$S-$estimators for functional principal component analysis*

Graciela Boente and Matías Salibián Barrera

Abstract

Principal components analysis is a widely used technique that provides an optimal lower-dimensional approximation to multivariate or functional data sets. These approximations can be very useful in identifying potential outliers among high–dimensional or functional observations. In this paper, we propose a new class of estimators for principal components based on robust scale estimators. For a fixed dimension q, we robustly estimate the $q-$dimensional linear space that provides the best prediction for the data, in the sense of minimizing the sum of robust scale estimators of the coordinates of the residuals. The extension to the infinite-dimensional case is also studied. In analogy to the linear regression case, we call this proposal $S-$estimators. Our method is consistent for elliptical random vectors, and is Fisher-consistent for elliptically distributed random elements on arbitrary Hilbert spaces. Numerical experiments show that our proposal is highly competitive when compared

*Graciela Boente (Email: gboente@dm.uba.ar) is Full Professor at the Departamento de Matemáticas from the Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Ciudad Universitaria, Pabellón 2, Buenos Aires 1428, Argentina (Tel./FAX: 54-11-45763375). She also has a researcher position at the CONICET. Matías Salibián Barrera (Email: matias@stat.ubc.ca) is Associate Professor at the Department of Statistics from the University of British Columbia, 3182 Earth Sciences Building, 2207 Main Mall, Vancouver, BC, V6T 1Z4, Canada (Tel./FAX: 1-604-8223410). This research was partially supported by Grants PIP 112-201101-00339 from CONICET, PICT 0397 from ANPCYT and W276 from the Universidad de Buenos Aires at Buenos Aires, Argentina (G. Boente) and Discovery Grant of the Natural Sciences and Engineering Research Council of Canada (M. Salibián Barrera). The authors wish to thank the Associate Editor and three anonymous referees for valuable comments which led to an improved version of the original paper.

with other existing methods when the data are generated both by finite- or infinite-rank stochastic processes. We also illustrate our approach using two real functional data sets, where the robust estimator is able to discover atypical observations in the data that would have been missed otherwise.

Key Words: Functional Data Analysis, Principal Components, Robust estimation, $S-$estimator, Sparse Data.

AMS Subject Classification: MSC 62G35, 62H25

1 Introduction

Principal components analysis (PCA) is a widely used method to obtain a lower-dimensional approximation to multivariate data. This approximation is optimal in the sense of minimizing the mean squared loss between the original observations and the resulting approximations. Estimated principal components can be a valuable tool to explore the data visually, and are also useful to describe some characteristics of the data (e.g. directions of high variability). Thanks to the ever reducing cost of collecting data, many data sets in current applications are both large and complex, sometimes with a very high number of variables. The chance of having outliers or other type of imperfections in the data increases both with the number of observations and their dimension. Thus, detecting these outlying observations is an important step, even when robust estimates are used, either as a pre-processing step or because there is some specific interest in finding anomalous observations. However, it is well known that detecting outliers or other anomalies in multivariate data can be difficult (Rousseeuw and van Zomeren, 1990; Becker and Gather, 1999, 2001), and one has to rely on robust statistical methodologies.

As a motivation, consider the problem of identifying days with an atypical concentration of ground level ozone (O3) in the air. Ground level ozone forms as a result of the reaction between sunlight, nitrogen oxide (NOx) and volatile organic compounds (VOC). It is an important air pollutant, present around urban areas, with higher concentrations in suburban or rural locations downwind from major sources of NOx and VOC, such as industries, gasoline vapours, and motor vehicle exhaust emissions (Sillman, 1993). Ground level ozone is a major irritant of the airways, and exposure to it can lead to an increased risk of developing cardiovascular disease and several respiratory conditions (U.S. Environmental Protection Agency, 2008). Its intensity is affected by several meteorological and topographical factors (such as temperature and wind direction), which affect the distribution of its precursors (Ainslie and Steyn, 2007). We obtained hourly average concentration of ground level ozone at a monitoring station in Richmond, BC (a few kilometres south of the city Vancouver, BC). The data comes from the Ministry of Environment of the province of British Columbia, and is available on line at `http://envistaweb.env.gov.bc.ca`. Since ground level ozone pollution is most severe in Summer, we focus on the month of August. Our data includes observations for the years 2004 to 2012. Figure 1 displays the data. Each line corresponds to the evolution of the hourly average concentration (in ppb) of ground level ozone for one day. The Canadian National Ambient Air Quality Objectives sets a maximum desired level of 50 ppb. It is easy to see that a few days exceeded the maximum desired level threshold, but also that there may be other days exhibiting a different pattern of hourly average concentration of O3. We are interested in identifying days with atypical hourly O3 trajectories.

In this paper, we study robust low–dimensional approximations for high−(or infinite−) dimensional data that can be used to identify poorly fitted observations as potential outliers. The earliest and probably most immediate approach to obtain robust estimates for the principal components consists in using the eigenvalues and eigenvectors of a robust scatter estimator (Devlin *et al.*, 1981; Campbell, 1980; Boente, 1987; Naga and Antille, 1990; Croux and Haesbroeck, 2000). A different approach was proposed by Locantore *et al.* (1999) based on using the covariance matrix of the data projected onto the unit sphere. Since principal component directions are also those that provide projections with the largest variability, robust PCA estimators can alternatively be obtained as the directions that maximize a robust estimator of scale of the projected data. This approach

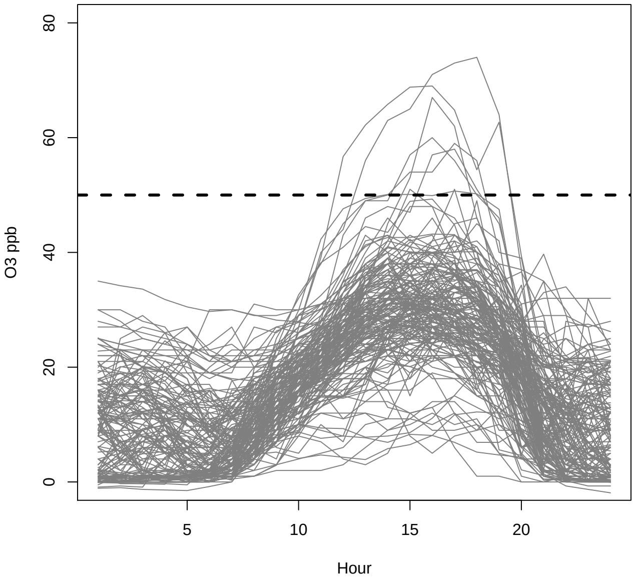


Figure 1: Hourly mean concentration (in ppb) of ground level ozone in Richmond, BC, Canada, for the month of August in years 2004 to 2012. Each line corresponds to one day. The darker dashed horizontal line at 50 ppb is the current maximum desired level set by the Canadian National Ambient Air Quality Objectives. The maximum acceptable level is 80 ppb.

is known in the literature as "projection pursuit" and has been studied by Li and Chen (1985), Croux and Ruiz–Gazen (1996, 2005), Hubert *et al.* (2002) and Hubert *et al.* (2005).

It is well–known that, for finite–dimensional observations with finite second moments, when using mean squared errors the best lower–dimensional approximation is given by the projections onto the linear space spanned by the eigenvectors of the covariance matrix corresponding to its largest eigenvalues. Several robust proposals exist in the literature exploiting this characterization of PCA. They amount to replacing the squared residuals with a different loss function. Liu *et al.* (2003) used the absolute value of the residuals, and McCoy and Tropp (2011) proposed a randomized algorithm to find an approximate solution to this L_1 minimization problem. Not surprisingly this approach may not work well when entire observations are atypical (corresponding to "high–leverage" points in linear regression models). Croux *et al.* (2003) proposed a weighted version of this procedure that reduces the effect of high–leverage points. Verboon and Heiser (1994) and De la Torre and Black (2001) used a bounded loss function applied to column–wise standardized residuals. Later, Maronna and Yohai (2008) proposed a similar loss function, but modified in such a way that the method reduces to the classical PCA when one uses a squared loss function. Maronna (2005) also considered best–estimating lower–dimensional subspaces directly, but his approach cannot be easily extended to infinite–dimensional settings because there may be infinitely many minimum eigenvalues.

There has been recent attention paid to a similar problem in the Engineering and Computer Science literature. The main assumption in their approach is that a proportion of the observations lies on a proper lower-dimensional subspace, and that there may be a sparse amount of arbitrary additive "noise" present. The objective is to fully recover the low-rank part of the data. Chandrasekaran *et al.* (2011), Candès *et al.* (2011), McCoy and Tropp (2011) and Xu *et al.* (2012) study different convex relaxations of the problem of finding an exact representation of the data matrix as the sum of a low-rank one and a sparse one. Lerman *et al.* (2012) and Zhang and Lerman (2014) also consider convex relaxations of this problem. The focus of these proposals is on obtaining fast algorithms, and they derive sufficient conditions to guarantee that the solution to the surrogate convex optimization problem is the lower dimensional subspace that properly contains the "non-outlying" points. These conditions can be interpreted as follows: a proportion of the data points needs to lie on a low-dimensional proper subspace of the sample space, and the corresponding low-rank matrix should be "diffuse" (for example, its row and column spaces cannot be aligned with the coordinate axes).

Our approach relies on a probabilistic model and assumes that our observations follow an elliptical distribution. We are interested in studying best lower-dimensional approximations, in the sense of minimizing the expected prediction error over the distribution of the random vector. These approximations need not fit exactly any subset of the data.

Moreover, our goal is to obtain robust alternatives for estimating principal spaces in infinite–dimensional settings. We use finite (or high-)dimensional estimators as a step towards achieving that purpose. Nevertheless, our proposal provides consistent estimators of the best lower–dimensional subspace when applied to multivariate data that follow an elliptical distribution, even if second moments do not exist. Furthermore, our approach is Fisher consistent for the case of infinite-dimensional observations. Few robust principal components estimates for functional data (FPCA) have been proposed in the literature. Gervini (2008) studied spherical principal components, and Hyndman and Ullah (2007) discuss a projection–pursuit approach using smoothed trajectories, but without studying their properties in detail. More recently, Sawant *et al.* (2012) adapted the BACONPCA method to detect outliers and to provide robust estimators of the functional components, while Bali *et al.* (2011) proposed robust projection–pursuit FPCA estimators and showed that they are consistent to the eigenfunctions and eigenvalues of the underlying process.

The rest of the paper is organized as follows. Section 2 tackles the problem of providing robust estimators for a $q-$dimensional approximation for Euclidean data. Section 3 discusses extending this methodology to accommodate functional data, and its use to detect outliers is described in Section 4. In Section 5 we report the results of a simulation study conducted to study the performance of the proposed procedure for functional data. Some real data sets are analysed in Section 6, where the advantage of the proposed procedure to detect possible influential observations is illustrated. Finally, Section 7 provides some further discussion and recommendations. Proofs are relegated to the Appendix.

2 $S-$estimators of the principal components in $\mathbb{R}^p$

Consider the problem of finding a lower–dimensional approximation to a set of observations $\mathbf{x}_i$, $1 \leq i \leq n$, in $\mathbb{R}^p$. More specifically, we look for $q < p$ vectors $\mathbf{b}^{(l)} \in \mathbb{R}^p$, $1 \leq l \leq q$, whose spanned linear sub–space provides a good approximation to the data. Let $\mathbf{B} \in \mathbb{R}^{p\times q}$ be the matrix given by $\mathbf{B} = \left(\mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(q)}\right)$ and denote $\mathbf{b}_j^{\mathrm{T}}$ the jth row of $\mathbf{B}$. Furthermore, let $\boldsymbol{\mu} \in \mathbb{R}^p$. The corresponding "fitted values" are $\widehat{\mathbf{x}}_i = \boldsymbol{\mu} + \mathbf{B}\,\mathbf{a}_i$, $1 \leq i \leq n$, where $\mathbf{a}_i \in \mathbb{R}^q$. We can also write $\widehat{x}_{ij} = \mu_j + \mathbf{a}_i^{\mathrm{T}}\mathbf{b}_j$. With this notation, principal components can be defined as minimizers, over matrices $\mathbf{A} \in \mathbb{R}^{n\times q}$, $\mathbf{B} \in \mathbb{R}^{p\times q}$ and vectors $\boldsymbol{\mu} \in \mathbb{R}^p$, of

$$L_2(\mathbf{A}, \mathbf{B}, \boldsymbol{\mu}) \;=\; \sum_{i=1}^{n} \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|^2_{\mathbb{R}^p} \;=\; \sum_{i=1}^{n}\sum_{j=1}^{p} r_{ij}^2 \,, \tag{1}$$

where the i-th row of the matrix $\mathbf{A} \in \mathbb{R}^{n\times q}$ is $\mathbf{a}_i$, $r_{ij} = x_{ij} - \widehat{x}_{ij}$ and $\|\cdot\|_{\mathbb{R}^p}$ denotes the usual Euclidean norm in $\mathbb{R}^p$. Furthermore, this optimization problem can be solved using alternating regression iterations. Note that if we restrict $\mathbf{B}$ to satisfy $\mathbf{B}^{\mathrm{T}}\mathbf{B} = \mathbf{I}_q$, i.e., if the columns $\mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(q)}$ are orthonormal, then the vectors $\mathbf{a}_i$, $1 \leq i \leq n$, correspond to the scores of the sample on the orthonormal basis $\mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(q)}$.

Our approach is based on noting that $L_2(\mathbf{A}, \mathbf{B}, \boldsymbol{\mu})$ in (1) is proportional to $\sum_{j=1}^{p} s_j^2$ where s_j^2 is the sample variance of the residuals' jth coordinate: $r_{1j}, r_{2j}, \ldots, r_{nj}$. To reduce the influence of atypical observations we propose to use robust scale estimates instead of sample variances. Our robustly estimated $q-$dimensional subspace best approximating the data is defined as the linear space spanned by the columns of the matrix $\mathbf{B}$ where $(\mathbf{A}, \mathbf{B}, \boldsymbol{\mu})$ minimizes

$$L_{\mathrm{S}}(\mathbf{A}, \mathbf{B}, \boldsymbol{\mu}) = \sum_{j=1}^{p} \widehat{\sigma}_j^2 \,, \tag{2}$$

and $\widehat{\sigma}_j$ denotes a robust scale estimator of the residuals $r_{ij} = x_{ij} - \widehat{x}_{ij}$, $1 \leq i \leq n$. As mentioned before, if $\mathbf{B}^{\mathrm{T}}\mathbf{B} = \mathbf{I}_q$, the vectors $\mathbf{a}_i$, $1 \leq i \leq n$, correspond to the robust scores of the ith observation in the orthonormal basis $\mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(q)}$. Note that if we use the sample variance s_j^2 instead of $\widehat{\sigma}_j^2$, then the objective function in (2) reduces to the classical one in (1).

Scale estimators measure the spread of a sample and are invariant under translations and equivariant under scale transformations (see, for example, Maronna *et al.* 2006). Although any robust scale estimator can be used in (2), to fix ideas we focus our presentation on $M-$estimators of scale (see Huber and Ronchetti, 2009). As in Maronna *et al.*

(2006), let $\rho : \mathbb{R} \to \mathbb{R}_+$ be a $\rho-$function, that is, an even function, non–decreasing on $|x|$, increasing for $x > 0$ when $\rho(x) < \lim_{t\to+\infty} \rho(t)$ and such that $\rho(0) = 0$. Given residuals $r_{ij}(\mathbf{A}, \mathbf{B}, \boldsymbol{\mu}) = x_{ij} - \widehat{x}_{ij}(\mathbf{A}, \mathbf{B}, \boldsymbol{\mu})$ with $\widehat{x}_{ij}(\mathbf{A}, \mathbf{B}, \boldsymbol{\mu}) = \mu_j + \mathbf{a}_i^{\mathrm{T}} \mathbf{b}_j$, the $M-$estimator of scale of the residuals $\widehat{\sigma}_j = \widehat{\sigma}_j(\mathbf{A}, \mathbf{B}, \boldsymbol{\mu})$ satisfies

$$\frac{1}{n} \sum_{i=1}^{n} \rho_c \left(\frac{r_{ij}(\mathbf{A}, \mathbf{B}, \boldsymbol{\mu})}{\widehat{\sigma}_j} \right) = b \,, \tag{3}$$

where $\rho_c(u) = \rho(u/c)$, and $c > 0$ is a user–chosen tuning constant. When $\rho(y) = \min\left(3y^2 - 3y^4 + y^6, 1\right)$, (Tukey's biweight function) with $c = 1.54764$ and $b = 1/2$, the estimator is Fisher–consistent at the normal distribution and has breakdown point 50%. In general, if $\|\rho\|_\infty = 1$, then the breakdown point of the $M-$scale estimator solving (3) is $\min(b, 1 - b)$.

We can write our estimator in a slightly more general way as follows. Given a matrix $\mathbf{B} \in \mathbb{R}^{p\times q}$, let $\mathcal{L}_{\mathbf{B}}$ be the $q-$dimensional linear space spanned by its columns $\mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(q)}$. Similarly, let $\pi(\mathbf{y}, \mathcal{L}_{\mathbf{B}})$ denote the orthogonal projection of $\mathbf{y}$ onto $\mathcal{L}_{\mathbf{B}}$. To simplify the presentation, assume that $\boldsymbol{\mu}$ is known. For each observation $\mathbf{x}_i \in \mathbb{R}^p$, $1 \le i \le n$, let $\mathbf{r}_i(\mathcal{L}_{\mathbf{B}}) = \mathbf{x}_i - \boldsymbol{\mu} - \pi(\mathbf{x}_i - \boldsymbol{\mu}, \mathcal{L}_{\mathbf{B}}) = (r_{i1}(\mathcal{L}_{\mathbf{B}}), \ldots, r_{ip}(\mathcal{L}_{\mathbf{B}}))^{\mathrm{T}}$ denote the corresponding vector of residuals and $\widehat{\sigma}_{j,\mathcal{L}_{\mathbf{B}}} = \widehat{\sigma}(r_{1j}(\mathcal{L}_{\mathbf{B}}), \ldots, r_{nj}(\mathcal{L}_{\mathbf{B}}))$ the scale estimator of the jth coordinate of the residuals. We define the $S-$estimator of the best $q-$dimensional approximation to the data as the linear space $\widehat{\mathcal{L}} = \mathcal{L}_{\widehat{\mathbf{B}}}$ that minimizes the sum of the $M-$estimators of scale of the coordinates of the residuals over all linear spaces $\mathcal{L}_{\mathbf{B}}$ of dimension q:

$$\mathcal{L}_{\widehat{\mathbf{B}}} = \underset{\dim(\mathcal{L}_{\mathbf{B}})=q}{\operatorname{argmin}} \widehat{\Psi}_n \left(\mathcal{L}_{\mathbf{B}}\right) \,, \tag{4}$$

where $\widehat{\Psi}_n \left(\mathcal{L}_{\mathbf{B}}\right) = \sum_{j=1}^{p} \widehat{\sigma}^2_{j,\mathcal{L}_{\mathbf{B}}}$.

To study the asymptotic properties of robust estimators it is convenient to think of them as functionals of the empirical distribution of the sample (Huber and Ronchetti, 2009). For example, $M-$scale estimators in (3) correspond to the functional $\sigma_{\mathrm{R}} : \mathcal{D} \to \mathbb{R}_+$ defined for each distribution function $F \in \mathcal{D}$ as the solution $\sigma_{\mathrm{R}}(F)$ to the equation $\int \rho_c(t/\sigma_{\mathrm{R}}(F))\, dF(t) = b$. Here $\mathcal{D}$ is a subset of all the univariate distributions that contains all the empirical ones. Given a sample $y_1, \ldots, y_n$ with empirical distribution F_n, we can write $\widehat{\sigma}(y_1, \ldots, y_n) = \sigma_{\mathrm{R}}(F_n)$.

In what follows we will assume that $\mathbf{x}_i \in \mathbb{R}^p$, $1 \le i \le n$ are independent and identically distributed random vectors with distribution P. The independence condition may be relaxed, for instance, requiring stationarity and a mixing condition or just ergodicity, since we only need the strong law of large numbers to hold in order to guarantee the consistency results given below. For a random vector $\mathbf{x}$ with distribution P, the functional $\mathcal{L}(P)$ corresponding to the $S-$estimators defined in (4) is the linear space of dimension q that satisfies

$$\mathcal{L}(P) \ = \mathop{\mathrm{argmin}}_{\dim(\mathcal{L})=q} \Psi(\mathcal{L})\,, \tag{5}$$

where $\Psi(\mathcal{L}) = \sum_{j=1}^{p} \sigma_{j,\mathcal{L}}^2$, $\sigma_{j,\mathcal{L}} = \sigma_{\mathrm{R}}(F_j(\mathcal{L}_{\mathbf{B}}))$ and $F_j(\mathcal{L}_{\mathbf{B}})$ denotes the distribution of $r_j(\mathcal{L}_{\mathbf{B}})$ with $\mathbf{r}(\mathcal{L}_{\mathbf{B}}) = \mathbf{x} - \boldsymbol{\mu} - \pi(\mathbf{x} - \boldsymbol{\mu}, \mathcal{L}_{\mathbf{B}}) = (r_1(\mathcal{L}_{\mathbf{B}}), \ldots, r_p(\mathcal{L}_{\mathbf{B}}))^{\mathrm{T}}$. Proposition 2.1 below shows that this functional is Fisher–consistent for elliptical random vectors.

Recall that a random vector is said to have a spherical distribution if its distribution is invariant under orthogonal transformations. In particular, the characteristic function of a spherically distributed $\mathbf{x} \in \mathbb{R}^p$ is of the form $\varphi_{\mathbf{x}}(\mathbf{t}) = \phi(\mathbf{t}^{\mathrm{T}}\mathbf{t})$ for $\mathbf{t} \in \mathbb{R}^p$ where $\phi : \mathbb{R} \to \mathbb{R}$ is the generator of the characteristic function. We write $\mathbf{x} \sim \mathcal{S}_p(\phi)$. For a $p \times p$ matrix $\mathbf{B}$ and a vector $\boldsymbol{\mu} \in \mathbb{R}^p$, the distribution of $\mathbf{x} = \mathbf{B}\mathbf{z} + \boldsymbol{\mu}$ when $\mathbf{z} \sim \mathcal{S}_p(\psi)$ is said to have an elliptical distribution, denoted by $\mathbf{x} \sim \mathcal{E}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \phi)$, where $\boldsymbol{\Sigma} = \mathbf{B}\mathbf{B}^{\mathrm{T}}$. The characteristic function of $\mathbf{x}$ is $\varphi_{\mathbf{x}}(\mathbf{t}) = \exp(i\mathbf{t}^{\mathrm{T}}\boldsymbol{\mu})\phi(\mathbf{t}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{t})$.

Proposition 2.1 *Let $\mathbf{x} \sim \mathcal{E}_p(\mathbf{0}, \boldsymbol{\Sigma}, \phi)$ be a random vector elliptically distributed with location $\mathbf{0}$ and scale $\boldsymbol{\Sigma}$ such that $\boldsymbol{\Sigma} = \boldsymbol{\beta}\boldsymbol{\Lambda}\boldsymbol{\beta}^{\mathrm{T}}$ where $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$, $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_p$, and $\boldsymbol{\beta}$ is an orthonormal matrix with columns $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p$. Assume that $\lambda_q > \lambda_{q+1}$. Then, if $\mathcal{L}_q$ is the linear space spanned by $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_q$, we have that $\mathcal{L}_q$ is the unique solution of (5), that is, $\mathcal{L}(P)$ is a Fisher–consistent functional at $P = \mathcal{E}_p(\mathbf{0}, \boldsymbol{\Sigma}, \phi)$.*

As mentioned before, this approach can also be used with any robust scale estimator.

For example, we can define $\tau-$estimators by considering a $\tau-$scale. Define as above, $r_{ij}(\mathbf{A},\mathbf{B},\boldsymbol{\mu}) = x_{ij} - \widehat{x}_{ij}(\mathbf{A},\mathbf{B},\boldsymbol{\mu})$ where $\widehat{x}_{ij}(\mathbf{A},\mathbf{B},\boldsymbol{\mu}) = \mu_j + \mathbf{a}_i^{\mathrm{T}}\mathbf{b}_j$ and let ρ and ρ_1 be two $\rho-$functions such that $\rho \le \rho_1$. The $\tau-$best lower dimensional approximations are given by the minimizers of

$$L_\tau(\mathbf{A},\mathbf{B},\boldsymbol{\mu}) = \sum_{j=1}^{p} \widehat{\sigma}_j^2 \sum_{i=1}^{n} \rho_1\left(\frac{r_{ij}(\mathbf{A},\mathbf{B},\boldsymbol{\mu})}{\widehat{\sigma}_j}\right),$$

where $\widehat{\sigma}_j = \widehat{\sigma}_j(\mathbf{A},\mathbf{B},\boldsymbol{\mu})$ is a robust scale estimator of $r_{ij}(\mathbf{A},\mathbf{B},\boldsymbol{\mu})$ computed as in (3) with the $\rho-$function ρ. When an iterative procedure is used to find the minimizer of (4), the p scale estimators $\widehat{\sigma}_j$ need to be updated at each stop of the algorithm.

Consistency of projection–pursuit principal component estimators for random vectors was derived in Cui *et al.* (2003) requiring uniform convergence over the unit ball of the projected data scale estimators to the scale functional. This condition was generalized in Bali *et al.* (2011) to the functional case. A natural extension of this condition for $q > 1$ is

$$\sup_{\dim(\mathcal{L})=q} |\widehat{\Psi}_n(\mathcal{L}) - \Psi(\mathcal{L})| \xrightarrow{a.s.} 0\,. \tag{6}$$

Note that this condition is easily verified when using a robust scale functional with finite–dimensional random vectors since the Stiefel manifold $\mathcal{V}_{p\times q} = \{\mathbf{B} \in \mathbb{R}^{p\times q} : \mathbf{B}^{\mathrm{T}}\mathbf{B} = \mathbf{I}_q\}$ is a compact set. Furthermore, the following proposition shows that this condition is sufficient to obtain consistency of the $S-$estimators in (4).

Proposition 2.2 *Assume that $\mathcal{L}(P)$ is unique and that* (6) *holds. Then, the estimators $\widehat{\mathcal{L}} = \mathcal{L}_{\widehat{\mathbf{B}}}$ obtained minimizing $\widehat{\Psi}_n(\mathcal{L})$ in* (4) *over linear spaces $\mathcal{L}$ of dimension q, are consistent to the linear space $\mathcal{L}(P)$ defined in (5). In other words, with probability one $\pi(\mathbf{x},\widehat{\mathcal{L}})$ converges to $\pi(\mathbf{x},\mathcal{L}(P))$, for almost all* $\mathbf{x}$.

2.1 Algorithm for $S-$estimators

The optimization problem defining our estimator is generally non-convex, and typically difficult to solve. In this Section, we show that first–order conditions for a critical point of the objective function in (4) naturally suggests iterative re–weighted least squares iterations. Once such iterations are available, a standard strategy used in the Statistical literature to compute this type of estimators (e.g. Rousseeuw and van Driessen, 1999; Maronna, 2005; Salibian-Barrera and Yohai, 2006) is to iterate a large number of random initial points, and select the best visited local minimum as the estimator.

Note that although $S-$scale estimators are only defined implicitly, explicit first–order conditions can be obtained differentiating both sides of (3). More specifically, let $\widehat{\sigma}_j$, $j = 1, \ldots, p$ be an $M-$estimator of scale of the residuals $x_{ij} - \widehat{x}_{ij}$, $i = 1, \ldots, n$. In other words, $\widehat{\sigma}_j$ satisfies $(1/n) \sum_{i=1}^n \rho\left((x_{ij} - \mu_j - \mathbf{a}_i^{\mathrm{T}} \mathbf{b}_j)/\widehat{\sigma}_j\right) = b$, where we have absorbed the constant c into the loss function ρ. The derivatives with respect to $\mathbf{a}_i$, $i = 1, \ldots, n$ are given by

$$\frac{\partial}{\partial \mathbf{a}_i} \left(\sum_{j=1}^p \widehat{\sigma}_j^2 \right) = \sum_{j=1}^p 2\, \widehat{\sigma}_j \frac{\partial \widehat{\sigma}_j}{\partial \mathbf{a}_i} = -2 \sum_{j=1}^p \widehat{\sigma}_j\, h_j^{-1} \rho' \left(\frac{r_{ij}}{\widehat{\sigma}_j} \right) \mathbf{b}_j\,, \quad i = 1, \ldots, n\,,$$

where $h_j = \sum_{i=1}^n \rho'\left(r_{ij}/\widehat{\sigma}_j\right) r_{ij}/\widehat{\sigma}_j$. Similarly, the other first–order conditions are

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{b}_s} \left(\sum_{j=1}^p \widehat{\sigma}_j^2 \right) &= \sum_{j=1}^p 2\, \widehat{\sigma}_j \frac{\partial \widehat{\sigma}_j}{\partial \mathbf{b}_s} = -2\, \widehat{\sigma}_s\, h_s^{-1} \sum_{i=1}^n \rho' \left(\frac{r_{is}}{\widehat{\sigma}_s} \right) \mathbf{a}_i\,, \quad s = 1, \ldots, p \\
\frac{\partial}{\partial \mu_\ell} \left(\sum_{j=1}^p \widehat{\sigma}_j^2 \right) &= \sum_{j=1}^p 2\, \widehat{\sigma}_j \frac{\partial \widehat{\sigma}_j}{\partial \mu_\ell} = -2\, \widehat{\sigma}_\ell\, h_\ell^{-1} \sum_{i=1}^n \rho' \left(\frac{r_{i\ell}}{\widehat{\sigma}_\ell} \right), \quad \ell = 1, \ldots, p\,.
\end{aligned}$$

Setting these to zero, we obtain a system of equations that can be re–expressed as re–weighted least–squares problems as follows: let $w_{ij} = \widehat{\sigma}_j\, h_j^{-1} r_{ij}^{-1}\, \rho'(r_{ij}/\widehat{\sigma}_j)$, then we need to solve

$$\begin{aligned}
\sum_{j=1}^p w_{ij}\, (x_{ij} - \mu_j) \mathbf{b}_j &= \left(\sum_{j=1}^p w_{ij}\, \mathbf{b}_j \mathbf{b}_j^{\mathrm{T}} \right) \mathbf{a}_i\,, \quad 1 \leq i \leq n\,, \\
\sum_{i=1}^n w_{ij}\, (x_{ij} - \mu_j) \mathbf{a}_i &= \left(\sum_{i=1}^n w_{ij}\, \mathbf{a}_i \mathbf{a}_i^{\mathrm{T}} \right) \mathbf{b}_j\,, \quad 1 \leq j \leq p\,, \\
\sum_{i=1}^n w_{ij}\, (x_{ij} - \mathbf{a}_i^{\mathrm{T}} \mathbf{b}_j) &= \sum_{i=1}^n w_{ij}\, \mu_j\,, \quad 1 \leq j \leq p\,.
\end{aligned}$$

This formulation suggests the usual iterative re–weighted least squares (IRWLS) algorithm. Given initial estimates $\mathbf{b}_j^{(0)}$, $1 \leq j \leq p$ and $\boldsymbol{\mu}^{(0)}$, compute the scores $\mathbf{a}_i^{(0)}$, $i = 1, \ldots, n$, the weights $w_{ij}^{(0)}$ and obtain updated values for $\mathbf{a}_i^{(1)}$, $\mathbf{b}_j^{(1)}$, $1 \leq i \leq n$, $1 \leq j \leq p$ and $\boldsymbol{\mu}^{(1)}$. We repeat these steps until the objective function changes less than a chosen tolerance value. The best $q-$dimensional linear space approximation is spanned by $\{\widehat{\mathbf{b}}^{(1)}, \cdots, \widehat{\mathbf{b}}^{(q)}\}$, the final values obtained above. For interpretation purposes, we orthogonalize the set $\{\widehat{\mathbf{b}}^{(1)}, \cdots, \widehat{\mathbf{b}}^{(q)}\}$ and compute the scores $\widehat{\mathbf{a}}_i$ as the corresponding orthogonal projections.

For the initial location vector $\boldsymbol{\mu}^{(0)}$ we use the L^1-median, and adapt the strategy of Rousseeuw and van Driessen (1999) to select initial values for $\mathbf{B}$ and $\mathbf{A}$. More specifically, we generate N_1 random starts for the matrix $\mathbf{B}$ which are orthogonalized, each of them leading to an initial matrix $\mathbf{B}^{(0)}$. The columns of the matrix $\mathbf{A}$ are the scores of each observation on the basis given by the q columns of $\mathbf{B}^{(0)}$. For each of these initial values we run N_2 IRWLS iterations, or until a tolerance level is achieved. The initial values giving the best objective function after N_2 iterations are then iterated until convergence. This algorithm depends on the number of random starts N_1, the desired tolerance for sequential change in the objective function, and the number of iterations N_2 that is applied to each random candidate. In our experiments we used a tolerance of 10^{-6} and found that using $N_1 = 50$ random starts and $N_2 = 50$ partial IRWLS iterations for each of them was typically sufficient to find a good solution to (4), which is in line with the results of Maronna (2005).

An implementation of this algorithm in `R` is publicly available on-line from `http://www.stat.ubc.ca/~matias/soft.html`. Although a formal computational complexity analysis of this algorithm is beyond the scope of his paper, our numerical experiments reported in Section 5 show that the algorithm works very well. We tested the speed of our `R` code using these settings on an Intel i7 CPU (3.5GHz) machine running Windows 7. In Table 1 we report the average time in CPU minutes over 10 random samples for different combinations of the sample size (n), number of variables (p) and dimension of the subspace (q). Note that these times could be improved notably if the algorithm was implemented in `C` or a language with faster linear algebra operations.

	p											
	50			100			200			500		
n	$q=1$	$q=2$	$q=5$	$q=1$	$q=2$	$q=5$	$q=1$	$q=2$	$q=5$	$q=1$	$q=2$	$q=5$
50	4.2	4.3	3.9	8.7	8.9	8.1	17.8	17.9	16.8	53.5	53.8	53.3
70	4.4	4.4	4.0	8.9	8.5	8.1	17.7	17.8	17.3	55.0	56.9	60.1
100	5.0	4.9	4.8	9.9	10.0	8.9	20.5	22.6	23.6	69.0	67.2	70.8
200	5.6	6.0	5.8	11.5	12.2	10.8	25.8	28.1	25.0	97.6	108.5	116.3

Table 1: Average timing of the IRWLS algorithm in CPU minutes for different data and model configurations using $N_1 = N_2 = 50$ and tolerance equal to 10^{-6}.

2.2 Choosing the dimension of the approximating subspace

In some cases, the desired dimension of the linear subspace providing an approximation to the data is either known or chosen in advance (e.g. for visualization purposes). In many applications, however, this dimension is selected based on the resulting "proportion of unexplained variability".

Proposition 2.1 shows that for elliptically distributed random vectors $\mathbf{x} \sim \mathcal{E}_p(\mathbf{0}, \boldsymbol{\Sigma}, \phi)$, the functional $\Psi(\mathcal{L})$ is minimized when $\mathcal{L} = \mathcal{L}_q$ the subspace spanned by the first q eigenvectors of the scatter matrix and $\Psi(\mathcal{L}_q) = \sum_{j=q+1}^{p} \lambda_j$. Note that for $q = 0$ we have $\Psi(\mathcal{L}_0) = \sum_{j=1}^{p} \lambda_j = \operatorname{tr}(\boldsymbol{\Sigma}) = \sum_{j=1}^{p} \sigma_{j,0}^2$, where $\sigma_{j,0} = \sigma_{\mathrm{R}}(F_{j,0})$ with $F_{j,0}$ the distribution of $r_j(\boldsymbol{\mu}) = x_j - \mu_j$. Thus, the proportion of unexplained variability can be defined as $u_q = \Psi(\mathcal{L}_q)/\Psi(\mathcal{L}_0)$ and an estimator of u_q is given by $\widehat{u}_q = \widehat{\Psi}_n(\widehat{\mathcal{L}}_q)/\widehat{\Psi}_n(\widehat{\mathcal{L}}_0)$, where $\widehat{\mathcal{L}}_q$ is defined in (4) and $\widehat{\mathcal{L}}_0$ corresponds to minimizing $\widehat{\Psi}_n(\mathcal{L}_0) = \sum_{j=1}^{p} \widehat{\sigma}_{j,\mathcal{L}_0}^2$ with $\widehat{\sigma}_{j,\mathcal{L}_0} = \widehat{\sigma}(r_{1j}(\boldsymbol{\mu}), \ldots, r_{nj}(\boldsymbol{\mu}))$ the scale estimator of the jth coordinate of the residuals $\mathbf{r}_i(\boldsymbol{\mu}) = \mathbf{x}_i - \boldsymbol{\mu} = (r_{i1}(\boldsymbol{\mu}), \ldots, r_{ip}(\boldsymbol{\mu}))^{\mathrm{T}}$. Proposition 2.2 can be used to show the consistency of $\widehat{u}_q$ to u_q.

To avoid the high computational cost of solving (4) for different values of q we adapt the strategy of Maronna (2005). Let $u_{\max}$ be the maximum allowed proportion of unexplained variability, and a maximum dimension $q_{\max}$ of the approximating subspace. We look for the smallest q_0 such that $q_0 \leq q_{\max}$ and $\widehat{u}_{q_0} \leq u_{\max}$. We first verify that $\widehat{u}_{q_{\max}} \leq u_{\max}$ otherwise the problem cannot be solved and we need to modify our goals. The procedure starts with $q_1 = 1$. If $\widehat{u}_1 \leq u_{\max}$ we are done. Otherwise, assume that after j steps, we have $\widehat{u}_{q_j} \geq u_{\max}$, where $q_j = j$ dimension used in step j. Let $\widehat{\boldsymbol{\mu}}^{(q_j)}$ be the estimated center and $\widehat{\mathbf{B}}_{q_j} \in \mathbb{R}^{p\times q_j}$ the orthonormal basis of the best q_j-dimensional subspace, with columns $\widehat{b}_{q_j}^{(1)}, \ldots, \widehat{b}_{q_j}^{(q_j)}$. As before, let $\widehat{\mathbf{A}}_{q_j} \in \mathbb{R}^{n\times q_j}$ be the matrix of scores. Let $q_{j+1} = q_j + 1$ and define the matrices $\mathbf{B} = \left(\widehat{\mathbf{B}}_{q_j}, \boldsymbol{\beta}\right) \in \mathbb{R}^{p\times q_{j+1}}$, with $\boldsymbol{\beta} \in \mathbb{R}^p$, and $\mathbf{A} = \left(\widehat{\mathbf{A}}_{q_j}, \boldsymbol{\alpha}\right) \in \mathbb{R}^{n\times q_{j+1}}$ with $\boldsymbol{\alpha} \in \mathbb{R}^n$. Let $\mathbf{b}_1, \ldots, \mathbf{b}_{q_j+1}$ and $\mathbf{a}_1, \ldots, \mathbf{a}_n$ denote the columns of $\mathbf{B}$ and the rows of $\mathbf{A}$, respectively. We construct our predictions as $\widehat{x}_{i\ell}^{(q_{j+1})} = \widehat{\mu}_\ell^{(q_j)} + \mathbf{a}_i^{\mathrm{T}}\mathbf{b}_\ell$, and note that the residuals satisfy $r_{i\ell}^{(q_{j+1})} = r_{i\ell}^{(q_j)} - \alpha_i\beta_\ell$. Our problem is now to minimize $L_{\mathrm{S}}(\mathbf{A}, \mathbf{B}, \widehat{\boldsymbol{\mu}}^{(q_j)})$ over $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ such that $\widehat{\mathbf{B}}_{q_j}^{\mathrm{T}}\boldsymbol{\beta} = \mathbf{0}$, with $L_{\mathrm{S}}(\mathbf{A}, \mathbf{B}, \boldsymbol{\mu})$ given in (2). A system of equations analogous to that described in Section 2.1 can be derived to formulate an iterative re–weighted least squares algorithm. Once the optimal $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are found, we optimize $L_{\mathrm{S}}(\mathbf{A}, \mathbf{B}, \boldsymbol{\mu})$ over $\boldsymbol{\mu}$ to obtain $\widehat{\boldsymbol{\mu}}^{(q_{j+1})}$. This approach is much faster than solving (4) for $q = q_{j+1}$. Note that $\widetilde{u}_{q_{j+1}} = \widehat{\Psi}_n(\widetilde{\mathcal{L}}_{q_{j+1}})/\widehat{\Psi}_n(\widehat{\mathcal{L}}_0)$ is typicaly larger than $\widehat{u}_{q_{j+1}}$, so that if $\widetilde{u}_{q_{j+1}} \leq u_{\max}$, we select $q = q_{j+1}$, and otherwise increase j and continue.

3 $S-$estimators in the functional setting

In this section, we discuss extensions of the estimators defined in Section 2 to accommodate functional data. The most common situation corresponds to the case when the observations correspond to realizations of a stochastic process $X \in L^2(\mathcal{I})$ with $\mathcal{I}$ an interval of the real line, which can be assumed to be $\mathcal{I} = [0, 1]$. A more general setup that can accommodate applications where observations are images, for example, is to consider realizations of a random element on a separable Hilbert space $\mathcal{H}$ with inner product

$\langle\cdot,\cdot\rangle_{\mathcal{H}}$ and norm $\|\cdot\|_{\mathcal{H}}$. Note that principal components for functional data (defined via the Karhunen–Loève decomposition of the covariance function of the process X) also have the property of providing best lower–dimensional approximations, in the L^2 sense. Recently, a stochastic best lower–dimensional approximation for elliptically distributed random elements on separable Hilbert spaces, such as those considered when dealing with multivariate data, was obtained by Boente *et al.* (2012). This optimality property does not require second moment conditions.

However, even in the simplest situation when $X \in L^2([0,1])$, one rarely observes entire curves. The functional datum for replication i usually corresponds to a finite set of discrete values $x_{i\,1},\dots,x_{i\,m_i}$ with $x_{ij} = X_i(t_{ij})$, $1 \le j \le m_i$. Depending on the characteristics of the grid of points t_{ij} where observations were obtained, one can employ different strategies to analyze these data.

The easiest situation is when observations were made at common design points. In this case, we have $p = m_1 = m_i$ and $t_{ij} = \tau_j$, for all $1 \le i \le n$ and $1 \le j \le p$. Defining $\mathbf{x}_i = (x_{i\,1},\dots,x_{i\,p})^{\mathrm{T}}$ a purely multivariate approach can be used as in Section 2 to obtain a $q-$dimensional linear space $\widehat{\mathcal{L}}$ spanned by orthonormal vectors $\widehat{\mathbf{b}}^{(1)},\cdots,\widehat{\mathbf{b}}^{(q)}$. An associated basis in $L^2([0,1])$ can be defined as $\widehat{\phi}_\ell(\tau_j) = a_\ell \widehat{b}_{\ell\,j}$, for $1 \le \ell \le q$, $1 \le j \le p$, where a_ℓ is a constant to ensure that $\|\widehat{\phi}_\ell\|_{L^2} = 1$ and $\widehat{\mathbf{b}}^{(\ell)} = (b_{\ell\,1},\cdots,b_{\ell\,p})^{\mathrm{T}}$. Smoothing over the observed data points one can recover the complete trajectory. This approach provides a consistent estimator for the best approximating linear space and the corresponding "fitted trajectories" $\pi(X_i,\widehat{\mathcal{L}})$, $1 \le i \le n$.

In many cases, however, trajectories are observed at different design points t_{ij}, $1 \le j \le m_i$, $1 \le i \le n$. In what follows we will assume that as the sample size n increases, so does the number of points where each trajectory is observed and that, in the limit, these points cover the interval $[0,1]$. Our approach consists of using a sequence of finite–dimensional functional spaces that increases with the sample size. The basic idea is to identify each observed point in $\mathcal{H}$ with the vector formed by its coordinates on a finite–dimensional basis that increases with the sample size. The procedure of Section 2 can be applied to these vectors to obtain a $q-$dimensional approximating subspace, which can then be mapped back onto $\mathcal{H}$.

More specifically, let $\{\delta_i\}_{i\geq 1}$ be an orthonormal basis of $\mathcal{H}$ and, for each $n \geq 1$, let $\mathcal{H}_{p_n}$ be the linear space spanned by $\delta_1, \dots, \delta_{p_n}$. To simplify the notation we write $p = p_n$. Let $x_{ij} = \langle X_i, \delta_j \rangle_{\mathcal{H}}$ be the coefficient of the ith trajectory on the jth element of the basis, $1 \leq j \leq p$, and form the $p-$dimensional vector $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^{\mathrm{T}}$. When, $\mathcal{H} = L^2([0,1])$, the inner products $\langle X_i, \delta_j \rangle_{\mathcal{H}}$ can be numerically computed using a Riemann sum over the design points for the ith trajectory $\{t_{ij}\}_{1\leq j\leq m_i}$. We apply the procedure described in Section 2 to the multivariate observations $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^p$ to obtain a $q-$dimensional linear space $\widehat{\mathcal{L}}$ spanned by orthonormal vectors $\widehat{\mathbf{b}}^{(1)}, \cdots, \widehat{\mathbf{b}}^{(q)}$ and the corresponding "predicted values" $\widehat{\mathbf{x}}_i = \widehat{\boldsymbol{\mu}} + \sum_{\ell=1}^{q} \widehat{a}_{i\ell} \widehat{\mathbf{b}}^{(\ell)}$, with $\widehat{\boldsymbol{\mu}} = (\widehat{\mu}_1, \dots, \widehat{\mu}_p)^{\mathrm{T}}$. It is now easy to find the corresponding approximation in the original space $\mathcal{H}$. The location parameter is $\widehat{\mu}_{\mathcal{H}} = \sum_{j=1}^{p} \widehat{\mu}_j \delta_j$, and the associated $q-$dimensional basis in $\mathcal{H}$ is $\widehat{\phi}_\ell = \sum_{j=1}^{p} \widehat{b}_{\ell j} \delta_j / \| \sum_{j=1}^{p} \widehat{b}_{\ell j} \delta_j \|_{\mathcal{H}}$, for $1 \leq \ell \leq q$. Furthermore, the "fitted values" in $\mathcal{H}$ are $\widehat{X}_i = \widehat{\mu}_{\mathcal{H}} + \sum_{\ell=1}^{q} \widehat{a}_{i\ell} \widehat{\phi}_\ell$. Moreover, since $\|\mathbf{x}_i - \widehat{\mathbf{x}}_i\|_{\mathbb{R}^p} \simeq \|X_i - \widehat{X}_i\|_{\mathcal{H}}$, we can also use squared residual norms to detect atypical observations.

As in Section 2, we will derive the Fisher–consistency of this Sieves–approach for observations generated by an elliptically distributed random object, which is a natural generalization of elliptical random vectors to an infinite–dimensional setup. The following definition was given in Bali and Boente (2009).

Definition 3.1. *Let X be a random element in a separable Hilbert space $\mathcal{H}$. We will say that X has an elliptical distribution with parameters $\mu_{\mathcal{H}} \in \mathcal{H}$ and $\boldsymbol{\Gamma} : \mathcal{H} \to \mathcal{H}$, where $\boldsymbol{\Gamma}$ is a self–adjoint, positive semi–definite and compact operator, if and only if for any linear and bounded operator $A : \mathcal{H} \to \mathbb{R}^d$ we have that the vector $A\,X$ has a $d-$variate elliptical distribution with location parameter $A\,\mu_{\mathcal{H}}$, shape matrix $A\,\Gamma\,A^*$ and characteristic generator ϕ, that is, $A\,X \sim \mathcal{E}_d(A\,\mu_{\mathcal{H}}, A\,\Gamma\,A^*, \phi)$ where $A^* : \mathbb{R}^d \to \mathcal{H}$ denotes the adjoint operator of A. We write $X \sim \mathcal{E}(\mu_{\mathcal{H}}, \Gamma, \phi)$.*

To study the Fisher–consistency of our Sieves approach, we need to introduce some notation in order to define the corresponding functional. Let $\otimes$ denote the tensor product in $\mathcal{H}$, i.e., for any two elements u, $v \in \mathcal{H}$ the operator $u \otimes v : \mathcal{H} \to \mathcal{H}$ is defined as $(u \otimes v)w = \langle v, w \rangle u$ for $w \in \mathcal{H}$. To simplify the presentation, assume that the location parameter $\mu_{\mathcal{H}}$ equals 0. Let $\mathbf{x} \in \mathbb{R}^p$ be the random vector defined by $\mathbf{x} = A_p X$ with $A_p : \mathcal{H} \to \mathbb{R}^p$ defined by

$$A_p = \sum_{j=1}^{p} \mathbf{e}_j \otimes \delta_j \, , \tag{7}$$

where $\mathbf{e}_j$, $1 \le j \le p$, denote the elements of the canonical basis of $\mathbb{R}^p$ and $\{\delta_i\}_{i \ge 1}$ is a fixed orthonormal basis of $\mathcal{H}$. In other words, $A_p X$ consists of the p coefficients of X on the basis $\delta_1, \ldots, \delta_p$. For a matrix $\mathbf{B} \in \mathbb{R}^{p \times q}$ with $\mathbf{B}^{\mathrm{T}}\mathbf{B} = \mathbf{I}_q$, let $\mathcal{L}_{\mathbf{B}}$ denote the linear space spanned by its columns. As in Section 2, define the objective function

$$\Psi_p(\mathcal{L}_{\mathbf{B}}) = \sum_{j=1}^{p} \sigma^2_{j,\mathcal{L}_{\mathbf{B}}} \, , \tag{8}$$

where $\sigma_{j,\mathcal{L}_{\mathbf{B}}} = \sigma_{\mathrm{R}}(F_j(\mathcal{L}_{\mathbf{B}}))$ and $F_j(\mathcal{L}_{\mathbf{B}})$ denotes the distribution of the jth coordinate $r_j(\mathcal{L}_{\mathbf{B}})$ of the vector of residuals $\mathbf{r}(\mathcal{L}_{\mathbf{B}}) = \mathbf{x} - \pi(\mathbf{x}, \mathcal{L}_{\mathbf{B}}) = (\mathbf{I} - \mathbf{B}\mathbf{B}^{\mathrm{T}})\mathbf{x} = (r_1(\mathcal{L}_{\mathbf{B}}), \ldots, r_p(\mathcal{L}_{\mathbf{B}}))^{\mathrm{T}}$. The subscript p in the function defined in (8) emphasizes the fact that we have transformed the random object X into the $p-$dimensional random vector $\mathbf{x}$. Let $\mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(q)}$ denote the columns of the matrix $\mathbf{B}$ and let $\phi_\ell(\mathbf{B}) \in \mathcal{H}$ be given by

$$\phi_\ell(\mathbf{B}) \;=\; \sum_{j=1}^{p} b_{\ell j} \delta_j \;=\; \Big(\sum_{j=1}^{p} \delta_j \otimes \mathbf{e}_j \Big) \, \mathbf{b}^{(\ell)} \, , \quad 1 \le \ell \le q \, . \tag{9}$$

We denote as $\mathcal{H}_{\mathbf{B}}$ the linear space spanned by the orthonormal elements $\phi_1(\mathbf{B}), \ldots, \phi_q(\mathbf{B})$.

In what follows, and without loss of generality, we will assume that $\mu_{\mathcal{H}} = 0$. Let X be an elliptical random element $X \sim \mathcal{E}(0, \Gamma, \phi)$, with Γ the self–adjoint, positive semi–definite and compact scale operator. Consider the spectral value decomposition of the scale operator $\Gamma = \sum_{j=1}^{\infty} \lambda_j \, \phi_j \otimes \phi_j$, where λ_j denotes the jth largest eigenvalue with associated eigenfunction ϕ_j, $j \ge 1$. The next proposition shows that, as p tends to infinity, the lowest value of $\Psi_p(\mathcal{L}_{\mathbf{B}})$ converges to $\sum_{j \ge q+1} \lambda_j$, the trace of the operator

$(\mathbb{I}_{\mathcal{H}} - P)\Gamma(\mathbb{I}_{\mathcal{H}} - P)^*$ where $P = \sum_{j=1}^{q} \phi_j \otimes \phi_j$, and $\mathbb{I}_{\mathcal{H}}$ is the identity operator in $\mathcal{H}$. This is the infinite–dimensional counterpart of the classical optimal property of principal components for random vectors. Together with Proposition A1 in Boente *et al.* (2012), the following result shows that the proposed estimators are Fisher–consistent for elliptically distributed random elements on a separable Hilbert space $\mathcal{H}$.

Proposition 3.1. *Let $X \sim \mathcal{E}(0, \Gamma, \phi)$ be an elliptically distributed random element on a separable Hilbert space $\mathcal{H}$ with location 0 and positive semi–definite, self–adjoint and compact scale operator Γ. Let $\lambda_1 \geq \lambda_2 \geq \dots$ be the eigenvalues of Γ with associated eigenfunctions ϕ_j, $j \geq 1$. If $\sum_{j\geq 1} \lambda_j < \infty$ and $\lambda_q > \lambda_{q+1}$, then*

$$\lim_{p\to\infty} \min_{\mathbf{B}\in\mathbb{R}^{p\times q}, \mathbf{B}^{\mathrm{T}}\mathbf{B}=\mathbf{I}_q} \Psi_p(\mathcal{L}_{\mathbf{B}}) = \sum_{j\geq q+1} \lambda_j \,. \tag{10}$$

Let $\mathbf{B}_{0,p}$ be the minimizer of (8) *over $\{\mathbf{B} \in \mathbb{R}^{p\times q}, \mathbf{B}^{\mathrm{T}}\mathbf{B} = \mathbf{I}_q\}$. Then, as $p \to \infty$, the sequence of linear spaces $\mathcal{H}_{\mathbf{B}_{0,p}}$ converges to the linear space spanned by the eigenfunctions $\phi_1, \dots, \phi_q$ associated with the q largest eigenvalues of Γ.*

3.1 Algorithm for functional data

In this section we give details on how to compute our $S-$estimators for functional principal components. The basic idea is given above and consists of applying the algorithm of Section 2.1 to the coordinates of the observed data on a sufficiently rich orthonormal basis of the Hilbert space, and then transforming back the result to the original variables.

To fix ideas, consider the case where the data consist of functions X_i, $1 \leq i \leq n$, observed at points $t_1, \dots, t_m$. We approximate the L^2 inner product with a Riemann sum over the grid of points: $\langle \alpha \,, \beta \rangle_{\mathcal{H}} = \int \alpha(t)\beta(t)\, dt \approx \sum_{\ell=2}^{m} \alpha(t_\ell)\beta(t_\ell)(t_\ell - t_{\ell-1})$. Let $\nu_1, \dots, \nu_p$ be a B-spline basis. We orthonormalize $\nu_1, \dots, \nu_p$ using the approximated inner product to obtain orthonormal elements $\delta_1, \dots, \delta_p$. Let $\mathbf{\Delta} \in \mathbb{R}^{m\times p}$ be the matrix of the functions δ_j evaluated at the points t_i: $\mathbf{\Delta} = (\boldsymbol{\delta}_1, \, \boldsymbol{\delta}_2 \dots \boldsymbol{\delta}_p)$, where

$\boldsymbol{\delta}_j = (\delta_j(t_1), \delta_j(t_2), \ldots, \delta_j(t_m))^{\mathrm{T}}$. Then, if $\mathbf{X} \in \mathbb{R}^{n\times m}$ is the matrix of observed trajectories (one in each row), the coordinates of each X_i on each element δ_j of the spline basis is denoted as $\widetilde{\mathbf{x}}_{i,j} = \sum_{\ell=2}^{m} X_i(t_\ell)\delta_j(t_\ell)(t_\ell - t_{\ell-1}) \approx \langle X_i\,, \delta_j\rangle_{\mathcal{H}}$, $1 \le i \le n$, $1 \le j \le p$. We now apply the algorithm given in Section 2.1 to the "data" matrix $\widetilde{\mathbf{X}} \in \mathbb{R}^{n\times p}$ of the coordinates of our observations on the B–spline basis. We obtain the centre vector $\widetilde{\boldsymbol{\mu}} \in \mathbb{R}^p$, an orthonormal basis $\widetilde{\mathbf{B}} \in \mathbb{R}^{p\times q}$ of the best q dimensional subspace and the matrix of scores $\widetilde{\mathbf{A}} \in \mathbb{R}^{n\times q}$. The matrix $\widehat{\widetilde{\mathbf{X}}} = \mathbb{I}_n\, \widetilde{\boldsymbol{\mu}}^{\mathrm{T}} + \widetilde{\mathbf{A}}\widetilde{\mathbf{B}}^{\mathrm{T}}$ provides the q dimensional approximation to our functional data written in the B–splines basis. Finally, we express our solution in the original variables $\widehat{\mathbf{X}} = \widehat{\widetilde{\mathbf{X}}}\,\boldsymbol{\Delta}^{\mathrm{T}}$. Note that $\widehat{\mathbf{X}} = \mathbb{I}_n\,(\boldsymbol{\Delta}\,\widetilde{\boldsymbol{\mu}})^{\mathrm{T}} + \widetilde{\mathbf{A}}(\boldsymbol{\Delta}\,\widetilde{\mathbf{B}})^{\mathrm{T}}$. In other words, $\boldsymbol{\Delta}\,\widetilde{\boldsymbol{\mu}} \in \mathbb{R}^m$ is the vector of the centre function $\hat{\mu}_{\mathcal{H}}$ evaluated at the points $t_1, \ldots, t_m$, and $\boldsymbol{\Delta}\,\widetilde{\mathbf{B}} \in \mathbb{R}^{m\times q}$ is the matrix of q orthonormal functions $\widehat{\phi}_\ell$ spanning the best lower approximation space in $\mathcal{H}$, evaluated on the same points.

4 Outlier detection

An important use of robust estimators for multivariate data is the detection of potentially atypical observations in the data, see, for example, Rousseeuw and Van Zomeren (1990), Becker and Gather (2001), Pison and van Aelst (2004) and Hardin and Rocke (2005). Unfortunately, these approaches to outlier detection do not extend naturally to the functional case.

Alternatively, one can consider the PCA residuals as indicators of outlyingness. Given a sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ in $\mathbb{R}^p$ and the estimated subspace $\widehat{\mathcal{L}} = \mathcal{L}_{\widehat{\mathbf{B}}}$ in (4), one can construct the corresponding "best $q-$dimensional" approximations $\widehat{\mathbf{x}}_i = \widehat{\boldsymbol{\mu}} + \pi(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}, \mathcal{L}_{\widehat{\mathbf{B}}}) = \widehat{\boldsymbol{\mu}} + \widehat{\mathbf{B}}\widehat{\mathbf{B}}^{\mathrm{T}}(\mathbf{x}_i - \widehat{\boldsymbol{\mu}})$, $1 \le i \le n$. We expect outlying or otherwise atypical observations to be poorly fitted and thus to have a relatively large residual $R_i = \|\mathbf{r}_i(\mathcal{L}_{\widehat{\mathbf{B}}})\|_{\mathbb{R}^p} = \|(\mathbf{I} - \widehat{\mathbf{B}}\widehat{\mathbf{B}}^{\mathrm{T}})(\mathbf{x}_i - \widehat{\boldsymbol{\mu}})\|_{\mathbb{R}^p}$, $1 \le i \le n$. Exploring the norm of these residuals sometimes provides sufficient information to detect abnormal points in the data. It is worth noticing that the distribution of the residuals squared norm R_i^2 is unknown, but typically skewed to the right because they are bounded by 0 from below. Following the approach of Hubert and Vandervieren (2008), we propose to flag an observation as atypical if its squared residual norm exceeds the upper whisker of an skewed-adjusted boxplot.

Another way to use principal components to look for potential outliers considers the scores of each point on the estimated principal eigenvectors. The solution to (4) provides an estimated basis $\widehat{\mathbf{b}}^{(j)}$, $1 \leq j \leq q$ (the columns of $\widehat{\mathbf{B}}$) for the optimal $q-$dimensional linear space spanned by the first q eigenvectors, but the $\widehat{\mathbf{b}}^{(j)}$'s themselves need not be estimates of the principal directions. However, we can use an approach similar to "projection pursuit" to sequentially search for vectors in $\widehat{\mathcal{L}}_{\widehat{\mathbf{B}}}$ that maximize a robust scale estimate of the corresponding projections of the data. Specifically, for each $\boldsymbol{\gamma} \in \widehat{\mathcal{L}}_{\widehat{\mathbf{B}}}$, let $F_n[\boldsymbol{\gamma}]$ be the empirical distribution of the projected observations $\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{x}_1, \ldots, \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{x}_n$, and $\sigma_{\mathrm{R}}(F_n[\boldsymbol{\gamma}])$ the corresponding scale estimator. The estimated first principal direction is obtained maximizing $\sigma_{\mathrm{R}}(F_n[\boldsymbol{\gamma}])$ over unitary vectors in $\widehat{\mathcal{L}}_{\widehat{\mathbf{B}}}$. Subsequent principal directions are similarly computed with the additional condition of being orthogonal to the previous ones. The scores of each observation on the estimated principal directions can be used to screen for atypical data points.

Both of these last two approaches have natural counterparts for functional data and can be used with the estimators defined in Section 3. Hyndman and Shang (2010) define two detection rules based on the scores of a robust two–dimensional fit and compare them with a residuals-based PCA procedure introduced in Hyndman and Ullah (2007). Our simulation study in Section 5 includes these methods as well those based on functional depth proposed by Febrero *et al.* (2007, 2008).

As in the finite–dimensional case, to find potential outliers one may consider looking for curves X_i that are poorly predicted by the $S-$estimator using the squared prediction errors $R^2_{i,\mathcal{H}} = \|X_i - \widehat{X}_i\|^2_{\mathcal{H}}$, $i = 1, \ldots, n$. As in the finite–dimensional case, the distribution of these prediction residuals is unknown and difficult to estimate. Hyndman and Ullah (2007) proposed to use a normal approximation to the residual squared norm, which they called the integrated squared error, to define a threshold. Our approach is more data analytic and does not depend on the underlying distribution of the process even if we always have in mind that the uncontaminated process has a symmetric distribution. For that reason, we mimic the proposal given in the finite–dimensional case and to decide whether an observation may be flagged as a potential outlier, we used the adjusted boxplot of Hubert and Vandervieren (2008) on the residuals $R^2_{i,\mathcal{H}}$, identifying as an atypical observation a value exceeding the upper whisker of the adjusted boxplot. We use this approach in the examples and in our simulation study discussed below.

5 Simulation

In this section we present the results of a a simulation study performed to investigate the finite–sample properties of our robust sieve proposal. In all cases, we generated 500 samples of size $n = 70$ where each trajectory was observed at $m = 100$ equidistant points in the interval $[0, 1]$. We used a cubic $B-$spline basis of dimension $p = 50$, which is sufficiently rich to represent the data well. This choice represents a realistic situation where the sample size is similar to the dimension of the problem. Other reasonable choices for the dimension of the spline basis (even with $n < p$) yielded very similar results and lead to the same conclusions in our numerical experiments.

5.1 Simulation settings

The following three different models constructed from finite– and infinite–range processes were used to generate the data. In two of them we included a relatively small proportion of measurement errors, as is usual in many applications.

Model 1 This model corresponds to the case where most of the curves follow a smooth trajectory, but some of them may display sudden vertical jumps at a few time points. In this setup, the non–contaminated observations $X_i \sim X$, $1 \leq i \leq n$, with $X(t_s) \sim 10 + \mu(t_s) + \xi_1\phi_1(t_s) + \xi_2\phi_2(t_s) + z_s,\ s = 1, \ldots, 100,$ where the additive errors z_s are i.i.d $N(0,1)$, the scores $\xi_1 \sim N(0,25/4)$, $\xi_2 \sim N(0,1/4)$, ξ_1 and ξ_2 are independent and independent of z_s. The mean function is $\mu(t)$=5+10 sin$(4\pi t)$ exp$(-2t)$+5 sin$(\pi t/3)$+2 cos$(\pi t/2)$ and $\phi_1(t) = \sqrt{2}\cos(2\pi t)$ and $\phi_2(t) = \sqrt{2}\sin(2\pi t)$ correspond to the Fourier basis.

We also generated contaminated trajectories $X_i^{(c)}$ as realizations of the process $X^{(c)}$ defined by $X^{(c)}(t_s) = X(t_s) + V\,Y(t_s)$, $s = 1,\ldots,100$, where $V \sim Bi(1,\epsilon_1)$ is independent of X and Y, $Y(t_s) = W_s\,\widetilde{z}_s$ with $W_s \sim Bi(1,\epsilon_2)$, $\widetilde{z}_s \sim N(\mu^{(c)}, 0.01)$, W_s and $\widetilde{z}_s$ are all independent. In other words, a trajectory is contaminated with probability ϵ_1, and at any point t_s the contaminated function is shifted with probability ϵ_2. The shift is random but tightly distributed around the constant $\mu^{(c)} = 30$. Samples without outliers correspond to $\epsilon_1 = 0$. To investigate the influence of different outlier configurations our estimator, we considered the settings: $\epsilon_1 = 0.10$ and $\epsilon_1 = 0.20$, with $\epsilon_2 = 0.30$ in both cases.

Model 2 This situation corresponds to a similar case as in Model 1, but with some curves starting on a different trajectory that joins smoothly with the one that most curves follow. In this case, non–contaminated observations $X_i \sim X$ were generated as $X(t_s) \sim 150 - 2\mu(t_s) + \xi_1\phi_1(t_s) + \xi_2\phi_2(t_s) + z_s$, $s = 1,\ldots,100$, where z_s, ξ_1, ξ_2, μ, ϕ_1 and ϕ_2 are as in the previous model. However, contaminated trajectories are only perturbed in a specific part of their range. The atypical observations satisfy $X_i^{(c)} \sim X^{(c)}$ where $X^{(c)}(t_s) = X(t_s) + V\,Y(t_s)$ for $t_s < 0.4$ and $X^{(c)}(t_s) = X(t_s)$ for $t_s \geq 0.4$, where $V \sim Bi(1,\epsilon_1)$ is independent of X and Y, $Y(t_s) = W_s\widetilde{z}_s$ with $W_s \sim Bi(1,\epsilon_2)$, $\widetilde{z}_s \sim N(\mu^{(c)}(t_s), 0.01)$, with $\mu^{(c)}(t_s) = -5 - 2\mu(ts)$, and W_s and $\widetilde{z}_s$ are all independent. In this model we used $\epsilon_1 = 0.10$ and $\epsilon_1 = 0.20$, and in both cases we set $\epsilon_2 = 0.90$.

Model 3 This setting corresponds to functions that follow an infinite-rank stochastic process. Contamination is present in terms of short sudden vertical shifts. Curves were generated from a Gaussian process with covariance kernel $\gamma_X(s,t) = 10\min(s,t)$. The eigenfunctions of the covariance operator equal $\phi_j(t) = \sqrt{2}\sin\left((2j-1)(\pi/2)t\right)$, $j \geq 1$, with associated eigenvalues $\lambda_j = 10\left(2/\left[d(2j-1)\pi\right]\right)^2$. As in Sawant *et al.* (2012), the contaminated observations $X_i^{(c)}$ are defined as $X_i^{(c)}(s) = X_i(s) + V_i\, D_i\, M\, \mathbb{I}_{\{T_i<s<T_i+\ell\}}$, where $V_i \sim Bi(1,\epsilon)$, $\mathbb{P}(D_i = 1) = \mathbb{P}(D_i = -1) = 1/2$, $T_i \sim \mathcal{U}(0, 1-\ell)$, $\ell < 1/2$ and V_i, X_i, D_i and T_i are independent. We choose $\ell = 1/15$, $M = 30$ and $\epsilon = 0.1$ and 0.2.

5.2 The estimators

We computed the classical principal components estimator (LS) as well as the robust one defined in (2), using an $M-$scale estimator, with function ρ_c in Tukey's bi–square family with tuning constants $c = 1.54764$ and $b = 0.50$. We also considered the choice $c = 3.0$ and $b = 0.2426$, which we expect to yield more efficiency. The robust estimators are labelled as S (1.5) and S (3) in the tables. As mentioned in Section 2.1, after obtaining the robust $q-$dimensional linear space, we orthonormalize its basis and compute the scores $\widehat{\mathbf{a}}_i$ as the corresponding orthogonal projections. We also computed the sieve projection–pursuit approach proposed in Bali *et al.* (2011), which is called "PP" in our Tables below. For comparison purposes, we have also calculated the mean squared prediction errors obtained with the true best $q-$dimensional linear space for uncontaminated data. This benchmark is indicated as "True" in all Tables.

Since trajectories following Models 1 and 2 were generated using a two–dimensional scatter operator (i.e. the underlying process had only 2 non–zero eigenvalues) plus measurement errors, we used $q = 1$ with our estimator. For Model 3, we used $q = 4$, which results in 95% of explained variance.

5.3 Simulation results

To summarize the results of our simulation study, for each replication we consider mean squared prediction errors in the original space, i.e., based on $\|X_i - \widehat{X}_i\|^2_{\mathcal{H}}$. The conclusions that can be reached using the finite–dimensional residuals squared prediction error $\|\mathbf{x}_i - \widehat{\mathbf{x}}_i\|^2_{\mathbb{R}^p}$ are the same as those discussed below, and hence are not reported here. We report the average mean squared error for outlying and non–outlying trajectories separately, as a way to quantify how the procedures fit the bulk of the data. More specifically, let $\gamma_i = 1$ when X_i is an outlier and $\gamma_i = 0$ otherwise, then

$$\mathrm{PE}_{\mathcal{H},\mathrm{OUT}} = \frac{1}{n}\sum_{i=1}^{n} \gamma_i \|X_i - \widehat{X}_i\|_{\mathcal{H}}^2 \quad \text{and} \quad \mathrm{PE}_{\mathcal{H},\mathrm{CLEAN}} = \frac{1}{n}\sum_{i=1}^{n} (1-\gamma_i) \|X_i - \widehat{X}_i\|_{\mathcal{H}}^2 \,. \tag{11}$$

Note that the total prediction error equals $\mathrm{PE}_{\mathcal{H}} = (1/n)\sum_{i=1}^{n} \|X_i - \widehat{X}_i\|_{\mathcal{H}}^2 = \mathrm{PE}_{\mathcal{H},\mathrm{OUT}} + \mathrm{PE}_{\mathcal{H},\mathrm{CLEAN}}$. We also report the mean PE over contaminated and clean trajectories separately:

$$\overline{\mathrm{PE}}_{\mathcal{H},\mathrm{OUT}} = \frac{\sum_{i=1}^{n} \gamma_i \|X_i - \widehat{X}_i\|_{\mathcal{H}}^2}{\sum_{i=1}^{n} \gamma_i} \,, \tag{12}$$

and

$$\overline{\mathrm{PE}}_{\mathcal{H},\mathrm{CLEAN}} = \frac{\sum_{i=1}^{n} (1-\gamma_i) \|X_i - \widehat{X}_i\|_{\mathcal{H}}^2}{\sum_{i=1}^{n} (1-\gamma_i)} \,. \tag{13}$$

We also compute the prediction squared errors of the actual best lower dimensional predictions $\widehat{X}_i^0$, obtained with the first q true eigenfunctions (recall that we used $q = 1$ in Models 1 and 2, and $q = 4$ in Model 3). The results for this "estimator" are tabulated in the row labelled "True". The averages over the 500 replications of $\mathrm{PE}_{\mathcal{H},\mathrm{OUT}}$, $\mathrm{PE}_{\mathcal{H},\mathrm{CLEAN}}$, $\overline{\mathrm{PE}}_{\mathcal{H},\mathrm{OUT}}$ and $\overline{\mathrm{PE}}_{\mathcal{H},\mathrm{CLEAN}}$ are labelled "Out", "Clean" , "$\overline{\text{Out}}$" and "$\overline{\text{Clean}}$", respectively.

	$\epsilon_1 = \epsilon_2 = 0.00$	$\epsilon_1 = 0.10$				$\epsilon_1 = 0.20$			
Method	Clean	Out	Clean	$\overline{\text{Out}}$	$\overline{\text{Clean}}$	Out	Clean	$\overline{\text{Out}}$	$\overline{\text{Clean}}$
True	1.266	26.930	1.138	269.316	1.264	53.780	1.013	269.685	1.265
LS	1.246	18.961	5.065	193.372	5.679	37.429	5.682	187.461	7.104
S (3)	1.253	26.922	1.126	269.245	1.252	53.425	1.081	268.453	1.361
S (1.5)	1.308	26.872	1.270	268.937	1.417	53.241	1.464	267.400	1.850
PP	1.335	26.536	1.335	265.791	1.486	51.845	1.559	260.972	1.972

Table 2: Mean prediction errors over 500 replications for Model 1.

As expected, when no outliers are present all procedures are comparable, with a small loss for the robust procedures. The $S-$estimator with $c = 3$ had the second smallest mean

	$\epsilon_1 = \epsilon_2 = 0.00$	$\epsilon_1 = 0.10$				$\epsilon_1 = 0.20$			
Method	Clean	Out	Clean	$\overline{\text{Out}}$	$\overline{\text{Clean}}$	Out	Clean	$\overline{\text{Out}}$	$\overline{\text{Clean}}$
True	1.359	10.063	1.222	100.589	1.358	20.054	1.087	100.598	1.358
LS	1.339	1.597	4.032	19.528	4.512	1.840	4.482	9.505	5.610
S (3)	1.346	9.839	1.380	99.230	1.541	12.427	2.357	69.919	3.035
S (1.5)	1.401	9.638	2.047	97.207	2.296	17.916	2.891	90.648	3.645
PP	1.428	8.922	1.427	90.696	1.589	14.865	1.618	76.535	2.039

Table 3: Mean prediction errors over 500 replications for Model 2.

	$\epsilon = 0.00$	$\epsilon = 0.10$				$\epsilon = 0.20$			
Method	Clean	Out	Clean	$\overline{\text{Out}}$	$\overline{\text{Clean}}$	Out	Clean	$\overline{\text{Out}}$	$\overline{\text{Clean}}$
True	0.304	4.411	0.274	44.163	0.304	8.842	0.243	44.088	0.304
LS	0.285	2.074	0.660	18.457	0.736	5.599	0.711	27.363	0.893
S (3)	0.301	4.412	0.269	44.148	0.299	8.846	0.237	44.113	0.297
S (1.5)	0.354	4.465	0.318	44.674	0.354	8.931	0.284	44.535	0.355
PP	0.385	4.439	0.355	44.397	0.394	8.913	0.321	44.430	0.402

Table 4: Mean prediction errors over 500 replications for Model 3.

squared prediction error, after the LS. When samples were contaminated, the classical procedure based on least squares tries to compromise between outlying and non–outlying trajectories and this is reflected on the values of $\text{PE}_{\mathcal{H},\text{OUT}}$ and $\text{PE}_{\mathcal{H},\text{CLEAN}}$ in (11), and also on the average prediction error of the contaminated and non–contaminated trajectories in (12) and (13) appearing in the columns labelled "Out" and "Clean". With contaminated samples the $S-$estimator had the best performance overall. Its mean squared prediction was closest to the "True" one, and it also provided better fits to the non–contaminated samples (and worse predictions for the contaminated trajectories). This last observation can be seen comparing the columns labelled "$\overline{\text{Out}}$" and "$\overline{\text{Clean}}$", which correspond to (12) and (13), respectively. The only case when the sieves projection–pursuit estimator performed slightly better than the $S-$estimator is for Model 1 with $\epsilon_1 = 0.10$ and $\epsilon_2 =$

0.60. The advantage of the $S-$estimator was more notable in all the other cases of Model 1, Model 2 and Model 3.

We also compared the performance of different outlier detection methods for functional data. As described in Section 4, we used the squared prediction errors $R^2_{i,\mathcal{H}} = \|X_i - \widehat{X}_i\|^2_{\mathcal{H}}$, $i = 1, \ldots, n$, to find curves X_i that are poorly predicted. Those with squared prediction errors exceeding the upper whisker of the adjusted boxplot will be flagged as outliers. We used the same approach with predictors $\widehat{X}_i$ obtained using the other estimators mentioned before.

In addition, we included other outlier–detection methods for functional data that appeared in the literature. We considered the functional high-density region and the functional bagplots of Hyndman and Shang (2010) with a 99% coverage, denoted HDR and BAG, respectively, as well as the integrated squared error method defined in Hyndman and Ullah (2007), denoted HU. The first two methods are based on the scores of a two–dimensional robust projection–pursuit fit. To keep the comparison fair, for HU we chose a $q-$dimensional robust fit with $q = 1$ under Models 1 and 2 and $q = 4$ under Model 3. Furthermore, we also compared our detection rule with the proposals based on a likelihood-ratio-type statistic given in Febrero *et al.* (2007) and on the modal depth, using both trimmed and weighted bootstrap estimates for the threshold as proposed in Febrero *et al.* (2008). These methods are denoted LRT, DTR and DWE, respectively. These detection rules are implemented in the `R` package `rainbow`.

For each model and each outlier detection method, in Tables 5 to 7 we report the average sensitivity and specificity over the 500 samples. Sensitivity is the proportion of actual outliers that are correctly flagged as such, while specificity is the proportion of non–outlying curves correctly identified as not atypical. An ideal method will simultaneously maintain high sensitivity and specificity.

For Model 1, we note that DRT, DWE and HU identify too many curves as outliers (resulting in a high sensitivity but low specificity). On the other hand, LRT, HDR and BAG consistently miss most of the outliers (low sensitivity), as does LS when the proportion of outliers is 20%. Using prediction residuals based on $S-$ and the projection pursuit estimators offers the best overall performance. When the data follow Model 2, LS, HDR, LRT, DTR and DWE fail to detect most of the outliers, as does BAG for $\epsilon = 0.20$. Again, HU flags too many curves as outlying. The relatively low specificity of DTR and DWE (and to some extent BAG) seems to indicate that the few observations flagged as outliers are not the truly atypical ones. Once again the approach based on S- and projection pursuit estimators works best. Note that although the S(1.5) appears to miss around half of the outliers for $\epsilon_1 = 0.20$, those flagged as atypical are correctly identified. The results for Model 3 are very similar to those for Model 1. Overall, for the three scenarios considered here, the clear best method to detect functional outliers is to use the squared prediction residuals based on a robust principal components estimator.

					Sensitivity					
ϵ_1	LS	PP	S(3)	S(1.5)	HDR	BAG	LRT	DTR	DWE	HU
0.10	0.914	1.000	1.000	0.998	0.155	0.597	0.305	1.000	1.000	1.000
0.20	0.295	0.835	0.856	0.833	0.074	0.224	0.018	1.000	1.000	1.000
					Specificity					
0.00	0.982	0.982	0.981	0.982	0.986	0.983	0.978	0.802	0.802	0.782
0.10	0.999	0.997	0.996	0.997	0.999	0.982	1.000	0.839	0.839	0.792
0.20	1.000	1.000	1.000	1.000	0.999	0.989	1.000	0.897	0.897	0.808

Table 5: Average sensitivity and specificity over 500 random samples following Model 1

6 Examples

6.1 Ground level Ozone concentrations

These data contains hourly average measurements of ground level ozone (O3) concentration from a monitoring station in Richmond, BC, Canada. Ozone at ground level is a

					Sensitivity					
ϵ_1	LS	PP	S(3)	S(1.5)	HDR	BAG	LRT	DTR	DWE	HU
0.10	0.178	0.996	0.979	0.915	0.135	0.774	0.059	0.350	0.353	1.000
0.20	0.020	0.708	0.637	0.474	0.053	0.079	0.005	0.239	0.239	1.000
					Specificity					
0.00	0.980	0.980	0.980	0.980	0.986	0.982	0.978	0.803	0.803	0.782
0.10	0.996	0.997	0.997	0.997	0.997	0.958	1.000	0.817	0.817	0.774
0.20	0.994	1.000	0.997	1.000	0.994	0.988	0.999	0.815	0.815	0.770

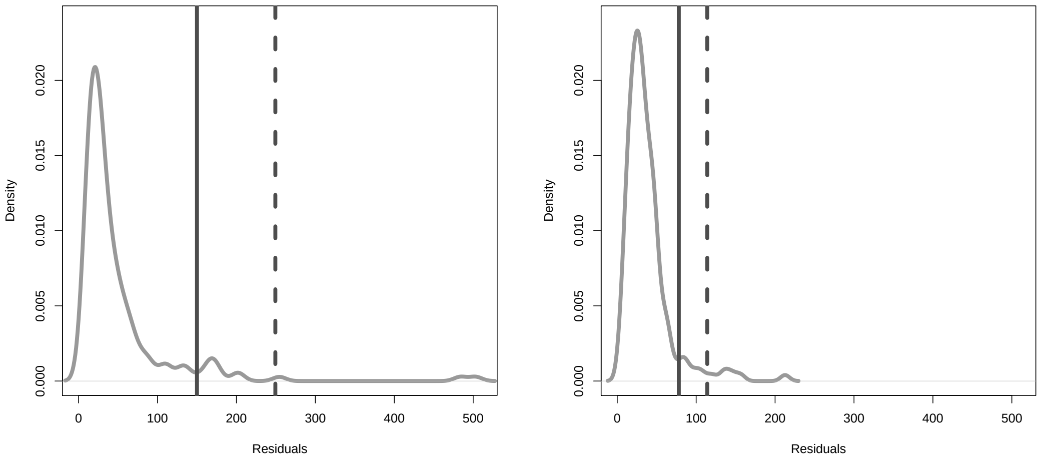
Table 6: Average sensitivity and specificity over 500 random samples following Model 2

					Sensitivity					
ϵ_1	LS	PP	S(3)	S(1.5)	HDR	BAG	LRT	DTR	DWE	HU
0.10	0.936	1.000	1.000	1.000	0.148	0.489	0.124	0.982	0.988	1.000
0.20	0.603	0.848	0.850	0.848	0.071	0.418	0.063	0.922	0.977	1.000
					Specificity					
0.00	0.987	0.986	0.987	0.987	0.986	0.983	0.990	0.804	0.804	0.849
0.10	0.998	0.997	0.998	0.998	0.999	0.988	1.000	0.838	0.837	0.869
0.20	1.000	1.000	1.000	1.000	0.999	0.991	1.000	0.863	0.886	0.896

Table 7: Average sensitivity and specificity over 500 random samples following Model 3

serious air pollutant and its presence typically peaks in summer months. We focus on the month of August, and obtained data for the years 2004 to 2012. We have 176 days with hourly average O3 measurements. Our purpose is to identify days in which the temporal pattern of O3 concentration appears different from the others. Based on the strong pattern observed in the data, we consider $1-$dimensional approximations. We use an $S-$estimator with tuning constant $c = 3$ applying the approach described in Section 3 with a cubic $B-$spline basis of dimension $p = 10$. To find potentially outlying curves we use as threshold the upper whisker of the adjusted boxplot of Hubert and Vandervieren (2008) applied to the squared prediction errors using the LS and S-estimators. Figure

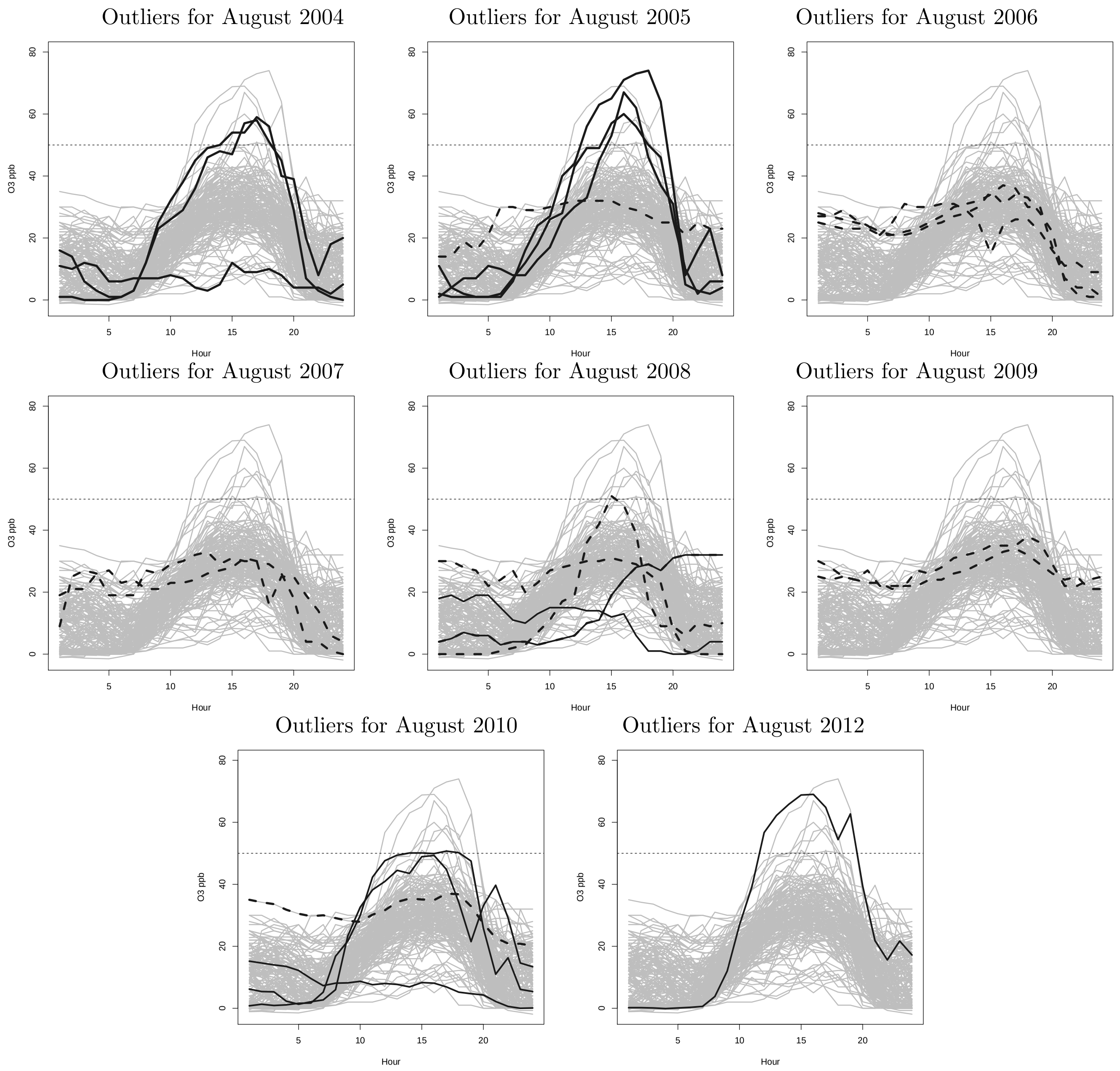
2 contains the estimated density of the L^2 norm of the residuals for each of the 176 curves when we compute predictions using our S-estimators (panel a) and the classical LS ones (panel b). The dashed line in Figure 2 corresponds to the threshold suggested by `adjbox()`. While there are a few extreme outliers at the right tail of each plot, both plots also show a relatively heavy tail that suggests the presence of moderate outliers. The solid line indicates approximately the beginning of this heavy tail, and is the cut-off used in our analysis.



(a) Residuals using a robust best fit (b) Residuals using classical PCA

Figure 2: Estimated density of the squared prediction errors for the ground level ozone data with (a) the S-estimator and (b) the classical one. The dashed line corresponds to the threshold suggested by `adjbox()` while the solid one indicates the beginning of a relatively heavy tail

To make the visualization of the results easier, each panel in Figure 3 shows the observations detected as outliers on one year, both by the robust estimator (solid lines) and the classical approach (dashed lines). The thin gray lines in the background show all the available observations, and are included as a visual reference, while the light dashed horizontal line at 50 ppb is the current maximum recommended level. We see that the robust fit identifies as outliers all of the days with relatively high peaks of O3 concentration, but also some days that exhibit a “flat” profile.



Outliers for August 2004

Outliers for August 2005

Outliers for August 2006

Outliers for August 2007

Outliers for August 2008

Outliers for August 2009

Outliers for August 2010

Outliers for August 2012

Figure 3: Hourly mean concentration (in ppb) of ground level ozone in Richmond, BC, Canada. Thin gray lines show all the available data. Solid lines correspond to potential outliers identified by the robust estimator, while those identified by the classical analysis are displayed with dashed lines.

Since ground level ozone is produced by the reaction between sunlight and other compounds in the air, we use temperature data to verify whether the potential outliers identified above correspond to atypical days. Figure 4 shows maximum daily temperature for the months of August between 2004 and 2012 together with the daily amount of rain. Days for which O3 data is not available are indicated with white circles. A day identified as having an atypical O3 profile by the robust fit is marked with a large solid circle. Potential outliers identified by the classical approach are indicated with a solid triangle. We see that the outliers identified by the robust fit correspond to days with either a very high or low temperature. Furthermore, outlying days with a “flat” O3 profile are those with a low maximum temperature, while days with a sharp O3 peak correspond to particularly hot days. On the other hand, days flagged as possible outliers by LS generally do not show any pattern with respect to temperature. This analysis shows that the robust method is able to identify potential outliers that correspond to extreme values of an unobserved but closely associated meteorological variable (temperature). In other words, the robust method is able to uncover outliers that correspond to actual atypical days.

6.2 Mortality data

In this example we explore human mortality data, available on–line from the Human Mortality Database (Human Mortality Database, 2013). We restrict our attention to death rates per age group for men in France. For each year, we use the logarithm of the death rate of people between the ages of 0 and 99. Panel (a) in Figure 5 shows the mortality curves for the years between 1816 to 2010. Dark lines correspond to years after 1945. We observe a clear difference in the pattern of male mortality curves in France before and after the Second World War. This phenomenon is sometimes attributed to technological advances and quality of life changes in Europe after 1945. We also note that there is a $3-$year transitional period (1946–1948) in which the mortality curves lie

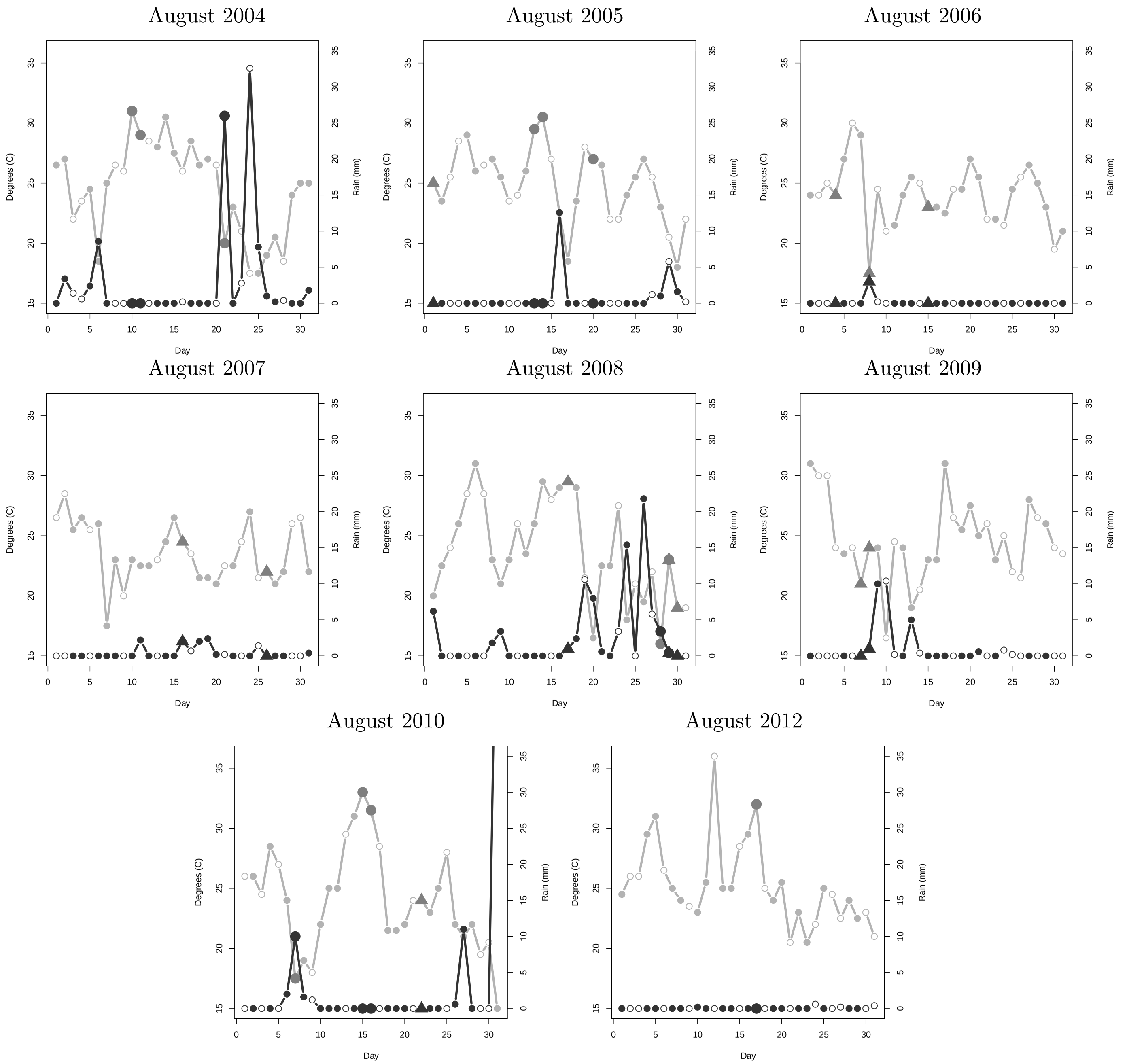
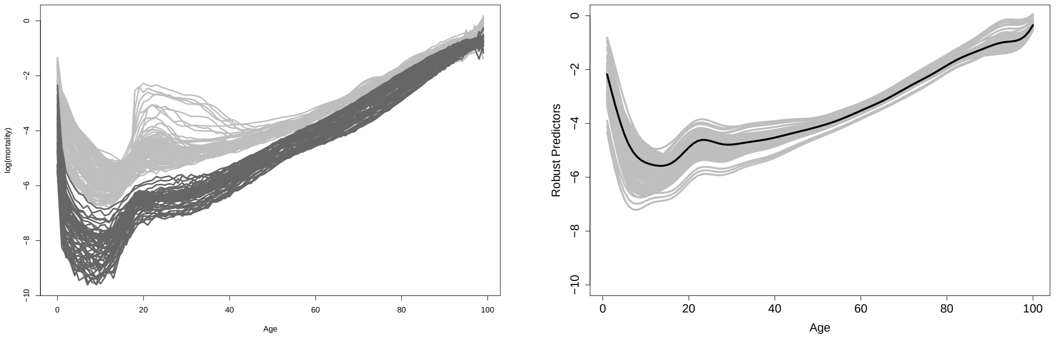


Figure 4: Maximum daily temperature profile (in black) and rain levels (in gray) for the month of August. Days with an atypical O3 profile as flagged by the robust method are indicated with large solid circles. Those identified as outlying by the classical approach are marked with a large triangle.



(a) Mortality data for the years 1816–2010 (b) Robust predictions for the years 1816–1948

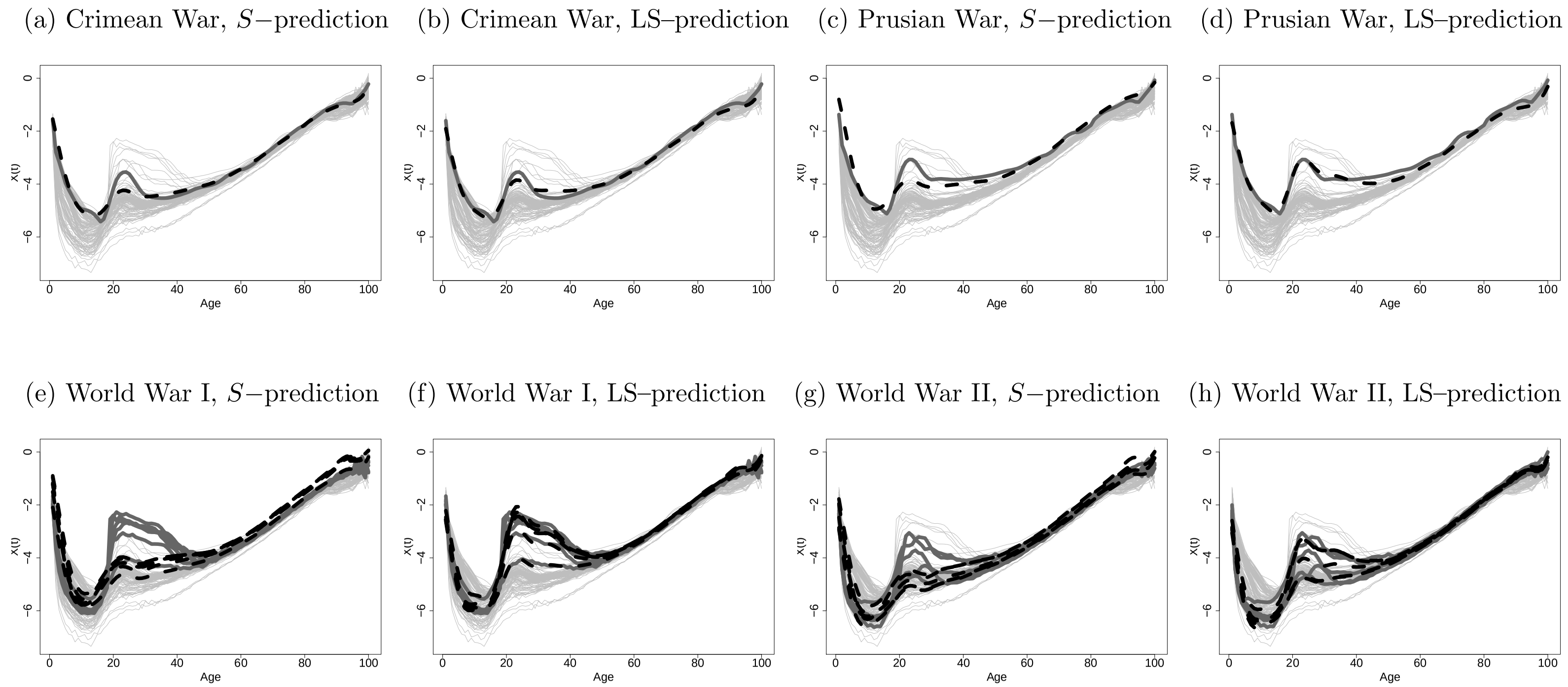
Figure 5: Mortality data. Panel (a) contains the curves for the years 1816 to 2010. Darker lines correspond to years after 1945. Panel (b) depicts the predicted trajectories corresponding to the $2-$dimensional subspace that best approximates the curves before the post–war years (1816–1948), estimated using our $S-$estimator. The black line is the estimated central curve.

between the two main groups. In this analysis we focus on the period 1816–1948, that includes the pre–war and the "transition" periods. The purpose of this analysis is to detect years in which the pattern of mortality across age groups is noticeably different from the majority of curves in the data. We computed an $S-$estimator with tuning constant $c = 3$ to find the best $2-$dimensional subspace approximating these curves. We used the approach described in Section 3 with a cubic $B-$spline basis of dimension $p = 20$. A two–dimensional fit was also considered in Hyndman and Ullah (2007) and Hyndman and Shang (2010). Figure 5 (b) contains the estimated central curve plotted over the original data, and also over the robustly predicted curves.

As described in Section 4, we looked for outlying curves by means of the adjusted boxplot of Hubert and Vandervieren (2008) on the squared prediction errors using the LS and $S-$estimators. With the robust fit we identified the following years as atypical: 1855, 1871, 1914–1919, 1940, and 1942–1948, while the LS fit only identifies the periods 1914–1915 and 1943–1948. It is interesting to note that in 1855 France was involved in the Crimean War, and in 1871 in the Prussian War. The period 1914 to 1919 corresponds to World War I and the Spanish Flu epidemic. France falls to German occupation in 1940 and after a relatively calm year in 1941, sees more action in the period 1942 to 1944. Figure 6 contains the curves corresponding to these four events (the Crimean and Prussian Wars, and the 2 World Wars), along with the predictions resulting from the S and LS estimators. Note that the predicted curves based on the robust estimator do not fit well the mortality profiles for these atypical years, which allows us to detect them as outlying. On the other hand, it is interesting to note that the LS estimator is not able to detect the Crimean and Prussian Wars, neither the early World War II casualties in France (1940 and 1942). Both estimators properly identify the post–war years as atypical.

7 Concluding Remarks

In this paper, we propose a robust estimator for the subspace spanned by the first q principal components. We show that our method is consistent and can be used in general settings, including functional data applications. In this case, our method works well when the observations can be well represented in an sufficiently rich but arbitrary basis. Moreover, the resulting robust predictions can be used to detect atypical observations in the data. This is confirmed in our simulation study, where this outlier detection method compares very favourably to other proposals in the literature. Our estimators are defined via a non-convex optimization problem which is difficult to solve. As it is done for similar problems arising in other contexts (robust linear regression and multivariate location and scatter estimators, for example) we use first order conditions to derive an iterative re-weighted least squares-type algorithm. Extensive numerical experiments show that this algorithm provides estimators with good statistical properties. It would be interesting, but beyond the scope of this work, to study whether a convex relaxation of the optimization problem (2) can provide a more scalable algorithm with comparable robustness and statistical properties.



(a) Crimean War, S−prediction (b) Crimean War, LS–prediction (c) Prusian War, S−prediction (d) Prusian War, LS–prediction

(e) World War I, S−prediction (f) World War I, LS–prediction (g) World War II, S−prediction (h) World War II, LS–prediction

Figure 6: Years 1855 and 1871 were detected as outliers by the robust estimator, during the Crimean and Prusian War, respectively (plots a) to d)). Curves for the years 1914–1919 (plots e) and f)) and 1940, 1942–1945 (plots g) and h) row). The curves of mortality rates for these periods, corresponding to both World Wars, were identified as outliers by the robust estimator. Solid thick gray lines correspond to the observed curve, while the black dashed lines correspond to the predicted trajectory.

References

Ainslie, B. and Steyn, D. G. (2007). “Spatio–temporal trends in episodic ozone pollution in the Lower Fraser Valley, British Columbia, in relation to mesoscale atmospheric circulation patterns and emissions”. *Journal of Applied Meteorology and Climatology*, **46**:10, 1631-1644.

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.

Bali, L., Boente, G., Tyler, D. and Wang, J. L. (2011). “Robust functional principal components: a projection-pursuit approach”. *Annals of Statistics*, **39**, 2852-2882.

Becker, C. and Gather, U. (1999). “The masking breakdown point of multivariate outlier identification rules”. *Journal of the American Statistical Association*, **94**, 947-955.

Becker, C. and Gather, U. (2001). “The largest nonidentifiable outliers: A comparison of multivariate simultaneous outliers identification rules”. *Computational Statistics and Data Analysis*, **36**, 119-127.

Boente, G. (1987). “Asymptotic theory for robust principal components”. *Journal of Multivariate Analysis*, **21**, 67-78.

Boente, G., Salibian–Barrera, M. and Tyler, D. (2012). “A characterization of elliptical distributions and some optimality properties of principal components for functional data”. Technical report. Available at `http://www.stat.ubc.ca/~matias/Property_FPCA.pdf`

Campbell, N.A. (1980). “Robust procedures in multivariate analysis I: robust covariance estimation”. *Applied Statistics*, **29**, 231-237.

Candès, E.J., Li, X., Ma, Y. and Wright, J. (2011) “Robust principal component analysis?”. *Journal of the ACM (JACM)*, **58**(3), 1-37. DOI: 10.1145/1970392.1970395

Chandrasekaran, V., Sanghavi, S., Parrilo, P. and Willsky, A.S. (2011). “Rank-sparsity incoherence for matrix decomposition”. *SIAM Journal of Optimization*, **21**(2), 572-596.

Croux, C. and Haesbroeck, G. (2000). “Principal component analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies”. *Biometrika*, **87**, 603-618.

Croux, C., Filzmoser, P., Pison, G. and Rousseeuw, P.J. (2003). “Fitting multiplicative models by robust alternating regressions”. *Statistics and Computing*, **13**, 23-36.

Croux, C. and Ruiz-Gazen, A. (1996). “A fast algorithm for robust principal components based on projection pursuit”. In *Compstat: Proceedings in computational statistics*, ed. A. Prat, Heidelberg: Physica-Verlag, pp. 211-216.

Croux, C. and Ruiz-Gazen, A. (2005). “High-breakdown estimators for principal components: the projection-pursuit approach revisited”. *Journal of Multivariate Analysis*, **95**, 206-226.

Cui, H., He, X. and Ng, K. W. (2003). “Asymptotic Distribution of Principal Components Based on Robust Dispersions”. *Biometrika*, **90**, 953-966.

De la Torre, F. and Black, M. J. (2001). “Robust principal components analysis for computer vision”. In *Proceedings of the 8th International Conference on Computer Vision*, **1**, pp.362-369. doi: 10.1109/ICCV.2001.937541

Devlin, S.J., Gnanadesikan, R. and Kettenring, J.R. (1981). “Robust estimation of dispersion matrices and principal components”. *Journal of the American Statistical Association*, **76**, 354-362.

Dunford, N. and Schwartz, J. (1963). *Linear Operators. II: Spectral Theory, Selfadjoint operators in Hilbert spaces.* Interscience, New York.

Febrero, M., Galeano, P. and Gonzalez-Manteiga, W. (2007). “A functional analysis of NOx levels: location and scale estimation and outlier detection”, *Computational Statistics*, **22**(3), 411-427.

Febrero, M., Galeano, P. and Gonzalez-Manteiga, W. (2008) “Outlier detection in functional data by depth measures, with application to identify abnormal NOx levels”, *Environmetrics*, **19**(4), 331-345.

Gervini, D. (2008). “Robust functional estimation using the spatial median and spherical principal components”. *Biometrika*, **95**, 587-600.

Huber P.J. and Ronchetti E.M. (2009). *Robust Statistics.* Wiley, New York, 2nd edition.

Hubert, M. and Vandervieren, E. (2008). “An adjusted boxplot for skewed distributions ”. *Computational Statistics & Data Analysis*, **52**(12), 5186-5201.

Hubert, M., Rousseeuw, P.J. and Vanden Branden, K. (2005). “ROBPCA: a new approach to robust principal component analysis”. *Technometrics*, **47**, 64-79.

Hubert, M., Rousseeuw, P. and Verboven, S. (2002). “A fast method for robust principal components with applications to chemometrics”. *Chemometrics and intelligent laboratory systems*, **60**, 101-111.

Human Mortality Database. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available at www.mortality.org or www.humanmortality.de (data downloaded on 8 Feb 2013).

Hyndman, R. J. and Ullah, S. (2007). "Robust forecasting of mortality and fertility rates: A functional data approach". *Computational Statistics and Data Analysis*, **51**, 4942-4956.

Hyndman, R. J. and Shang, H.L. (2010) "Rainbow plots, bagplots, and boxplots for functional data". *Journal of Computational and Graphical Statistics*, **19**(1), 29-45.

Lerman, G., McCoy, M., Tropp, J.A. and Zhang, T. (2012) "Robust computation of linear models, or how to find a needle in a haystack". Unpublished manuscript. arXiv:1202.4044v1 [cs.IT].

Li, G. and Chen, Z. (1985). "Projection pursuit approach to robust dispersion matrices and principal components: Primary theory and Monte Carlo". *Journal of the American Statistical Association*, **80**, 759-766.

Liu, L., Hawkins, D., Ghosh, S. and Young, S. (2003). "Robust singular value decomposition analysis of microarray data". In *Proceedings of the National Academy of Sciences*, **100**, 13167-13172.

Locantore, N., Marron, J.S., Simpson, D.G., Tripoli, N., Zhang, J.T., and Cohen, K.L. (1999). "Robust principal components for functional data". *Test*, **8**, 1-28.

Maronna, R. (2005). "Principal components and orthogonal regression based on robust scales". *Technometrics*, **47**, 264-273.

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

Maronna, R. and Yohai, V. (2008). "Robust lower–rank approximation of data matrices with element–wise contamination". *Technometrics*, **50**, 295-304.

McCoy, M. and Tropp, J.A. (2011). "Two proposals for robust PCA using semidefinite programming". *Electronic Journal of Statistics*, **5**, 1123-1160.

Naga, R. and Antille, G. (1990). "Stability of robust and non–robust principal component analysis". *Computational Statistics and Data Analysis*, **10**, 169-174.

Pison, G. and van Aelst, S. (2004). "Diagnostic plots for robust multivariate methods". *Journal of Computational and Graphical Statistics*, **13**, 1-20.

Rousseeuw, P.J., and Van Zomeren, B.C. (1990). "Unmasking multivariate outliers and leverage points". *Journal of the American Statistical Association*, **85**, 633-651.

Rousseeuw, P.J., and van Driessen, K. (1999). "A fast algorithm for the Minimum Covariance Determinant Estimator". *Technometrics*, **41**, 212-223.

Osborn, J. (1975). "Spectral approximation for compact operators". *Mathematics of Computation*, **29**, 712-725.

Salibian-Barrera, M. and Yohai, V.J. (2006). "A fast algorithm for S-regression estimates". *Journal of Computational and Graphical Statistics*, **15**, 414-427.

Sawant, P., Billor, N. and Shin, H. (2012). "Functional outlier detection with robust functional principal component analysis". *Computational Statistics*, **27**, 83-102.

Seber, G. (1984). *Multivariate Observations.* Wiley, New York.

Sillman, S. (1993). "Tropospheric Ozone: The debate over control strategies". *Annual Review of Energy and the Environment*, **18**: 31-56.

U.S. Environmental Protection Agency. (2008) *National Air Quality: Status and Trends through 2007.* Office of Air Quality Planning and Standards, Air Quality Assessment Division, Research Triangle Park, North Carolina. Report EPA-454/R-08-006. Available on-line at http://www.regulations.gov/#!documentDetail;D=EPA-HQ-OAR-2009-0171-11674.

Verboon, P., and Heiser, W. J. (1994). "Resistant lower–rank approximation of matrices by Iterative majorization". *Computational Statistics and Data Analysis*, **18**, 457-467.

Xu, H., Caramanis, C. and Sanghavi, S. (2012). "Robust PCA via outlier pursuit". *IEEE Transactions on Information Theory*, **58**(5), 3047-3064.

Zhang, T. and Lerman, G. (2014). "A novel M-estimator for robust PCA". *Journal of Machine Learning Research*, **15**, 749-808.

A Appendix

PROOF OF PROPOSITION 2.1. Note that since $\mathbf{x} \sim \mathcal{E}_p(\mathbf{0}, \boldsymbol{\Sigma}, \phi)$, $\mathbf{z} = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$ is spherically distributed, so that all its components have the same distribution G. Without loss of generality, assume that $\sigma_{\mathrm{R}}(G) = 1$. Let $\mathcal{L}$ be a linear space of dimension q, with orthonormal basis $\mathbf{b}^{(1)}, \dots, \mathbf{b}^{(q)}$. If we arrange this basis as columns of a matrix $\mathbf{B} \in \mathbb{R}^{p\times q}$ we have that $\mathbf{r}(\mathcal{L}) = (r_1(\mathcal{L}), \dots, r_p(\mathcal{L}))^{\mathrm{T}} = \mathbf{x} - \pi(\mathbf{x}, \mathcal{L}) = (\mathbf{I} - \mathbf{B}\mathbf{B}^{\mathrm{T}})\mathbf{x} \sim \mathcal{E}_p(\mathbf{0}, \boldsymbol{\Sigma}_{\mathcal{L}}, \phi)$, with

$\boldsymbol{\Sigma}_{\mathcal{L}} = (\mathbf{I} - \mathbf{B}\mathbf{B}^{\mathrm{T}})\boldsymbol{\Sigma}(\mathbf{I} - \mathbf{B}\mathbf{B}^{\mathrm{T}})^{\mathrm{T}} = \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^{\mathrm{T}}$, with $\mathbf{C} = (\mathbf{I} - \mathbf{B}\mathbf{B}^{\mathrm{T}})$. Since $\mathbf{x} = \boldsymbol{\beta}\boldsymbol{\Lambda}^{1/2}\mathbf{z}$, we see that $\mathbf{r}(\mathcal{L})$ can be written as $\mathbf{C}\boldsymbol{\beta}\boldsymbol{\Lambda}^{1/2}\mathbf{z}$. Therefore, the characteristic function of $\mathbf{r}(\mathcal{L})$ is given by $\varphi_{\mathbf{r}(\mathcal{L})}(\mathbf{t}) = \varphi_{\mathbf{z}}(\boldsymbol{\Lambda}^{1/2}\boldsymbol{\beta}^{\mathrm{T}}\mathbf{C}^{\mathrm{T}}\mathbf{t}) = \phi(\mathbf{t}^{\mathrm{T}}\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^{\mathrm{T}}\mathbf{t})$, where ϕ denotes the generator of the characteristic function of $\mathbf{z}$. Hence, for the jth coordinate of the vector of residuals we have $\varphi_{r_j(\mathcal{L})}(t) = \varphi_{\mathbf{r}(\mathcal{L})}(t\mathbf{e}_j) = \phi(t^2\,\mathbf{e}_j^{\mathrm{T}}\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^{\mathrm{T}}\mathbf{e}_j) = \phi(t^2\,\mathbf{c}_j^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{c}_j) = \varphi_{z_1}(t^2\,\mathbf{c}_j^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{c}_j)$. It follows that $r_j(\mathcal{L}) \sim \xi_j z_1$ where $z_1 \sim G$ and $\xi_j^2 = \mathbf{c}_j^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{c}_j$. This implies that $\sigma^2_{j,\mathcal{L}} = \sigma^2_{\mathrm{R}}(F_j(\mathcal{L})) = \mathbf{c}_j^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{c}_j$. Hence, $\sum_{j=1}^{p}\sigma^2_{j,\mathcal{L}} = \sum_{j=1}^{p}\mathbf{c}_j^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{c}_j = \mathrm{tr}(\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^{\mathrm{T}})$. This last expression is minimized when $\mathbf{B} = (\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_q)$ (see Seber, 1984, Theorem 5.3) and the solution is unique since $\lambda_q > \lambda_{q+1}$. □

Proof of Proposition 2.2. Let $a_n = \sup_{\dim(\mathcal{L})=q} |\widehat{\Psi}_n(\mathcal{L}) - \Psi(\mathcal{L})|$ and note that $\widehat{\Psi}_n(\widehat{\mathcal{L}}) \leq \widehat{\Psi}_n(\mathcal{L}(P)) = \Psi(\mathcal{L}(P)) + a_n$ and similarly $\Psi(\mathcal{L}(P)) \leq \Psi(\widehat{\mathcal{L}}) \leq \widehat{\Psi}_n(\widehat{\mathcal{L}}) + a_n$. Hence $\widehat{\Psi}_n(\widehat{\mathcal{L}}) \geq \Psi(\mathcal{L}(P)) - a_n$ and we obtain $\widehat{\Psi}_n(\widehat{\mathcal{L}}) \xrightarrow[n\to\infty]{\text{a.s.}} \Psi(\mathcal{L}(P))$ and $\Psi(\widehat{\mathcal{L}}) \xrightarrow[n\to\infty]{\text{a.s.}} \Psi(\mathcal{L}(P))$. Standard arguments now imply the convergence of the linear spaces since $\mathcal{L}(P)$ is unique. Hence, one can choose an orthonormal basis of $\widehat{\mathcal{L}}$ converging with probability one to a basis of $\mathcal{L}(P)$. □

Proof of Proposition 3.1. To illustrate the main idea of the proof, we start with the (easy) case where the orthonormal basis $\{\delta_j\}$ is the basis ϕ_j of eigenfunctions of $\boldsymbol{\Gamma}$. Assume that $m = m_n$ is such that $m_n > q$ and $\{\phi_1, \dots, \phi_q\} \subset \{\delta_1, \delta_2, \dots, \delta_{m_n}\}$. Without loss of generality, assume that $\delta_j = \phi_j$, for $1 \leq j \leq q$ and that $\delta_j = \phi_{\ell_j}$ for $q+1 \leq j \leq m_n$ with $q < \ell_{q+1} < \cdots < \ell_{m_n}$. Then, $\mathbf{x} = AX \sim \mathcal{E}_p(\mathbf{0}, \boldsymbol{\Sigma}, \phi)$ where A is defined in (7) and $\boldsymbol{\Sigma} = A\Gamma A^* = \mathrm{diag}(\lambda_1, \dots, \lambda_q, \lambda_{\ell_{q+1}}, \dots, \lambda_{\ell_m})$ where $\lambda_q > \lambda_{\ell_{q+1}} > \cdots > \lambda_{\ell_m}$. Then, using Proposition 2.1, for any $\mathbf{B} \in \mathbb{R}^{m\times q}$ such that $\mathbf{B}^{\mathrm{T}}\mathbf{B} = \mathbf{I}_q$, we have $\Psi_m(\mathcal{L}_{\mathbf{B}}) = \sum_{j=1}^{m}\sigma^2_{j,\mathcal{L}_{\mathbf{B}}} \geq \sum_{j=1}^{m}\sigma^2_{j,\mathcal{L}_{\mathbf{B}_{0,m}}} = \sum_{s=q+1}^{m}\lambda_{\ell_s}$, where $\mathbf{B}_{0,m} = (\mathbf{e}_1, \dots, \mathbf{e}_q)$. Hence, using that $\lim_{m\to\infty}\sum_{s=q+1}^{m}\lambda_{\ell_s} = \sum_{s\geq q+1}\lambda_s = \mathrm{tr}(\Gamma) - \sum_{j=1}^{q}\lambda_j = \mathrm{tr}\left((\mathbb{I}_{\mathcal{H}} - P)\Gamma(\mathbb{I}_{\mathcal{H}} - P)^*\right)$. Note that, in this case, $\phi_j(\mathbf{B}_{0,m}) = \phi_j$, where $\phi_j(\mathbf{B})$ is defined in (9). Hence, $\mathcal{H}_{\mathbf{B}_{0,m}}$ is the linear space spanned by $\phi_1, \dots, \phi_q$, which shows Fisher–consistency.

Let us now consider the general situation. As before, we have $\mathbf{x} = AX \sim \mathcal{E}_p(\mathbf{0}, \boldsymbol{\Sigma}, \phi)$ where A is defined in (7) and $\boldsymbol{\Sigma} = A\Gamma A^*$. Recall that $A^* = \sum_{j=1}^{m}\delta_j \otimes \mathbf{e}_j$, so that for any $\mathbf{y} \in \mathbb{R}^m$, $A^*\mathbf{y} = \sum_{j=1}^m y_j \delta_j$. Let $\mathcal{H}_m$ be the linear subspace spanned by $\{\delta_1, \dots, \delta_m\}$ and $\Pi_m : \mathcal{H} \to \mathcal{H}_m$ be the projection operator over $\mathcal{H}_m$, that is, $\Pi_m = \sum_{j=1}^m \delta_j \otimes \delta_j$. We have that Π_m is self–adjoint and $\Pi_m \nu = \nu$ for $\nu \in \mathcal{H}_m$. Moreover, $\Pi_m \to \mathbb{I}_{\mathcal{H}}$ in the strong operator topology, where $\mathbb{I}_{\mathcal{H}}$ is the identity operator in $\mathcal{H}$, that is, $\Pi_m x \to x$ for any $x \in \mathcal{H}$, as $m \to \infty$. It follows that for any compact operator Υ, $\Pi_m \Upsilon \to \Upsilon$ as $m \to \infty$ in the norm operator topology.

It is easy to show that, if $\mathbf{u} \in \mathbb{R}^m$ is an eigenvector of $\mathbf{\Sigma}$ related to an eigenvalue α, then $\nu = A^*\mathbf{u}$ is an eigenfunction of $\Upsilon_m = \Pi_m \Gamma \Pi_m^*$ associated to α. Similarly, if ν is an eigenfunction of Υ_m with eigenvalue α, then $A\nu$ is an eigenvector of $\mathbf{\Sigma}$ with the same eigenvalue α. Hence, the $m-$largest eigenvalues of Υ_m are those of $\mathbf{\Sigma}$ with the relation among eigenvectors and eigenfunctions just described. Note that since the range of Υ_m is m, Υ_m has at most m non–null eigenvalues. Let $\mathbf{B}_{0,m} \in \mathbb{R}^{m \times q}$ be a matrix containing the eigenvectors of $\mathbf{\Sigma}$ related to its m largest eigenvalues as columns. In other words, $\mathbf{B}_{0,m} = (\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_q)$ where $\boldsymbol{\beta}_j$ is the eigenvector of $\mathbf{\Sigma}$ related to its jth largest eigenvalue denoted α_j. Then, $\alpha_j = \lambda_j(\Upsilon_m)$, where $\lambda_j(\Upsilon)$ denotes the jth largest eigenvalue of the operator Υ.

Using Proposition 2.1 we get that, for any $\mathbf{B} \in \mathbb{R}^{m \times q}$ such that $\mathbf{B}^{\mathrm{T}}\mathbf{B} = \mathbf{I}_q$, $\Psi_m(\mathcal{L}_{\mathbf{B}}) = \sum_{j=1}^m \sigma_{j,\mathcal{L}_{\mathbf{B}}}^2 \geq \sum_{j=1}^m \sigma_{j,\mathcal{L}_{\mathbf{B}_{0,m}}}^2 = \sum_{s=q+1}^m \alpha_s = \sum_{s=q+1}^m \lambda_s(\Upsilon_m)$, Noting that $\mathrm{tr}(\mathbf{\Sigma}) = \mathrm{tr}(\Upsilon_m) = \sum_{s=1}^m \lambda_s(\Upsilon_m)$, we obtain the bound $\Psi_m(\mathcal{L}_{\mathbf{B}}) \geq \Psi_m(\mathcal{L}_{\mathbf{B}_{0,m}}) = \mathrm{tr}(\Upsilon_m) - \sum_{s=1}^q \lambda_s(\Upsilon_m)$, that is

$$\min_{\mathbf{B} \in \mathbb{R}^{m \times q}, \mathbf{B}^{\mathrm{T}}\mathbf{B} = \mathbf{I}_q} \Psi_m(\mathcal{L}_{\mathbf{B}}) = \Psi_m(\mathcal{L}_{\mathbf{B}_{0,m}}) = \mathrm{tr}(\Upsilon_m) - \sum_{s=1}^q \lambda_s(\Upsilon_m) \,. \tag{A.1}$$

As noted above, we have $\|\Upsilon_m - \Gamma\| \to 0$ as $m \to \infty$. By the continuity of the eigenvalues with respect to the operators norm (see for instance, Osborn, 1975), we have that, for each fixed k, $\lambda_k(\Upsilon_m) \to \lambda_k(\Gamma) = \lambda_k$ as $m \to \infty$. Hence, $\lim_{m\to\infty} \sum_{s=1}^q \lambda_s(\Upsilon_m) = \sum_{s=1}^q \lambda_s$.

It remains to show that $\lim_{m\to\infty} \mathrm{tr}(\Upsilon_m) = \mathrm{tr}(\Gamma)$. First note that, Proposition A.1 in Boente *et al.* (2012) shows that $\lambda_k(\Upsilon_m) \leq \lambda_k$, hence $\mathrm{tr}(\Upsilon_m) = \sum_{s=1}^m \lambda_s(\Upsilon_m) \leq \sum_{s=1}^m \lambda_s \leq \mathrm{tr}(\Gamma)$. Therefore, we only have to show that, for any $\epsilon > 0$, there exists m_0 such that for $m \geq m_0$, we have $\mathrm{tr}(\Upsilon_m) \geq \mathrm{tr}(\Gamma) - \epsilon$. Since $\mathrm{tr}(\Gamma) < \infty$, there exists $N \in \mathbb{N}$ such that $N > q$ and $0 \leq \mathrm{tr}(\Gamma) - \sum_{j=1}^{N} \lambda_j < \epsilon/2$. Using that $\lim_{m\to\infty} \sum_{j=1}^{N} \lambda_j(\Upsilon) = \sum_{j=1}^{N} \lambda_j$, choose m_0 such that for $m \geq m_0$, $|\sum_{j=1}^{N} \lambda_j(\Upsilon_m) - \sum_{j=1}^{N} \lambda_j| \leq \epsilon/2$. Now, for $m \geq \max\{m_0, N\}$ we have $\mathrm{tr}(\Upsilon_m) = \sum_{j=1}^{m} \lambda_j(\Upsilon_m) \geq \sum_{j=1}^{N} \lambda_j(\Upsilon_m) \geq \sum_{j=1}^{N} \lambda_j - \epsilon/2 \geq \mathrm{tr}(\Gamma) - \epsilon$, as desired. Hence, from (A.1), we have that $\lim_{m\to\infty} \min_{\mathbf{B}\in\mathbb{R}^{m\times q}, \mathbf{B}^{\mathrm{T}}\mathbf{B}=\mathbf{I}_q} \Psi_m(\mathcal{L}_{\mathbf{B}}) = \lim_{m\to\infty} \Psi_m(\mathcal{L}_{\mathbf{B}_{0,m}}) = \mathrm{tr}(\Gamma) - \sum_{s=1}^{q} \lambda_s$, concluding the proof of (10). Finally, note that the linear space $\mathcal{H}_{\mathbf{B}_{0,m}}$ is spanned by $\phi_1(\mathbf{B}_{0,m}), \ldots,\ \phi_q(\mathbf{B}_{0,m})$, where $\phi_j(\mathbf{B}_{0,m}) = A^*\boldsymbol{\beta}_j$. Then, we have $\phi_j(\mathbf{B}_{0,m}) = \phi_j(\Upsilon_m)$. Using again that, $\|\Upsilon_m - \Gamma\| \to 0$ as $m \to \infty$ and the fact that $\lambda_q > \lambda_{q+1}$, we see that the linear space spanned by $\phi_1(\Upsilon_m), \ldots, \phi_q(\Upsilon_m)$ converges to that spanned by $\phi_1, \ldots, \phi_q$, (see for instance, Osborn 1975 or Dunford and Schwartz, 1963), concluding the proof. □