a real data set the usefulness of the robust common functional principal direction estimators to detect observations with a different pattern and to discriminate between groups is illustrated.

Acknowledgements

The authors wish to thank the Associate Editor and two anonymous referees for valuable comments which led to an improved version of the original paper. This research was partially supported by Grants PIP 112-201101-00339 from CONICET, PICT 2014-0351 from ANPCYT and 20120130100279BA from the Universidad de Buenos Aires at Buenos Aires, Argentina.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.csda.2016.08.017.

References

Bali, J.L., 2012. Métodos Robustos de estimación de componentes principales funcionales y el modelo de componentes principales comunes (Ph.D. thesis), Universidad de Buenos Aires (in spanish). Available on http://digital.bl.fcen.uba.ar/Download/Tesis/Tesis_5153_Bali.pdf.

Bali, J.L., Boente, G., 2009. Principal points and elliptical distributions from the multivariate setting to the functional case. Statist. Probab. Lett. 79, 1858–1865.

Ball, J.L., Boente, G., 2014. Consistency of a numerical approximation to the first principal component projection pursuit estimator. Statist. Probab. Lett. 94, 181–191.

Bali, J.L., Boente, G., Tyler, D.E., Wang, J.L., 2011. Robust functional principal components: a projection-pursuit approach. Ann. Statist. 39, 2852–2882.

Beaton, A.E., Tukey, J.W., 1974. The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. Technometrics 16, 147–185. Benko, M., Härdle, W., 2005. Common functional implied volatility analysis. In: Čížek, Pavel, Weron, Rafal, Härdle, Wolfgang (Eds.), Statistical Tools for

Finance and Insurance. Springer, pp. 115–134.

Benko, M., Härdle, W., Kneip, A., 2009. Common functional principal components. Ann. Statist. 37, 1–34.

Boente, G., Orellana, L.C., 2001. A robust approach to common principal components. In: Statistics in Genetics and in Environmental Sciences. pp. 117–146.

Boente, G., Pires, A., Rodrigues, I., 2006. General projection-pursuit estimators for the common principal components model: influence functions and Monte Carlo study. J. Multivariate Anal. 97, 124–147.

Boente, G., Rodriguez, D., Sued, M., 2010. Inference under functional proportional and common principal components models. J. Multivariate Anal. 101, 464–475.

Boente, G., Salibian-Barrera, M., Tyler, D., 2014. A characterization of elliptical distributions and some optimality properties of principal components for functional data. J. Multivariate Anal. 131, 254–264.

Campbell, N., 1978. The influence function as an aid in outlier detection in discriminant analysis. Appl. Stat. 27, 251–258.

Coffey, N., Harrison, A.J., Donaghue, O.A., Hayes, K., 2011. Common functional principal components analysis: A new approach to analyzing human movement data. Hum. Mov. Sci. 30, 1144–1166.

Critchley, F., 1985. Influence in principal components analysis. Biometrika 72, 627-636.

Critchley, F., Vitiello, C., 1991. The influence of observations on misclassification probability estimates in linear discriminant analysis. Biometrika 78, 677-690.

Croux, C., Dehon, C., 2001. Robust linear discriminant analysis using S-estimators. Canad. J. Statist. 29, 473-492.

Croux, C., Filzmoser, P., Joossens, K., 2008. Classification efficiencies for robust linear discriminant analysis. Statist, Sinica 18, 581–599.

Croux, C., Haesbroeck, G., 1999. Empirical influence functions for robust principal component analysis. In: Proceedings of the Statistical Computing Section of the American Statistical Association. Am. Statist. Assoc., Alexandria, VA, pp. 201-206.

Croux, C., Joossens, K., 2005. Influence of observations on the misclassification probability in guadratic discriminant analysis. J. Multivariate Anal. 96, 384-403

Croux, C., Ruiz-Gazen, A., 1996. A fast algorithm for robust principal components based on projection pursuit. In: Prat, A. (Ed.), Compstat: Proceedings in Computational Statistics. Physica–Verlag, Heidelberg, pp. 211–217.

Croux, C., Ruiz-Gazen, A., 2005. High breakdown estimators for principal components: the projection-pursuit approach revisited. J. Multivariate Anal. 95. 206-226

Cui, H., He, X., Ng, K., 2003. Asymptotic distributions of principal components based on robust dispersions. Biometrika 90, 953–966. Dauxois, J., Pousse, A., Romain, Y., 1982. Asymptotic theory for the principal component analysis of a vector random function: Some applications to statistical

inference. J. Multivariate Anal. 12, 136-154.

Donoghue, O., Harrison, A.J., Coffey, N., Hayes, K., 2008. Functional data analysis of running kinematics in chronic Achilles tendon injury. Med. Sci. Sports Exerc. 40, 1323-1335

Fengler, M., Härdle, W., Villa, P., 2003. The dynamics of implied volatilities: A common principal components approach. Rev. Deriv. Res. 6, 179–202.

Flury, B., 1984. Common principal components in k groups. J. Amer. Statist. Assoc. 79, 892-898.

Gervini, D., 2008. Robust functional estimation using the spatial median and spherical principal components. Biometrika 95, 587–600.

Gervini, D., 2009. Detecting and handling outlying trajectories in irregularly sampled functional datasets. Ann. Appl. Stat. 3, 1758–1775.

Hampel, F., 1971. A general qualitative definition of robustness. Ann. Math. Statist. 42, 1887–1896.

Hössier, O., Croux, C., 1995. Generalizing univariate signed rank statistics for testing and estimating a multivariate location parameter. Nonparametr. Stat. 4, 293–308. Huber, P.J., 1981. Robust Statistics. John Wiley and Sons.

Hubert, M., Van Driessen, K., 2004. Fast and robust discriminant analysis. Comput. Statist. Data Anal. 45, 301-320.

Hyndman, R.J., Ullah, S., 2007. Robust forecasting of mortality and fertility rates: A functional data approach. Comput. Statist. Data Anal. 51, 4942–4956. Jaupi, L., Saporta, G., 1993. Using the influence function in robust principal components analysis. In: Morgenthaler, S., Ronchetti, E., Stahel, W. (Eds.), New Directions in Statistical Data Analysis and Robustness. Birkhauser, Basel, pp. 147–156.

Joossens, K., Croux, C., 2004. Empirical comparison of the classification performance of robust linear and quadratic discriminant analysis. In: Hubert, M., Pison, G., Struyf, A., Van Aelst, S. (Eds.), Theory and Applications of Recent Robust Methods. Birkhauser, Basel, pp. 131-140.

Kraus, D., Panaretos, V.M., 2012. Dispersion operators and resistant second-order analysis of functional data. Biometrika 99, 813–832.

Lee, S., Shin, H., Billor, N., 2013. M-type smoothing splines estimators for principal functions. Comput. Statist. Data Anal. 66, 89-100.

Locantore, N., Marron, J.S., Simpson, D.G., Tripoli, N., Zhang, J.T., Cohen, K.L., 1999. Robust principal components for functional data (with Discussion). TEST 8.1-73

Maronna, R., Martin, R.D., Yohai, V., 2006. Robust Statistics: Theory and Methods. John Wiley and Sons.

Pires, A.M., Branco, J.A., 2010. Projection-pursuit approach to robust linear discriminant analysis. J. Multivariate Anal. 101, 2464–2485.

Ramsay, J.O., Silverman, B.W., 2002. Applied Functional Data Analysis. Springer-Verlag, New York,

Ramsay, J.O., Silverman, B.W., 2005. Functional Data Analysis, second ed. Springer-Verlag.

Sawant, P., Billor, N., Shin, H., 2012. Functional outlier detection with robust functional principal component analysis. Comput. Statist. 27, 83–102.

Shi, L., 1997. Local influence in principal components analysis. Biometrika 84, 175–186.

Todorov, V.K., 2006. rrcov: Scalable robust estimators with high breakdown point, R package version 0.3-05.

Supplementary file: Robust estimators under a functional common principal components model

Juan Lucas Bali^a, Graciela Boente^{b,*}

^aCONICET, CITEDEF, Universidad de Buenos Aires and Universidad Nacional de San Martín, Argentina ^bIMAS, CONICET and Universidad de Buenos Aires, Argentina

Abstract

This supplementary file contains some additional results to ensure the proper definition of the projection– pursuit functionals, the proofs of Lemmas 1 to 4 and that of Theorems 1 and 2.

S. Preliminary results

We first state two lemmas that show that the functionals considered in Section 2.1 are properly defined.

Lemma S.1. If $f : [0, \infty) \to \mathbb{R}$ is a continuous function and σ_i is weakly continuous, then $\sup_{\|\alpha\|=1} \varsigma_f(\alpha)$ is reached for some $\alpha \in S_1$ and so the functional $\phi_{f,1}(P)$ is well defined.

PROOF. Using that f is a continuous function and σ_i is weakly continuous, we get easily that $f \circ \sigma_i : \mathcal{H} \to \mathbb{R}$ and $\varsigma_f : \mathcal{H} \to \mathbb{R}$ are continuous functions with respect to the weak topology in \mathcal{H} . The result follows now from the fact that the unit ball $\{\|\alpha\| = 1\}$ is weakly-compact, since any continuous function reaches its maximum over a compact set. \Box

Similar arguments to those considered in the proof of Lemma S.1 allow to show that the conclusion of Lemma S.1 still holds when considering $\sup_{\|\alpha\|=1,\alpha\in\mathcal{B}_{f,m}}\varsigma_f(\alpha)$.

The following Lemma ensures the existence of $\phi_{f,1}(P)$ when $f = \log$.

Lemma S.2. Let $f : (0, \infty) \to \mathbb{R}$ be a continuous function strictly increasing and such that $f(0) = \lim_{t\to 0^+} f(t) = -\infty$. If σ_i is a weakly-continuous function and there exists α_0 such that $\sigma_i(\alpha_0) > 0$ for all i, then $\sup_{\|\alpha\|=1} \varsigma_f(\alpha)$ will be reached for some $\alpha \in S_1$ and so the functional $\phi_{f,1}(P)$ is well defined.

As above, the same ideas used in the proof can be considered to obtain that $\sup_{\|\alpha\|=1,\alpha\in\mathcal{B}_{f,m}}\varsigma_f(\alpha)$ is attained if there exists $\alpha\in\mathcal{B}_{f,m}$ such that $\sigma_i(\alpha)>0$ for all *i*.

^{*}Corresponding Author

Departamento de Matemáticas, FCEyN, UBA

Ciudad Universitaria, Pabellón 1 Buenos Aires, C1428EHA, Argentina

Buenos Aires, C1428E

FAX 54-11-45763335

Email addresses: lbali@dm.uba.ar (Juan Lucas Bali), gboente@dm.uba.ar (Graciela Boente)

PROOF. Let $\alpha_n \in S_1$ be a sequence such that $\varsigma_f(\alpha_n) \to \sup_{\|\alpha\|=1} \varsigma_f(\alpha)$. Let us begin by showing that lim $\inf_{n\to\infty} \sigma_i(\alpha_n) = d_i > 0$ for all i = 1, ..., k. Assume that this assertion does not hold, i.e., that there exists an i and a subsequence α_{n_j} such that $\sigma_i(\alpha_{n_j}) \to 0$ as $n_j \to \infty$. Then, $\varsigma_f(\alpha_{n_j}) \to \sum_{i=1}^k q_i f\{\sigma_i^2(0)\} = -\infty$ which implies that $\sup_{\|\alpha\|=1} \varsigma_f(\alpha) = -\infty$. On the other hand, using that there exists α_0 such that $\sigma_i(\alpha_0) > 0$ for all i, we get $-\infty < \varsigma_f(\alpha_0/\|\alpha_0\|) \le \sup_{\|\alpha\|=1} \varsigma_f(\alpha)$ leading to a contradiction.

Hence, $\liminf_{n\to\infty} \sigma_i(\alpha_n) = d_i > 0$ for all *i*. Without loss of generality, assume that $\sigma_i(\alpha_n) \to d_i$. Therefore, there exists n_0 such that for $n \ge n_0$, we have that $\sigma_i(\alpha_n) > d_i/2 > 0$. After relabelling the sequence, we can assume that $\sigma_i(\alpha_n) > A > 0$ for all *i* and *n*.

Using that S_1 is weakly compact, we have that the exists a subsequence α_{n_m} converging to $\beta \in \mathcal{H}$ and $\|\beta\| \leq 1$. Let us show that $\|\beta\| \neq 0$. If $\beta = 0$, we have that $\sigma_i(\beta) = 0$. However, the weak continuity of σ_i entails that $\sigma_i(\beta) = \lim_m \sigma_i(\alpha_{n_m}) > A > 0$ leading to a contradiction. Hence $\|\beta\| \neq 0$. Then, $\alpha_{n_m} \to \beta$ and $\sigma_i(\alpha_{n_m}) > A$ for all i and n. Since $f : [A, \infty) \to \mathbb{R}$ is a continuous function, we get that $\varsigma_f(\alpha_{n_m}) \to \varsigma_f(\beta)$. On the other hand, $\varsigma_f(\alpha_{n_m}) \to \sup_{\|\alpha\|=1} \varsigma_f(\alpha)$, thus $\sup_{\|\alpha\|=1} \varsigma_f(\alpha) = \varsigma_f(\beta)$. It remains to show that $\|\beta\| = 1$. Assume that $\|\beta\| < 1$ and define $\gamma = \beta/\|\beta\|$. Using that f is strictly increasing, σ_i is a scale functional and $\|\beta\| < 1$ we obtain

$$\varsigma_f(\gamma) = \sum_{i=1}^k q_i f\left\{\frac{\sigma_i^2(\beta)}{\|\beta\|^2}\right\} > \sum_{i=1}^k q_i f\left\{\sigma_i^2(\beta)\right\} = \varsigma_f(\beta) = \sup_{\|\alpha\|=1} \varsigma_f(\alpha) ,$$

leading to a contradiction. Therefore, $\|\beta\| = 1$ and the supremum is reached at β .

We now state some results that will be used in the sequel. From now on, we denote $d_{PR}(P,Q)$ the Prohorov metric between the probability measures P and Q. Thus, P_n converges weakly to P if and only if $d_{PR}(P_n, P) \to 0$.

The following lemma, which generalizes the requirement in assumption **C3** to deal with general score functions and sequences of weights \hat{q}_i converging to q_i , shows that $\sup_{\|\alpha\|=1} |\widehat{\varsigma}(\alpha) - \varsigma_f(\alpha)| \xrightarrow{a.s.} 0$, for general continuous score functions defined at 0. It excludes, however, the logarithm which will be treated separately.

Lemma S.3. Let $\sigma_{\mathbb{R}}$ be a continuous scale functional and let $f : [0, +\infty) \to \mathbb{R}$ be a strictly increasing function such that $f : [0, +\infty) \to \mathbb{R}$ is a continuous function.

- a) Let $\{P_n\}_{n\in\mathbb{N}}$ and P be probability measures defined on a separable Hilbert space \mathcal{H} , such that $d_{\mathrm{PR}}(P_n, P) \to 0$. Then, $\sup_{\|\alpha\|=1} |f\{\sigma_{\mathrm{R}}^2(P_n[\alpha])\} f\{\sigma_{\mathrm{R}}^2(P[\alpha])\}| \longrightarrow 0$.
- b) Let $\{P_{i,n_i}\}_{n_i \in \mathbb{N}}$ and P_i , i = 1, ..., k, be probability measures defined on a separable Hilbert space \mathcal{H} , such that $d_{\mathrm{PR}}(P_{i,n_i}, P_i) \to 0$ and let q_{i,n_i} be such that $0 \leq q_{i,n_i}$ and $q_{i,n_i} \to q_i$ with $0 \leq q_i \leq 1$, $\sum_{i=1}^k q_i = 1$. Then, $\sup_{\|\alpha\|=1} |\sum_{i=1}^k q_{i,n_i}f\{\sigma_{\mathrm{R}}^2(P_{i,n_i}[\alpha])\} - \sum_{i=1}^k q_if\{\sigma_{\mathrm{R}}^2(P_i[\alpha])\}| \longrightarrow 0$.

PROOF. a) Note that there exists a metric d generating the weak topology in \mathcal{H} , the closed ball $\mathcal{V}_r = \{\alpha : \|\alpha\| \leq r\}$ is weakly compact and so, compact with respect to d. On the other hand, $\sigma(\alpha) = \sigma_{\mathrm{R}}(P[\alpha])$ is a

weakly continuous function of α in \mathcal{H} , hence continuous with respect to d. These facts entail that the set $\mathcal{A} = \{\sigma^2(\alpha) : \|\alpha\| \le 1\}$ is compact in $[0, +\infty)$, so bounded. Let us assume that $\mathcal{A} \subset [0, A] \subset \mathbb{R}$. The fact that f is continuous in $[0, \infty)$ implies that it is uniformly continuous in [0, A+1]. Hence, for any $\epsilon > 0$ there exists $\delta > 0$ such that $u, v \in [0, A+1]$, $|u-v| \le \delta$ entail $|f(u) - f(v)| < \epsilon$.

Theorem 6.2 in Bali et al. [2] implies that $\sup_{\|\alpha\|=1} |\sigma_{\mathbb{R}}(P_n[\alpha]) - \sigma_{\mathbb{R}}(P[\alpha])| \to 0$. Hence, there exist $n_o \in \mathbb{N}$ such that $\sup_{\|\alpha\|=1} |\sigma_{\mathbb{R}}(P_n[\alpha]) - \sigma_{\mathbb{R}}(P[\alpha])| \leq \min(\delta/(2(A+1)), \eta)$, for any $n \geq n_o$, where $\eta \leq \min(1/\sqrt{A}, 1)/4$. Thus, using that $\sigma_{\mathbb{R}}(P[\alpha]) \in [0, \sqrt{A}]$, for any $\alpha \in S_1$, we get that, for any $\alpha \in S_1$, $u_n = \sigma_{\mathbb{R}}^2(P_n[\alpha]) \in [0, A+1]$, $v = \sigma_{\mathbb{R}}^2(P[\alpha]) \in [0, A+1]$ and $|u_n - v| \leq \delta$, which entails that $|f\{\sigma_{\mathbb{R}}^2(P_n[\alpha])\} - f\{\sigma_{\mathbb{R}}^2(P[\alpha])\}| < \epsilon$, for any $\alpha \in S_1$, concluding the proof of a).

b) Using a) and the fact that $q_{i,n_i} \leq 2$, for n_i large enough, we easily get that

$$\sup_{\|\alpha\|=1} |\sum_{i=1}^{k} q_{i,n_{i}} \left[f\{\sigma_{\mathbf{R}}^{2}(P_{i,n_{i}}[\alpha])\} - f\{\sigma_{\mathbf{R}}^{2}(P_{i}[\alpha])\} \right] | \to 0.$$

It remains to show that $\sup_{\|\alpha\|=1} |\sum_{i=1}^{k} (q_{i,n_i} - q_i) f\{\sigma_{\mathbb{R}}^2(P_i[\alpha])\}| \longrightarrow 0$. Noting again that the closed ball $\mathcal{V}_1 = \{\alpha : \|\alpha\| \le 1\}$ is weakly compact and $g_i(\alpha) = f\{\sigma_i^2(\alpha)\} = f\{\sigma_{\mathbb{R}}^2(P_i[\alpha])\}$ are weakly continuous functions of α in \mathcal{H} , we get that the sets $\mathcal{B}_i = \{f\{\sigma_i^2(\alpha)\} : \|\alpha\| \le 1\}$ are compact sets and therefore bounded, which together with the fact that $q_{i,n_i} \to q_i$ concludes the proof of b). \Box

Using Lemma S.3, we get the following result

Corollary S.1. Let $\sigma_{\mathbb{R}}$ be a continuous scale functional and let $f : [0, +\infty) \to \mathbb{R}$ be a strictly increasing function such that $f : [0, +\infty) \to \mathbb{R}$ is a continuous function.

- a) Given P be a probability measure in a separable Hilbert space \mathcal{H} and P_n be the empirical measure of a random sample X_1, \ldots, X_n with $X_i \sim P$, we have that $\sup_{\|\alpha\|=1} |f\{\sigma_{\mathrm{R}}^2(P_n[\alpha])\} f\{\sigma_{\mathrm{R}}^2(P[\alpha])\}| \xrightarrow{a.s.} 0$.
- b) Given probability measures P_i , for i = 1, ..., k, defined on a separable Hilbert space \mathcal{H} and $\{P_{i,n_i}\}_{n_i \in \mathbb{N}}$ the empirical measures of independent random samples $X_{i,1}, ..., X_{i,n_i}$ with $X_{i,1} \sim P_i$, we have that $\sup_{\|\alpha\|=1} |\sum_{i=1}^k q_{i,n_i} f\{\sigma_{\mathbb{R}}^2(P_{i,n_i}[\alpha])\} - \sum_{i=1}^k q_i f\{\sigma_{\mathbb{R}}^2(P_i[\alpha])\}| \xrightarrow{a.s.} 0$, for any sequence q_{i,n_i} such that $0 \leq q_{i,n_i}$ and $q_{i,n_i} \xrightarrow{a.s.} q_i$, with $0 \leq q_i \leq 1$, $\sum_{i=1}^k q_i = 1$.

The following lemma will be used to derive the results stated in Section 4 when considering general continuous score functions defined at 0. Its proof is omitted since it follows using analogous arguments to those considered in the proof of Lemma S.3.

Lemma S.4. Let $\sigma_{\mathbb{R}}$ be a continuous scale functional and let $f : [0, +\infty) \to \mathbb{R}$ be a strictly increasing function such that $f : [0, +\infty) \to \mathbb{R}$ is a continuous function. Let $\{P_{i,n_i}\}_{n_i \in \mathbb{N}}$ and P_i be probability measures for $i = 1, \ldots, k$, defined on a separable Hilbert space \mathcal{H} , such that $\sup_{\alpha \in \mathcal{A}_N} |\sigma_{\mathbb{R}}^2(P_{i,n_i}[\alpha]) - \sigma_{\mathbb{R}}^2(P_i[\alpha])| \longrightarrow 0$, where $\mathcal{A}_N \subset \mathcal{V}_1 = \{\alpha : ||\alpha|| \le 1\}$, and let q_{i,n_i} be such that $0 \le q_{i,n_i}$ and $q_{i,n_i} \to q_i$ with $0 \le q_i \le 1$, $\sum_{i=1}^k q_i = 1$. Then, $\sup_{\alpha \in \mathcal{A}_N} |\sum_{i=1}^k q_{i,n_i} f\{\sigma_{\mathbb{R}}^2(P_{i,n_i}[\alpha])\} - \sum_{i=1}^k q_i f\{\sigma_{\mathbb{R}}^2(P_i[\alpha])\}| \longrightarrow 0$.

In order to prove Theorem 2, we will need the following Lemma, which is a particular case of Lemma A.1 in Bali and Boente [1].

Lemma S.5. Let $X \in \mathcal{H}$, where \mathcal{H} is a separable Hilbert space, be such that $\mathbb{P}(X = 0) = 0$. Assume that $X \sim \mathcal{E}(0, \Gamma)$ with Γ a compact operator such that $\Gamma = \sum_{i \geq 1} \lambda_i \phi_i \otimes \phi_i$, with $\lambda_1 \geq \lambda_2 \geq \ldots$ its ordered eigenvalues and ϕ_i the related eigenfunctions. Given $0 < \epsilon < 1$, let $p_n = \mathbb{P}(\langle X_1/||X_1||, \phi_1\rangle^2 < 1 - \epsilon)$. Then, there exist $n_0 \in \mathbb{N}$ and 0 < q < 1 such that $p_n \leq q < 1$ for all $n \geq n_0$.

T. Proofs of Lemmas 1 to 4

PROOF OF LEMMA 1. The fact that P_1, \ldots, P_k are partially weakly-FCPC of order $s \ge 1$ under $\sigma_{\mathbb{R}}$ entails that $\phi_{\mathbb{R},j}(P_i) = \phi_{\mathbb{R},j}(P_m) = \phi_{\mathbb{R},j}$ for all $j = 1, \ldots, s$ and $i, m = 1, \ldots, k$. So, for any $\alpha \in S_1$, we have that $\sigma_{\mathbb{R}}^2(P_i[\phi_{\mathbb{R},1}]) \ge \sigma_{\mathbb{R}}^2(P_i[\alpha])$ which together with the fact that f is strictly increasing entails that $f\{\sigma_{\mathbb{R}}^2(P_i[\phi_{\mathbb{R},1}])\} \ge f\{\sigma_{\mathbb{R}}^2(P_i[\alpha])\}$. Hence, $\varsigma_f(\alpha) \le \varsigma_f(\phi_{\mathbb{R},1})$, which implies that $\phi_{f,1} = \phi_{\mathbb{R},1}$.

The proof follows now by an induction argument. Assume that $\phi_{f,j}(P) = \phi_{R,j}(P_1)$ for $j = 1, \ldots, \ell < s$, we want to show that $\phi_{f,\ell+1}(P) = \phi_{R,\ell+1}(P_1)$. First note that $\mathcal{B}_{\ell+1} = \mathcal{B}_{f,\ell+1}$, so for any $\alpha \in S_1 \cap \mathcal{B}_{f,\ell+1}$, we have that $\sigma_i(\alpha) \leq \sigma_i(\phi_{R,\ell+1})$, so $\varsigma_f(\alpha) \leq \varsigma_f(\phi_{R,\ell+1})$, concluding the proof. \Box

PROOF OF LEMMA 2. Let us begin by showing the result for j = 1. Note that since P_1, \ldots, P_k are partially strongly-FCPC of order s they are partially weakly-FCPC of the same order under σ_R . Thus, for any $\alpha \in S_1$, we have that $\sigma_R^2(P_i[\phi_1]) = \sigma_i^2(\phi_1) = c_i\lambda_{i,1} \ge c_i\langle\alpha, \Gamma_i\alpha\rangle = \sigma_i^2(\alpha) = \sigma_R^2(P_i[\alpha])$ and the inequality is strict when $i = i_0$ and $\alpha \ne \phi_1$. Hence, $f\{\sigma_i^2(\phi_1)\} \ge f\{\sigma_i^2(\alpha)\}$ for any $i = 1, \ldots, k$ and $f\{\sigma_{i_0}^2(\phi_1)\} > f\{\sigma_{i_0}^2(\alpha)\}$ since f is strictly increasing which together with the fact that $q_i \ge 0$ and $q_{i_0} > 0$ imply that $\varsigma_f(\phi_1)) > \varsigma_f(\alpha)$ for any $\alpha \in S_1$. Thus, $\phi_{f,1}(P) = \phi_1$.

The proof follows easily using an induction argument. Assume that $\phi_{f,j}(P) = \phi_j$, for $j = 1, \ldots, m-1$, $m \leq \min(s,d)$, we want to show that $\phi_{f,m}(P) = \phi_m$. Now the set $\mathcal{B}_{f,m}$ equals $\{\alpha : \langle \alpha, \phi_j \rangle = 0, j = 1, \ldots, m-1\}$, hence, for any $\alpha \in S_1 \cap \mathcal{B}_{f,m}$, we have that $\langle \alpha, \Gamma_i \alpha \rangle \leq \langle \phi_m, \Gamma_i \phi_m \rangle = \lambda_{i,m}$ with strict inequality when $i = i_0$ and $\alpha \neq \phi_m$. This implies that for any $\alpha \in S_1 \cap \mathcal{B}_{f,m}$, we have $f\{\sigma_i^2(\phi_m)\} = f\{c_i\langle \phi_m, \Gamma_i \phi_m \rangle \geq f(c_i\langle \alpha, \Gamma_i \alpha \rangle) = f\{\sigma_i^2(\alpha)\}$ for any $i = 1, \ldots, k$ and $f\{\sigma_{i_0}^2(\phi_m)\} > f\{\sigma_{i_0}^2(\alpha)\}$ which entails that $\varsigma_f(\phi_m) > \varsigma_f(\alpha)$, so $\phi_{f,m}(P) = \phi_m$, concluding the proof.

The result regarding the eigenvalues follow easily since $\lambda_{f,i,j} = \sigma_{\mathbf{R}}^2 (P_i[\phi_{f,j}]) = \sigma_i^2(\phi_j) = c_i \langle \phi_j, \Gamma_i \phi_j \rangle = c_i \lambda_{i,j}$. \Box

PROOF OF LEMMA 3. The proof follows the same lines as those used in the proof of Proposition 1 in Boente et al. [3]. Given $\alpha \in \mathcal{H}$ such that $\|\alpha\| = 1$, we have that $\alpha = \sum_{j\geq 1} a_j \phi_j + b_i \phi_i^*$ where $\sum_{j\geq 1} a_j^2 + b_i^2 = 1$ and ϕ_i^* is a norm one element of the kernel of Γ_i when ker(Γ_i) $\neq \{0\}$. If ker(Γ_i) = $\{0\}$, we have $\alpha = \sum_{j\geq 1} a_j \phi_j$ and $\sum_{j\geq 1} a_j^2 = 1$. Then, $\sigma_{\mathbb{R}}^2(P_i[\alpha]) = c_i \sum_{j\geq 1} \lambda_{i,j} a_j^2$. Note that since f is continuous, convex and f(0) = 0,

we have that $f(\sum_{j=1}^{m} c_i \lambda_{i,j} a_j^2) = f(\sum_{j=1}^{m} a_j^2 c_i \lambda_{i,j} + (1 - \sum_{j=1}^{m} a_j^2) \cdot 0) \le \sum_{j=1}^{m} a_j^2 f(c_i \lambda_{i,j})$ which entails that

$$\varsigma_f(\alpha) = \sum_{i=1}^{k} q_i f\{\sigma_{\mathbf{R}}^2(P_i[\alpha])\} = \sum_{i=1}^{k} q_i f(c_i \sum_{j \ge 1} \lambda_{i,j} a_j^2) \le \sum_{j \ge 1} a_j^2 \sum_{i=1}^{k} q_i f(c_i \lambda_{i,j}) = \sum_{j \ge 1} a_j^2 \nu_j.$$

Therefore, $\varsigma_f(\alpha) \leq \varsigma_f(\phi_1) = \nu_1$ and the inequality is strict unless $\alpha = \phi_1$. Similarly, for $2 \leq r \leq d$ and α such that $\langle \alpha, \phi_j \rangle = 0$ for $1 \leq j \leq r-1$, we have that $\varsigma_f(\alpha) \leq \sum_{j \geq r} a_j^2 \nu_j < \nu_r$, which concludes the proof. \Box

From now on, $o_{\text{a.s.}}(1)$ stands for a term which converges to 0 almost surely.

PROOF OF LEMMA 4. a) Let $\mathcal{N} = \{\omega : \varsigma_f(\widehat{\phi}_1(\omega)) \not\to \varsigma_f(\phi_{f,1})\}$ and fix $\omega \notin \mathcal{N}$, then $\varsigma_f(\widehat{\phi}_1(\omega)) \to \varsigma_f(\phi_{f,1})$. Using $\mathcal{V}_1 = \{\|\alpha\| \leq 1\}$ is weakly compact, we have that for any subsequence γ_ℓ of the sequence $\widehat{\phi}_1(\omega)$ there exists a subsequence γ_{ℓ_s} such that $\gamma_{\ell_s} \to \gamma \in \mathcal{H}$ such that that $\|\gamma\| \leq 1$. Besides, using that $\varsigma_f(\widehat{\phi}_1(\omega)) \to \varsigma_f(\phi_{f,1})$, we get that $\varsigma_f(\gamma_{\ell_s}) \to \varsigma_f(\phi_{f,1})$.

Let us show that $\varsigma_f(\gamma) = \varsigma_f(\phi_{f,1})$ and $\gamma \neq 0$. We will consider the situation in which f is continuous at 0 and when $f(0) = -\infty$.

i) We will begin by considering $|f(0)| < +\infty$ and f is continuous at 0. Hence, using **C1** we have that $\varsigma_f : \mathcal{H} \to \mathbb{R}$ is a weakly uniformly continuous function on \mathcal{V}_1 which entails that $\varsigma_f(\gamma_{\ell_s}) \to \varsigma_f(\gamma)$, as $s \to \infty$. Hence, $\varsigma_f(\gamma) = \varsigma_f(\phi_{f,1})$ which entails that $\gamma \neq 0$. Effectively, assume that $\gamma = 0$. Then, we have that $\sigma_{\mathrm{R}}(P_i[\gamma]) = \sigma_{\mathrm{R}}(P_i[0]) = 0$ which implies that $\varsigma_f(\gamma) = f(0)$, since $\sum_{i=1}^k q_i = 1$. Therefore, $\varsigma_f(\phi_{f,1}) = f(0)$ and $\varsigma_f(\phi_{f,1}) = \sum_{i=1}^k q_i f(\lambda_{f,i,1})$. Using that f is strictly increasing and the fact that **C0** implies that there exist i such that $\lambda_{f,i,1} > 0$, we get that $\sum_{i=1}^k q_i f(\lambda_{f,i,1}) > f(0)$ leading to a contradiction. Hence, $\gamma \neq 0$.

ii) Consider now the case $f(0) = -\infty$. The proof is quite similar to the previous one, avoiding the problems caused by the singularity at 0. Again, **C1** imply that $\sigma_i : \mathcal{H} \to \mathbb{R}$ is a weakly uniformly continuous function on \mathcal{V}_1 which entails that $\sigma_i(\gamma_{\ell_s}) \to \sigma_i(\gamma)$, as $s \to \infty$, for $i = 1, \ldots, k$. If there exist $i = 1, \ldots, k$ such that $\sigma_i(\gamma) = 0$, (which includes the situation $\gamma = 0$), the fact that $q_i > 0$ implies that $\varsigma_f(\gamma_{\ell_s}) \to -\infty$ and so, $\varsigma_f(\phi_{f,1}) = -\infty$ which contradicts the fact that $\varsigma_f(\phi_{f,1}) = \sum_{i=1}^k q_i f(\lambda_{f,i,1})$ and $\lambda_{f,i,1} > 0$ for all $i = 1, \ldots, k$. Thus $\sigma_i(\gamma) \neq 0$, for all $i = 1, \ldots, k$, (which entails that $\gamma \neq 0$), the continuity of f in $(0, +\infty)$ implies that $\varsigma_f(\gamma_{\ell_s}) \to \varsigma_f(\gamma)$ and so, $\varsigma_f(\gamma) = \varsigma_f(\phi_{f,1})$.

So, in both cases we have that $\varsigma_f(\gamma) = \varsigma_f(\phi_{f,1})$ and $\gamma \neq 0$. Assume that $\|\gamma\| < 1$ and let $\tilde{\gamma} = \gamma/\|\gamma\|$, then $\tilde{\gamma} \in S_1$ which implies that $\varsigma_f(\tilde{\gamma}) \leq \varsigma_f(\phi_{f,1})$. On the other hand, using that σ_R is a scale functional, $\|\gamma\| < 1$ and f is strictly increasing, we get

$$\varsigma_f(\widetilde{\gamma}) = \sum_{i=1}^k q_i f\left\{\sigma_{\mathbf{R}}^2\left(P_i\left[\widetilde{\gamma}\right]\right)\right\} = \sum_{i=1}^k q_i f\left\{\frac{\sigma_{\mathbf{R}}^2\left(P_i[\gamma]\right)}{\|\gamma\|^2}\right\} > \sum_{i=1}^k q_i f\left\{\sigma_{\mathbf{R}}^2\left(P_i[\gamma]\right)\right\} = \varsigma_f(\gamma) = \varsigma_f(\phi_{f,1})$$

which contradicts the fact that $\varsigma_f(\phi_{f,1}) = \max_{\|\alpha\|=1} \varsigma_f(\alpha)$. Hence, $\|\gamma\| = 1$ and **C0** implies that $\gamma = \phi_{f,1}$ except maybe for a sign change, that is, $\langle \gamma, \phi_{f,1} \rangle^2 = 1$. Thus, any subsequence of $\widehat{\phi}_1(\omega)$ will have a limit converging either to $\phi_{R,1}$ or $-\phi_{R,1}$, concluding the proof of a).

b) Write $\hat{\phi}_m$ as $\hat{\phi}_m = \sum_{j=1}^{m-1} \hat{a}_j \phi_{f,j} + \hat{\gamma}_m$, with $\langle \hat{\gamma}_m, \phi_{f,j} \rangle = 0$ for $j = 1, \ldots, m-1$. To obtain b) we only have to show that $\langle \hat{\gamma}_m, \phi_{f,m} \rangle^2 \xrightarrow{a.s.} 1$. Note that $\langle \hat{\phi}_m, \hat{\phi}_j \rangle \xrightarrow{a.s.} 0$, for $j \neq m$, implies that $\hat{a}_j = \langle \hat{\phi}_m, \phi_{f,j} \rangle = \langle \hat{\phi}_m, \phi_{f,j} - \hat{\phi}_j \rangle + \langle \hat{\phi}_m, \hat{\phi}_j \rangle = \langle \hat{\phi}_m, \phi_{f,j} - \hat{\phi}_j \rangle + o_{a.s.}(1)$. Thus, using that $\hat{\phi}_j \xrightarrow{a.s.} \phi_{f,j}$, for $j = 1, \ldots, m-1$, and $\|\hat{\phi}_m\| \xrightarrow{a.s.} 1$, we get that $\hat{a}_j \xrightarrow{a.s.} 0$, for $j = 1, \ldots, m-1$. Therefore, $\|\hat{\phi}_m - \hat{\gamma}_m\|^2 \xrightarrow{a.s.} 0$. Moreover, using that $\|\hat{\phi}_m\|^2 = \sum_{j=1}^{m-1} \hat{a}_j^2 + \|\hat{\gamma}_m\|^2$ and $\|\hat{\phi}_m\|^2 \xrightarrow{a.s.} 1$, we get that $\|\hat{\gamma}_m\|^2 \leq 1$ and $\|\hat{\gamma}_m\|^2 \xrightarrow{a.s.} 1$, which implies that $\|\hat{\phi}_m - \hat{\gamma}_m\| \xrightarrow{a.s.} 0$, where $\tilde{\gamma}_m = \hat{\gamma}_m/\|\hat{\gamma}_m\|$.

Let us show that $\varsigma_f(\tilde{\gamma}_m) \xrightarrow{a.s.} \varsigma_f(\phi_{f,m})$. We will again consider the situation in which f is continuous at 0 and when $f(0) = -\infty$.

i) We will begin by considering $|f(0)| < +\infty$ and f is continuous at 0 Using now that $\varsigma_f(\alpha)$ is a weakly uniformly continuous function of α in \mathcal{V}_1 , we obtain that $\varsigma_f(\widetilde{\gamma}_m) - \varsigma_f(\widehat{\phi}_m) \xrightarrow{a.s.} 0$ which together with the fact that $\varsigma_f(\widehat{\phi}_m) \xrightarrow{a.s.} \varsigma_f(\phi_{f,m})$ implies that $\varsigma_f(\widetilde{\gamma}_m) \xrightarrow{a.s.} \varsigma_f(\phi_{f,m})$.

ii) Consider now the case $f(0) = -\infty$. Using now that, for all i = 1, ..., k, $\sigma_i^2(\alpha)$ is a weakly uniformly continuous function of α in \mathcal{V}_1 , we obtain that $\sigma_i^2(\tilde{\gamma}_m) - \sigma_i^2(\hat{\phi}_m) = o_{\text{a.s.}}(1)$. It is enough to show that, for any subsequence $\tilde{\gamma}_{m,\ell}$ of $\tilde{\gamma}_m$, we have that $\varsigma_f(\tilde{\gamma}_{m,\ell}) \xrightarrow{a.s.} \varsigma_f(\phi_{f,m})$ as $\ell \to \infty$. For simplicity, we will show that the result holds for the original sequence.

Let $\mathcal{N}_m = \bigcup_{i=1}^k \{\omega : \sigma_i^2\{\widetilde{\gamma}_m(\omega)\} - \sigma_i^2\{\widehat{\phi}_m(\omega)\} \neq 0\} \cup \{\omega : \varsigma_f\{\widehat{\phi}_m(\omega)\} \neq \varsigma_f(\phi_{f,m})\}$ and fix $\omega \notin \mathcal{N}_m$. Then $\varsigma_f\{\widehat{\phi}_m(\omega)\} \to \varsigma_f(\phi_{f,m})$ and $\sigma_i^2\{\widetilde{\gamma}_m(\omega)\} - \sigma_i^2\{\widehat{\phi}_m(\omega)\} \to 0$, for all $1 \leq i \leq k$. As in a) it is easy to see that there exists a subsequence ν_{ℓ_s} of $\{\widehat{\phi}_m(\omega)\}$ such that $\nu_{\ell_s} \to \nu \in \mathcal{H}$ such that that $\nu \neq 0$ and $\varsigma_f(\phi_{f,m}) = \varsigma_f(\nu)$. Using that $\sigma_i : \mathcal{H} \to \mathbb{R}$ is a weakly uniformly continuous function on \mathcal{V}_1 , we get that $\sigma_i(\nu_{\ell_s}) \to \sigma_i(\nu)$, as $s \to \infty$, for $i = 1, \ldots, k$, so that $\sigma_i^2\{\widetilde{\gamma}_{m,\ell_s}(\omega)\}$ converges to $\sigma_i^2(\nu)$, where $\widetilde{\gamma}_{m,\ell_s}(\omega)$ is the subsequence of $\{\widetilde{\gamma}_m(\omega)\}$ related to ν_{ℓ_s} . Therefore, $\varsigma_f\{\widetilde{\gamma}_{m,\ell_s}(\omega)\}$ converges to $\varsigma_f(\nu) = \varsigma_f(\phi_{f,m})$.

Hence, in both situations we have that $\varsigma_f(\widetilde{\gamma}_m) \xrightarrow{a.s.} \varsigma_f(\phi_{f,m})$. The proof follows now as in a) using the fact that $\widetilde{\gamma}_m \in \mathcal{C}_m$, with $\mathcal{C}_m = \{\alpha \in \mathcal{S}_1 : \langle \alpha, \phi_{f,j} \rangle = 0, \ j = 1, \dots, m-1\}$ and $\phi_{f,m}$ is the unique maximizer of $\varsigma_f(\alpha)$ over \mathcal{C}_m . \Box

U. Proof of Theorem 1

We will need first some auxiliary definitions. Denote by \mathcal{L}_k and $\widehat{\mathcal{L}}_k$ the linear spaces spanned by $\{\phi_{f,1}, \ldots, \phi_{f,k}\}$ and by $\widehat{\phi}_1, \ldots, \widehat{\phi}_k$, respectively. Let π_k and $\widehat{\pi}_k$ the orthogonal projection onto \mathcal{L}_k^{\perp} and $\widehat{\mathcal{L}}_k^{\perp}$, respectively, that is, that is, $\pi_k(\alpha) = \alpha - \sum_{j=1}^k \langle \alpha, \phi_{f,j} \rangle \phi_{f,j}$ and $\widehat{\pi}_k(\alpha) = \alpha - \sum_{j=1}^k \langle \alpha, \widehat{\phi}_j \rangle \widehat{\phi}_j$.

The proof will be done by deriving the following assertions

- a) $\varsigma_N(\widehat{\phi}_1) \xrightarrow{a.s.} \varsigma_f(\phi_{f,1})$ and $\varsigma_f(\widehat{\phi}_1) \xrightarrow{a.s.} \varsigma_f(\phi_{f,1})$.
- b) $\langle \widehat{\phi}_1, \phi_{f,1} \rangle^2 \xrightarrow{a.s.} 1$ and $\widehat{\lambda}_{i,1} \xrightarrow{a.s.} \lambda_{f,i,1}$.

- c) For any m = 2, ..., d, if $\widehat{\phi}_{\ell} \xrightarrow{a.s.} \phi_{f,\ell}$ for $\ell = 1, ..., m-1$, then, $\varsigma_N(\widehat{\phi}_m) \xrightarrow{a.s.} \varsigma_f(\phi_{f,m})$ and $\varsigma_f(\widehat{\phi}_m) \xrightarrow{a.s.} \varsigma_f(\phi_{f,m})$.
- d) For m = 1, ..., d, $\langle \hat{\phi}_m, \phi_{f,m} \rangle^2 \xrightarrow{a.s.} 1$ and $\hat{\lambda}_{i,m} \xrightarrow{a.s.} \sigma_i^2(\phi_{f,m})$.
- a) The proof of $\varsigma_N(\widehat{\phi}_1) \xrightarrow{a.s.} \varsigma_f(\phi_{f,1})$ will follow if we show that

$$\varsigma_f(\phi_{f,1}) \geq \varsigma_N(\widehat{\phi}_1) + o_{\text{a.s.}}(1) , \qquad (S.1)$$

$$\varsigma_f(\phi_{f,1}) \leq \varsigma_N(\widehat{\phi}_1) + o_{\text{a.s.}}(1) .$$
(S.2)

Denote $\hat{a}_N = \varsigma_N(\hat{\phi}_1) - \varsigma_f(\hat{\phi}_1)$ and $\hat{b}_N = \varsigma_N(\phi_{f,1}) - \varsigma_f(\phi_{f,1})$ and assume that

$$\widehat{a}_N = \varsigma_N(\widehat{\phi}_1) - \varsigma_f(\widehat{\phi}_1) = o_{\text{a.s.}}(1) , \qquad (S.3)$$

$$\widehat{b}_N = \varsigma_N(\phi_{f,1}) - \varsigma_f(\phi_{f,1}) = o_{\text{a.s.}}(1)$$
(S.4)

hold. Let us show that (S.4) implies (S.2) while (S.3) entails (S.1).

Effectively, if (S.4) hold, using that $\hat{\phi}_1 = \operatorname{argmax}_{\|\alpha\|=1} \varsigma_N(\alpha)$ we get that $\varsigma_N(\hat{\phi}_1) \ge \varsigma_N(\phi_{f,1}) = \varsigma_f(\phi_{f,1}) + \hat{b}_N$ concluding the proof of (S.2).

On the other hand, if (S.3) hold using that $\varsigma_f(\phi_{f,1}) = \sup_{\|\alpha\|=1} \varsigma_f(\alpha)$ and $\|\widehat{\phi}_1\| = 1$, we obtain easily that $\varsigma_f(\phi_{f,1}) \ge \varsigma_f(\widehat{\phi}_1) = \varsigma_N(\widehat{\phi}_1) - \widehat{a}_N$, concluding the proof of (S.1).

Hence, $\varsigma_N(\widehat{\phi}_1) \xrightarrow{a.s.} \varsigma_f(\phi_{f,1})$. On the other hand, using again (S.3) and since $\varsigma_f(\widehat{\phi}_1) - \varsigma_f(\phi_1) = \varsigma_N(\widehat{\phi}_1) - \varsigma_f(\phi_1) - \varsigma_f(\phi_1) - \widehat{a}_N$, we get that $\varsigma_f(\widehat{\phi}_1) \xrightarrow{a.s.} \varsigma_f(\phi_{f,1})$, concluding the proof of a).

Therefore, we only have to show that (S.3) and (S.4) hold. We will distinguish the case in which f is defined at 0 and when $f(0) = -\infty$.

a.i) Consider first the case $f : [0, \infty) \to \mathbb{R}$ a strictly increasing and continuous function. Using that $\sigma_{\mathbf{R}}$ is a functional of scale and the convergence given in assumption ii), we get easily from Lemma S.4 in the online supplement that

$$\sup_{\|\alpha\| \le 1} |\varsigma_N(\alpha) - \varsigma_f(\alpha)| = o_{\text{a.s.}}(1) , \qquad (S.5)$$

which entails (S.4) and (S.3).

a.ii) Suppose now that $f(0) = -\infty$.

We begin by showing that (S.4) holds. The convergence given in assumption ii) implies that $\hat{b}_{N,i} = \sigma_{i,n_i}^2(\phi_{f,1}) - \sigma_i^2(\phi_{f,1}) = o_{\text{a.s.}}(1)$. Then, using that $\sigma_i^2(\phi_{f,1}) > 0$ and the continuity of f, we get that $f(\sigma_{i,n_i}^2(\phi_{f,1})) \xrightarrow{a.s.} f(\sigma_i^2(\phi_{f,1}))$, which concludes the proof of (S.4). As mentioned above, (S.4) entails (S.2).

To derive (S.3), note that $\hat{a}_{N,i} = \sigma_{i,n_i}^2(\hat{\phi}_1) - \sigma_i^2(\hat{\phi}_1) = o_{\text{a.s.}}(1)$ since $\|\hat{\phi}_1\| = 1$. Let $\mathcal{N} = \bigcup_{i=1}^k \{\omega : \hat{a}_{N,i}(\omega) \neq 0\}$, that is, the set of probability 0 where the almost sure convergence of $\hat{a}_{N,i}$ to 0 does not hold for $1 \leq i \leq k$. For any $\omega \notin \mathcal{N}$, let us show that

$$\liminf \sigma_i^2\{\widehat{\phi}_1(\omega)\} > 0 \quad \text{for} \quad i = 1, \dots, k .$$
(S.6)

Effectively, assume there exists i_0 such that $\liminf \sigma_{i_0}^2 \{ \widehat{\phi}_1(\omega) \} = 0$. Then, there exists a subsequence of $\gamma_N = \widehat{\phi}_1(\omega)$ such that $\sigma_{i_0}^2(\gamma_{N_\ell}) \to 0$. Using that $\widehat{a}_{N,i_0}(\omega) = \sigma_{i_0,n_{i_0}}^2 \{ \widehat{\phi}_1(\omega) \} - \sigma_{i_0}^2 \{ \widehat{\phi}_1(\omega) \} \to 0$, we get that $\sigma_{i_0,n_{i_0,\ell}}^2(\gamma_{N_\ell}) \to 0$, so $\varsigma_{N_\ell}(\gamma_{N_\ell}) \to -\infty$. Thus, since (S.2) holds, we get that $\varsigma_f(\phi_{f,1}) = -\infty$ which contradicts the fact that $\lambda_{f,i,1} > 0$, concluding the proof of (S.6).

Hence, there exists $\epsilon > 0$ such that $\sigma_i^2\{\widehat{\phi}_1(\omega)\} \in [\epsilon, +\infty)$, for all $i = 1, \ldots, k$ and $n \in \mathbb{N}$. Using that σ_i is weakly continuous and the unit ball is weakly compact, we have that the set $\{\sigma_i(\alpha) : \alpha \in \mathcal{V}_1\}$ is bounded, which entails that there exists A > 0 such that $\sigma_i^2\{\widehat{\phi}_1(\omega)\} \in [\epsilon, A]$, for any $i = 1, \ldots, k$ and $n \in \mathbb{N}$. Since $\widehat{a}_{N,i}(\omega) = \sigma_{i,n_i}^2\{\widehat{\phi}_1(\omega)\} - \sigma_i^2\{\widehat{\phi}_1(\omega)\} \to 0$, we obtain that, for N large enough, $\sigma_{i,n_i}^2\{\widehat{\phi}_1(\omega)\} \in [\epsilon/2, 2A]$ which together with the uniform continuity of the function f on $[\epsilon/2, 2A]$ entails that (S.3) holds concluding the proof of a).

b) Lemma 4 entails immediately that $\langle \hat{\phi}_1, \phi_{f,1} \rangle^2 \xrightarrow{a.s.} 1$. Without loss of generality, we can assume that $\hat{\phi}_1 \xrightarrow{a.s.} \phi_{f,1}$, since both σ_{i,n_i}^2 and σ_i^2 are invariant under sign changes.

The convergence given in assumption ii) implies that $\hat{\lambda}_{i,1} - \sigma_i^2(\hat{\phi}_1) = \sigma_{i,n_i}^2(\hat{\phi}_1) - \sigma_i^2(\hat{\phi}_1) = o_{\text{a.s.}}(1)$. On the other hand, the fact that $\hat{\phi}_1 \xrightarrow{a.s.} \phi_{f,1}$, together with the weak continuity of σ_i imply that $\sigma_i^2(\hat{\phi}_1) \xrightarrow{a.s.} \sigma_i^2(\phi_{f,1}) = \lambda_{f,i,1}$, concluding the proof of b).

c) The almost sure convergence of $\hat{\phi}_j$ to $\phi_{f,j}$, for $j = 1, \ldots, m-1$, implies

$$\sup_{|\alpha||=1} \|\widehat{\pi}_{m-1}\alpha - \pi_{m-1}\alpha\| \xrightarrow{a.s.} 0, \qquad (S.7)$$

which will be used in the sequel. As in a), to show that $\varsigma_N(\widehat{\phi}_m) \xrightarrow{a.s.} \varsigma_f(\phi_{f,m})$ it will be enough to prove that

$$\varsigma_f(\phi_{f,m}) \ge \varsigma_N(\widehat{\phi}_m) + o_{\text{a.s.}}(1), \tag{S.8}$$

$$\varsigma_f(\phi_{f,m}) \le \varsigma_N(\widehat{\phi}_m) + o_{\text{a.s.}}(1).$$
(S.9)

Recall that $\varsigma_f(\phi_{f,m}) = \sup_{\alpha \in S_1 \cap \mathcal{T}_{m-1}} \varsigma_f(\alpha)$. Therefore, since σ_R is a scale functional and f is increasing, we have $\sup_{\alpha \in \{\|\alpha\|=1\} \cap \mathcal{T}_{m-1}} \varsigma_f(\alpha) \ge \sup_{\|\alpha\|=1} \varsigma_f(\pi_{m-1}\alpha)$, which implies

$$\varsigma_f(\phi_{f,m}) \ge \varsigma_f(\pi_{m-1}\widehat{\phi}_m) = \varsigma_N(\widehat{\phi}_m) - \widehat{b}_m,$$

where $\hat{b}_m = \varsigma_N(\hat{\pi}_{m-1}\hat{\phi}_m) - \varsigma_f(\pi_{m-1}\hat{\phi}_m) = \varsigma_N(\hat{\phi}_m) - \varsigma_f(\pi_{m-1}\hat{\phi}_m)$. Hence, to derive (S.8) it is enough to show

$$\widehat{b}_m = \varsigma_N(\widehat{\pi}_{m-1}\widehat{\phi}_m) - \varsigma_f(\pi_{m-1}\widehat{\phi}_m) = o_{\text{a.s.}}(1).$$
(S.10)

Similarly, using that $\sigma_{\rm R}$ is a scale functional and f is increasing, we have

$$\varsigma_N(\widehat{\phi}_m) = \sup_{\|\alpha\|=1, \alpha \in \widehat{\mathcal{T}}_{m-1}} \varsigma_N(\alpha) \ge \varsigma_N(\widehat{\pi}_{m-1}\phi_{f,m}) = \varsigma_f(\phi_{f,m}) + \widehat{c}_m ,$$

where $\hat{c}_m = \varsigma_N(\hat{\pi}_{m-1}\phi_{f,m}) - \varsigma_f(\pi_{m-1}\phi_{f,m}) = \varsigma_N(\hat{\pi}_{m-1}\phi_{f,m}) - \varsigma_f(\phi_{f,m})$. Hence, to obtain (S.9) it is enough to show that

$$\widehat{c}_m = \varsigma_N(\widehat{\pi}_{m-1}\phi_{f,m}) - \varsigma_f(\pi_{m-1}\phi_{f,m}) = o_{\text{a.s.}}(1).$$
(S.11)

Finally, defining $\widehat{a}_N = \varsigma_N(\widehat{\phi}_m) - \varsigma_f(\widehat{\phi}_m)$, we have that

$$\varsigma_f(\widehat{\phi}_m) - \varsigma_f(\phi_{f,m}) = \varsigma_N(\widehat{\phi}_m) - \varsigma_f(\phi_{f,m}) - \widehat{a}_N,$$

so that the proof of c) will be concluded if we show (S.10), (S.11) and

$$\widehat{a}_N = \varsigma_N(\widehat{\phi}_m) - \varsigma_f(\widehat{\phi}_m) = o_{\text{a.s.}}(1).$$
(S.12)

As in a), we will distinguish the situation f defined at 0 and $f(0) = -\infty$.

c.i) Assume that $f:[0,\infty)\to\mathbb{R}$ is continuous and let us show that

$$\sup_{\|\alpha\|=1} |\varsigma_N(\widehat{\pi}_{m-1}\alpha) - \varsigma_f(\pi_{m-1}\alpha)| = o_{\text{a.s.}}(1) .$$
(S.13)

Note that

$$\sup_{\|\alpha\|=1} |\varsigma_N(\widehat{\pi}_{m-1}\alpha) - \varsigma_f(\pi_{m-1}\alpha)| \leq \sup_{\|\alpha\|=1} |\varsigma_N(\widehat{\pi}_{m-1}\alpha) - \varsigma_f(\widehat{\pi}_{m-1}\alpha)| + \sup_{\|\alpha\|=1} |\varsigma_f(\pi_{m-1}\alpha) - \varsigma_f(\widehat{\pi}_{m-1}\alpha)|.$$

Using (S.5) and the fact that $\|\alpha\| = 1$ implies $\|\widehat{\pi}_{m-1}\alpha\| \leq 1$, we get that the first term on the right hand side of the above inequality converges to zero almost surely. Hence, we only have to prove that $\sup_{\|\alpha\|=1} |\varsigma_f(\pi_{m-1}\alpha) - \varsigma_f(\widehat{\pi}_{m-1}\alpha)| \xrightarrow{a.s.} 0$. The fact that ς_f is weakly uniformly continuous over \mathcal{V}_1 and (S.7) implies that $\sup_{\|\alpha\|=1} |\varsigma_f(\pi_{m-1}\alpha) - \varsigma_f(\widehat{\pi}_{m-1}\alpha)| = o_{a.s.}(1)$, concluding the proof of (S.13).

Using (S.13) and the fact that $\|\widehat{\phi}_m\| = 1$, we obtain that (S.10) and (S.11) hold. On the other hand, the fact that $\|\widehat{\phi}_m\| = 1$ and (S.5) entail that (S.12), which concludes the proof of c) when f is continuous at 0.

c.ii) Suppose now that $f(0) = -\infty$. Denote $\hat{a}_{N,i} = \sigma_{i,n_i}^2(\hat{\phi}_m) - \sigma_i^2(\hat{\phi}_m), \hat{b}_{N,i} = \sigma_{i,n_i}^2(\hat{\pi}_{m-1}\hat{\phi}_m) - \sigma_i^2(\pi_{m-1}\hat{\phi}_m) = \sigma_{i,n_i}^2(\hat{\phi}_m) - \sigma_i^2(\pi_{m-1}\hat{\phi}_m), \hat{c}_{N,i} = \sigma_{i,n_i}^2(\hat{\pi}_{m-1}\phi_{f,m}) - \sigma_i^2(\phi_{f,m}).$ Note that (S.7) and the fact that $\|\hat{\phi}_m\| = \|\phi_{f,m}\| = 1$ entail that $\pi_{m-1}\hat{\phi}_m - \hat{\phi}_m$ and $\hat{\pi}_{m-1}\phi_{f,m} - \phi_{f,m}$ converge to 0 almost surely. Therefore, using that $\pi_{m-1}\hat{\phi}_m \in \mathcal{V}_1$ and $\hat{\phi}_m \in \mathcal{V}_1$ and the fact that σ_i is uniformly weakly continuous in \mathcal{V}_1 , we obtain that $\sigma_i^2(\pi_{m-1}\hat{\phi}_m) - \sigma_i^2(\hat{\phi}_m) \xrightarrow{a.s.} 0$. Similarly, we have that $\sigma_i^2(\hat{\pi}_{m-1}\phi_{f,m}) - \sigma_i^2(\phi_{f,m}) \xrightarrow{a.s.} 0$. Hence, the convergences given in assumption ii) allow to conclude that $\hat{a}_{N,i}, \hat{b}_{N,i}$ and $\hat{c}_{N,i}$ converge to 0 almost surely.

Arguing as in a), and using that $\sigma_i^2(\phi_{f,m}) > 0$, for all $i = 1, \ldots, k$, we get that (S.12), (S.10) and (S.11) hold.

d) We have already proved that when m = 1 the result holds. We proceed by induction and assume that $\langle \hat{\phi}_{\ell}, \phi_{f,\ell} \rangle^2 \xrightarrow{a.s.} 1$, for $\ell = 1, \ldots, m-1$, to show that $\langle \hat{\phi}_m, \phi_{f,m} \rangle^2 \xrightarrow{a.s.} 1$. Without loss of generality, we can assume that $\hat{\phi}_{\ell} \xrightarrow{a.s.} \phi_{f,\ell}$, for $\ell = 1, \ldots, m-1$. Using c) we have that $\varsigma_f(\hat{\phi}_m)$ and $\|\hat{\phi}_m\|$ converge to $\varsigma_f(\phi_{f,m})$ and 1 almost surely, respectively. Hence, from Lemma 4 we get that $\langle \hat{\phi}_m, \phi_{f,m} \rangle^2 \xrightarrow{a.s.} 1$. Without loss of generality we can assume that $\hat{\phi}_m \xrightarrow{a.s.} \phi_{f,m}$, since σ_{i,n_i}^2 and σ_i^2 are invariant under sign changes. Hence, as in

b) using that $\|\widehat{\phi}_m\| = 1$ and assumption ii), we get that $\widehat{\lambda}_{i,m} - \sigma_i^2(\widehat{\phi}_m) = \sigma_{i,n_i}^2(\widehat{\phi}_m) - \sigma_i^2(\widehat{\phi}_m) = o_{a.s.}(1)$. On the other hand, the fact that $\widehat{\phi}_m \xrightarrow{a.s.} \phi_{f,m}$ together with the weak continuity of σ_i implies that $\sigma_i^2(\widehat{\phi}_m) \xrightarrow{a.s.} \sigma_i^2(\phi_{f,m}) = \lambda_{f,i,m}$, concluding the proof of d). \Box

V. Proof of Theorem 2

Using that $X_{i,j}$ is elliptically distributed, we have that $\sigma_i^2(\alpha) = c_i \langle \alpha, \Gamma_i \alpha \rangle$ where $\phi_{f,1}(P) = \phi_1$. On the other hand, since $\sigma_i^2(\alpha) = c_i \langle \alpha, \Gamma_i \alpha \rangle$ and Γ_i is a compact operator, **C1** holds. Finally, since $\mathbb{P}(X_{i,1} = 0) = 0$ then $\lambda_{i,1} > 0$, and so we have that $\lambda_{f,i,1}(P) = \sigma_i^2(\phi_1) = c_i\lambda_{i,1} > 0$ since **C0** holds.

Define $A_N = \{X_{i,j} | || X_{i,j} ||, 1 \le i \le k, 1 \le j \le n_i\}$ and let

$$\widetilde{\phi}_1 = \widetilde{\phi}_{1,N} = \operatorname*{argmin}_{\alpha \in \mathcal{A}_N} 1 - \langle \alpha, \phi_1 \rangle^2$$

that is, $\langle \widetilde{\phi}_1, \phi_1 \rangle^2 = \max_{1 \le i \le k, 1 \le j \le n_i} \langle X_{i,j} / \| X_{i,j} \|, \phi_1 \rangle^2$.

We will begin proving that $\langle \tilde{\phi}_1, \phi_1 \rangle^2 \xrightarrow{a.s.} 1$. Since $\langle \tilde{\phi}_1, \phi_1 \rangle^2 \leq 1$, it is enough to show that, for all $\epsilon > 0$, $\lim_{M \to \infty} \mathbb{P} \left[\bigcup_{N \geq M} \{ (1 - \langle \tilde{\phi}_{1,N}, \phi_1 \rangle^2) > \epsilon \} \right] = 0$, which will follow from the fact that $\sum_{N \geq 1} \mathbb{P}(1 - \langle \tilde{\phi}_{1,N}, \phi_1 \rangle^2 > \epsilon) < \infty$. We have that

$$\begin{split} \mathbb{P}(1-\langle \widetilde{\phi}_1, \phi_1 \rangle^2) > \epsilon) &= \mathbb{P}\left(\max_{1 \le i \le k, 1 \le j \le n_i} \langle \frac{X_{i,j}}{\|X_{i,j}\|}, \phi_1 \rangle^2 < 1-\epsilon\right) \\ &= \prod_{i=1}^k \prod_{j=1}^{n_i} \mathbb{P}\left(\langle \frac{X_{i,j}}{\|X_{i,j}\|}, \phi_1 \rangle^2 < 1-\epsilon\right) = \prod_{i=1}^k \mathbb{P}\left(\langle \frac{X_{i,1}}{\|X_{i,1}\|}, \phi_1 \rangle^2 < 1-\epsilon\right)^{n_i} \\ &= \left\{\prod_{i=1}^k \mathbb{P}\left(\langle \frac{X_{i,1}}{\|X_{i,1}\|}, \phi_1 \rangle^2 < 1-\epsilon\right)^{q_{i,N}}\right\}^N. \end{split}$$

Let i_1 be such that $\mathbb{P}(X_{i_1,1}=0)=0$ and denote $p_{i_1}=\mathbb{P}\left(\langle X_{i_1,1}/||X_{i_1,1}||, \phi_1\rangle^2 < 1-\epsilon\right)$. Then, Lemma S.5 implies that $p_{i_1}<1$. Hence, using that

$$\mathbb{P}(1 - \langle \widetilde{\phi}_1, \phi_1 \rangle^2 > \epsilon) = \left(\prod_{i=1}^k p_i^{q_{i,N}}\right)^N = \left(\prod_{i \neq i_1}^k p_i^{q_{i,N}}\right)^N (p_{i_1}^{q_{i_1,N}})^N \le (p_{i_1}^{q_{i_1}})^{N\frac{q_{i_1,N}}{q_{i_1}}}$$

the fact that $0 < q_{i_1} < 1$ implies that $\langle \tilde{\phi}_1, \phi_1 \rangle^2 \xrightarrow{a.s.} 1$. Without loss of generality, we will assume that $\tilde{\phi}_1 \xrightarrow{a.s.} \phi_1$.

The continuity of σ_i entails that $\sigma_i^2(\widetilde{\phi}_1) - \sigma_i^2(\phi_1) \xrightarrow{a.s.} 0$. Besides, **C3** implies that $s_{i,n_i}^2(\widetilde{\phi}_1) - \sigma_i^2(\widetilde{\phi}_1) \xrightarrow{a.s.} 0$, hence using that Lemma 2 entails that $\lambda_{f,1}(P) = c_i \lambda_{i,1}$, for some $c_i > 0$, we get

$$s_{i,n_i}^2(\widetilde{\phi}_1) \xrightarrow{a.s.} \sigma_i^2(\phi_1) = \lambda_{f,i,1}(P) > 0 .$$
(S.14)

Using that f is a continuous function on $(0, \infty)$ and $\lambda_{f,i,1}(P) > 0$, we get that (S.14) implies that $\widehat{\varsigma}(\widetilde{\phi}_1) \xrightarrow{a.s.} \varsigma_f(\phi_1)$, so that

$$\widehat{b}_N = \widehat{\varsigma}(\widetilde{\phi}_1) - \varsigma_f(\phi_1) \xrightarrow{a.s.} 0.$$
(S.15)

On the other hand, since $\tilde{\phi}_1 \in \mathcal{A}_N$, $\hat{\phi}_{f,CR,1} = \operatorname{argmax}_{\alpha \in \mathcal{A}_N} \hat{\varsigma}(\alpha)$, $\phi_1 = \operatorname{argmax}_{\|\alpha\|=1} \varsigma_f(\alpha)$ and $\|\hat{\phi}_{f,CR,1}\| = 1$, we have the following inequalities

$$\widehat{\varsigma}(\widehat{\phi}_1) \leq \widehat{\varsigma}(\widehat{\phi}_{f,\mathrm{CR},1}), \qquad (S.16)$$

$$\varsigma_f(\phi_1) \leq \varsigma_f(\phi_1),$$
 (S.17)

$$\varsigma_f(\phi_{f,\mathrm{CR},1}) \leq \varsigma_f(\phi_1).$$
 (S.18)

Denote $\hat{a}_N = \hat{\varsigma}(\hat{\phi}_{f,CR,1}) - \varsigma_f(\hat{\phi}_{f,CR,1})$. From (S.16) and (S.18), we obtain

$$\varsigma_f(\phi_1) = \widehat{\varsigma}(\widetilde{\phi}_1) - \widehat{b}_N \le \widehat{\varsigma}(\widehat{\phi}_{f,\mathrm{CR},1}) - \widehat{b}_N = \varsigma_f(\widehat{\phi}_{f,\mathrm{CR},1}) + \widehat{a}_N - \widehat{b}_N \le \varsigma_f(\phi_1) + \widehat{a}_N - \widehat{b}_N \,. \tag{S.19}$$

Assume that we have shown that

$$\widehat{a}_N = o_{\text{a.s.}}(1) . \tag{S.20}$$

Then using (S.19) and (S.15), we get that $\varsigma_f(\widehat{\phi}_{f,CR,1}) \xrightarrow{a.s.} \varsigma_f(\phi_1)$ and the result follows from Lemma 4.

In order to prove (S.20), and as in the proof of Theorem 1, we will distinguish the case in which f is defined at 0 and when $f(0) = -\infty$.

i) If $|f(0)| < \infty$ and f is continuous at 0, Corollary S.1 implies that $\sup_{\|\alpha\|=1} |\widehat{\varsigma}(\alpha) - \varsigma_f(\alpha)| \xrightarrow{a.s.} 0$, so that (S.20) holds.

ii) Consider now the case $f(0) = -\infty$. Let \mathcal{N}_1 the 0-measure set where **C3** does not hold, $\mathcal{N}_2 = \{\omega \in \Omega : \widehat{b}_N \neq 0\}$ and $\mathcal{N}_3 = \{\omega \in \Omega : \widetilde{\phi}_1 \neq \phi_1\}$. Define $\mathcal{N} = \bigcup_{i=1}^3 \mathcal{N}_i$. Therefore, if $\omega \notin \mathcal{N}$, we have that

$$\sup_{\|\alpha\|=1} \left| s_{i,n_i}^2(\alpha) - \sigma_i^2(\alpha) \right| \to 0 , \quad \widehat{b}_N \to 0 , \quad \widetilde{\phi}_1 \to \phi_1 .$$
(S.21)

Noting that $\varsigma_f(\phi_1) + \widehat{b}_N \leq \widehat{\varsigma}(\widehat{\phi}_{f,CR,1})$, we obtain that

$$\liminf_{N \to \infty} \widehat{\varsigma}\{\widehat{\phi}_{f,\mathrm{CR},1}(\omega)\} \ge \varsigma_f(\phi_1) . \tag{S.22}$$

The fact that $\lambda_{f,i,1} = \sigma_i^2(\phi_1) > 0$, for all $1 \le i \le k$ implies that $|\varsigma_f(\phi_1)| < \infty$ which allows us to conclude that, for all $1 \le i \le k$, $\liminf_{N\to\infty} \sigma_i^2\{\widehat{\phi}_{f,CR,1}(\omega)\} > 0$. Indeed, suppose that this does not hold, then, there exist i_0 such that $\liminf_{N\to\infty} \sigma_{i_0}^2\{\widehat{\phi}_{f,CR,1}(\omega)\} = 0$, that is, it exists a subsequence γ_{N_ℓ} of $\gamma_N = \widehat{\phi}_{f,CR,1}(\omega)$ such that $\sigma_{i_0}^2(\gamma_{N_\ell}) \to 0$. Since (S.21) holds, we have that $s_{i_0,q_{i_0}N_\ell}^2(\gamma_{N_\ell}) \to 0$ which implies that $\widehat{\varsigma}_{N_\ell}(\gamma_{N_\ell}) \to -\infty$, which contradicts (S.22). Thus, for all $1 \le i \le k$, $\liminf_{N\to\infty} \sigma_i^2\{\widehat{\phi}_{f,CR,1}(\omega)\} > 0$. Using analogous arguments to those considered in the proof of Theorem 1, it follows that (S.20) holds concluding the proof. \Box

References

- Bali, J. L. and Boente, G. (2014). Consistency of a numerical approximation to the first principal component projection pursuit estimator. Statist. Probab. Lett., 94, 181-191.
- [2] Bali, J. L., Boente, G., Tyler, D. E. & Wang, J. L. (2011). Robust Functional Principal Components: a projectionpursuit approach. Ann Statist., 39,2852-2882.
- [3] Boente G., Pires A. and Rodrigues I. (2006) General projection-pursuit estimators for the common principal components model: influence functions and Monte Carlo study. J. Multivar. Anal., 97, 124-147.