ROBUST FUNCTIONAL PRINCIPAL COMPONENTS: A PROJECTION-PURSUIT APPROACH

By Juan Lucas Bali[1], Graciela Boente[1,4],
David E. Tyler[2] and Jane-Ling Wang[3,4]

Universidad de Buenos Aires and CONICET, Universidad de Buenos Aires and CONICET, Rutgers University and University of California at Davis

In many situations, data are recorded over a period of time and may be regarded as realizations of a stochastic process. In this paper, robust estimators for the principal components are considered by adapting the projection pursuit approach to the functional data setting. Our approach combines robust projection-pursuit with different smoothing methods. Consistency of the estimators are shown under mild assumptions. The performance of the classical and robust procedures are compared in a simulation study under different contamination schemes.

1. Introduction. Analogous to classical principal components analysis (PCA), the projection-pursuit approach to robust PCA is based on finding projections of the data which have maximal dispersion. Instead of using the variance as a measure of dispersion, a robust scale estimator s_n is used for the maximization problem. This approach was introduced by Li and Chen (1985), who proposed estimators based on maximizing (or minimizing) a robust scale. In this way, given a sample $\mathbf{x}_i \in \mathbb{R}^d$, $1 \le i \le n$, the first robust principal component vector is defined as

$$\widehat{\mathbf{a}} = \underset{\{\mathbf{a} \in \mathbb{R}^d \,:\, \mathbf{a}^{\mathrm{T}}\mathbf{a}=1\}}{\arg\max} \; s_n(\mathbf{a}^{\mathrm{T}}\mathbf{x}_1, \ldots, \mathbf{a}^{\mathrm{T}}\mathbf{x}_n).$$

The subsequent principal component vectors are obtained by imposing orthogonality conditions. In the multivariate setting, Li and Chen (1985) argue that the breakdown point for this projection-pursuit based procedure is the same as that

of the scale estimator s_n. Later on, Croux and Ruiz-Gazen (2005) derived the influence functions of the resulting principal components, while their asymptotic distribution was studied in Cui, He and Ng (2003). A maximization algorithm for obtaining $\hat{\mathbf{a}}$ was proposed in Croux and Ruiz-Gazen (1996) and adapted for high-dimensional data in Croux, Filzmoser and Oliveira (2007).

The aim of this paper is to adapt the projection pursuit approach to the functional data setting. We focus on functional data that are recorded over a period of time and regarded as realizations of a stochastic process, often assumed to be in the L^2 space on a real interval. Various choices of robust scales, including the median of the absolute deviation about the median (MAD) and M-estimates of scale are considered and compared.

Classical functional PCA uses the eigenvalues and eigenfunctions of the sample covariance operator. Dauxois, Pousse and Romain (1982) have studied the asymptotic properties of these sample functional principal components. Rice and Silverman (1991) proposed to smooth the principal components by imposing an additive roughness penalty to the sample variance. The consistency of this method was subsequently studied by Pezzulli and Silverman (1993). Another approach to smoothing the principal components, proposed in Silverman (1996) and reviewed in Ramsay and Silverman (2005), is based on penalizing the norm rather than the sample variance, while Boente and Fraiman (2000) considered a kernel-based approach. More recent work on estimation of the principal components and the covariance function includes Gervini (2006), Hall and Hosseini-Nasab (2006), Hall, Müller and Wang (2006) and Yao and Lee (2006).

The literature on robust principal components in the functional data setting, though, is rather sparse. To our knowledge, the first attempt to provide estimators of the principal components that are less sensitive to anomalous observations was due to Locantore et al. (1999), although their approach is multivariate in nature. Gervini (2008) studied 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). Hyndman and Shahid Ullah (2007) give an application of a robust projection-pursuit approach, applied to smoothed trajectories, but did not study the properties of their method in detail.

In this paper, we introduce several robust estimators of the principal components in the functional data setting. Our approach uses a robust projection-pursuit combined with various smoothing methods. A primary focus of this paper is to provide a rigorous theoretical foundation for this approach to robust functional PCA. In particular, we establish under very general conditions the strong consistency of the our proposed estimators.

In Section 3, the robust estimators of the principal components, based on both the raw and smoothed approaches, are introduced. Consistency results and the asymptotic robustness of the procedure are established in Section 4, while Fisher-consistency of the related functionals is studied in Section 5. Section 6 provides conditions under which one of the crucial assumptions hold. Selection of the

smoothing parameters for the smooth principal components is discussed in Section 7. The results of a Monte Carlo study are reported in Section 8. Finally, Section 9 contains some concluding remarks. Most proofs are relegated to the Appendix and to the technical supplementary material available online; see Bali et al. (2011a). We begin the next section with notation and a review of some basic concepts which are utilized in later sections.

2. Preliminaries.

2.1. *Functional principal components analysis.* Principal components analysis, which was originally developed for multivariate data, has been successfully extended to accommodate functional data, and is usually referred to as functional PCA. Principal components analysis for general Hilbert spaces can be described as follows.

Let $X \in \mathcal{H}$ be a random element of a Hilbert space $\mathcal{H}$ defined in $(\Omega, \mathcal{A}, P)$. Denote by $\langle \cdot, \cdot \rangle$ the inner product in $\mathcal{H}$ and by $\|\alpha\|^2 = \langle \alpha, \alpha \rangle$. Assume that X has finite second moment, that is, $\mathbb{E}(\|X\|^2) < \infty$. The bilinear operator $a_X : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ defined as $a_X(\alpha, \beta) = \mathrm{cov}(\langle \alpha, X \rangle, \langle \beta, X \rangle)$ leads to a continuous operator. The Riesz representation theorem then implies that there exists a bounded operator, $\Gamma_X : \mathcal{H} \to \mathcal{H}$, such that $a_X(\alpha, \beta) = \langle \alpha, \Gamma_X \beta \rangle$. The operator Γ_X is called the covariance operator of X and is linear, self-adjoint and continuous.

Although the general situation in which $X \in \mathcal{H}$ is treated in this paper, to help simplify the basic ideas, we first consider the case $X \in L^2(\mathcal{I})$ where $\mathcal{I} \subset \mathbb{R}$ is a finite interval. We take the usual inner product for $L^2(\mathcal{I})$, that is, $\langle \alpha, \beta \rangle = \int_{\mathcal{I}} \alpha(t) \beta(t) \, dt$ and denote the covariance function of X by $\gamma_X(t, s) = \mathrm{cov}(X(t), X(s))$. The corresponding covariance operator $\Gamma_X : L^2(\mathcal{I}) \to L^2(\mathcal{I})$ is such that $\Gamma_X(\alpha)(t) = \int_{\mathcal{I}} \gamma_X(t, s) \alpha(s) \, ds$. It is assumed the covariance function satisfies $\int_{\mathcal{I}} \int_{\mathcal{I}} \gamma_X^2(t, s) \, dt \, ds < \infty$. Consequently, Γ_X is a Hilbert–Schmidt operator.

A Hilbert–Schmidt operator has a countable number of eigenvalues, all of which are real. $\mathcal{F}$ will stand for the Hilbert space of such operators with inner product defined by $\langle \Gamma_1, \Gamma_2 \rangle_{\mathcal{F}} = \sum_{j=1}^{\infty} \langle \Gamma_1 u_j, \Gamma_2 u_j \rangle$, where $\{u_j : j \geq 1\}$ is any orthonormal basis of $L^2(\mathcal{I})$. Furthermore, since the covariance operator Γ_X is also positive semi-definite, its eigenvalues are nonnegative. As with symmetric matrices in finite-dimensional Euclidean spaces, one can choose the eigenfunctions of a Hilbert–Schmidt operator so that they form an orthonormal basis for $L^2(\mathcal{I})$. Let $\{\phi_j : j \geq 1\}$ and $\{\lambda_j : j \geq 1\}$ be respectively an orthonormal basis of eigenfunctions and their corresponding eigenvalues for the covariance operator Γ_X, with $\lambda_j \geq \lambda_{j+1}$. The spectral value decomposition for Γ_X can then be expressed as $\Gamma_X = \sum_{j=1}^{\infty} \lambda_j \phi_j \otimes \phi_j$, with $\otimes$ being the tensor product, or equivalently $\gamma_X(t, s) = \sum_{j=1}^{\infty} \lambda_j \phi_j(t) \phi_j(s)$, with $\sum_{j=1}^{\infty} \lambda_j^2 = \|\Gamma_X\|_{\mathcal{F}}^2 = \int_{\mathcal{I}} \int_{\mathcal{I}} \gamma_X^2(t, s) \, dt \, ds$. The jth principal component variable is then defined as $Z_j = \langle \phi_j, X \rangle$, which leads to the Karhunen–Loève expansion $X(t) = \mu(t) + \sum_{j=1}^{\infty} Z_j \phi_j(t)$, with $\mu(t) =$

$\mathbb{E}(X(t))$ and the Z_j's being uncorrelated and having variances λ_j in descending order.

In general, for $Y = \langle \alpha, X \rangle$, which is a linear function of the process $\{X(s)\}$, we have $\text{var}(Y) = \langle \alpha, \mathbf{\Gamma}_X \alpha \rangle$. An important optimality property of the first principal component variable is that it can be defined as the variable $Z_1 = \langle \alpha_1, X \rangle$ such that

$$(2.1) \qquad \text{var}(Z_1) = \sup_{\{\alpha \,:\, \|\alpha\|=1\}} \text{var}(\langle \alpha, X \rangle) = \sup_{\{\alpha \,:\, \|\alpha\|=1\}} \langle \alpha, \mathbf{\Gamma}_X \alpha \rangle.$$

Any solution to (2.1), that is, any α for which the supremum is obtained, corresponds to an eigenfunction associated with the largest eigenvalue of the covariance operator $\mathbf{\Gamma}_X$, that is, $\alpha_1 = \phi_1$ and $\text{var}(Z_1) = \lambda_1$. If $\lambda_1 > \lambda_2$, then α_1 is unique up to a sign change. As in the multivariate setting, the other principal components can be obtained successively via (2.1), but under the orthogonality condition that $\langle \alpha_j, \alpha_k \rangle = 0$ for $j < k$.

2.2. *Scale functionals and estimates.* The basic idea underlying our approach is to view principal components as in (2.1), but with the variance replaced by a robust scale functional. We first recall the definition of a scale functional. Denote by $\mathcal{G}$ the set of all univariate distributions. A scale functional $\sigma_R : \mathcal{G} \to [0, +\infty)$ is one which is location invariant and scale equivariant, that is, if $G_{a,b}$ stands for the distribution of $aY + b$ when $Y \sim G$, then, $\sigma_R(G_{a,b}) = |a|\sigma_R(G)$, for all real numbers a and b. Two well-known examples of scale functionals are the standard deviation, $\text{SD}(G) = \{\mathbb{E}(Y - \mathbb{E}(Y))^2\}^{1/2}$, where $Y \sim G$, and the median absolute deviation about the median, $\text{MAD}(G) = c\,\text{median}(|Y - \text{median}(Y)|)$. The normalization constant c, used in the MAD, can be chosen so that its empirical or sample version is consistent for a scale parameter of interest. Typically, one chooses $c = 1/\Phi^{-1}(0.75)$ so that the MAD equals the standard deviation at a normal distribution.

The breakdown points, a measure of global robustness, for the standard deviation and the MAD are 0 and $1/2$, respectively. The MAD, however, has a discontinuous influence function, which reflects some local instability. Furthermore, the empirical version of the MAD is known to be fairly inefficient at the normal and other distributions; see Huber (1981). In the finite-dimensional setting, as reported in Cui, He and Ng (2003) the impact of a discontinuous influence function on the efficiency of the estimators of the principal directions is even more dramatic covariance function.

A broader class of robust scale functionals, which includes as special cases both the SD and the MAD, are the M-scale functionals. An M-scale functional with a bounded and continuous score function can have both a high breakdown point and a continuous and bounded influence function. Also, their empirical versions, the M-estimates of scale, can be tuned to have good efficiency over a broad range of distributions. Given a location parameter μ, an M-scale functionals $\sigma_M(G)$ with a

continuous score function $\chi : \mathbb{R} \to \mathbb{R}$ can be defined to be a solution to the equation

$$(2.2) \qquad \mathbb{E}\left[\chi\left(\frac{Y - \mu}{\sigma_{\mathrm{R}}(G)} \right) \right] = \delta.$$

Given a location statistic $\widehat{\mu}_n$, the corresponding M-estimate of scale is then a solution $\widehat{\sigma}_n$ to the M-estimating equation

$$(2.3) \qquad \frac{1}{n} \sum_{i=1}^{n} \chi\left(\frac{Y_i - \widehat{\mu}_n}{\widehat{\sigma}_n} \right) = \delta.$$

If the score function is discontinuous, as is the case with the MAD, then a slight modification to (2.2) and (2.3) is needed; see Martin and Zamar (1993).

Typically, the score function χ is even with $\chi(0) = 0$, nondecreasing on $\mathbb{R}_+$ and with $0 < \sup_{x \in \mathbb{R}} \chi(x) = \chi(+\infty) = \lim_{x \to +\infty} \chi(x)$. When $\chi(+\infty) = 2\delta$, the M-estimate of scale has a 50% breakdown point, and by choosing χ properly one can also obtain a highly efficient estimate; see Croux (1994). One such popular choice, and the one we use in our Monte Carlo study, is the score function introduced by Beaton and Tukey (1974), namely

$$(2.4) \qquad \chi_c(y) = \min\big(3(y/c)^2 - 3(y/c)^4 + (y/c)^6, 1\big)$$

with c being a tuning constant chosen so that the corresponding M-estimator of scale is consistent for a scale parameter of interest. For example, the choice $c = 1.56$ when $\delta = 1/2$ ensures that the M-scale functional is Fisher-consistent at the normal distribution and has a 50% breakdown point.

For continuous and nondecreasing score functions χ, the solutions to (2.2) and (2.3) are unique, and the simple re-weighting algorithm

$$\{\widehat{\sigma}_n^{(k+1)}\}^2 = \frac{1}{n\delta} \sum_{i=1}^{n} w\left(\frac{Y_i - \widehat{\mu}}{\widehat{\sigma}_n^{(k)}} \right) (Y_i - \widehat{\mu})^2,$$

where $w(y) = \chi(y)/y^2$ for $y \neq 0$ and $w(0) = \chi''(0)$, is known to always converge to the unique solution of (2.3) regardless of the initial value $\widehat{\sigma}_n^{(0)}$. In practice, the initial value $\widehat{\sigma}_n^{(0)}$ is usually taken to be the MAD. A discussion on the convergence of the algorithm can be found in Maronna, Martin and Yohai (2006).

For a bounded score function χ, if the solution $\sigma_{\mathrm{R}}(G_0)$ of (2.2) is unique, as it is the case when χ is continuous and nondecreasing, then the functional σ_{R} is weakly continuous at G_0. Weakly continuity of σ_{R} at G_0, that is, continuity with respect to the weak topology in $\mathcal{G}$ which is given by the Prohorov metric, and consistency in a neighborhood of G_0 entails robustness at G_0. For details, see Huber (1981) and Hampel (1971).

3. The estimators. We consider several robust approaches in this section and define them on a separable Hilbert space $\mathcal{H}$, keeping in mind that the main application will be $\mathcal{H} = L^2(\mathcal{I})$. From now on and throughout the paper, $\{X_i : 1 \leq i \leq n\}$ denote realizations of the stochastic process $X \sim P$ in a separable Hilbert space $\mathcal{H}$. Thus, $X_i \sim P$ are independent stochastic processes that follow the same law. This independence condition could be relaxed, since we only need the strong law of large numbers to hold in order to guarantee the results in this paper.

3.1. *Raw robust projection-pursuit approach.* Based on property (2.1) of the first principal component and given $\sigma_R(F)$ a robust scale functional, the raw (meaning unsmoothed) robust functional principal component directions are defined as

$$(3.1) \quad \begin{cases} \phi_{R,1}(P) = \underset{\|\alpha\|=1}{\arg\max}\, \sigma_R(P[\alpha]), \\ \phi_{R,m}(P) = \underset{\|\alpha\|=1, \alpha \in \mathcal{B}_m}{\arg\max}\, \sigma_R(P[\alpha]), \qquad 2 \leq m, \end{cases}$$

where $P[\alpha]$ stands for the distribution of $\langle \alpha, X \rangle$ when $X \sim P$ and $\mathcal{B}_m = \{\alpha \in \mathcal{H} : \langle \alpha, \phi_{R,j}(P) \rangle = 0, 1 \leq j \leq m - 1\}$. We will denote the mth largest principal value by

$$(3.2) \quad \lambda_{R,m}(P) = \sigma_R^2(P[\phi_{R,m}]) = \underset{\|\alpha\|=1, \alpha \in \mathcal{B}_m}{\max}\, \sigma_R^2(P[\alpha]).$$

Since the unit ball is weakly compact, the maximum above is attained if the scale functional σ_R is (weakly) continuous.

Next, denote by $s_n^2 : \mathcal{H} \to \mathbb{R}$ the function $s_n^2(\alpha) = \sigma_R^2(P_n[\alpha])$, where $\sigma_R(P_n[\alpha])$ stands for the functional σ_R computed at the empirical distribution of $\langle \alpha, X_1 \rangle, \ldots, \langle \alpha, X_n \rangle$. Analogously, the mapping $\sigma : \mathcal{H} \to \mathbb{R}$ stands for $\sigma(\alpha) = \sigma_R(P[\alpha])$. The components in (3.1) will be estimated empirically by

$$(3.3) \quad \begin{cases} \widehat{\phi}_{RAW,1} = \underset{\|\alpha\|=1}{\arg\max}\, s_n(\alpha), \\ \widehat{\phi}_{RAW,m} = \underset{\alpha \in \widehat{\mathcal{B}}_m}{\arg\max}\, s_n(\alpha), \qquad 2 \leq m, \end{cases}$$

where $\widehat{\mathcal{B}}_m = \{\alpha \in \mathcal{H} : \|\alpha\| = 1, \langle \alpha, \widehat{\phi}_{RAW,j} \rangle = 0, \forall 1 \leq j \leq m - 1\}$. The estimators of the principal values are then computed as

$$(3.4) \quad \widehat{\lambda}_{RAW,m} = s_n^2(\widehat{\phi}_{RAW,m}), \qquad 1 \leq m.$$

3.2. *Smoothed robust principal components.* Sometimes instead of raw functional principal components, smoothed ones are of interest. The advantages of smoothed functional PCA are well documented; see, for instance, Rice and Silverman (1991) and Ramsay and Silverman (2005). One compelling argument is that

smoothing is a regularization tool that might reveal more interpretable and interesting features of the modes of variation for functional data. As noted in the Introduction, Rice and Silverman (1991) and Silverman (1996) proposed two smoothing approaches by penalizing the variance and the norm, respectively.

To be more specific, Rice and Silverman (1991) estimate the first principal component by maximizing over $\|\alpha\| = 1$, the objective function $\widehat{\mathrm{var}}(\langle \alpha, X \rangle) - \rho \lceil \alpha, \alpha \rceil$, where $\widehat{\mathrm{var}}$ stands for the sample variance and $\lceil \alpha, \beta \rceil = \int_0^1 \alpha''(t)\beta''(t)\,dt$. Consistency for these estimators was established by Pezzulli and Silverman (1993).

Another regularization method proposed by Silverman (1996) is to penalize the roughness through a norm defined via the penalized inner product, $\langle \alpha, \beta \rangle_\tau = \langle \alpha, \beta \rangle + \tau \lceil \alpha, \beta \rceil$. The smoothed first direction $\widehat{\phi}_1$ is then the one that maximizes $\widehat{\mathrm{var}}(\langle \alpha, X \rangle)$ over $\|\alpha\|_\tau = 1$. Consistency of these estimators is also established in Silverman (1996) under the assumption that ϕ_j has finite roughness, that is, $\lceil \phi_j, \phi_j \rceil < \infty$.

Clearly the smoothing parameters ρ and τ need to converge to 0 in order to get consistency results.

Let us consider $\mathcal{H}_\mathrm{S}$, the subset of "smooth elements" of $\mathcal{H}$. In order to obtain consistency results, we will assume that $\phi_{\mathrm{R},j}(P) \in \mathcal{H}_\mathrm{S}$. Let $D : \mathcal{H}_\mathrm{S} \to \mathcal{H}$ be a linear operator, which we will refer to as the "differentiator." Using D, we define the symmetric positive semidefinite bilinear form $\lceil \cdot, \cdot \rceil : \mathcal{H}_\mathrm{S} \times \mathcal{H}_\mathrm{S} \to \mathbb{R}$, where $\lceil \alpha, \beta \rceil = \langle D\alpha, D\beta \rangle$. The "penalization operator" is then defined as $\Psi : \mathcal{H}_\mathrm{S} \to \mathbb{R}$, $\Psi(\alpha) = \lceil \alpha, \alpha \rceil$, and the penalized inner product as $\langle \alpha, \beta \rangle_\tau = \langle \alpha, \beta \rangle + \tau \lceil \alpha, \beta \rceil$. Therefore, $\|\alpha\|_\tau^2 = \|\alpha\|^2 + \tau \Psi(\alpha)$. As in Pezzulli and Silverman (1993), we will assume that the bilinear form is closable.

REMARK 3.1. The most common setting for functional data is to choose $\mathcal{H} = L^2(\mathcal{I})$, $\mathcal{H}_\mathrm{S} = \{\alpha \in L^2(\mathcal{I}), \alpha \text{ is twice differentiable, and } \int_{\mathcal{I}} (\alpha''(t))^2 \, dt < \infty\}$, $D\alpha = \alpha''$ and $\lceil \alpha, \beta \rceil = \int_{\mathcal{I}} \alpha''(t)\beta''(t)\,dt$ so that $\Psi(\alpha) = \int_{\mathcal{I}} (\alpha''(t))^2 \, dt$.

Let $\sigma_\mathrm{R}(F)$ be a robust scale functional and define $s_n(\alpha)$ and $\sigma(\alpha)$ as in Section 3.1. Then we can adapt the classical procedure by defining the smoothed robust functional principal direction estimators either:

(a) by penalizing the norm as

$$(3.5) \quad \begin{cases} \widehat{\phi}_{\mathrm{PN},1} = \underset{\|\alpha\|_\tau = 1}{\arg\max}\, s_n^2(\alpha) = \underset{\beta \neq 0}{\arg\max}\, \dfrac{s_n^2(\beta)}{\langle \beta, \beta \rangle + \tau \Psi(\beta)}, \\ \widehat{\phi}_{\mathrm{PN},m} = \underset{\alpha \in \widehat{\mathcal{B}}_{m,\tau,\mathrm{PN}}}{\arg\max}\, s_n^2(\alpha), \qquad 2 \leq m, \end{cases}$$

where $\widehat{\mathcal{B}}_{m,\tau,\mathrm{PN}} = \{\alpha \in \mathcal{H} : \|\alpha\|_\tau = 1, \langle \alpha, \widehat{\phi}_{\mathrm{PN},j} \rangle_\tau = 0, \forall 1 \leq j \leq m-1\}$,

(b) or by penalizing the scale as

(3.6)
$$\begin{cases} \widehat{\phi}_{\mathrm{PS},1} = \underset{\|\alpha\|=1}{\arg\max}\{s_n^2(\alpha) - \rho\Psi(\alpha)\}, \\ \widehat{\phi}_{\mathrm{PS},m} = \underset{\alpha\in\widehat{\mathcal{B}}_{m,\mathrm{PS}}}{\arg\max}\{s_n^2(\alpha) - \rho\Psi(\alpha)\}, \qquad 2 \le m, \end{cases}$$

where $\widehat{\mathcal{B}}_{m,\mathrm{PS}} = \{\alpha \in \mathcal{H} : \|\alpha\| = 1, \langle\alpha, \widehat{\phi}_{\mathrm{PS},j}\rangle = 0, \forall 1 \le j \le m-1\}$.

The corresponding principal value estimators are respectively defined as

(3.7)
$$\widehat{\lambda}_{\mathrm{PS},m} = s_n^2(\widehat{\phi}_{\mathrm{PS},m}) \quad \text{and} \quad \widehat{\lambda}_{\mathrm{PN},m} = s_n^2(\widehat{\phi}_{\mathrm{PN},m}).$$

3.3. *Sieve approach for robust functional principal components.* Another approach, motivated by using B-splines as a smoothing tool, is to consider the method of sieves. The method of sieves involves approximating an infinite-dimensional parameter space Θ by a sequence of finite-dimensional parameter spaces Θ_n, which depend on the sample size n, and then estimate the parameters on the spaces Θ_n rather than Θ.

Let $\{\delta_i\}_{i\ge 1}$ be a basis of $\mathcal{H}$ and define $\mathcal{H}_{p_n}$ as the linear space spanned by $\delta_1, \ldots, \delta_{p_n}$ and by $\mathcal{S}_{p_n} = \{\alpha \in \mathcal{H}_{p_n} : \|\alpha\| = 1\}$, that is,

$$\mathcal{H}_{p_n} = \left\{\alpha \in \mathcal{H} : \alpha = \sum_{j=1}^{p_n} a_j\delta_j\right\}$$

and $\mathcal{S}_{p_n} = \{\alpha \in \mathcal{H} : \alpha = \sum_{j=1}^{p_n} a_j\delta_j$, such that $\|\alpha\|^2 = \sum_{j=1}^{p_n}\sum_{s=1}^{p_n} a_j a_s \langle\delta_j, \delta_s\rangle = 1\}$. Note that $\mathcal{S}_{p_n}$ approximates the unit sphere $\mathcal{S} = \{\alpha \in \mathcal{H} : \|\alpha\| = 1\}$. When $\{\delta_i\}_{i\ge 1}$ is an orthonormal basis, $\|\alpha\|^2 = \sum_{j=1}^{p_n} a_j^2 = \mathbf{a}^{\mathrm{T}}\mathbf{a}$ where $\mathbf{a} = (a_1, \ldots, a_{p_n})^{\mathrm{T}}$, hence, the norm of α equals the Euclidean norm of the vector $\mathbf{a}$. Define the robust sieve estimators of the principal components as

(3.8)
$$\begin{cases} \widehat{\phi}_{\mathrm{SI},1} = \underset{\alpha\in\mathcal{S}_{p_n}}{\arg\max}\, s_n(\alpha), \\ \widehat{\phi}_{\mathrm{SI},m} = \underset{\alpha\in\widehat{\mathcal{B}}_{n,m}}{\arg\max}\, s_n(\alpha), \qquad 2 \le m, \end{cases}$$

where $\widehat{\mathcal{B}}_{n,m} = \{\alpha \in \mathcal{S}_{p_n} : \langle\alpha, \widehat{\phi}_{\mathrm{SI},j}\rangle = 0, \forall 1 \le j \le m-1\}$, and let the principal value estimators be

(3.9)
$$\widehat{\lambda}_{\mathrm{SI},m} = s_n^2(\widehat{\phi}_{\mathrm{SI},m}).$$

Some of the frequently used bases for functional data are the Fourier, polynomial, spline and wavelet bases; see, for instance, Ramsay and Silverman (2005).

To the best of our knowledge, the above sieve approach is new to functional principal component analysis, even if one considers the classical sieve estimators, that is, when σ_{R} in (3.8) is the standard deviation.

3.4. *A unified formulation for the robust projection pursuit approaches.* To help formulate a unified approach to the different estimators considered in Sections 3.2, 3.2 and 3.3, let the products $\rho \Psi(\alpha)$ or $\tau \Psi(\alpha)$ be defined as 0 when $\rho = 0$ or $\tau = 0$, respectively, even when $\alpha \notin \mathcal{H}_S$ for which case $\Psi(\alpha) = \infty$. Moreover, when $p_n = \infty$, define $\mathcal{H}_{p_n} = \mathcal{H}$. All the projection pursuit estimators considered in the previous subsections then can be viewed as special cases of the following general class of estimators:

$$(3.10) \qquad \begin{cases} \widehat{\phi}_1 = \underset{\alpha \in \mathcal{H}_{p_n}, \|\alpha\|_\tau = 1}{\arg\max} \{s_n^2(\alpha) - \rho \Psi(\alpha)\}, \\ \widehat{\phi}_m = \underset{\alpha \in \widehat{\mathcal{B}}_{m,\tau}}{\arg\max} \{s_n^2(\alpha) - \rho \Psi(\alpha)\}, \qquad 2 \le m, \end{cases}$$

where $\widehat{\mathcal{B}}_{m,\tau} = \{\alpha \in \mathcal{H}_{p_n} : \|\alpha\|_\tau = 1, \langle \alpha, \widehat{\phi}_j \rangle_\tau = 0, \forall 1 \le j \le m - 1\}$.

With this definition and by taking $p_n = \infty$, the raw estimators are obtained when $\rho = \tau = 0$, while $\widehat{\phi}_{\mathrm{PN},m}$ and $\widehat{\phi}_{\mathrm{PS},m}$ correspond to $\rho = 0$ and $\tau = 0$, respectively. On the other hand, the sieve estimators correspond a finite choice for p_n and $\tau = \rho = 0$.

4. Consistency results.

In this section, we show that under mild conditions the functionals $\phi_{\mathrm{R},m}(P)$ and $\lambda_{\mathrm{R},m}(P)$ defined through (3.1) and (3.2) are weakly continuous. Moreover, we state conditions that guarantee the consistency of the estimators defined in Section 3. Proofs for this section can be found in the Appendix and in the supplemental article [Bali et al. (2011a)].

To derive the consistency of the proposed estimators, we need the following assumptions:

(S0) For some $q \ge 2$ and $1 \le j \le q$, $\phi_{\mathrm{R},j}(P)$ are unique up to a sign change.

(S1) $\sigma : \mathcal{H} \to \mathbb{R}$ is a weakly continuous function, that is, continuous with respect to the weak topology in $\mathcal{H}$.

(S2) $\sup_{\|\alpha\|=1} |s_n^2(\alpha) - \sigma^2(\alpha)| \xrightarrow{\text{a.s.}} 0$.

Note that condition (S0) holds if and only if $\lambda_{\mathrm{R},1}(P) > \cdots > \lambda_{\mathrm{R},q+1}(P)$.

Some remarks. (i) (S1) holds when the scale functional σ_{R} is a continuous functional (with respect to the weak topology under the Prohorov distance). This is because α_k converges weakly to α, which implies $\langle \alpha_k, X \rangle \xrightarrow{\omega} \langle \alpha, X \rangle$ and hence $\sigma_{\mathrm{R}}(P[\alpha_k]) \to \sigma_{\mathrm{R}}(P[\alpha])$. For the case when the scale functional is the standard deviation, and the underlying probability P has a compact covariance operator $\boldsymbol{\Gamma}_X$, we see from the relationship $\sigma^2(\alpha) = \langle \alpha, \boldsymbol{\Gamma}_X \alpha \rangle$ that condition (S1) holds, even though the standard deviation itself is not a weakly continuous functional.

(ii) Since there exists a metric d generating the weak topology in $\mathcal{H}$ and that the closed ball $\mathcal{V}_r = \{\alpha : \|\alpha\| \le r\}$ is weakly compact, we see that (S1) implies that $\sigma(\alpha)$ is uniformly continuous with respect to the metric d and hence, with respect to the weak topology, over $\mathcal{V}_r$.

(iii) Assumption (S2) holds for the classical estimators based on the sample variance since the empirical covariance operator, $\widehat{\Gamma}$, is consistent in the unit ball. Indeed, as shown in Dauxois, Pousse and Romain (1982), $\|\widehat{\Gamma} - \Gamma_X\| \xrightarrow{\text{a.s.}} 0$, which entails that $\sup_{\|\alpha\|=1} |s_n^2(\alpha) - \sigma^2(\alpha)| \leq \|\widehat{\Gamma} - \Gamma_X\| \xrightarrow{\text{a.s.}} 0$. However, this assumption may seem harder to verify for other scale functionals since the unit sphere $\mathcal{S} = \{\|\alpha\| = 1\}$ is not compact, and $s_n^2(\alpha)$ is usually not defined through a covariance operator estimator. To be more precise, even under some conditions to be considered in Section 5, there exists a compact operator Γ such that $\sigma(\alpha) = \langle \alpha, \Gamma\alpha \rangle$, $s_n^2(\alpha)$ cannot be expressed as $\langle \alpha, \Gamma_n\alpha \rangle$ for some consistent estimator Γ_n of Γ. Corollary 6.1 in Section 6 establishes that (S2) holds for any scale functional σ_R that is continuous with respect to the weak topology.

The following lemma, whose proof can be found in Section B of the technical supplemental article given in Bali et al. (2011a), is useful for deriving the consistency and continuity of the principal direction estimators. In this lemma and in the subsequent theorems, it should be noted that $\langle \widehat{\phi}, \phi \rangle^2 \to 1$ implies, under the same mode of convergence, that the sign of $\widehat{\phi}$ can be chosen so that $\widehat{\phi} \to \phi$.

For the sake of simplicity, denote by $\lambda_{R,j} = \lambda_{R,j}(P)$ and $\phi_{R,j} = \phi_{R,j}(P)$.

LEMMA 4.1. *Let $\widehat{\phi}_m \in \mathcal{S}$ be such that $\langle \widehat{\phi}_m, \widehat{\phi}_j \rangle \xrightarrow{\text{a.s.}} 0$ for $j \neq m$ and assume that* (S0) *and* (S1) *hold. Then*:

(a) *If $\sigma^2(\widehat{\phi}_1) \xrightarrow{\text{a.s.}} \sigma^2(\phi_{R,1})$, then $\langle \widehat{\phi}_1, \phi_{R,1} \rangle^2 \xrightarrow{\text{a.s.}} 1$.*

(b) *Given $2 \leq m \leq q$, if $\sigma^2(\widehat{\phi}_m) \xrightarrow{\text{a.s.}} \sigma^2(\phi_{R,m})$ and $\widehat{\phi}_s \xrightarrow{\text{a.s.}} \phi_{R,s}$, for $1 \leq s \leq m-1$, then $\langle \widehat{\phi}_m, \phi_{R,m} \rangle^2 \xrightarrow{\text{a.s.}} 1$.*

Let $d_{PR}(P, Q)$ stand for the Prohorov distance between the probability measures P and Q. Thus, $P_n \xrightarrow{\omega} P$ if and only if $d_{PR}(P_n, P) \to 0$. Theorem 4.1 below establishes the continuity of the functionals defined in (3.1) and (3.2), 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. As it will be shown in Section 6, the uniform convergence required in assumption (ii) below is satisfied, for instance, if σ_R is a continuous scale functional when $P_n \xrightarrow{\omega} P$.

To accommodate data driven smoothing parameters a more general framework is considered in Theorem 4.1, which allows for the smoothing parameters τ_n and ρ_n to be random, such that $\tau_n \xrightarrow{\text{a.s.}} 0$ and $\rho_n \xrightarrow{\text{a.s.}} 0$.

THEOREM 4.1. *Let P_n be a sequence of probability measures and $\tau = \tau_n \geq 0$, $\rho = \rho_n \geq 0$ be random smoothing parameters. Denote by $\sigma_n^2(\alpha) = \sigma_R^2(P_n[\alpha])$ and*

define $\widehat{\lambda}_m = \sigma_n^2(\widehat{\phi}_m)$ *with*

(4.1)
$$\begin{cases} \widehat{\phi}_1 = \underset{\|\alpha\|_\tau = 1}{\arg\max} \{\sigma_n^2(\alpha) - \rho\Psi(\alpha)\}, \\ \widehat{\phi}_m = \underset{\alpha \in \widehat{\mathcal{B}}_{m,\tau}}{\arg\max} \{\sigma_n^2(\alpha) - \rho\Psi(\alpha)\}, \qquad 2 \le m, \end{cases}$$

where $\widehat{\mathcal{B}}_{m,\tau} = \{\alpha \in \mathcal{H} : \|\alpha\|_\tau = 1, \langle \alpha, \widehat{\phi}_j \rangle_\tau = 0, \forall 1 \le j \le m-1\}$. *Let* P *be a probability measure satisfying* (S0). *Assume that*:

  (i) (S1) *holds*;
  (ii) $\sup_{\|\alpha\|=1} |\sigma_n^2(\alpha) - \sigma_R^2(P[\alpha])| \xrightarrow{a.s.} 0$;
  (iii) $\tau_n \xrightarrow{a.s.} 0$ *and* $\rho_n \xrightarrow{a.s.} 0$;
  (iv) *moreover, if* $\tau_n > 0$ *or* $\rho_n > 0$, *for all* $n \ge n_0$, *assume that* $\phi_{R,j} \in \mathcal{H}_S$, *that is,* $\Psi(\phi_{R,j}) < \infty$, *for all* $1 \le j \le q$.

Then:

  (a) $\widehat{\lambda}_1 \xrightarrow{a.s.} \lambda_{R,1}$ *and* $\sigma^2(\widehat{\phi}_1) \xrightarrow{a.s.} \sigma^2(\phi_{R,1})$. *Moreover,* $\rho\Psi(\widehat{\phi}_1) \xrightarrow{a.s.} 0$ *and* $\tau\lceil\widehat{\phi}_1, \widehat{\phi}_1\rceil \xrightarrow{a.s.} 0$, *and so* $\|\widehat{\phi}_1\| \xrightarrow{a.s.} 1$;
  (b) $\langle\widehat{\phi}_1, \phi_{R,1}\rangle^2 \xrightarrow{a.s.} 1$;
  (c) *for any* $2 \le m \le q$, *if* $\widehat{\phi}_\ell \xrightarrow{a.s.} \phi_{R,\ell}$, $\tau\Psi(\widehat{\phi}_\ell) \xrightarrow{a.s.} 0$ *and* $\rho\Psi(\widehat{\phi}_\ell) \xrightarrow{a.s.} 0$ *for* $1 \le \ell \le m-1$, *then* $\widehat{\lambda}_m \xrightarrow{a.s.} \sigma^2(\phi_{R,m})$ *and* $\sigma^2(\widehat{\phi}_m) \xrightarrow{a.s.} \sigma^2(\phi_{R,m})$. *Moreover,* $\rho\Psi(\widehat{\phi}_m) \xrightarrow{a.s.} 0$, $\tau\Psi(\widehat{\phi}_m) \xrightarrow{a.s.} 0$ *and so,* $\|\widehat{\phi}_m\| \xrightarrow{a.s.} 1$;
  (d) *for* $1 \le m \le q$, $\langle\widehat{\phi}_m, \phi_{R,m}\rangle^2 \xrightarrow{a.s.} 1$.

Note that assumption (ii) corresponds to (S2) when P_n is the empirical probability measure. On the other hand, when $\sigma_R(\cdot)$ is a continuous scale functional, Theorem 6.2 implies that (ii) holds whenever $d_{PR}(P_n, P) \xrightarrow{a.s.} 0$. Moreover, if $\sigma_R(\cdot)$ is a continuous scale functional and P satisfies (S0), Theorem 4.1 entails the continuity of the functionals $\phi_{R,j}(\cdot)$ and $\lambda_{R,j}(\cdot)$ at P, for $1 \le j \le q$, and so the proposed estimators are qualitatively robust and consistent. In particular, the estimators are robust at any elliptical distribution $\mathcal{E}(\mu, \mathbf{\Gamma})$, as defined in Section 5, such that the largest $q+1$ eigenvalues of the operator $\mathbf{\Gamma}$ are all distinct.

Theorem 4.1 establishes the consistency of the raw estimators of the principal components under (S0) to (S2) by taking $\rho = \tau = 0$. It also shows that proposals (3.5) and (3.6) give consistent estimators if $\phi_{R,j} \in \mathcal{H}_S$, $1 \le j \le q$. In Bali et al. (2010), it is shown that the estimators $\widehat{\phi}_{PN,m}$ and $\widehat{\lambda}_{PN,m}$ defined in (3.5) and (3.7) are still consistent if $\phi_{R,j} \in \overline{\mathcal{H}_S}$, $1 \le j \le q$, where $\overline{\mathcal{H}_S}$ stands for the closure of $\mathcal{H}_S$. The condition $\phi_{R,j} \in \overline{\mathcal{H}_S}$ generalizes the assumption of smoothness, $\phi_{R,j} \in \mathcal{H}_S$, required in Silverman (1996) and holds, for example, when $\mathcal{H}_S$ is a dense subset of $\mathcal{H}$.

Theorem 4.2 establishes the consistency of the estimators of the principal directions defined through the sieve approach given in (3.8). Below we give a separate

statement for the consistency of the sieve estimators to avoid imposing additional burdensome assumptions, such as smoothness conditions for the basis elements, whenever either $\tau \neq 0$ or $\rho \neq 0$ in (3.10). Its proof is relegated to the Section C of the technical supplement [Bali et al. (2011a)].

THEOREM 4.2. *Let $\widehat{\phi}_{\mathrm{SI},m}$ and $\widehat{\lambda}_{\mathrm{SI},m}$ be the estimators defined in* (3.8) *and* (3.9), *respectively. Under* (S0) *to* (S2), *if $p_n \to \infty$, then:*

(a) $\widehat{\lambda}_{\mathrm{SI},1} \xrightarrow{a.s.} \sigma^2(\phi_{\mathrm{R},1})$ *and* $\sigma^2(\widehat{\phi}_{\mathrm{SI},1}) \xrightarrow{a.s.} \sigma^2(\phi_{\mathrm{R},1})$.

(b) *Given* $2 \leq m \leq q$, *if* $\widehat{\phi}_{\mathrm{SI},\ell} \xrightarrow{a.s.} \phi_{\mathrm{R},\ell}$, *for* $1 \leq \ell \leq m-1$, *then* $\widehat{\lambda}_{\mathrm{SI},m} \xrightarrow{a.s.} \sigma^2(\phi_{\mathrm{R},m})$ *and* $\sigma^2(\widehat{\phi}_{\mathrm{SI},m}) \xrightarrow{a.s.} \sigma^2(\phi_{\mathrm{R},m})$.

(c) *For* $1 \leq m \leq q$, $\langle \widehat{\phi}_{\mathrm{SI},m}, \phi_{\mathrm{R},m} \rangle^2 \xrightarrow{a.s.} 1$.

5. Fisher-consistency under elliptical distributions. The results in Section 4 ensure that, under mild conditions, the estimates of the principal directions converge almost surely to $\phi_{\mathrm{R},m}$ defined in (3.1). An important point to highlight is what the functions $\phi_{\mathrm{R},m}$ represent, at least in some particular situations. This section focuses on showing that, for the functional elliptical families defined in Bali and Boente (2009), the functionals $\phi_{\mathrm{R},m}(P)$ and $\lambda_{\mathrm{R},m}(P)$ have a simple interpretation. In particular, our results hold for the functional elliptical family, but are not restricted to it. We recall here their definition for the sake of completeness.

Let X be a random element in a separable Hilbert space $\mathcal{H}$ and $\mu \in \mathcal{H}$. Let $\Gamma : \mathcal{H} \to \mathcal{H}$ be a self-adjoint, positive semidefinite and compact operator. We will say that X has an elliptical distribution with parameters (μ, Γ), denoted as $X \sim \mathcal{E}(\mu, \Gamma)$, if for any linear and bounded operator $A : \mathcal{H} \to \mathbb{R}^d$, AX has a multivariate elliptical distribution with parameters $A\mu$ and $A\Gamma A^*$, that is, $AX \sim \mathcal{E}_d(A\mu, A\Gamma A^*)$, 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 $a \in \mathbb{R}$.

The elliptical distributions in $\mathcal{H}$ include the Gaussian distributions. Other elliptical distributions can be obtained from the following construction. Let V_1 be a Gaussian element in $\mathcal{H}$ with zero mean and covariance operator Γ_{V_1}, and let Z be a random variable independent of V_1. Given $\mu \in \mathcal{H}$, define $X = \mu + ZV_1$. Then, X has an elliptical distribution $\mathcal{E}(\mu, \Gamma)$ with the operator Γ being proportional to Γ_{V_1}. Note that Γ exist even if the second moment of X do not exist. For random elements which admit a finite Karhunen–Loève expansion, that is, $X(t) = \mu(t) + \sum_{j=1}^{q} \lambda_j^{1/2} U_j \phi_j(t)$, the assumption that X has an elliptical distribution is analogous to assuming that $\mathbf{U} = (U_1, \ldots, U_q)^{\mathrm{T}}$ has a spherical distribution.

Lemma 5.1 below states the Fisher-consistency of the functionals defined through (3.1) under the following assumption:

(S3) There exists a constant $c > 0$ and a self-adjoint, positive semidefinite and compact operator Γ_0, such that for any $\alpha \in \mathcal{H}$, we have $\sigma^2(\alpha) = c\langle \alpha, \Gamma_0 \alpha \rangle$.

Its proof follows immediately and is thus omitted. Note that (S3) entails that the function $\sigma : \mathcal{H} \to \mathbb{R}$ defined as $\sigma(\alpha) = \sigma_R(P[\alpha])$ is weakly continuous, hence (S1) holds. Besides, as a consequence of Lemma 5.1, (S0) holds under (S3) if the q largest eigenvalues of $\boldsymbol{\Gamma}_0$ are distinct.

LEMMA 5.1. *Let $\phi_{R,m}$ and $\lambda_{R,m}$ be the functionals defined in* (3.1) *and* (3.2), *respectively. Let $X \sim P$ be a random element such that* (S3) *holds. Denote by $\lambda_1 \geq \lambda_2 \geq \cdots$ the eigenvalues of $\boldsymbol{\Gamma}_0$ and by ϕ_j the eigenfunction of $\boldsymbol{\Gamma}_0$ associated to λ_j. Assume that for some $q \geq 2$, and for all $1 \leq j \leq q$, $\lambda_1 > \lambda_2 > \cdots > \lambda_q > \lambda_{q+1}$. Then, we have that $\phi_{R,j}(P) = \phi_j$ and $\lambda_{R,j}(P) = c\lambda_j$.*

For any distribution possessing finite second moments, if the scale functional is taken to be the standard deviation, then (S3) holds with $\boldsymbol{\Gamma}_0 = \boldsymbol{\Gamma}_X$. When considering a robust scale functional, (S3) holds if X has an elliptical distribution $\mathcal{E}(\mu, \boldsymbol{\Gamma})$ taking $\boldsymbol{\Gamma}_0 = \boldsymbol{\Gamma}$, and so Lemma 5.1 entails that the functionals $\phi_{R,j}(P)$ defined through (3.1) are Fisher-consistent. As in the finite-dimensional setting, the scale functional σ_R can be calibrated to attain Fisher-consistency of the principal values.

Assumption (S3) ensures that we are estimating the target directions. It may seem restrictive since it is difficult to verify outside the family of elliptical distributions except when the scale is taken to be the standard deviation. However, even in the finite-dimensional case, asymptotic properties have been derived only under similar restrictions when considering projection-pursuit estimators. For instance, both Li and Chen (1985) and Croux and Ruiz-Gazen (2005) assume an underlying elliptical distribution in order to obtain consistency results and influence functions, respectively. Also, in Cui, He and Ng (2003) the influence function of the projected data is assumed to be of the form $h(\mathbf{x}, \mathbf{a}) = 2\sigma(F[\mathbf{a}]) \mathrm{IF}(\mathbf{x}, \sigma_{\mathbf{a}}; F_0)$, where $F[\mathbf{a}]$ stands for the distribution of $\mathbf{a}^T\mathbf{x}$ when $\mathbf{x} \sim F$. This condition, though, primarily holds when the distribution is elliptical.

REMARK 5.1. An alternative to the robust projection pursuit approach for functional principal components is to consider the spectral value decomposition of a robust covariance or scatter operator. The spherical principal components, noted in the Introduction, which were proposed by Locantore et al. (1999) and further developed by Gervini (2008), apply this approach using the spatial covariance operator. The spatial covariance operator is defined to be

$$\mathbf{V} = \mathbb{E}\big((X - \eta) \otimes (X - \eta)/\|X - \eta\|^2\big)$$

with η being the spatial median, that is,

(5.1) $$\eta = \underset{\theta \in \mathcal{H}}{\arg\min}\, \mathbb{E}(\|X - \theta\| - \|X\|).$$

The spatial median is sometimes referred to as the multivariate L^1 median, but this is a misnomer since the the norm in (5.1) is the L^2 norm. Note that when the norm is replaced by the square of the norm in (5.1), the resulting parameter is the mean.

Gervini (2008) proved the Fisher-consistency of the eigenfunctions of the spatial covariance operator, but under the additional assumption that X has a finite Karhunen–Loève expansion so that $\mathbf{V}$ has only a finite number of nonzero eigenvalues, which is essentially the multivariate setting. Unlike the projection pursuit approach, though, under an elliptical model the eigenvalues of $\mathbf{V}$ are not proportional to the eigenvalues of the shape parameter $\mathbf{\Gamma}$. Consequently, as discussed, for example, in Marden (1999), Boente and Fraiman (1999) and Visuri, Koivunen and Oja (2000), this implies that even if the second moments exist, the amount of variance explained by a principal component variable is not equivalent to the ratio of the eigenvalue to the sum of all the eigenvalues. That is, $\widetilde{\lambda}_k / \sum_{j=1}^{\infty} \widetilde{\lambda}_j$ is not the same as the explained proportion $\lambda_k / \sum_{j=1}^{\infty} \lambda_j$, where $\widetilde{\lambda}_k$ and λ_k are the kth largest eigenvalue of $\mathbf{V}$ and $\mathbf{\Gamma}$ respectively.

In the multivariate setting, it is also known that the eigenvectors obtained from the sample spatial covariance matrix can be extremely inefficient estimates whenever the eigenvalues differ greatly; see Croux (1999). Intuitively, the reason for this inefficiency is that the spatial covariance matrix down-weights observations according to their Euclidean distance from the center. This seems reasonable when the distribution is close to being spherical, but less so when there are strong dependencies in the variables. In some sense, this is the antithesis of PCA, since one is usually interested in principal components when one suspects the latter.

As noted in Maronna, Martin and Yohai (2006), there is a vast literature on robust estimates for covariance matrices, such as M-estimates, S-estimates and the *MCD*, among others. These estimates downweight observations relative to the shape of the data cloud. It may seem reasonable then to try to extend these estimates to the functional setting. An important feature of these estimates, though, is that they are affine equivariant, and as shown in Tyler (2010), this implies that, when the sample size is no greater than the dimension plus one, such estimates are simply proportional to the sample covariance matrix. In the functional data setting, the sample size is always less than the dimension, which is infinite. Thus, at this time, we view the robust projection-pursuit approach as more viable.

6. Some uniform convergence results. In this section, we show that when the scale functional is continuous with respect to the Prohorov distance, (S2) and more generally, condition (ii) in Theorem 4.1 hold whenever $P_n \xrightarrow{\omega} P$. To derive these results, we will first state some properties regarding the weak convergence of empirical measures that hold not only in $L^2(\mathcal{I})$ but in any complete and separable metric space. These properties may be useful in other settings. The proofs for the theorems in this section are relegated to Section D of the supplemental article [Bali et al. (2011a)].

Let $\mathcal{M}$ be a complete and separable metric space (Polish space) and $\mathcal{B}$ the Borel σ-algebra of $\mathcal{M}$. Lemma 6.1, which is a restatement of Theorem 3 in Varadarajan (1958), ensures that the empirical measures converge weakly almost surely on a Polish space to the probability measure generating the observations.

LEMMA 6.1. *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and $X_n : \Omega \to \mathcal{M}$, $n \in \mathbb{N}$, be a sequence of independent and identically distributed random elements such that $X_i \sim P$. Assume that $\mathcal{M}$ is a Polish space, and denote by P_n the the empirical probability measure, that is, $P_n(A) = (1/n) \sum_{i=1}^{n} I_A(X_i)$ with $I_A(X_i) = 1$ if $X_i \in A$ and 0 elsewhere. Then, $P_n \xrightarrow{\omega} P$ almost surely, that is, $d_{\mathrm{PR}}(P_n, P) \xrightarrow{a.s.} 0$.*

Let P be a probability measure in $\mathcal{M}$, a separable Banach space, and let $\mathcal{M}^*$ denote the dual space. For a given $f \in \mathcal{M}^*$, define $P[f]$ as the real measure of the random variable $f(X)$, with $X \sim P$.

THEOREM 6.1. *Let $\{P_n\}_{n \in \mathbb{N}}$ and P be probability measures defined on $\mathcal{M}$ such that $d_{\mathrm{PR}}(P_n, P) \to 0$. Then, $\sup_{\|f\|_*=1} d_{\mathrm{PR}}(P_n[f], P[f]) \to 0$.*

When the Banach space $\mathcal{M}$ above is a separable Hilbert space $\mathcal{H}$, the Riesz representation theorem implies that for $f \in \mathcal{H}^*$ with $\|f\|_* = 1$, there exists $\alpha \in \mathcal{H}$ such that $f(X) = \langle \alpha, X \rangle$. The following result states that when σ_{R} is a continuous scale functional, uniform convergence can be attained and so, assumption (ii) in Theorem 4.1 is satisfied.

THEOREM 6.2. *Let $\{P_n\}_{n \in \mathbb{N}}$ and P be probability measures defined on a separable Hilbert space $\mathcal{H}$, such that $d_{\mathrm{PR}}(P_n, P) \to 0$. Let σ_{R} be a continuous scale functional. Then, $\sup_{\|\alpha\|=1} |\sigma_{\mathrm{R}}(P_n[\alpha]) - \sigma_{\mathrm{R}}(P[\alpha])| \longrightarrow 0$.*

Using Lemma 6.1 and Theorem 6.2, we get the following result that shows that (S2) holds if σ_{R} is a continuous scale functional.

COROLLARY 6.1. *Let P be a probability measure in a separable Hilbert space $\mathcal{H}$, P_n be the empirical measure of a random sample $X_1, \ldots, X_n$ with $X_i \sim P$, and σ_{R} be a continuous scale functional. Then we have that $\sup_{\|\alpha\|=1} |\sigma_{\mathrm{R}}(P_n[\alpha]) - \sigma_{\mathrm{R}}(P[\alpha])| \xrightarrow{a.s.} 0$.*

7. Selection of the smoothing parameters. The selection of the smoothing parameters is an important practical issue. The most popular general approach to address such a selection problem is to use the cross-validation methods. In nonparametric regression, the sensitivity of L^2 cross-validation methods to outliers has been pointed out by Wang and Scott (1994) and by Cantoni and Ronchetti (2001), among others. The latter also proposed more robust alternatives to L^2 cross-validation. The idea of robust cross-validation can be adapted to the present situation. Assume for the moment that we are interested in a fixed number, ℓ, of components. We propose to proceed as follows:

(CV1) Center the data. That is, define $\widetilde{X}_i = X_i - \widehat{\mu}$ where $\widehat{\mu}$ is a robust location estimator, such as the functional spatial median defined in Gervini (2008).

(CV2) For the penalized roughness approaches and for each m in the range $1 \leq m \leq \ell$ and $\tau > 0$, let $\widehat{\phi}_{m,\tau}^{(-j)}$ denote the robust estimator of the mth principal direction computed without the jth observation.

(CV3) Define $X_j^{\perp}(\tau) = \widetilde{X}_j - \pi(\widetilde{X}_j; \widehat{\mathcal{L}}_\ell^{(-j)})$, for $1 \leq j \leq n$, where $\pi(X; \mathcal{L})$ stands for the orthogonal projection of X onto the linear (closed) space $\mathcal{L}$, and $\widehat{\mathcal{L}}_\ell^{(-j)}$ stands for the linear space spanned by $\widehat{\phi}_{1,\tau}^{(-j)}, \ldots, \widehat{\phi}_{\ell,\tau}^{(-j)}$.

(CV4) Let $\mathrm{RCV}_\ell(\tau) = \sigma_n^2(\|X_1^{\perp}(\tau)\|, \ldots, \|X_n^{\perp}(\tau)\|)$, where σ_n is a robust measure of scale about zero. We then choose τ_n to be the value of τ which minimizes $\mathrm{RCV}_\ell(\tau)$.

By a robust measure of scale about zero, we mean that no location estimator is applied to center the data. For instance, in the classical setting, one takes $\sigma_n^2(z_1, \ldots, z_n) = (1/n) \sum_{i=1}^n z_i^2$, while in the robust situation, one might choose $\sigma_n(z_1, \ldots, z_n) = \mathrm{median}(|z_1|, \ldots, |z_n|)$ or to be an M-estimator satisfying equation (2.3) when setting $\widehat{\mu}_n = 0$.

For large sample sizes, it is well understood that cross-validation methods can be computationally prohibitive. In such cases, K-fold cross-validation provides a useful alternative. In the following, we briefly describe a robust K-fold cross-validation procedure suitable for our proposed estimates.

(K1) First center the data as above, using $\widetilde{X}_i = X_i - \widehat{\mu}$.

(K2) Partition the centered data set $\{\widetilde{X}_i\}$ randomly into K disjoint subsets of approximately equal sizes with the jth subset having size $n_j \geq 2$, $\sum_{j=1}^K n_j = n$. Let $\{\widetilde{X}_i^{(j)}\}_{1 \leq i \leq n_j}$ be the elements of the jth subset, and $\{\widetilde{X}_i^{(-j)}\}_{1 \leq i \leq n-n_j}$ denote the elements in the complement of the jth subset. The set $\{\widetilde{X}_i^{(-j)}\}_{1 \leq i \leq n-n_j}$ will be the training set and $\{\widetilde{X}_i^{(j)}\}_{1 \leq i \leq n_j}$ the validation set.

(K3) Similar to step (CV2) but with leaving the jth validation subset $\{\widetilde{X}_i^{(j)}\}_{1 \leq i \leq n_j}$ out instead of the jth observation.

(K4) Define $X_j^{(j)\perp}(\tau)$ the same way as in step (CV3), using the validation set. For instance, $X_i^{(j)\perp}(\tau) = \widetilde{X}_i^{(j)} - \pi(\widetilde{X}_i^{(j)}; \widehat{\mathcal{L}}_\ell^{(-j)})$, $1 \leq i \leq n_j$, where $\widehat{\mathcal{L}}_\ell^{(-j)}$ stands for the linear space spanned by $\widehat{\phi}_{1,\tau}^{(-j)}, \ldots, \widehat{\phi}_{\ell,\tau}^{(-j)}$.

(K5) Let $\mathrm{RCV}_{\ell,\mathrm{KCV}}(\tau) = \sum_{j=1}^K \sigma_n^2(\|X_1^{(j)\perp}(\tau)\|, \ldots, \|X_{n_j}^{(j)\perp}(\tau)\|)$, and choose τ_n to be the value of τ which minimizes $\mathrm{RCV}_{\ell,\mathrm{KCV}}(\tau)$.

A similar approach can be developed to choose p_n for the sieve estimators.

8. Monte Carlo study. The results of Section 4 established under general conditions the consistency of the various robust projection pursuit approaches to functional principal components analysis. The classical approach to functional principal components analysis also yields consistent estimates, provided second moment exists. A study of the influence functions and the asymptotic distributions

of the various procedures would be useful to compare them. We leave these important and challenging theoretical problems for future research. For now, to help illuminate possible differences in the various approaches, we present in this section the results of a Monte Carlo study.

8.1. *Algorithms.* All the computational methods to be considered here are modifications of the basic CR algorithm proposed by Croux and Ruiz-Gazen (1996) for the computation of principal components using projection-pursuit. The basic algorithm applies to multivariate data, say m-dimensional, and requires a search over projections in $\mathbb{R}^m$. The GRID algorithm described in Croux, Filzmoser and Oliveira (2007) can also be considered, in particular, when the number of variables m is larger than the sample size n. For the sake of completeness, we briefly recall the CR algorithm.

Let $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_n)$ be the sample in $\mathbb{R}^m$, and let $\widehat{\boldsymbol{\mu}}_n(\mathbf{Y})$ be a location estimate computed from this sample. Let $1 \leq q \leq m$ be the desired number of components to be computed and denote by ξ_n the univariate projection index to be maximized. In the multivariate setting, the index ξ_n corresponds to a robust univariate scale statistic.

(CR1) For $k = 1$, set $\mathbf{y}_i^{(1)} = \mathbf{y}_i - \widehat{\boldsymbol{\mu}}_n(\mathbf{Y})$. Let the set of candidate directions for the first principal direction be $\mathcal{A}_{n,1}(\mathbf{Y}) = \{\mathbf{y}_i^{(1)}/\nu_i, 1 \leq i \leq n\}$ where $\nu_i^2 = \mathbf{y}_i^{(1)\mathrm{T}}\mathbf{y}_i^{(1)}$. We then define $\mathbf{v}_1 = \arg\max_{\mathbf{a} \in \mathcal{A}_{n,1}(\mathbf{Y})} \xi_n(\mathbf{a}^{\mathrm{T}}\mathbf{y}_1, \dots, \mathbf{a}^{\mathrm{T}}\mathbf{y}_n)$.

(CR2) For $2 \leq k \leq q$, define recursively $z_i^{(k-1)} = \mathbf{v}_{k-1}^{\mathrm{T}}\mathbf{y}_i$ and $\mathbf{y}_i^{(k)} = \mathbf{y}_i^{(k-1)} - z_i^{(k-1)}\mathbf{v}_{k-1} = \mathbf{y}_i^{(1)} - \pi_{\mathcal{V}_{k-1}}(\mathbf{y}_i^{(1)})$, where $\pi_{\mathcal{V}_{k-1}}(\mathbf{y})$ stands for the orthogonal projection of $\mathbf{y}$ over the linear space $\mathcal{V}_{k-1}$ spanned by $\mathbf{v}_1, \dots, \mathbf{v}_{k-1}$. Let the set of candidate directions for the kth principal direction be $\mathcal{A}_{n,k}(\mathbf{Y}) = \{\mathbf{y}_i^{(k)}/\nu_i, 1 \leq i \leq n\}$ where $\nu_i^2 = \mathbf{y}_i^{(1k)\mathrm{T}}\mathbf{y}_i^{(k)}$, and define $\mathbf{v}_k = \arg\max_{\mathbf{a} \in \mathcal{A}_{n,k}(\mathbf{Y})} \xi_n(\mathbf{a}^{\mathrm{T}}\mathbf{y}_1, \dots, \mathbf{a}^{\mathrm{T}}\mathbf{y}_n)$.

The vectors $\mathbf{v}_k$ then yield approximations to the kth principal direction, for $1 \leq k \leq q$, and approximate scores for the kth principal variable are given by $z_i^{(k)} = \mathbf{v}_k^{\mathrm{T}}\mathbf{y}_i$, for $1 \leq i \leq n$. As mentioned in Croux and Ruiz-Gazen (1996), the CR algorithm makes no smoothness assumptions on the index ξ_n, is simple and fast, and requires only $O(n)$ computing space.

To apply the algorithm to functional data when considering either the raw or a penalized approach, we first discretize the domain of the observed function over m equally spaced points in $\mathcal{I} = [-1, 1]$. The resulting multivariate observations are then $\mathbf{y}_i = (X_i(t_1), \dots, X_i(t_m))^{\mathrm{T}}$, where $t_0 = -1 < t_1 < \cdots < t_m < t_{m+1} = 1$. The index ξ_n in the algorithm depends on the approach being used. For instance, for the raw robust estimate and for those penalizing the norm the index is a robust scale. On the other hand, for the robust penalized scale approach the index is the square of the robust scale plus the penalization term. Also, for the penalized norm approach the orthogonal projection $\pi_{\mathcal{V}_{k-1}}(\mathbf{y})$ in step (CR2) is with respect to

the inner product $\langle \cdot, \cdot \rangle_\tau$ so, the finite-dimensional inner product is modified as in Silverman (1996). The resulting directions $\mathbf{v}_k$ then give numerical approximations for $\{\widehat{\phi}_k(t_1), \ldots, \widehat{\phi}_k(t_m)\}$. One can then interpolate or use smoothing methods to obtain $\widehat{\phi}_k(t)$ for $t \in \mathcal{I}$.

For the sieve approach, let $\widetilde{\delta}_1, \ldots, \widetilde{\delta}_{p_n}$ be an orthonormal basis for $\mathcal{H}_{p_n}$, which can be obtained by using Gramm–Schmidt on the original basis. For $\alpha \in \mathcal{H}_{p_n}$, we then have $\langle \alpha, X_i \rangle = \mathbf{a}^{\mathrm{T}} \mathbf{y}_i$, where $\alpha = \sum_{j=1}^{p_n} a_j \widetilde{\delta}_j$, $\mathbf{a} = (a_1, \ldots, a_{p_n})^{\mathrm{T}}$, $\mathbf{y}_i = (\langle X_i, \widetilde{\delta}_1 \rangle, \ldots, \langle X_i, \widetilde{\delta}_{p_n} \rangle)^{\mathrm{T}}$. Consequently, we can take $m = p_n$ and apply the CR algorithm to the inner scores $\mathbf{y}_i$. The inner scores are computed numerically by approximating the integrals over a grid of 50 points. A numerical approximation for $\widehat{\phi}_k(t)$ is then given by $\sum_{j=1}^{p_n} v_{k,j} \widetilde{\delta}_j(t)$ with $\mathbf{v}_k = (v_{k,1}, \ldots, v_{k,p_n})^{\mathrm{T}}$.

8.2. *The estimators.* There are three main characteristics that distinguish the different estimators: the method of centering in the first step of the CR algorithm, the scale function being used and the type of smoothing method.

Centering: For classical procedures, that is, those based on the standard deviation, we use the sample mean as the centering point for the trajectories. For the robust procedures, that is, those based on MAD or M-SCALE, we center the data by using either (i) the component-wise sample median, that is, the median at each time point, or (ii) the sample spatial median; see (5.1). It turns out that the two robust centering methods produced similar results, so only the results for the spatial median are reported.

Scale function: Three scale estimators are considered here: the classical standard deviation (SD), the median absolute deviation (MAD) and an M-estimator of scale (M-SCALE). For the M-estimator, we use the score function (2.4) introduced by Beaton and Tukey (1974), as discussed in Section 2.2, with $c = 1.56$, $\delta = 1/2$.

Smoothing parameters τ and ρ: For both the classical and robust procedures defined in Section 3.2, a penalization depending on the L^2 norm of the second derivative, multiplied by a smoothing factor, is included, that is, $\Psi(\alpha) = \int_{-1}^{1} (\alpha''(t))^2 \, dt$. Again the integral is computed over the same grid of points $t_1, \ldots, t_m$, and the second derivative of α at t_i is approximated by $\{\alpha(t_{i+1}) - 2\alpha(t_i) + \alpha(t_{i-1})\}/(t_{i+1} - t_i)^2$, since we choose an equidistant grid of points. Note that when $\rho = \tau = 0$, the raw estimators defined in Section 3.1 are obtained.

Sieve: Two different sieve basis are considered: the Fourier basis obtained by taking δ_j to be the Fourier basis functions, and the cubic B-spline basis functions. The Fourier basis used in the sieve method is the same basis used to generate the data.

In all tables, the estimators corresponding to each scale choice are labeled as SD, MAD, M-SCALE. For each scale, we consider four estimators, the raw estimators

where no smoothing is used, the estimators obtained by penalizing the scale function defined in (3.6), those obtained by penalizing the norm defined in (3.5) and the sieve estimators defined in (3.8). In all tables, as in Section 3, the jth principal direction estimators related to each method are labeled as $\widehat{\phi}_{\mathrm{RAW},j}$, $\widehat{\phi}_{\mathrm{PS},j}$, $\widehat{\phi}_{\mathrm{PN},j}$ and $\widehat{\phi}_{\mathrm{SI},j}$, respectively.

When using the penalized estimators, several values for the penalizing parameters τ and ρ were chosen. Since large values of the smoothing parameters make the penalizing term the dominant component regardless of the amount of contamination considered, we choose τ and ρ equal to $an^{-\alpha}$ for $\alpha = 3$ and 4 and a equal to 0.05, 0.10, 0.25, 0.5, 0.75, 1, 1.5 and 2.

For the sieve estimators based on the Fourier basis, ordered as $\{1, \cos(\pi x), \sin(\pi x), \ldots, \cos(q_n \pi x), \sin(q_n \pi x), \ldots\}$, the values $p_n = 2q_n$ with $q_n = 5$, 10 and 15 are used, while for the sieve estimators based on the B-splines, the dimension of the linear space considered is selected as $p_n = 10$, 20 and 50. The basis for the B-splines is generated from the R function *cSplineDes*, with the knots being equally spaced in the interval $[-1, 1]$ and the number of knots equal to $p_n + 1$. To conserve space, we only report here the results corresponding to $p_n = 30$ and $p_n = 50$ for the Fourier and B-spline basis, respectively. More extensive simulation results are listed in the technical report [Bali et al. (2010)].

8.3. *Simulation settings.* The sample was generated using a finite Karhunen–Loève expansion with the functions, $\phi_i : [-1, 1] \to \mathbb{R}$, $i = 1, 2, 3$, where $\phi_1(x) = \sin(4\pi x)$, $\phi_2(x) = \cos(7\pi x)$ and $\phi_3(x) = \cos(15\pi x)$. It is worth noticing that, when considering the sieve estimators based on the Fourier basis, the third component cannot be detected when $q_n < 15$, since in this case $\phi_3(x)$ is orthogonal to the estimating space. Likewise, the second component cannot be detected when $q_n < 7$.

We performed $NR = 1{,}000$ replications generating independent samples $\{X_i\}_{i=1}^n$ of size $n = 100$ following the model $X_i = Z_{i1}\phi_1 + Z_{i2}\phi_2 + Z_{i3}\phi_3$, where $\mathbf{Z}_i = (Z_{i1}, Z_{i2}, Z_{i3})^{\mathrm{T}}$ are independent vectors whose distribution will depend on the situation to be considered. The central model, denoted C_0, corresponds to Gaussian samples. We also consider four contaminations of the central model, labeled C_2, $C_{3,a}$, $C_{3,b}$ and C_{23} depending on the components to be contaminated. Contamination models are commonly considered in robust statistics since they tend be the more difficult models to be robust against and are the basis for the concept of bias robustness, see Maronna, Martin and Yohai (2006) for further discussion. In all these situations Z_{i1}, Z_{i2} and Z_{i3} are also independent. For each of the models, we took $\sigma_1 = 4$, $\sigma_2 = 2$ and $\sigma_3 = 1$. The central model and the contaminations can be described as follows:

C_0: $Z_{i1} \sim N(0, \sigma_1^2)$, $Z_{i2} \sim N(0, \sigma_2^2)$ and $Z_{i3} \sim N(0, \sigma_3^2)$.

C_2: Z_{i2} are independent and identically distributed as $0.8N(0, \sigma_2^2) + 0.2N(10, 0.01)$, while $Z_{i1} \sim N(0, \sigma_1^2)$ and $Z_{i3} \sim N(0, \sigma_3^2)$. This contamination corresponds

to a strong contamination on the second component and changes the mean value of the generated data Z_{i2} and also the first principal component. Note that $\text{var}(Z_{i2}) = 19.202$.

$C_{3,a}$: $Z_{i1} \sim N(0, \sigma_1^2)$, $Z_{i2} \sim N(0, \sigma_2^2)$ and $Z_{i3} \sim 0.8N(0, \sigma_3^2) + 0.2N(15, 0.01)$. This contamination corresponds to a strong contamination on the third component. Note that $\text{var}(Z_{i3}) = 36.802$.

$C_{3,b}$: $Z_{i1} \sim N(0, \sigma_1^2)$, $Z_{i2} \sim N(0, \sigma_2^2)$ and $Z_{i3} \sim 0.8N(0, \sigma_3^2) + 0.2N(6, 0.01)$. This contamination corresponds to a strong contamination on the third component. Note that $\text{var}(Z_{i3}) = 6.562$.

C_{23}: Z_{ij} are independent and such that $Z_{i1} \sim N(0, \sigma_1^2)$, $Z_{i2} \sim 0.9N(0, \sigma_2^2) + 0.1N(15, 0.01)$ and $Z_{i3} \sim 0.9N(0, \sigma_3^2) + 0.1N(20, 0.01)$. This contamination corresponds to a mild contamination on the last two components. Note that $\text{var}(Z_{i2}) = 23.851$ and $\text{var}(Z_{i3}) = 36.901$.

We also include a long-tailed model, namely a Cauchy model, labeled C_{Cauchy}, which is defined by taking $\mathbf{Z}_i \sim \mathcal{C}_3(0, \mathbf{\Sigma})$ with $\mathbf{\Sigma} = \text{diag}(\sigma_1^2, \sigma_2^2, \sigma_3^2)$, where $\mathcal{C}_p(0, \mathbf{\Sigma})$ stands for the p-dimensional elliptical Cauchy distribution centered at 0 with scatter matrix $\mathbf{\Sigma}$. For this situation, the covariance operator does not exist, and thus the classical principal components are not defined.

It is worth noting that the directions ϕ_1, ϕ_2 and ϕ_3 correspond to the classical principal components for the case C_0, but not necessarily for the other cases. For instance, when $\sigma_R^2 = \text{VAR}$, $C_{3,a}$ interchanges the order between ϕ_1 and ϕ_3, that is, $\phi_3 = \phi_{R,1}(C_{3,a})$ as defined in (3.1), and so it corresponds to the first principal component of the covariance operator, while ϕ_1 is the second and ϕ_2 is the third one.

8.4. *Simulation results.* For each situation, we compute the estimators of the first three principal directions and the square distance between the true and the estimated direction (normalized to have L^2 norm 1), that is,

$$D_j = \left\| \frac{\widehat{\phi}_j}{\|\widehat{\phi}_j\|} - \phi_j \right\|^2.$$

Note that all the estimators except those penalizing the norm, are such that $\|\widehat{\phi}_j\| = 1$. Tables 4 to 9 of the supplementary material [Bali et al. (2011b)] report the means of D_j over replications, which hereafter is referred to as mean square error. To help understand the influence of the grid size m on the estimators, Tables 3, 4 and 5 give the mean squared errors for $m = 50, 100, 150, 250$ and 250, under C_0 for various values of the penalizing parameters. As can be seen, for the first two components some slight improvement is observed when using $m = 250$ as opposed to $m = 50$ points, but at the expense of increasing the computing time about 2.6 fold. On the other hand, for the third principal direction, taking $m = 100$ compared to $m = 50$ reduces the mean square error by at least a half for the penalized estimators, while the gain is not so prominent for the raw estimates. The

size $m = 50$ is selected for presentation in the remainder of our study since it provides a reasonable a compromise between the performance of the estimators and the computational time.

To help understand the effect of penalization, consider Table 6. This table shows results for the raw and penalized estimators under C_0 for different choices of the penalizing parameters. From this table, we see that a better performance is achieved in most cases when $\alpha = 3$ is used. To be more precise, the results in Table 6 show that the best choice for $\widehat{\phi}_{\mathrm{PS},j}$ is $\rho = 2n^{-3}$ for all j. Note that $\rho = 1.5n^{-3}$ gives fairly similar results when using the M-scale, reducing the mean squared error relative to the raw estimate by about a half and a third for $j = 2$ and 3, respectively.

For the norm penalized estimators, $\widehat{\phi}_{\mathrm{PN},j}$, the best choice for the penalizing parameter seems to depend upon both the component to be estimated and the scale estimator used. For instance, when using the standard deviation, the best choice is $\tau = 0.10n^{-3}$, for $j = 1$ and 2 while for $j = 3$ a smaller order is needed to obtain an improvement over the raw estimators, with the value $\tau = 0.75n^{-4}$ leading to a small gain over the raw estimators. For the robust procedures, larger values are needed to see an advantage to using the penalized norm approach relative to the raw estimators. For example, for $j = 1$, the largest reduction is observed when $\tau = 2n^{-3}$ while for $j = 2$, the best choices correspond to $\tau = 0.5n^{-3}$ and $\tau = 0.25n^{-3}$ when using the MAD and M-scale, respectively. When using the M-scale, a good compromise is to choose $\tau = 0.75n^{-3}$, which gives a reduction of around 30% and 50% for the first and second principal directions, respectively, although smaller values of τ are again better for estimating the third component.

Based upon the above observations, we report here only the results corresponding to $\rho = 1.5n^{-3}$ and $\tau = 0.75n^{-3}$ for the penalized estimators $\widehat{\phi}_{\mathrm{PN},j}$ and $\widehat{\phi}_{\mathrm{PS},j}$, respectively, under the contamination models and the Cauchy model. Results for other choices of ρ and τ are given in Bali et al. (2010).

The simulation study confirms the expected inadequate behavior of the classical estimators in the presence of outliers. Under contamination, the classical estimators of the principal directions do not estimate the target directions very accurately. This is also the case when considering the Cauchy distribution. Curiously, though, the principal directions, under the Cauchy model, do not seem to be totally arbitrary and they partially recover ϕ_1, ϕ_2 and ϕ_3 when the standard deviation is used, although not as well as when using a robust scale, even though the covariance operator does not exist, nor do the population principal directions as defined in (2.1).

The robust estimators of the first principal directions are not heavily affected by any of the contaminations, while the estimates of the second and third principal directions appear to be most affected under model $C_{3,a}$. In particular, for the third direction, the projection-pursuit estimators based on an the MAD seems to be most affected by this type of contamination when penalizing the norm, although much less so than the classical methods. With respect to the contamination

model $C_{3,a}$, the estimators $\widehat{\phi}_{PN,j}$, which are the robust penalized norm estimators, tend to have the best performance among all the robust competitors for the first two components, and, in particular, when using the M-scale; see Table 8. It is worth noting that the classical estimators of the first component are not affected by this contamination when penalizing the norm since the penalization dominates the contaminated variances. The same phenomena is observed under $C_{3,b}$ when using the classical estimators for the selected amount of penalization. For the raw estimators, the sensitivity of the classical estimators under this contamination can be observed in Table 7. We refer to Bali et al. (2010) for the behavior when other values of the smoothing parameters are chosen.

As noted in Silverman (1996), for the classical estimators, some degree of smoothing in the procedure based on penalizing the norm will give a better estimation of ϕ_j in the L^2 sense under mild conditions. In particular, both the procedure penalizing the norm and the scale provide some improvement with respect to the raw estimators if $\Psi(\phi_j) < \Psi(\phi_\ell)$, when $j < \ell$. This means that the principal directions are rougher as the eigenvalues decrease [see Pezzulli and Silverman (1993) and Silverman (1996)], which is also reflected in our simulation study. The advantages of the smooth projection pursuit procedures are most striking when estimating ϕ_2 and ϕ_3 with an M-scale and using the penalized scale approach.

As expected, when using the sieve estimators, the Fourier basis gives the best performance over all the methods under C_0, since our data were generated from this basis; see Table 9. The choice of the B-spline basis give similar results quite to those obtained with $\widehat{\phi}_{PS,j}$ when estimating the first direction, except under C_{Cauchy} where the penalized estimators show a better performance. For the second and third components, the estimators obtained with the B-spline basis show larger the mean square errors than the raw or penalized estimators.

8.5. *Kth fold simulation.* Table 1 reports the computing times in minutes for 1,000 replications and for a fixed value of τ or ρ, run on a computer Core Quad I7 930 (2.80 GHz) with 8 Gb of Ram memory. We also report the computing times when using the sieve approach with the Fourier basis and a fixed value of p_n. This

TABLE 1
Computing times in minutes for 1,000 *replications and a fixed value
of* τ, ρ *or* p_n (*when using the Fourier basis*)

	SD	MAD	*M*-SCALE
$\widehat{\phi}_{\text{RAW}}$	5.62	6.98	17.56
$\widehat{\phi}_{\text{PS}}$	7.75	9.00	20.18
$\widehat{\phi}_{\text{PN}}$	31.87	33.21	44.04
$\widehat{\phi}_{\text{SI}}$	0.5	5.22	16.08

suggests that the leave-one-out cross-validation may be difficult to perform, and so a K-fold approach is adopted instead. It is worth noticing that the robust procedures based on the MAD are much faster than those based on the M-scale, so they may be preferred in terms of computing time. However, as mentioned in Section 2, the main disadvantage of the MAD is its low efficiency and lack of smoothness, which is related to the discontinuity of its influence function. This is particularly important when estimating principal components in the finite-dimensional setting, since as it was pointed out by Cui, He and Ng (2003) and Croux and Ruiz-Gazen (2005) the variances of some elements of the estimated principal directions may blow up when using the MAD leading to highly inefficient estimators. As expected and mentioned in Section 8.4, Table 6 reveals a high loss of efficiency for the MAD, in our functional setting, for any choice of the smoothing parameter.

For the procedure which penalizes the scale or the norm, the smoothing parameters ρ and τ are selected using the procedure described in Section 7 with $K = 4$ and $\ell = 1$. Due to the extensive computing time, we have only performed 500 replications. The simulation results when penalizing the scale function, that is, for the estimators defined through (3.6), are reported in Table 2. Under C_0, when estimating the second and third principal directions, the robust estimators based on the M-scale combined with a penalization in the scale clearly have smaller mean square error than the raw estimators, while those penalizing the norm improve the performance of the raw estimators and also that of $\widehat{\phi}_{\mathrm{PS},j}$, on the first and second directions.

From the results in Table 2 we observe that the classical estimators are sensitive to the contaminations in the simulation settings, and, except for contaminations in the third component, the robust counterpart shows a clear advantage. Note that $C_{3,b}$ affects more the classical estimators when the smoothing parameter is selected by the robust K-fold cross-validation method than for the fixed values studied in the previous section. This can be explained by the fact that contamination $C_{3,b}$ is a mild contamination in the third component which has a large $\|\phi_3''\|^2$, and so the classical estimators are more sensitive to it, just as the raw estimators, if smaller values of the smoothing parameter are chosen. It is worth noticing that the penalized robust estimators based on the M-scale improve the performance of the raw estimators based on the M-scale, even under contaminations, when the penalizing parameter is selected using the K-fold approach. This advantage is more striking when penalizing the norm and when the two first principal components are considered.

Note that we choose $\ell = 1$, and so our focus was on the first principal component. To improve the observed performance, a different approach should be considered, maybe by selecting a different smoothing parameter for each principal direction.

TABLE 2

Mean values of $\|\widehat{\phi}_j/\|\widehat{\phi}_j\| - \phi_j\|^2$ when the penalizing parameter is selected using K-fold cross-validation

Model	Scale estimator	$\widehat{\phi}_{PS,j}$			$\widehat{\phi}_{PN,j}$		
		$j = 1$	$j = 2$	$j = 3$	$j = 1$	$j = 2$	$j = 3$
C_0	SD	0.0073	0.0094	0.0078	0.0075	0.0094	0.0360
	MAD	0.0662	0.0993	0.0634	0.0497	0.0660	0.2573
	M-scale	0.0225	0.0311	0.0172	0.0208	0.0271	0.0839
C_2	SD	1.2840	1.2837	0.0043	1.2076	1.2232	0.0301
	MAD	0.3731	0.3915	0.0504	0.3360	0.3770	0.2832
	M-scale	0.4261	0.4286	0.0153	0.3679	0.4049	0.1607
$C_{3,a}$	SD	1.7840	1.8901	1.9122	1.7795	1.8861	1.9134
	MAD	0.2271	0.5227	0.5450	0.0573	0.2289	0.9540
	M-scale	0.2176	0.4873	0.5437	0.0257	0.1187	0.8710
$C_{3,b}$	SD	0.0192	0.8350	0.8525	0.0173	0.5902	0.7502
	MAD	0.0986	0.3930	0.3820	0.0553	0.1417	0.5167
	M-scale	0.0404	0.2251	0.2285	0.0241	0.1080	0.3174
C_{23}	SD	1.7645	0.5438	1.6380	1.7537	0.6496	1.4305
	MAD	0.2407	0.3443	0.2064	0.1414	0.2214	0.6824
	M-scale	0.2613	0.3707	0.2174	0.1313	0.1870	0.5901
C_{Cauchy}	SD	0.3580	0.4835	0.2287	0.2862	0.3525	0.3435
	MAD	0.0788	0.1511	0.1082	0.0613	0.0855	0.3147
	M-scale	0.0444	0.0707	0.0434	0.0349	0.0463	0.1465

9. Concluding remarks. In this paper, we propose robust principal component analysis for functional data based on a projection-pursuit approach. The different procedures correspond to robust versions of the unsmoothed principal component estimators, to the estimators obtained penalizing the scale and to those obtained by penalizing the norm. A sieve approach based on approximating the elements of the unit ball by elements over finite-dimensional spaces is also considered. In particular, the procedures based on smoothing and sieves are new. A robust cross-validation procedure is introduced to select the smoothing parameters. Consistency results are derived for the four type of estimators. Finally, a simulation study confirms the expected inadequate behavior 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 a penalization in the norm, where the smoothing parameter is selected via a robust K-fold cross-validation, is recommended.

Among other contributions we highlight the following:

(a) We obtain the continuity of the principal directions and eigenvalue functionals, which implies the asymptotic qualitative robustness of the corresponding

estimators. This extends the results of Li and Chen (1985) from Euclidean spaces to infinite-dimensional Hilbert spaces, where the unit ball is not compact. Noncompactness poses technical challenges which we overcome with tools from functional analysis.

(b) Our results not only include the finite-dimensional case but also improve upon some of the results obtained in that situation for the projection pursuit estimators. For example, the assumptions in Li and Chen (1985) regarding the robust scale functional are stronger than ours. Also, to derive the consistency of the raw estimators, we only require uniform convergence over the unit ball of $s_n(\alpha)$ to $\sigma(\alpha)$, which holds if the scale functional σ_R is continuous. This improves upon the consistency results given in Cui, He and Ng (2003), who require a uniform Bahadur expansion for $s_n(\alpha)$ over the unit ball.

(c) A key step in proving the continuity of the projection pursuit functional is to show that weak convergence of probability measures over a Hilbert space implies uniform convergence of the laws of the projections of the stochastic processes, that is, Theorem 6.2. This uniform convergence result can be useful in other statistical problems where projection methods are considered.

(d) The proofs for the penalized estimators include, as particular cases, the estimators defined by Rice and Silverman (1991) and studied by Pezzulli and Silverman (1993), and those considered by Silverman (1996). Extending the results to scale estimators other than the standard deviation required more challenging arguments since, unlike the classical setting, the projection-pursuit index $s_n^2(\alpha)$ cannot be expressed in the simple form $\langle \alpha, \mathbf{\Gamma}_n \alpha \rangle$ for some compact operator $\mathbf{\Gamma}_n$.

APPENDIX

In this Appendix, we give the proofs of the results stated in Section 4. Some technicalities are omitted, and we refer to the technical report [Bali et al. (2010)] for details. Before presenting the proof, some additional notation is needed.

Denote by $\mathcal{L}_{m-1}$ the linear space spanned by $\{\phi_{R,1}, \ldots, \phi_{R,m-1}\}$, and let $\widehat{\mathcal{L}}_{m-1}$ be the linear space spanned by the first $m-1$ estimated principal directions, that is, by $\{\widehat{\phi}_{SI,1}, \ldots, \widehat{\phi}_{SI,m-1}\}$ or $\{\widehat{\phi}_1, \ldots, \widehat{\phi}_{m-1}\}$, where it will be clear in each case which linear space we are considering. The latter includes the situation of the linear spaces spanned by $\{\widehat{\phi}_{RAW,1}, \ldots, \widehat{\phi}_{RAW,m-1}\}$, $\{\widehat{\phi}_{PS,1}, \ldots, \widehat{\phi}_{PS,m-1}\}$ and $\{\widehat{\phi}_{PN,1}, \ldots, \widehat{\phi}_{PN,m-1}\}$. Finally, for any linear space $\mathcal{L}$, $\pi_{\mathcal{L}} : \mathcal{H} \to \mathcal{L}$ stands for the orthogonal projection onto the linear space $\mathcal{L}$, which exists if $\mathcal{L}$ is a closed linear space. In particular, $\pi_{\mathcal{L}_{m-1}}$, $\pi_{\widehat{\mathcal{L}}_{m-1}}$ and $\pi_{\mathcal{H}_{p_n}}$ are well defined.

Moreover, for the sake of simplicity, denote by $\mathcal{T}_k = \mathcal{L}_k^{\perp}$ the linear space orthogonal to $\phi_1, \ldots, \phi_k$ and by $\pi_k = \pi_{\mathcal{T}_k}$ the orthogonal projection with respect to the inner product defined in $\mathcal{H}$. On the other hand, let $\widehat{\pi}_{\tau,k}$ be the projection onto the linear space orthogonal to $\widehat{\phi}_1, \ldots, \widehat{\phi}_k$ in the space $\mathcal{H}_S$ in the inner product $\langle \cdot, \cdot \rangle_{\tau}$, that is, for any $\alpha \in \mathcal{H}_S$, $\widehat{\pi}_{\tau,k}(\alpha) = \alpha - \sum_{j=1}^k \langle \alpha, \widehat{\phi}_j \rangle_{\tau} \widehat{\phi}_j$. Moreover, let $\widehat{\mathcal{T}}_{\tau,k}$ stand

for the linear space orthogonal to $\widehat{\mathcal{L}}_k$ with the inner product $\langle \cdot, \cdot \rangle_\tau$. Thus, $\widehat{\pi}_{\tau,k}$ is the orthogonal projection onto $\widehat{\mathcal{T}}_{\tau,k}$ with respect to this inner product.

PROOF OF THEOREM 4.1. First note that the fact that σ_R is a scale functional entails that $\sigma_n(\alpha) = \|\alpha\|\sigma_n(\alpha/\|\alpha\|)$. Thus from assumption (ii) and the fact that $\|\alpha\| \leq \|\alpha\|_\tau$, we get that

(A.1) $$\sup_{\|\alpha\| \leq 1} |\sigma_n^2(\alpha) - \sigma^2(\alpha)| \xrightarrow{\text{a.s.}} 0 \quad \text{and} \quad \sup_{\|\alpha\|_\tau \leq 1} |\sigma_n^2(\alpha) - \sigma^2(\alpha)| \xrightarrow{\text{a.s.}} 0.$$

(a) To prove that $\widehat{\lambda}_1 \xrightarrow{\text{a.s.}} \sigma^2(\phi_{R,1})$ it is enough to show that

(A.2) $$\sigma^2(\phi_{R,1}) \geq \widehat{\lambda}_1 + o_{\text{a.s.}}(1),$$

(A.3) $$\sigma^2(\phi_{R,1}) \leq \widehat{\lambda}_1 + o_{\text{a.s.}}(1),$$

where $o_{\text{a.s.}}(1)$ stands for a term converging to 0 almost surely.

Note that from (A.1), we get that $a_{n,1} = \sigma_n^2(\widehat{\phi}_1) - \sigma^2(\widehat{\phi}_1) \xrightarrow{\text{a.s.}} 0$ and $b_{n,1} = \sigma_n^2(\phi_{R,1}) - \sigma^2(\phi_{R,1}) \xrightarrow{\text{a.s.}} 0$. Using that σ is a scale functional and that $\sigma^2(\phi_{R,1}) = \sup_{\alpha \in \mathcal{S}} \sigma^2(\alpha)$, we obtain easily that

$$\sigma^2(\phi_{R,1}) \geq \sigma^2\left(\frac{\widehat{\phi}_1}{\|\widehat{\phi}_1\|}\right) = \frac{\sigma^2(\widehat{\phi}_1)}{\|\widehat{\phi}_1\|^2} \geq \sigma^2(\widehat{\phi}_1) = \sigma_n^2(\widehat{\phi}_1) - a_{n,1} = \widehat{\lambda}_1 + o_{\text{a.s.}}(1)$$

concluding the proof of (A.2).

To derive (A.3), note that since $\phi_{R,1} \in \mathcal{H}_S$, $\|\phi_{R,1}\|_\tau < \infty$ and $\|\phi_{R,1}\|_\tau \geq \|\phi_{R,1}\| = 1$, then, defining $\beta_1 = \phi_{R,1}/\|\phi_{R,1}\|_\tau$, we have that $\|\beta_1\|_\tau = 1$, which implies that $\widehat{\lambda}_1 = \sigma_n^2(\widehat{\phi}_1) \geq \sigma_n^2(\widehat{\phi}_1) - \rho\Psi(\widehat{\phi}_1) \geq \sigma_n^2(\beta_1) - \rho\Psi(\beta_1)$. Hence, using that σ_R is a scale functional and that $\Psi(a\alpha) = a^2\Psi(\alpha)$, for any $a \in \mathbb{R}$, we get

$$\widehat{\lambda}_1 \geq \sigma_n^2(\beta_1) - \rho\Psi(\beta_1) = \frac{\sigma_n^2(\phi_{R,1}) - \rho\Psi(\phi_{R,1})}{\|\phi_{R,1}\|_\tau^2} = \frac{\sigma^2(\phi_{R,1}) + b_{n,1} - \rho\Psi(\phi_{R,1})}{\|\phi_{R,1}\|_\tau^2}.$$

When $\rho = 0$, we have defined $\rho\Psi(\phi_{R,1}) = 0$ and similarly when $\tau = 0$. So from now on, we will assume that $\tau_n > 0$ and $\rho_n > 0$. Since $b_{n,1} = o_{\text{a.s.}}(1)$, $\rho \xrightarrow{\text{a.s.}} 0$ and $\tau \xrightarrow{\text{a.s.}} 0$, we have that $\rho\Psi(\phi_{R,1}) \xrightarrow{\text{a.s.}} 0$ and $\|\phi_{R,1}\|_\tau \xrightarrow{\text{a.s.}} \|\phi_{R,1}\| = 1$, concluding the proof of (A.3). Hence, $\widehat{\lambda}_1 \xrightarrow{\text{a.s.}} \sigma^2(\phi_{R,1})$.

From (A.1) and the fact that $\|\widehat{\phi}_1\| \leq 1$, we obtain that $\widehat{\lambda}_1 - \sigma^2(\widehat{\phi}_1) = \sigma_n^2(\widehat{\phi}_1) - \sigma^2(\widehat{\phi}_1) \xrightarrow{\text{a.s.}} 0$. Therefore, using that $\widehat{\lambda}_1 \xrightarrow{\text{a.s.}} \sigma^2(\phi_{R,1})$, we get that

(A.4) $$\sigma^2(\widehat{\phi}_1) \xrightarrow{\text{a.s.}} \sigma^2(\phi_{R,1}).$$

Moreover, the inequalities $\sigma^2(\phi_{R,1}) \geq \sigma^2(\widehat{\phi}_1/\|\widehat{\phi}_1\|) \geq \sigma^2(\widehat{\phi}_1)$ obtained above also imply that

(A.5) $$\sigma^2(\widehat{\phi}_1/\|\widehat{\phi}_1\|) \xrightarrow{\text{a.s.}} \sigma^2(\phi_{R,1}).$$

Using that $\|\widehat{\phi}_1\|_\tau = 1$, we get that $\tau\Psi(\widehat{\phi}_1) = 1 - \|\widehat{\phi}_1\|^2 = 1 - \sigma^2(\widehat{\phi}_1)/\sigma^2(\widehat{\phi}_1/\|\widehat{\phi}_1\|)$. Hence, (A.4) and (A.5) entail that $\tau\Psi(\widehat{\phi}_1) \xrightarrow{\text{a.s.}} 0$.

It only remains to show that $\rho\Psi(\widehat{\phi}_1) \xrightarrow{\text{a.s.}} 0$, which follows easily from the fact that $\widehat{\lambda}_1 \xrightarrow{\text{a.s.}} \sigma^2(\phi_{R,1})$, $\sigma_n^2(\phi_{R,1}) \xrightarrow{\text{a.s.}} \sigma^2(\phi_{R,1})$, $\rho \xrightarrow{\text{a.s.}} 0$ and $\|\phi_{R,1}\|_\tau \xrightarrow{\text{a.s.}} 1$ since $\widehat{\lambda}_1 \geq \sigma_n^2(\widehat{\phi}_1) - \rho\lceil\widehat{\phi}_1, \widehat{\phi}_1\rceil \geq (\sigma_n^2(\phi_{R,1}) - \rho\lceil\phi_{R,1}, \phi_{R,1}\rceil)/\|\phi_{R,1}\|_\tau^2$.

Note that we have not used the weak continuity of σ as a function of α to derive (a).

(b) Note that since $\|\widehat{\phi}_1\|_\tau = 1$, we have that $\|\widehat{\phi}_1\| \leq 1$. Moreover, from (a), $\|\widehat{\phi}_1\| \xrightarrow{\text{a.s.}} 1$. Let $\widetilde{\phi}_1 = \widehat{\phi}_1/\|\widehat{\phi}_1\|$, then $\widetilde{\phi}_1 \in \mathcal{S}$ and $\sigma(\widetilde{\phi}_1) = \sigma(\widehat{\phi}_1)/\|\widehat{\phi}_1\|$. Using that $\sigma^2(\widehat{\phi}_1) \xrightarrow{\text{a.s.}} \sigma^2(\phi_{R,1})$ and $\|\widehat{\phi}_1\| \xrightarrow{\text{a.s.}} 1$, we obtain that $\sigma^2(\widetilde{\phi}_1) \xrightarrow{\text{a.s.}} \sigma^2(\phi_{R,1})$, and thus the proof follows using Lemma 4.1.

(c) Let us show that $\widehat{\lambda}_m \xrightarrow{\text{a.s.}} \sigma^2(\phi_{R,m})$. The proof will be done in several steps by showing

$$(A.6) \qquad \sup_{\|\alpha\|_\tau \leq 1} |\sigma^2(\pi_{m-1}\alpha) - \sigma_n^2(\widehat{\pi}_{\tau,m-1}\alpha)| \xrightarrow{\text{a.s.}} 0,$$

$$(A.7) \qquad \sigma^2(\phi_{R,m}) \geq \widehat{\lambda}_m + o_{\text{a.s.}}(1),$$

$$(A.8) \qquad \sigma^2(\phi_{R,m}) \leq \widehat{\lambda}_m + o_{\text{a.s.}}(1).$$

Note that (A.6) corresponds to an extension of assumption (ii) while (A.7) and (A.8) are analogous to (A.2) and (A.3).

We begin by proving (A.6). Note that $\sup_{\|\alpha\|_\tau \leq 1}|\sigma_n^2(\pi_{m-1}\alpha) - \sigma^2(\widehat{\pi}_{\tau,m-1}\alpha)| \leq \sup_{\|\alpha\|_\tau \leq 1}|\sigma^2(\pi_{m-1}\alpha) - \sigma^2(\widehat{\pi}_{\tau,m-1}\alpha)| + \sup_{\|\alpha\|_\tau \leq 1}|\sigma_n^2(\widehat{\pi}_{\tau,m-1}\alpha) - \sigma^2(\widehat{\pi}_{\tau,m-1}\alpha)|$. Using (A.1) and the fact that if $\|\alpha\|_\tau \leq 1$, then $\|\widehat{\pi}_{\tau,m-1}\alpha\|_\tau \leq 1$, we get that the second term on the right-hand side converges to 0 almost surely. To complete the proof of (A.6), it remains to show that

$$(A.9) \qquad \sup_{\|\alpha\|_\tau \leq 1} |\sigma^2(\pi_{m-1}\alpha) - \sigma^2(\widehat{\pi}_{\tau,m-1}\alpha)| \xrightarrow{\text{a.s.}} 0.$$

Using that $\widehat{\phi}_j \xrightarrow{\text{a.s.}} \phi_{R,j}$ and that $\tau\Psi(\widehat{\phi}_j) = \tau\lceil\widehat{\phi}_j, \widehat{\phi}_j\rceil \xrightarrow{\text{a.s.}} 0$, for $1 \leq j \leq m-1$ and arguing as in Silverman (1996) [see Bali et al. (2010) for details], we get that, for $1 \leq j \leq m-1$, $\sup_{\|\alpha\|_\tau \leq 1}\|\langle\alpha, \phi_{R,j}\rangle\phi_{R,j} - \langle\alpha, \widehat{\phi}_j\rangle_\tau\widehat{\phi}_j\| \xrightarrow{\text{a.s.}} 0$, entailing that $\sup_{\|\alpha\|_\tau \leq 1}\|\widehat{\pi}_{\tau,m-1}\alpha - \pi_{m-1}\alpha\| \xrightarrow{\text{a.s.}} 0$. Therefore, using that σ is weakly uniformly continuous over the unit ball, we get easily that (A.9) holds, concluding the proof of (A.6).

As in (a), we will next show that (A.7) holds. Using again that σ is a scale functional, we get easily that $\sup_{\alpha \in \mathcal{S} \cap \mathcal{T}_{m-1}} \sigma^2(\alpha) = \sup_{\alpha \in \mathcal{S}} \sigma^2(\pi_{m-1}\alpha)$, so using again that $\|\widehat{\phi}_m\| \leq \|\widehat{\phi}_m\|_\tau = 1$, we obtain that $\sigma^2(\phi_{R,m}) = \sup_{\alpha \in \mathcal{S}} \sigma^2(\pi_{m-1}\alpha) \geq \sigma^2(\pi_{m-1}\widehat{\phi}_m/\|\widehat{\phi}_m\|) \geq \sigma^2(\pi_{m-1}\widehat{\phi}_m)$. From (A.6) and the fact that $\|\widehat{\phi}_m\|_\tau = 1$, we get that $b_m = \sigma^2(\pi_{m-1}\widehat{\phi}_m) - \sigma_n^2(\widehat{\pi}_{\tau,m-1}\widehat{\phi}_m) \xrightarrow{\text{a.s.}} 0$, and so since $\widehat{\pi}_{\tau,m-1}\widehat{\phi}_m = \widehat{\phi}_m$

and $\|\widehat{\phi}_m\| \leq 1$, we get that $\sigma^2(\phi_{R,m}) \geq \sigma^2(\pi_{m-1}\widehat{\phi}_m) = \sigma_n^2(\widehat{\pi}_{\tau,m-1}\widehat{\phi}_m) + o_{a.s.}(1) = \widehat{\lambda}_m + o_{a.s.}(1)$, completing the proof of (A.7).

We will show now (A.8). Note that $\phi_{R,m} \in \mathcal{H}_S$, so that $\|\phi_{R,m}\|_\tau < \infty$ and $\|\phi_{R,m}\|_\tau \to \|\phi_{R,m}\| = 1$. Using that σ_R is a scale functional, the fact that $\widehat{\lambda}_m = \sigma_n^2(\widehat{\phi}_m) \geq \sigma_n^2(\widehat{\phi}_m) - \rho\Psi(\widehat{\phi}_m) = \sup_{\|\alpha\|_\tau=1, \alpha\in\widehat{\mathcal{T}}_{\tau,m-1}}\{\sigma_n^2(\alpha) - \rho\Psi(\alpha)\}$ and that for any $\alpha \in \mathcal{H}_S$ such that $\|\alpha\|_\tau = 1$ we have that $\|\widehat{\pi}_{\tau,m-1}\alpha\|_\tau \leq 1$, we get easily that $\widehat{\lambda}_m \geq \sup_{\|\alpha\|_\tau=1}\{\sigma_n^2(\widehat{\pi}_{\tau,m-1}\alpha) - \rho\Psi(\widehat{\pi}_{\tau,m-1}\alpha)\}$, and so $\widehat{\lambda}_m \geq (\sigma_n^2(\widehat{\pi}_{\tau,m-1}\phi_{R,m}) - \rho\Psi(\widehat{\pi}_{\tau,m-1}\phi_{R,m}))/\|\phi_{R,m}\|_\tau^2$. From (A.6) we obtain that $d_m = \sigma_n^2(\widehat{\pi}_{\tau,m-1}\phi_{R,m}) - \sigma^2(\pi_{m-1}\phi_{R,m}) \xrightarrow{a.s.} 0$. Moreover, the fact that $\tau \xrightarrow{a.s.} 0$ entails that $\|\phi_{R,m}\|_\tau \xrightarrow{a.s.} \|\phi_{R,m}\| = 1$. On the other hand, using that $\rho\Psi(\widehat{\phi}_\ell) \xrightarrow{a.s.} 0$, $1 \leq \ell \leq m-1$, and the fact that $\rho \xrightarrow{a.s.} 0$ implies that $\rho\Psi(\phi_{R,m}) = o_{a.s.}(1)$, analogous arguments to those considered in Pezzulli and Silverman (1993) allow us to show that $\rho\Psi(\widehat{\pi}_{m-1}\phi_{R,m}) = \rho\lceil\widehat{\pi}_{m-1}\phi_{R,m}, \widehat{\pi}_{m-1}\phi_{R,m}\rceil \xrightarrow{a.s.} 0$. Hence, we get that

$$\widehat{\lambda}_m \geq \frac{\sigma^2(\pi_{m-1}\phi_{R,m}) + d_m - \rho\Psi(\widehat{\pi}_{\tau,m-1}\phi_{R,m})}{1 + o(1)}$$

$$\geq \frac{\sigma^2(\pi_{m-1}\phi_{R,m}) + d_m - o_{a.s.}(1)}{1 + o(1)}$$

$$= \sigma^2(\phi_{R,m}) + o_{a.s.}(1),$$

where the last equality follows from the fact that $\pi_{m-1}\phi_{R,m} = \phi_{R,m}$.

Therefore, from (A.7) and (A.8), we obtain that $\widehat{\lambda}_m \xrightarrow{a.s.} \sigma^2(\phi_{R,m})$.

On the other hand, (A.6) entails that $\widehat{\lambda}_m - \sigma^2(\widehat{\phi}_m) = \sigma_n^2(\widehat{\phi}_m) - \sigma^2(\widehat{\phi}_m) \xrightarrow{a.s.} 0$, which together with $\widehat{\lambda}_m \xrightarrow{a.s.} \sigma^2(\phi_{R,m})$ implies that $\sigma^2(\widehat{\phi}_m) \xrightarrow{a.s.} \sigma^2(\phi_{R,m})$.

To complete the proof of (c), it remains to show that $\tau\Psi(\widehat{\phi}_m) \xrightarrow{a.s.} 0$ and $\rho\Psi(\widehat{\phi}_m) \xrightarrow{a.s.} 0$. As in (a), we have that the following inequalities converge to equalities:

$$(A.10) \quad \sigma^2(\phi_{R,m}) \geq \sigma^2\left(\pi_{m-1}\frac{\widehat{\phi}_m}{\|\widehat{\phi}_m\|}\right) \geq \sigma^2(\pi_{m-1}\widehat{\phi}_m) = \widehat{\lambda}_m + o_{a.s.}(1).$$

Using that σ is a scale estimator and that $\|\widehat{\phi}_m\|_\tau = 1$, we get that $\tau\Psi(\widehat{\phi}_m) = 1 - \|\widehat{\phi}_m\|^2 = 1 - \sigma^2(\pi_{m-1}\widehat{\phi}_m)/\sigma^2(\pi_{m-1}\widehat{\phi}_m/\|\widehat{\phi}_m\|)$, which together with (A.10) entails that the second term on the right-hand side is $1 + o_{a.s.}(1)$ and so, $\tau\lceil\widehat{\phi}_m, \widehat{\phi}_m\rceil \xrightarrow{a.s.} 0$, entailing that $\|\widehat{\phi}_m\| \xrightarrow{a.s.} 1$.

On the other hand, we also have that

$$(A.11) \quad \widehat{\lambda}_m = \sigma_n^2(\widehat{\phi}_m) \geq \sigma_n^2(\widehat{\phi}_m) - \rho\Psi(\widehat{\phi}_m) \geq \sigma^2(\phi_{R,m}) + o_{a.s.}(1),$$

so using that $\widehat{\lambda}_m = \sigma_n^2(\widehat{\phi}_m) \xrightarrow{a.s.} \sigma^2(\phi_{R,m})$, we obtain that $\rho\Psi(\widehat{\phi}_m) \xrightarrow{a.s.} 0$, concluding the proof of (c).

(d) We have already proved that when $m = 1$ the result holds. We proceed by induction and assume that $\langle \widehat{\phi}_\ell, \phi_{R,\ell} \rangle^2 \to 1$, $\tau \Psi(\widehat{\phi}_\ell) \xrightarrow{\text{a.s.}} 0$ and $\rho \Psi(\widehat{\phi}_\ell) \xrightarrow{\text{a.s.}} 0$ for $1 \le \ell \le m - 1$, to show that $\langle \widehat{\phi}_m, \phi_{R,m} \rangle^2 \to 1$. Without loss of generality, we can assume that $\widehat{\phi}_\ell \xrightarrow{\text{a.s.}} \phi_{R,\ell}$, for $1 \le \ell \le m - 1$. Denote by $\widetilde{\phi}_j = \widehat{\phi}_j / \|\widehat{\phi}_j\|$. Then, for $1 \le \ell \le m - 1$, $\|\widehat{\phi}_\ell\| \to 1$, and so $\widetilde{\phi}_\ell \xrightarrow{\text{a.s.}} \phi_{R,\ell}$. It suffices to show that $\langle \phi_{R,m}, \widetilde{\phi}_m \rangle^2 \xrightarrow{\text{a.s.}} 1$.

Using (c) we have that $\sigma^2(\widehat{\phi}_m) \xrightarrow{\text{a.s.}} \sigma^2(\phi_{R,m})$ and that $\|\widehat{\phi}_m\| \xrightarrow{\text{a.s.}} 1$, and so $\sigma^2(\widetilde{\phi}_m) \xrightarrow{\text{a.s.}} \sigma^2(\phi_{R,m})$. The proof follows now from Lemma 4.1 if we show that $\langle \widetilde{\phi}_m, \widetilde{\phi}_\ell \rangle \xrightarrow{\text{a.s.}} 0$, $1 \le \ell \le m - 1$.

Using that $\tau \Psi(\widehat{\phi}_\ell) \xrightarrow{\text{a.s.}} 0$, for $1 \le \ell \le m - 1$, and that from (c) $\tau \Psi(\widehat{\phi}_m) \xrightarrow{\text{a.s.}} 0$ we get that $\tau \lceil \widehat{\phi}_\ell, \widehat{\phi}_m \rceil \xrightarrow{\text{a.s.}} 0$ for $1 \le \ell \le m - 1$. Therefore, the fact that $\langle \widehat{\phi}_m, \widehat{\phi}_\ell \rangle_\tau = 0$ entails that $\langle \widehat{\phi}_m, \widehat{\phi}_\ell \rangle = \langle \widehat{\phi}_m, \widehat{\phi}_\ell \rangle_\tau - \tau \lceil \widehat{\phi}_\ell, \widehat{\phi}_m \rceil \xrightarrow{\text{a.s.}} 0$, and so $\langle \widetilde{\phi}_m, \widetilde{\phi}_\ell \rangle \xrightarrow{\text{a.s.}} 0$, concluding the proof. $\square$

SUPPLEMENTARY MATERIAL

Supplement A: Robust functional principal components (DOI: 10.1214/11-AOS923SUPPA; .pdf). In this Supplement, we give the proof of some of the results stated in Sections 4 and 6.

Supplement B: Robust functional principal components (DOI: 10.1214/11-AOS923SUPPB; .pdf). In this Supplement, we report the results obtained in the Monte Carlo study for the raw estimators and for the penalized ones when the smoothing parameters are fixed.

REFERENCES

BALI, J. L. and BOENTE, G. (2009). Principal points and elliptical distributions from the multivariate setting to the functional case. *Statist. Probab. Lett.* **79** 1858–1865. MR2560484

BALI, J. L., BOENTE, G., TYLER, D. E. and WANG, J. L. (2010). Robust functional principal components: A projection-pursuit approach. Available at www.ic.fcen.uba.ar/preprints/baliboentetylerwang2010.pdf.

BALI, J. L., BOENTE, G., TYLER, D. E. and WANG, J. L. (2011a). Supplement A to "Robust functional principal components: A projection-pursuit approach." DOI:10.1214/11-AOS923SUPPA.

BALI, J. L., BOENTE, G., TYLER, D. E. and WANG, J. L. (2011b). Supplement B to "Robust functional principal components: A projection-pursuit approach." DOI:10.1214/11-AOS923SUPPB.

BEATON, A. E. and TUKEY, J. W. (1974). The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics* **16** 147–185.

BOENTE, G. and FRAIMAN, R. (1999). Discussion of "Robust principal components for functional data," by Locantore et al. *Test* **8** 28–35.

BOENTE, G. and FRAIMAN, R. (2000). Kernel-based functional principal components. *Statist. Probab. Lett.* **48** 335–345. MR1771495

CANTONI, E. and RONCHETTI, E. (2001). Resistant selection of the smoothing parameter for smoothing splines. *Stat. Comput.* **11** 141–146. MR1837133

CROUX, C. (1994). Efficient high-breakdown *M*-estimators of scale. *Statist. Probab. Lett.* **19** 371–379. MR1278674

CROUX, C. (1999). Discussion of "Robust principal components for functional data," by Locantore et al. *Test* **8** 41–46.

CROUX, C., FILZMOSER, P. and OLIVEIRA, M. R. (2007). Algorithms for projection-pursuit robust principal component analysis. *Chemometrics and Intelligent Laboratory Systems* **87** 218–225.

CROUX, C. and RUIZ-GAZEN, A. (1996). A fast algorithm for robust principal components based on projection pursuit. In *Compstat*: *Proceedings in Computational Statistics* (A. Prat, ed.) 211–217. Physica-Verlag, Heidelberg.

CROUX, C. and RUIZ-GAZEN, A. (2005). High breakdown estimators for principal components: The projection-pursuit approach revisited. *J. Multivariate Anal.* **95** 206–226. MR2164129

CUI, H., HE, X. and NG, K. W. (2003). Asymptotic distributions of principal components based on robust dispersions. *Biometrika* **90** 953–966. MR2024769

DAUXOIS, J., POUSSE, A. and 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. MR0650934

GERVINI, D. (2006). Free-knot spline smoothing for functional data. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68** 671–687. MR2301014

GERVINI, D. (2008). Robust functional estimation using the median and spherical principal components. *Biometrika* **95** 587–600. MR2443177

HALL, P. and HOSSEINI-NASAB, M. (2006). On properties of functional principal components analysis. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68** 109–126. MR2212577

HALL, P., MÜLLER, H.-G. and WANG, J.-L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *Ann. Statist.* **34** 1493–1517. MR2278365

HAMPEL, F. R. (1971). A general qualitative definition of robustness. *Ann. Math. Statist.* **42** 1887–1896. MR0301858

HUBER, P. J. (1981). *Robust Statistics*. Wiley, New York. MR0606374

HYNDMAN, R. J. and SHAHID ULLAH, M. (2007). Robust forecasting of mortality and fertility rates: A functional data approach. *Comput. Statist. Data Anal.* **51** 4942–4956. MR2364551

LI, G. and CHEN, Z. (1985). Projection-pursuit approach to robust dispersion matrices and principal components: Primary theory and Monte Carlo. *J. Amer. Statist. Assoc.* **80** 759–766.

LOCANTORE, N., MARRON, J. S., SIMPSON, D. G., TRIPOLI, N., ZHANG, J. T. and COHEN, K. L. (1999). Robust principal component analysis for functional data. *Test* **8** 1–73. MR1707596

MARDEN, J. I. (1999). Some robust estimates of principal components. *Statist. Probab. Lett.* **43** 349–359. MR1707944

MARONNA, R. A., MARTIN, R. D. and YOHAI, V. J. (2006). *Robust Statistics*: *Theory and Methods*. Wiley, Chichester. MR2238141

MARTIN, R. D. and ZAMAR, R. H. (1993). Bias robust estimation of scale. *Ann. Statist.* **21** 991–1017. MR1232529

PEZZULLI, S. and SILVERMAN, B. W. (1993). Some properties of smoothed principal components analysis for functional data. *Comput. Statist.* **8** 1–16. MR1220336

RAMSAY, J. O. and SILVERMAN, B. W. (2005). *Functional Data Analysis*, 2nd ed. Springer, New York. MR2168993

RICE, J. A. and SILVERMAN, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *J. Roy. Statist. Soc. Ser. B* **53** 233–243. MR1094283

SILVERMAN, B. W. (1996). Smoothed functional principal components analysis by choice of norm. *Ann. Statist.* **24** 1–24. MR1389877

TYLER, D. E. (2010). A note on multivariate location and scatter statistics for sparse data sets. *Statist. Probab. Lett.* **80** 1409–1413. MR2669781

VARADARAJAN, V. S. (1958). On the convergence of sample probability distributions. *Sankhyā* **19** 23–26. MR0094839

VISURI, S., KOIVUNEN, V. and OJA, H. (2000). Sign and rank covariance matrices. *J. Statist. Plann. Inference* **91** 557–575. MR1814801

WANG, F. T. and SCOTT, D. W. (1994). The L_1 method for robust nonparametric regression. *J. Amer. Statist. Assoc.* **89** 65–76. MR1266287

YAO, F. and LEE, T. C. M. (2006). Penalized spline models for functional principal component analysis. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68** 3–25. MR2212572

J. L. BALI
G. BOENTE
INSTITUTO DE CÁLCULO
FACULTAD DE CIENCIAS EXACTAS Y NATURALES
CIUDAD UNIVERSITARIA, PABELLÓN 2
BUENOS AIRES, 1428
ARGENTINA
E-MAIL: lbali@dm.uba.ar
        gboente@dm.uba.ar

D. E. TYLER
DEPARTMENT OF STATISTICS
HILL CENTER, BUSCH CAMPUS
RUTGERS UNIVERSITY
PISCATAWAY NEW JERSEY 08854
USA
E-MAIL: dtyler@rci.rutgers.edu

J.-L. WANG
DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
DAVIS, CALIFORNIA 95616
USA
E-MAIL: wang@wald.ucdavis.edu