$S-$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 1, Buenos Aires 1428, Argentina (Tel./FAX: 54-11-45763335). 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 112-200801-00216 and 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)

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). Monitoring the evolution of ground level ozone is useful to evaluate its impact on population health, and to understand its dynamics. 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 for the average concentration of ground level ozone over a 1-hour period. This is indicated with a dark dashed horizontal line in the plot. The corresponding maximum acceptable level is 80 ppb. The pattern observed in these data corresponds with our expectations: ozone levels peak in the early afternoon, when pollution emitted during the first part of the day reacts with the sunlight at the time of its highest intensity. 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

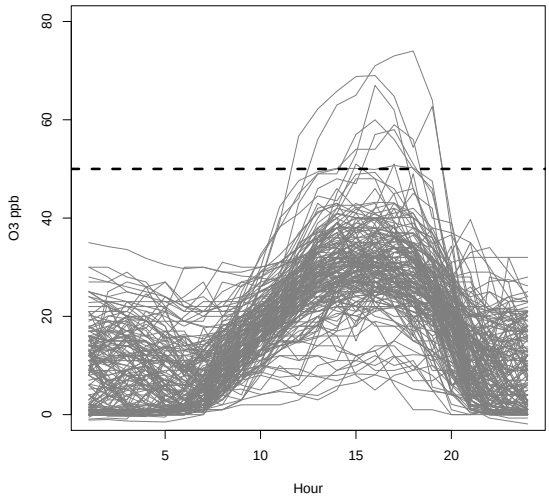


Figure 1: Hourly mean concentration (in ppb) of ground level ozone in Richmond, BC, Canada, for the month of August in years 2004 to 2012. Each line corresponds to one day. The darker dashed horizontal line at 50 ppb is the current maximum desired level set by the Canadian National Ambient Air Quality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

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 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 second finite–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. 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 while recovering the same best lower dimensional approximation properties mentioned above when second moments exist. Noting that the first principal components provide the solution when using the mean squared loss, several proposals for a robust estimator exist in the literature that exploit this characterization of PCA. They amount to replacing the squared residuals with a different loss function of them. Liu *et al.* (2003) used the absolute value of the residuals. 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.

In this paper, we propose to estimate the lower–dimensional linear space that minimizes a robust measure of dispersion of the resulting prediction errors. Our approach can also be extended to functional data. 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, 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. Our procedure provides Fisher–consistent estimators of the best lower–

dimensional subspace when applied to observations (either finite– or infinite–dimensional) with an elliptical distribution, even when second moments do not exist.

The rest of the paper is organized as follows. In Section 2, the problem of providing robust estimators for a $q-$dimensional approximation for euclidean data is described. Outlier detection and a description of the algorithm considered are described in Sections 2.1 and 2.2, respectively. Section 3 discusses some extensions of the estimators defined in Section 2 to accommodate functional data. Finally, Section 4 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 5, where the advantage of the proposed procedure to detect possible influential observations is illustrated. 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

$$\mathbf{B} = \begin{pmatrix} \mathbf{b}_1^{\mathrm{T}} \\ \vdots \\ \mathbf{b}_p^{\mathrm{T}} \end{pmatrix} = \left(\mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(q)}\right) ,$$

and 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 $\widehat{x}_{ij} = \mu_j + \mathbf{a}_i^{\mathrm{T}}\mathbf{b}_j$, where $\mathbf{b}_j$ is the jth row of $\mathbf{B}$ and $\mathbf{a}_i \in \mathbb{R}^q$ is the ith row of the matrix $\mathbf{A} \in \mathbb{R}^{n \times q}$

$$\mathbf{A} = \begin{pmatrix} \mathbf{a}_1^{\mathrm{T}} \\ \vdots \\ \mathbf{a}_n^{\mathrm{T}} \end{pmatrix} .$$

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 $r_{ij} = x_{ij} - \widehat{x}_{ij}$ 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 $\mathbf{a}_i$, $1 \le i \le n$, correspond to the scores of the sample on the orthonormal basis $\mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(q)}$.

Robust estimators motivated by (1) and the alternating regression algorithm have been considered by several authors. Croux *et al.* (2003) introduced a procedure, usually called RAR, that replaces the squared residuals with their absolute values and adds weights to protect against high–leverage points. Verboon and Heiser (1994) and De la Torre and Black (2001) proposed to replace r_{ij}^2 with $\rho(r_{ij}/\widehat{\sigma}_j)$, where $\widehat{\sigma}_j$ is a robust scale estimator of the jth column of the matrix of residuals. More recently, Maronna and Yohai (2008) replaced r_{ij}^2 with $\widehat{\sigma}_j^2\rho(r_{ij}/\widehat{\sigma}_j)$, so that when $\rho(u) = u^2$ their proposal reduces to the classical one.

In what follows we will consider a different way to robustify the criterion in (1) to obtain a $q-$dimensional approximation to the data. We will also show that our proposal is Fisher consistent for elliptically distributed random vectors, and the resulting estimators are consistent under standard regularity conditions. 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, 1981). 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 $\rho(0) = 0$. When ρ is bounded, it is assumed that $\sup_{u\in\mathbb{R}} \rho_c(u) = \|\rho\|_\infty = 1$. 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. To ensure consistency of the scale estimator when the data are normally distributed the tuning parameter c is chosen to satisfy $b = E_\Phi(\rho_c(Z))$ where Φ denotes the standard normal distribution. If, in addition, $b = \|\rho\|_\infty/2$ then, the $M-$estimate of scale in (3) has maximal breakdown point (50%). When $\rho(y) = \min(3y^2 - 3y^4 + y^6, 1)$, (Tukey's biweight function) the choices $c = 1.54764$ and $b = 1/2$ ensure that the estimator is Fisher–consistent at the normal distribution and has breakdown point 50%.

Since we are interested in estimating a $q-$dimensional subspace that approximates the data well, 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

$\mathbf{x}_i \in \mathbb{R}^p$, $1 \leq i \leq n$, 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:

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

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

In order to study the asymptotic properties of robust estimators it is sometimes convenient to think of them as functionals evaluated on 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 $\widehat{\sigma}(y_1, \ldots, y_n) = \sigma_{\mathrm{R}}(F_n)$.

For a random vector $\mathbf{x} \in \mathbb{R}^p$ 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

$$\mathcal{L}(P) = \underset{\dim(\mathcal{L})=q}{\operatorname{argmin}} \Psi(\mathcal{L}), \tag{5}$$

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

Recall that a random vector $\mathbf{x} \in \mathbb{R}^d$ is said to have a $p-$dimensional 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 $\mathbf{t} \in \mathbb{R}^p$ and some function $\phi : \mathbb{R} \to \mathbb{R}$. The function ϕ is called the generator of

the characteristic function, so that we write $\mathbf{x} \sim \mathcal{S}_d(\phi)$. As is well known, the elliptical distributions in $\mathbb{R}^p$ correspond to those distributions arising from affine transformations of spherically distributed random vectors in $\mathbb{R}^p$. 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 $\mathbf{x}$ has the simple form $\varphi_{\mathbf{x}}(\mathbf{t}) = \exp(i\mathbf{t}^{\mathrm{T}}\boldsymbol{\mu})\phi(\mathbf{t}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{t})$. For a fixed ϕ, the family of elliptical distributions $\mathcal{E}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \phi)$ forms a symmetric location–scatter class of distributions with location parameter $\boldsymbol{\mu}$ and symmetric positive semi–definite scatter parameter $\boldsymbol{\Sigma}$. If the first moment exists, then $\mathbb{E}(\mathbf{x}) = \boldsymbol{\mu}$, and if second moments exist, the covariance matrix of $\mathbf{x}$ is proportional to $\boldsymbol{\Sigma}$.

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 $\boldsymbol{\Lambda} = diag(\lambda_1, \ldots, \lambda_p)$, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$, 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 $\rho \leq \rho_1$. 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. Note the dependence of the scale estimate $\widehat{\sigma}_j$ on $(\mathbf{A}, \mathbf{B}, \boldsymbol{\mu})$. When an iterative procedure is used to find the minimizer of (4), one needs to update the p scale estimators $\widehat{\sigma}_j$ at each step of the algorithm. Analogous arguments to those considered

in Proposition 2.1 can be used to show that the corresponding functional will also be Fisher–consistent for elliptically distributed observations.

In the finite–dimensional setting, consistency of projection–pursuit principal component estimators was derived in Cui *et al.* (2003) requiring uniform convergence over the unit ball of the scale estimators of the projected data to the scale functional computed at the distribution of the projected random vector. This condition was generalized in Bali *et al.* (2011) to the functional case. A natural extension of this condition to the case $q > 1$ 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 Outlier detection

An important use of robust estimators is the detection of potentially atypical observations in the data. 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 fit and thus to have a relatively large residual $\|\mathbf{r}_i(\mathcal{L}_{\widehat{\mathbf{B}}})\|_{\mathbb{R}^p} = \|(\mathbf{I} - \widehat{\mathbf{B}}\widehat{\mathbf{B}}^{\mathrm{T}})(\mathbf{x}_i - \widehat{\boldsymbol{\mu}})\|_{\mathbb{R}^p}$, $1 \le i \le n$. Exploring the norm of these residuals sometimes provides sufficient information to detect abnormal points in the data.

A different way to use principal components to examine the data for potential outliers looks at the scores of each point on the estimated principal eigenvectors. Note that 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. More 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 $\widehat{\mathcal{L}}_{\widehat{\mathbf{B}}}$. Once an estimate for the first principal component is obtained, 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.

2.2 Algorithm for $S-$estimators

To compute the estimator defined in (4), we use an algorithm similar to the one used for linear regression $S-$estimators. Note that, although $S-$scale estimators are only defined implicitly, explicit first–order conditions can be obtained differentiating both sides of the equation (3). The resulting equations suggest re–weighted least squares iterations to find local extrema of the objective function (4) as a function of the vector $\boldsymbol{\mu}$ and the matrices $\mathbf{B}$ and $\mathbf{A}$. These iterations are started from a large number of random initial points and the critical point with the smallest value of the objective function is selected as the best solution.

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

$$\frac{1}{n} \sum_{i=1}^{n} \rho\left(\frac{x_{ij} - \mu_j - \mathbf{a}_i^{\mathrm{T}} \mathbf{b}_j}{\widehat{\sigma}_j}\right) \;=\; b\,,$$

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

$$h_j \;=\; \sum_{i=1}^{n} \rho'\left(\frac{r_{ij}}{\widehat{\sigma}_j}\right)\,\frac{r_{ij}}{\widehat{\sigma}_j}\,.$$

Similarly, the other first–order conditions are

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

and

$$\frac{\partial}{\partial \mu_\ell}\left(\sum_{j=1}^{p} \widehat{\sigma}_j^2\right) \;=\; \sum_{j=1}^{p} 2\,\widehat{\sigma}_j\,\frac{\partial\,\widehat{\sigma}_j}{\partial \mu_\ell} = \;-2\,\widehat{\sigma}_\ell\,h_\ell^{-1}\sum_{i=1}^{n} \rho'\left(\frac{r_{i\ell}}{\widehat{\sigma}_\ell}\right)\,, \qquad \ell = 1, \ldots, p\,.$$

Setting these to zero we obtain the following system of equations:

$$\begin{aligned}
\sum_{j=1}^{p} \widehat{\sigma}_j\,h_j^{-1}\,\rho'\left(\frac{r_{ij}}{\widehat{\sigma}_j}\right)\,\mathbf{b}_j \quad &= \quad \mathbf{0}\,, \quad 1 \le i \le n\,, \\
\sum_{i=1}^{n} \rho'\left(\frac{r_{is}}{\widehat{\sigma}_s}\right)\,\mathbf{a}_i \quad &= \quad \mathbf{0}\,, \quad 1 \le s \le p\,, \\
\sum_{i=1}^{n} \rho'\left(\frac{r_{i\ell}}{\widehat{\sigma}_\ell}\right) \quad &= \quad 0\,, \quad 1 \le \ell \le p\,.
\end{aligned}$$

These equations 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

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

This formulation suggests the usual iterative re–weighted least squares algorithm. Given initial estimates $\mathbf{b}_j^{(0)}$, $1 \le j \le p$ 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 $\mathbf{a}_i^{(1)}$, $\mathbf{b}_j^{(1)}$, $1 \le i \le n$, $1 \le j \le p$ 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, it is generally preferable to provide an orthonormal basis for the linear space, so 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. The "fitted values" corresponding to this last method will be called "scoring $S-$estimators".

`R` code implementing this algorithm can be downloaded from `http://www.stat.ubc.ca/~matias/soft.html`

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. Furthermore, Boente *et al.* (2012) recently extended this characterization to a best stochastically lower–dimensional approximation for elliptically distributed random elements on separable Hilbert spaces.

A potential practical complication arises when dealing with functional data. 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 $x_{i\,1}, \ldots, x_{i\,p_i}$ with $x_{ij} = X_i(t_{ij})$, $1 \leq j \leq p_i$. Depending on the characteristics of the grid of points t_{ij} where observations were obtained, one can employ different strategies to analyse these data.

The easiest situation is when observations were made at common design points. In this case we have $p_i = p$ and $t_{ij} = \tau_j$, for all $1 \leq i \leq n$ and $1 \leq j \leq p$. Defining $\mathbf{x}_i = (x_{i\,1}, \ldots, x_{i\,p})^{\mathrm{T}}$ 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 $1 \leq \ell \leq q$, $1 \leq j \leq p$, 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" $\pi(X_i, \widehat{\mathcal{L}})$, $1 \leq i \leq n$.

In many cases, however, trajectories are observed at different design points t_{ij}, $1 \leq j \leq p_i$, $1 \leq i \leq n$. 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. This method is sometimes called Sieves and we describe it here in the more general case where the observed random process

X takes values on an arbitrary separable Hilbert space $\mathcal{H}$. 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 $\mathcal{H}_{m_n}$ be the linear space spanned by $\delta_1, \ldots, \delta_{m_n}$. To simplify the notation we write $m = m_n$. Let $x_{ij} = \langle X_i, \delta_j \rangle_{\mathcal{H}}$ be the coeficient of the ith trajectory on the jth element of the basis, $1 \leq j \leq m$, and form the $m-$dimensional vector $\mathbf{x}_i = (x_{i1}, \ldots, x_{im})^{\mathrm{T}}$. 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 $\{t_{ij}\}_{1\leq j\leq p_i}$. We apply the procedure described in Section 2 to the multivariate observations $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^m$ 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 $\widehat{\boldsymbol{\mu}} = (\widehat{\mu}_1, \ldots, \widehat{\mu}_m)^{\mathrm{T}}$. It is now easy to find the corresponding approximation in the original space $\mathcal{H}$. The location parameter is $\widehat{\mu} = \sum_{j=1}^{m} \widehat{\mu}_j \delta_j$, and the associated $q-$dimensional basis in $\mathcal{H}$ is $\widehat{\phi}_\ell = \sum_{j=1}^{m} \widehat{b}_{\ell j} \delta_j / \| \sum_{j=1}^{m} \widehat{b}_{\ell j} \delta_j \|_{\mathcal{H}}$, for $1 \leq \ell \leq q$. Furthermore, the "fitted values" in $\mathcal{H}$ are $\widehat{X}_i = \widehat{\mu} + \sum_{\ell=1}^{q} \widehat{a}_{i\ell} \widehat{\phi}_\ell$. 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 $\mu \in \mathcal{H}$ and $\mathbf{\Gamma} : \mathcal{H} \to \mathcal{H}$, where $\mathbf{\Gamma}$ 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 $A\,\mu$, shape matrix $A\,\Gamma\,A^*$ and characteristic generator ϕ, that is, $A\,X \sim \mathcal{E}_d(A\,\mu, A\,\Gamma\,A^*, \phi)$ where $A^* : \mathbb{R}^d \to \mathcal{H}$ denotes the adjoint operator of A. We write $X \sim \mathcal{E}(\mu, \Gamma, \phi)$.

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 μ equals 0. Let $\mathbf{x} \in \mathbb{R}^m$ be the random vector defined by $\mathbf{x} = A_m X$ where $A_m : \mathcal{H} \to \mathbb{R}^m$ is defined by

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

and $\{\delta_i\}_{i\geq 1}$ is a fixed orthonormal basis of $\mathcal{H}$. In other words, $A_m X$ consists of the m coefficients of X on the basis $\delta_1, \ldots, \delta_m$. For a matrix $\mathbf{B} \in \mathbb{R}^{m\times q}$ 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

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

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 $\mathbf{r}(\mathcal{L}_{\mathbf{B}}) = \mathbf{x} - \pi(\mathbf{x}, \mathcal{L}_{\mathbf{B}}) = (\mathbf{I} - \mathbf{B}\mathbf{B}^{\mathrm{T}})\mathbf{x} = (r_1(\mathcal{L}_{\mathbf{B}}), \ldots, r_m(\mathcal{L}_{\mathbf{B}}))^{\mathrm{T}}$. The subscript m in the function defined in (8) emphasizes the fact that we have transformed the random object X into the $m-$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

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

The linear space spanned by the orthonormal elements $\phi_1(\mathbf{B}), \ldots, \phi_q(\mathbf{B})$ will be denoted by $\mathcal{H}_{\mathbf{B}}$.

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, $j \geq 1$. The next proposition shows that, as m tends to infinity, the lowest value of $\Psi_m(\mathcal{L}_{\mathbf{B}})$ converges to $\sum_{j \geq q+1} \lambda_j$, 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 these 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 $\lambda_1 \geq \lambda_2 \geq \ldots$ 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*

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

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

It is worth noting that, contrary to what happens in the finite–dimensional case, functional principal component analysis can not be easily defined directly through an alternating functional regression algorithm due to the ill–posed problem for estimating the regression operator.

4 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 $p_i = p = 100$ equidistant

points in the interval $[0, 1]$.

4.1 Simulation settings

The following four 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 In this setup, the non–contaminated observations $X_i \sim X$, $1 \leq i \leq n$, satisfy

$$X(t_s) \sim 10 + \mu(t_s) + \xi_1\phi_1(t_s) + \xi_2\phi_2(t_s) + z_s\,, \quad s = 1, \ldots, 100\,,$$

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 also considered the following four settings: $\epsilon_1 = 0.10$ with $\epsilon_2 = 0.30$ and $\epsilon_2 = 0.60$; $\epsilon_1 = 0.20$ with $\epsilon_2 = 0.30$, and $\epsilon_1 = 0.30$ with $\epsilon_2 = 0.30$.

Model 2 In this case, the non–contaminated observations $X_i \sim X$ were generated as

$$X(t_s) \sim 150 - 2\mu(t_s) + \xi_1\phi_1(t_s) + \xi_2\phi_2(t_s) + z_s\,, \qquad s = 1, \ldots, 100\,,$$

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 We also generate observations from a Wiener process contaminated. The uncontaminated observations are Gaussian 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. This contamination produces irregular trajectories introducing a peak contamination. We choose $\ell = 1/15$, $M = 30$ and $\epsilon = 0.1$ and 0.2.

Model 4 In this setting, the generated uncontaminated observations are Gaussian with covariance kernel $\gamma_X(s,t) = (1/2)(1/2)^{0.9|s-t|}$, which corresponds to the Ornstein Uhlenbeck process. The contamination is generated in the same way as in Model 3.

4.2 The estimators

For the four models considered, we have computed the best lower approximations using the functional data approach described in Section 3. We first project the observations taking as basis a cubic $B-$spline basis of dimension $m = 50$. With the finite–dimensional approximations, we then 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 suggested in Section 2.2, after obtaining the robust $q-$dimensional linear space, we selected an orthonormal basis of the found linear space, that is, we orthogonalize the set $\{\widehat{\mathbf{b}}^{(j)}\}$ and computed 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 Models 3 and 4, we used $q = 4$, which results in 95% of explained variance for Model 3. All our computations were performed using `R` code.

4.3 Simulation results

To summarize the results of our simulation study, for each replication we consider the mean squared prediction error

$$\mathrm{PE}_{\mathcal{H}} = \frac{1}{n}\sum_{i=1}^{n} \|X_i - \widehat{X}_i\|^2_{\mathcal{H}} \, . \tag{11}$$

The conclusions that can be reached using the finite–dimensional mean squared prediction error $(1/n)\sum_{i=1}^{n} \|\mathbf{x}_i - \widehat{\mathbf{x}}_i\|^2_{\mathbb{R}^p}$ are the same as those discussed below, and hence are not reported here.

We report the minimum, maximum and mean (Min, Max and Mean) as well as the median (Q2) and the 25% and 75% quantiles (Q1 and Q3) of these mean errors over the 500 replications. In addition, we also 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

$$\mathrm{PE}_{\mathcal{H},\mathrm{OUT}} = \frac{1}{n}\sum_{i=1}^{n}\gamma_i\|X_i - \widehat{X}_i\|^2_{\mathcal{H}} \text{ and } \mathrm{PE}_{\mathcal{H},\mathrm{CLEAN}} = \frac{1}{n}\sum_{i=1}^{n}(1-\gamma_i)\|X_i - \widehat{X}_i\|^2_{\mathcal{H}}\,. \quad (12)$$

We also report the mean PE over contaminated and clean trajectories separately:

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

and

$$\overline{\mathrm{PE}}_{\mathcal{H},\mathrm{CLEAN}} = \frac{\sum_{i=1}^{n}(1-\gamma_i)\|X_i - \widehat{X}_i\|^2_{\mathcal{H}}}{\sum_{i=1}^{n}(1-\gamma_i)}\,. \quad (14)$$

Finally, we also compute the prediction squared error 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 Models 3 and 4). The results for this "estimator" are tabulated in the row labelled "True".

Tables 1 to 5 report the results obtained under Model 1, Tables 6 to 8 correspond to Model 2, the results for Model 3 are in Tables 9 to 11 while those in Tables 12 to 14 are for the samples following Model 4. In all Tables, 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.

Method	Min	Q1	Q2	Mean	Q3	Max	Out	Clean	$\overline{\text{Out}}$	$\overline{\text{Clean}}$
True	1.119	1.232	1.264	1.266	1.299	1.430	0.000	1.266		1.266
LS	1.099	1.212	1.245	1.246	1.278	1.409	0.000	1.246		1.246
S (3)	1.105	1.219	1.252	1.253	1.285	1.417	0.000	1.253		1.253
S (1.5)	1.163	1.272	1.305	1.308	1.340	1.495	0.000	1.308		1.308
PP	1.199	1.290	1.327	1.335	1.375	1.549	0.000	1.335		1.335

Table 1: Model 1. No outliers.

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 squared prediction error, after the LS. When samples were contaminated, the classical

Method	Min	Q1	Q2	Mean	Q3	Max	Out	Clean	$\overline{\text{Out}}$	$\overline{\text{Clean}}$
True	1.265	20.810	27.370	28.070	34.380	70.440	26.930	1.138	269.316	1.264
LS	1.251	19.020	23.640	24.030	28.990	53.780	18.961	5.065	193.372	5.679
S (3)	1.258	20.780	27.340	28.050	34.360	70.430	26.922	1.126	269.245	1.252
S (1.5)	1.315	20.840	27.370	28.140	34.490	64.740	26.872	1.270	268.937	1.417
PP	1.307	20.760	27.210	27.870	34.050	68.990	26.536	1.335	265.791	1.486

Table 2: Model 1. $\epsilon_1 = 0.10$, $\epsilon_2 = 0.30$.

Method	Min	Q1	Q2	Mean	Q3	Max	Out	Clean	$\overline{\text{Out}}$	$\overline{\text{Clean}}$
True	1.265	40.570	53.570	55.110	69.060	127.800	53.975	1.138	539.206	1.264
LS	1.251	21.330	26.560	27.110	32.350	59.150	20.572	6.534	203.416	7.261
S (3)	1.258	40.750	53.680	53.150	64.790	99.870	51.043	2.109	520.484	2.388
S (1.5)	1.315	38.710	46.480	47.700	55.470	113.900	44.329	3.366	462.582	3.792
PP	1.307	40.260	52.920	54.050	67.420	120.800	52.558	1.494	527.444	1.668

Table 3: Model 1. $\epsilon_1 = 0.10$, $\epsilon_2 = 0.60$.

Method	Min	Q1	Q2	Mean	Q3	Max	Out	Clean	$\overline{\text{Out}}$	$\overline{\text{Clean}}$
True	17.910	46.050	53.970	54.790	62.850	105.700	53.780	1.013	269.685	1.265
LS	17.000	36.860	42.750	43.110	49.060	75.390	37.429	5.682	187.461	7.104
S (3)	17.870	46.060	53.900	54.510	63.080	91.530	53.425	1.081	268.453	1.361
S (1.5)	17.910	46.370	53.960	54.700	62.680	94.990	53.241	1.464	267.400	1.850
PP	18.210	45.420	52.990	53.400	61.510	89.560	51.845	1.559	260.972	1.972

Table 4: Model 1. $\epsilon_1 = 0.20$, $\epsilon_2 = 0.30$.

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 (12), and also on the average prediction error of the contaminated and non–contaminated trajectories in (13) and (14) appearing in the columns labelled "Out" and "Clean". With contaminated samples the $S-$estimator had the best performance overall. Its mean squared prediction

Method	Min	Q1	Q2	Mean	Q3	Max	Out	Clean	$\overline{\text{Out}}$	$\overline{\text{Clean}}$
True	36.530	71.260	81.010	81.290	90.600	135.400	80.399	0.888	269.717	1.265
LS	30.310	54.610	61.550	61.660	68.240	95.160	56.593	5.069	189.685	7.214
S (3)	36.400	70.430	78.280	76.750	84.030	103.600	75.067	1.685	254.757	2.503
S (1.5)	36.570	70.380	79.770	79.990	88.680	131.700	78.196	1.794	263.041	2.600
PP	36.030	68.180	76.550	76.060	84.870	106.900	73.853	2.206	249.538	3.222

Table 5: Model 1 - $\epsilon_1 = 0.30$, $\epsilon_2 = 0.30$

Method	Min	Q1	Q2	Mean	Q3	Max	Out	Clean	$\overline{\text{Out}}$	$\overline{\text{Clean}}$
True	1.234	1.325	1.355	1.359	1.391	1.530	0.000	1.359		1.359
LS	1.215	1.307	1.335	1.339	1.371	1.497	0.000	1.339		1.339
S (3)	1.220	1.313	1.342	1.346	1.378	1.504	0.000	1.346		1.346
S (1.5)	1.278	1.367	1.399	1.401	1.434	1.582	0.000	1.401		1.401
PP	1.289	1.385	1.422	1.428	1.467	1.639	0.000	1.428		1.428

Table 6: Model 2. No outliers.

Method	Min	Q1	Q2	Mean	Q3	Max	Out	Clean	$\overline{\text{Out}}$	$\overline{\text{Clean}}$
True	1.364	8.486	11.120	11.280	13.830	25.550	10.063	1.222	100.589	1.358
LS	1.350	5.113	5.622	5.629	6.135	8.463	1.597	4.032	19.528	4.512
S (3)	1.357	8.535	11.070	11.220	13.720	22.080	9.839	1.380	99.230	1.541
S (1.5)	1.414	8.731	11.490	11.690	14.060	25.360	9.638	2.047	97.207	2.296
PP	1.406	8.216	10.200	10.350	12.350	20.630	8.922	1.427	90.696	1.589

Table 7: Model 2. $\epsilon_1 = 0.10$, $\epsilon_2 = 0.90$.

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 (13) and (14), 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

Method	Min	Q1	Q2	Mean	Q3	Max	Out	Clean	$\overline{\text{Out}}$	$\overline{\text{Clean}}$
True	7.335	18.130	21.030	21.140	24.210	37.000	20.054	1.087	100.598	1.358
LS	4.034	5.710	6.256	6.322	6.844	9.625	1.840	4.482	9.505	5.610
S (3)	4.639	7.507	16.780	14.780	19.930	25.580	12.427	2.357	69.919	3.035
S (1.5)	7.481	18.190	20.840	20.810	23.560	35.390	17.916	2.891	90.648	3.645
PP	6.936	15.230	16.710	16.480	18.110	22.660	14.865	1.618	76.535	2.039

Table 8: Model 2. $\epsilon_1 = 0.20$, $\epsilon_2 = 0.90$.

Method	Min	Q1	Q2	Mean	Q3	Max	Out	Clean	$\overline{\text{Out}}$	$\overline{\text{Clean}}$
True	0.259	0.295	0.304	0.304	0.314	0.356	0.000	0.304		0.304
LS	0.244	0.276	0.284	0.285	0.295	0.336	0.000	0.285		0.285
S (3)	0.256	0.289	0.299	0.301	0.312	0.356	0.000	0.301		0.301
S (1.5)	0.301	0.340	0.352	0.354	0.367	0.444	0.000	0.354		0.354
PP	0.322	0.363	0.382	0.385	0.405	0.476	0.000	0.385		0.385

Table 9: Model 3. No outliers.

Method	Min	Q1	Q2	Mean	Q3	Max	Out	Clean	$\overline{\text{Out}}$	$\overline{\text{Clean}}$
True	0.864	3.431	4.675	4.684	5.668	9.818	4.411	0.274	44.163	0.304
LS	0.361	1.674	2.618	2.735	3.514	7.072	2.074	0.660	18.457	0.736
S (3)	0.781	3.440	4.653	4.681	5.693	9.869	4.412	0.269	44.148	0.299
S (1.5)	0.943	3.528	4.722	4.784	5.771	9.888	4.465	0.318	44.674	0.354
PP	0.895	3.533	4.720	4.794	5.849	10.240	4.439	0.355	44.397	0.394

Table 10: Model 3. $\epsilon_1 = 0.10$.

$S-$estimator was more notable in all the other cases of Model 1, Model 2 and Model 3. For Model 4, although the $S-$estimator still performed better than the projection–pursuit one, its advantage was smaller than in the other cases.

Method	Min	Q1	Q2	Mean	Q3	Max	Out	Clean	$\overline{\text{Out}}$	$\overline{\text{Clean}}$
True	3.959	7.653	8.968	9.085	10.440	15.460	8.842	0.243	44.088	0.304
LS	1.835	5.061	6.179	6.310	7.453	12.110	5.599	0.711	27.363	0.893
S (3)	3.647	7.637	8.960	9.083	10.530	15.380	8.846	0.237	44.113	0.297
S (1.5)	3.822	7.728	9.109	9.215	10.590	16.090	8.931	0.284	44.535	0.355
PP	3.543	7.807	9.149	9.234	10.610	15.890	8.913	0.321	44.430	0.402

Table 11: Model 3. $\epsilon_1 = 0.20$.

Method	Min	Q1	Q2	Mean	Q3	Max	Out	Clean	$\overline{\text{Out}}$	$\overline{\text{Clean}}$
True	0.018	0.020	0.021	0.021	0.021	0.024	0.000	0.021		0.021
LS	0.016	0.018	0.019	0.019	0.019	0.022	0.000	0.019		0.019
S (3)	0.017	0.019	0.020	0.020	0.021	0.024	0.000	0.020		0.020
S (1.5)	0.020	0.023	0.024	0.024	0.025	0.032	0.000	0.024		0.024
PP	0.021	0.025	0.026	0.026	0.028	0.034	0.000	0.026		0.026

Table 12: Model 4. No outliers.

Method	Min	Q1	Q2	Mean	Q3	Max	Out	Clean	$\overline{\text{Out}}$	$\overline{\text{Clean}}$
True	0.648	3.223	4.417	4.447	5.467	9.657	4.429	0.019	44.290	0.021
LS	0.061	0.746	1.705	1.909	2.705	6.490	1.680	0.229	14.265	0.256
S (3)	0.554	3.230	4.384	4.448	5.497	9.668	4.430	0.018	44.288	0.020
S (1.5)	0.023	3.305	4.340	4.409	5.439	8.879	4.385	0.023	44.344	0.026
PP	0.652	3.247	4.543	4.521	5.584	9.540	4.497	0.024	44.996	0.027

Table 13: Model 4. $\epsilon_1 = 0.10$.

5 Examples

In this section we present two data analyses where the interest is in identifying potentially atypical observations. We first analyse the ground level ozone concentration data discussed in the Introduction. Our second example deals with human mortality data per age group for French males between 1816 and 2010. We will see than in both cases

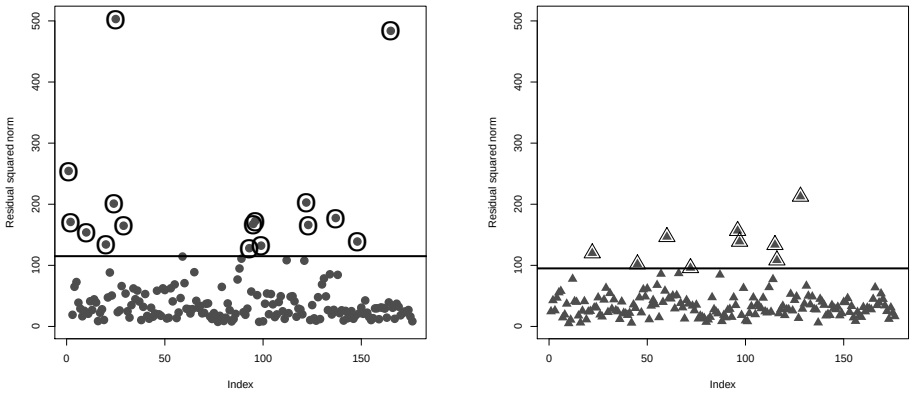
Method	Min	Q1	Q2	Mean	Q3	Max	Out	Clean	$\overline{\text{Out}}$	$\overline{\text{Clean}}$
True	3.665	7.498	8.894	8.908	10.250	15.290	8.892	0.017	44.339	0.021
LS	0.796	4.128	5.269	5.442	6.604	11.580	5.182	0.260	25.173	0.327
S (3)	3.442	7.461	8.841	8.900	10.260	15.150	8.884	0.016	44.307	0.020
S (1.5)	3.477	7.277	8.535	8.744	10.090	15.530	8.722	0.022	43.479	0.028
PP	3.683	7.623	8.925	9.044	10.430	15.360	9.022	0.022	44.984	0.028

Table 14: Model 4. $\epsilon_1 = 0.20$.

the differences between the observations and the robust predictions can easily be used to detect potentially outlying curves, and that in both examples the identified outliers correspond to "real anomalies" in the specific context of each problem. In both examples, the $S-$estimators were computed as in Section 3 with tuning constant $c = 3$.

5.1 Ground level Ozone concentrations

Our 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 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 (displayed in Figure 1). 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, which corresponds to what may be expected from the way O3 concentrations behave in summer days, we consider $1-$dimensional approximations. We use an $S-$estimator applying the approach described in Section 3 with a cubic $B-$spline basis of dimension $m = 10$. Figure 2 contains 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). Highlighted points correspond to residuals that are relatively larger than the majority, and will be considered as potential outliers. To make the visualization of the results easier, each panel in Figure 3 shows the



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

Figure 2: L_2 norms of the residuals obtained with the S-estimator (panel a) and the classical one (panel b) for the ground level ozone data. Highlighted points in each panel correspond to suspected atypical days.

observations detected as outliers on one year, both by the robust estimator (solid lines) and the classical approach (dotted lines). The thin gray lines in the background show all the available observations, and are included as a visual reference. 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. In total, the S-estimator finds 16 days as possibly atypical, while the classical approach identifies 9 days as outliers, of which only one is also classified as unusual by the robust method.

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 actually 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

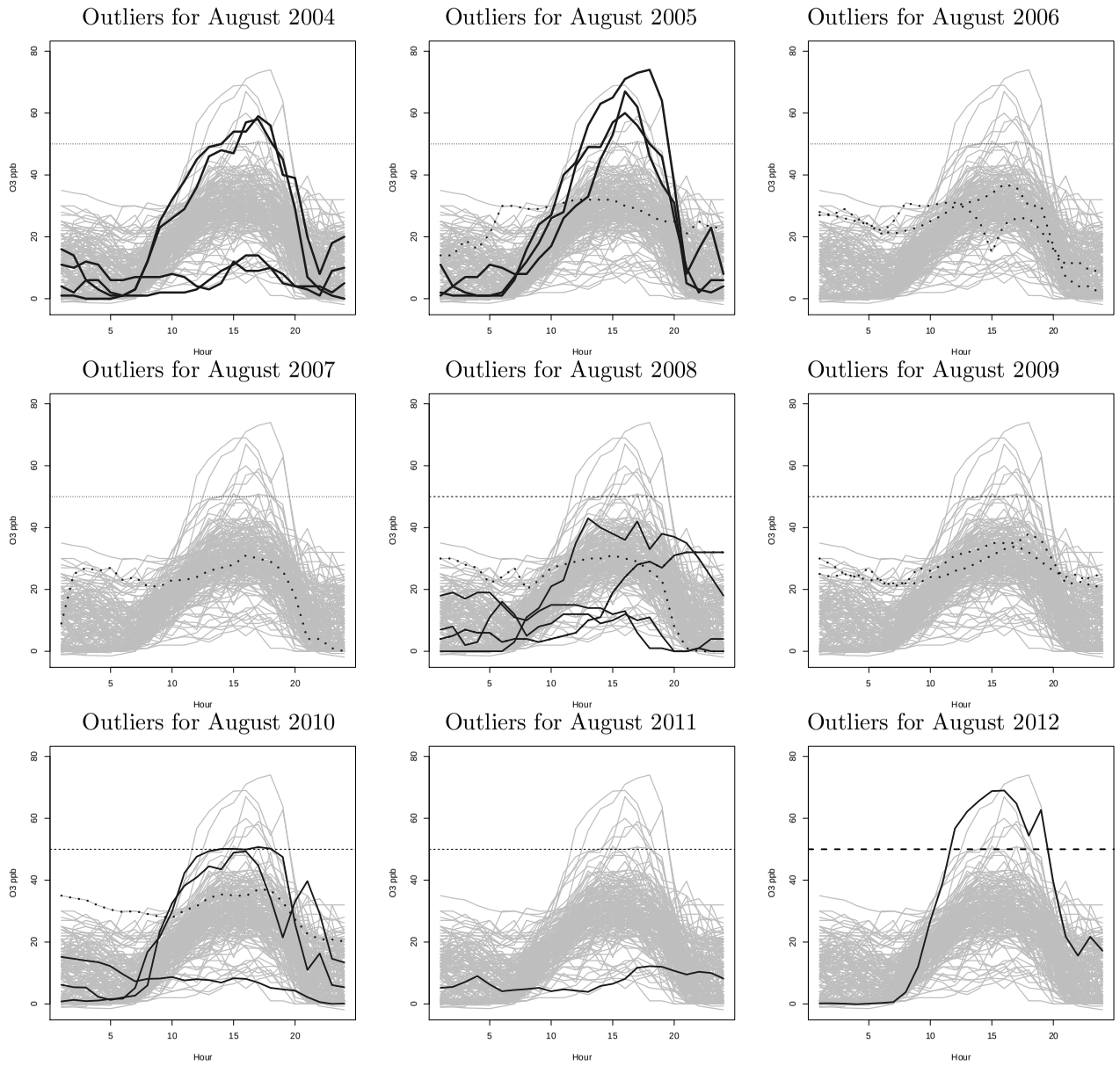


Figure 3: Hourly mean concentration (in ppb) of ground level ozone in Richmond, BC, Canada. Thin gray lines show all the available data. Solid lines correspond to potential outliers identified by the robust estimator, while those identified by the classical analysis are displayed with dotted lines.

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 (cold summer days), while days with a sharp O3 peak correspond to particularly hot days. On the other hand, days flagged as possible outliers by the least squares approach 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.

5.2 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 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 for the best 2−dimensional subspace approximating these curves using the approach described in Section 3 with a cubic $B-$spline basis of dimension $m = 20$. Figure 5 (b) contains the estimated central curve plotted over the original data, and also over the robustly predicted curves. To

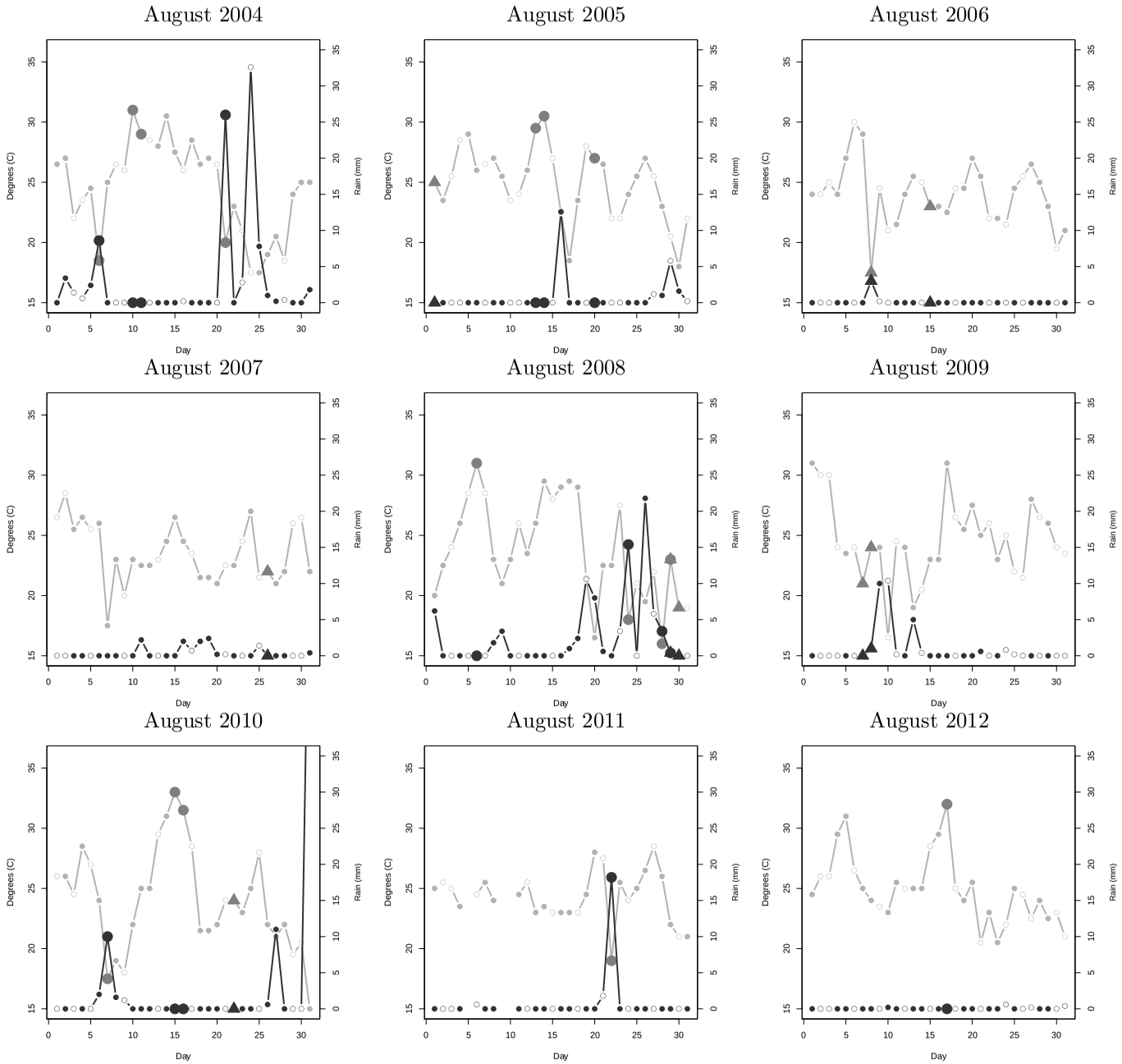
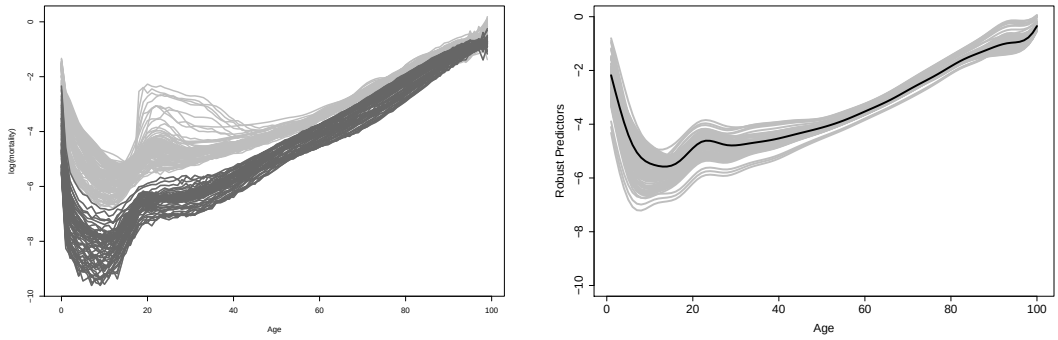
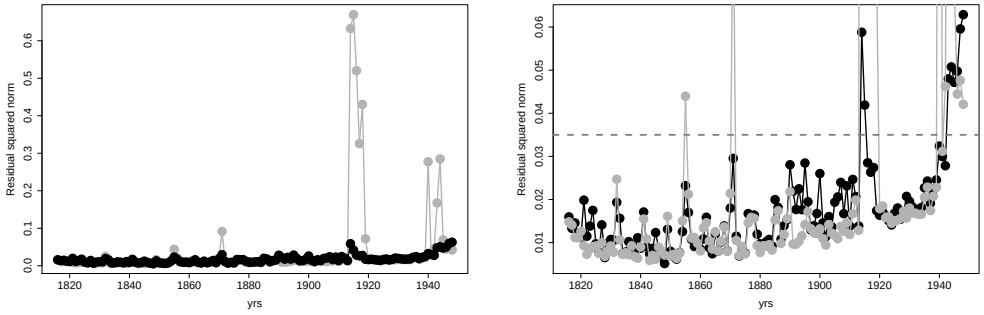


Figure 4: Maximum daily temperature profile and rain levels for the month of August. Days with an atypical O3 profile as flagged by the robust method are indicated with large solid circles. Those identified as outlying by the classical approach are marked with a large triangle.



(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.



(a) L^2 norm of the prediction residuals (b) Detail of the lower section of plot (a)

Figure 6: L^2 norm of the residuals corresponding to the predictions obtained with the $S-$estimator (light gray points) and the classical Least Squares estimator (black points). Panel (b) is a detail of the lower portion of panel (a).

determine whether any curve appears as atypical, in Figure 6 we plot the L^2 norm of the residuals for each of the 135 curves. Light gray points correspond to residuals from the predictions obtained with the $S-$estimator, while the residuals associated with the classical estimator are represented using black points. We see that the $S-$estimator

clearly identifies four periods of atypical observations. Panel (b) in Figure 6 zooms in the lower part of the plot to explore more carefully the residuals of the predictions based on the LS estimator. There seem to be two clear "peaks" in the LS–based residuals, that partially coincide with two of the 4 sets of large residuals found by the $S-$estimator. Using this plot we select the value 0.035 (the dashed horizontal line in panel (b) of Figure 6) as our threshold to identify potential outliers. The robust fit identifies 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 7 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. 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.

5.3 Lip movement data

Here we analyse the "lip movement" data. The observations correspond to the position of the lower lip as a function of time as a subject repeated the syllable "bob" 32 times. The available data have been pre–processed and the times standardized to 700 milliseconds. We have 501 observations for each of the 32 trajectories, which are available on–line from `http://www.stats.ox.ac.uk/~silverma/fdacasebook/lipemg.html`. More details can be found at the same URL and in Malfait and Ramsay (2003). This data set was also studied by Gervini (2008) who detected as atypical three trajectories with delayed second peaks, labelled as 24, 25 and 27.

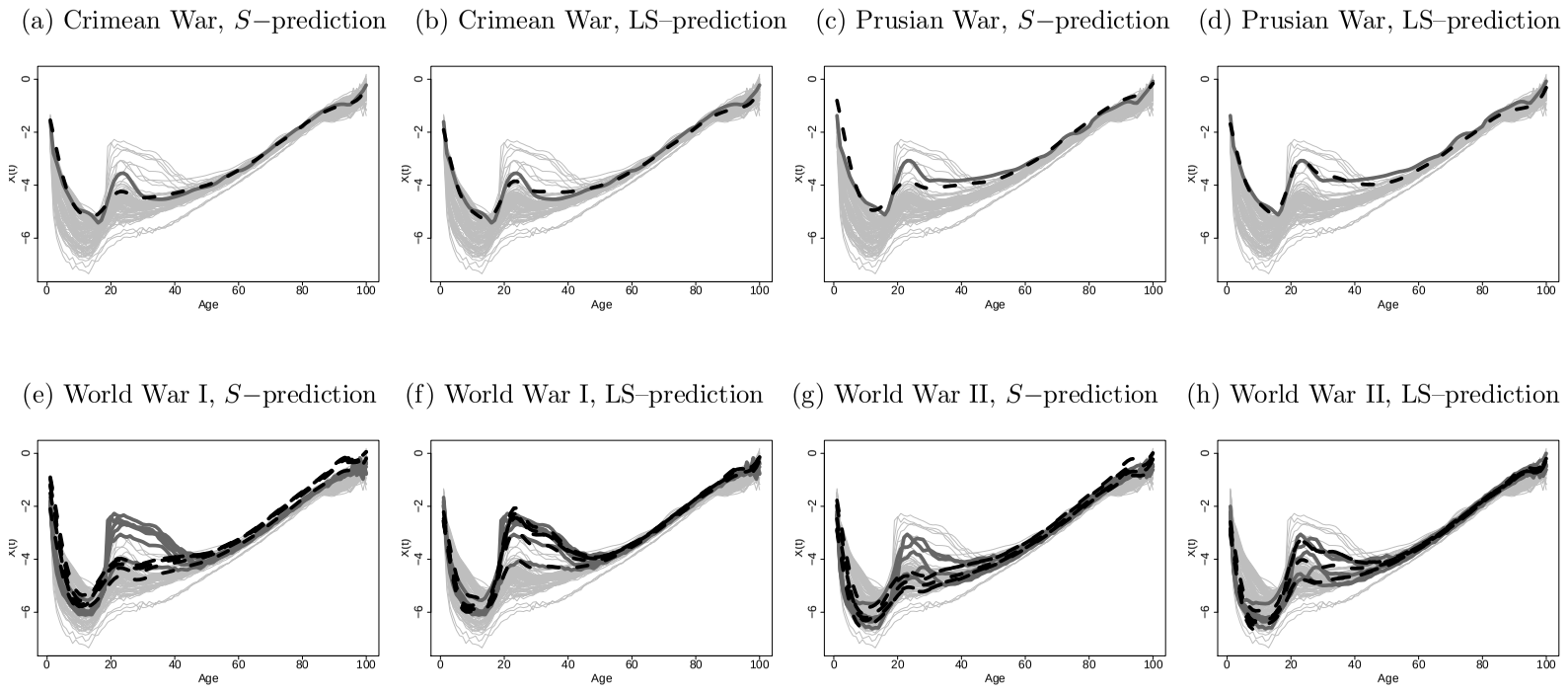
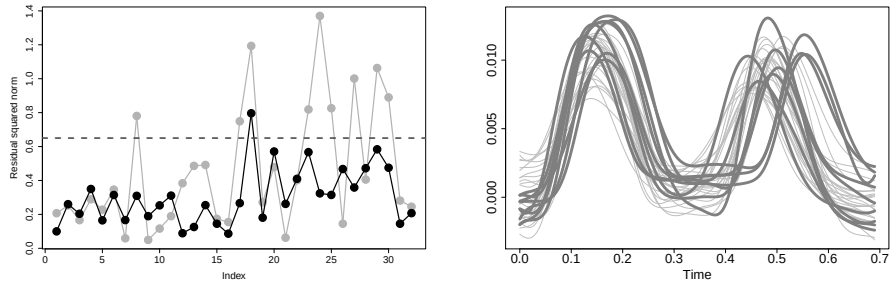


Figure 7: Years 1855 and 1871 were detected as outliers by the robust estimator, during the Crimean and Prusian War, respectively (plots a) to d)). Curves for the years 1914–1919 (plots e) and f)) and 1940, 1942–1945 (plots g) and h) row). The curves of mortality rates for these periods, corresponding 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.

Using a cubic $B-$spline basis of dimension 20 we obtained $5-$dimensional robust predicted trajectories for each observed curve. The L^2 norms of the residuals (both for the $S-$ and the classical LS estimators) are displayed in panel (a) of Figure 8. There are 7 curves that clearly stand out in terms of their prediction residuals when using the robust estimator. The residuals associated with the LS predictions show one curve that is fit slightly worse than the others. Panel (b) of Figure 8 displays the original data, with the seven suspected outliers highlighted with darker lines. It is interesting to note that, without having "extreme" observations in the data (for example, there are no unusually large response values), all the flagged curves do exhibit a behaviour that is rather peculiar. Namely, the lip stays in its lower position for a much longer period of time that during the other repetitions. At the same time, their "second peak" occurs either earlier or later than the rest of the curves. This is another example where the robust estimator is able to identify observations that are different from the bulk of the data. These curves will not have been found using a classical principal components analysis.



(a) Residuals based on the $S-$ (light) and LS–estimates (dark)

(b) Observations (light) and potential outliers (dark)

Figure 8: Lip movement data. Panel (a) shows the prediction residuals (times 10^6) obtained using the robust $S-$estimator (light gray points) and the LS estimator (black points). Panel (b) displays the data (light gray) with the suspected outliers corresponding to the dark lines.

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/m̃atias/Property_FPCA.pdf

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

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 International Conference on Computer Vision*, available at http://citeseer.ist.psu.edu/torre01robust.html.

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.

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

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

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

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

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.

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

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

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

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.

A Appendix

PROOF OF PROPOSITION 2.1. Note that since $\mathbf{x} \sim \mathcal{E}_p(\mathbf{0}, \mathbf{\Sigma}, \phi)$, $\mathbf{z} = \mathbf{\Lambda}^{-1/2}\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$ 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 $\mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(q)}$. If we arrange this basis as columns of a matrix $\mathbf{B} \in \mathbb{R}^{p\times q}$ we have that $\mathbf{r}(\mathcal{L}) = (r_1(\mathcal{L}), \ldots, r_p(\mathcal{L}))^{\mathrm{T}} = \mathbf{x} - \pi(\mathbf{x}, \mathcal{L}) = (\mathbf{I} - \mathbf{B}\mathbf{B}^{\mathrm{T}})\mathbf{x} \sim \mathcal{E}_p(\mathbf{0}, \mathbf{\Sigma}_{\mathcal{L}}, \phi)$, with $\mathbf{\Sigma}_{\mathcal{L}} = (\mathbf{I} - \mathbf{B}\mathbf{B}^{\mathrm{T}})\mathbf{\Sigma}(\mathbf{I} - \mathbf{B}\mathbf{B}^{\mathrm{T}})^{\mathrm{T}} = \mathbf{C}\mathbf{\Sigma}\mathbf{C}^{\mathrm{T}}$, with $\mathbf{C} = (\mathbf{I} - \mathbf{B}\mathbf{B}^{\mathrm{T}})$. Since $\mathbf{x} = \boldsymbol{\beta}\mathbf{\Lambda}^{1/2}\mathbf{z}$, we see that $\mathbf{r}(\mathcal{L})$ can be written as $\mathbf{C}\boldsymbol{\beta}\mathbf{\Lambda}^{1/2}\mathbf{z}$. Therefore, the characteristic function of $\mathbf{r}(\mathcal{L})$ is given by $\varphi_{\mathbf{r}(\mathcal{L})}(\mathbf{t}) = \varphi_{\mathbf{z}}(\mathbf{\Lambda}^{1/2}\boldsymbol{\beta}^{\mathrm{T}}\mathbf{C}^{\mathrm{T}}\mathbf{t}) = \phi(\mathbf{t}^{\mathrm{T}}\mathbf{C}\mathbf{\Sigma}\mathbf{C}^{\mathrm{T}}\mathbf{t})$, where ϕ denotes the generator of the characteristic function of $\mathbf{z}$. Hence, for the jth coordinate of the vector of residuals we have $\varphi_{r_j(\mathcal{L})}(t) = \varphi_{\mathbf{r}(\mathcal{L})}(t\mathbf{e}_j) = \phi(t^2\,\mathbf{e}_j^{\mathrm{T}}\mathbf{C}\mathbf{\Sigma}\mathbf{C}^{\mathrm{T}}\mathbf{e}_j) = \phi(t^2\,\mathbf{c}_j^{\mathrm{T}}\mathbf{\Sigma}\mathbf{c}_j) = \varphi_{z_1}(t^2\,\mathbf{c}_j^{\mathrm{T}}\mathbf{\Sigma}\mathbf{c}_j)$. It follows that $r_j(\mathcal{L}) \sim \xi_j z_1$ where $z_1 \sim G$ and $\xi_j^2 = \mathbf{c}_j^{\mathrm{T}}\mathbf{\Sigma}\mathbf{c}_j$. This implies that $\sigma_{j,\mathcal{L}}^2 = \sigma_{\mathrm{R}}^2(F_j(\mathcal{L})) = \mathbf{c}_j^{\mathrm{T}}\mathbf{\Sigma}\mathbf{c}_j$. Hence, $\sum_{j=1}^p \sigma_{j,\mathcal{L}}^2 = \sum_{j=1}^p \mathbf{c}_j^{\mathrm{T}}\mathbf{\Sigma}\mathbf{c}_j = \mathrm{tr}(\mathbf{C}\mathbf{\Sigma}\mathbf{C}^{\mathrm{T}})$. This last expression is minimized when $\mathbf{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_q)$ (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 $\mathbf{\Gamma}$. Assume that $m = m_n$ is such that $m_n > q$ and $\{\phi_1, \ldots, \phi_q\} \subset \{\delta_1, \delta_2, \ldots, \delta_{m_n}\}$. 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, $\mathbf{x} = AX \sim \mathcal{E}_p(\mathbf{0}, \mathbf{\Sigma}, \phi)$ where A is defined in (7) and $\mathbf{\Sigma} = A\Gamma A^* = \mathrm{diag}(\lambda_1, \ldots, \lambda_q, \lambda_{\ell_{q+1}}, \ldots, \lambda_{\ell_m})$ 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 $\mathbf{B}_{0,m} = (\mathbf{e}_1, \ldots, \mathbf{e}_q)$. 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 $\phi_1, \ldots, \phi_q$, which shows Fisher–consistency.

Let us now consider the general situation. As before, we have $\mathbf{x} = AX \sim \mathcal{E}_p(\mathbf{0}, \mathbf{\Sigma}, \phi)$ where A is defined in (7) and $\mathbf{\Sigma} = A\Gamma A^*$. 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 $\{\delta_1, \ldots, \delta_m\}$ 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, $\mathbf{B}_{0,m} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_q)$ 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 $\mathbf{B}^{\mathrm{T}}\mathbf{B} = \mathbf{I}_q$, $\Psi_m(\mathcal{L}_{\mathbf{B}}) = \sum_{j=1}^{m} \sigma^2_{j,\mathcal{L}_{\mathbf{B}}} \geq \sum_{j=1}^{m} \sigma^2_{j,\mathcal{L}_{\mathbf{B}_{0,m}}} = \sum_{s=q+1}^{m} \alpha_s = \sum_{s=q+1}^{m} \lambda_s(\Upsilon_m)$, 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. □