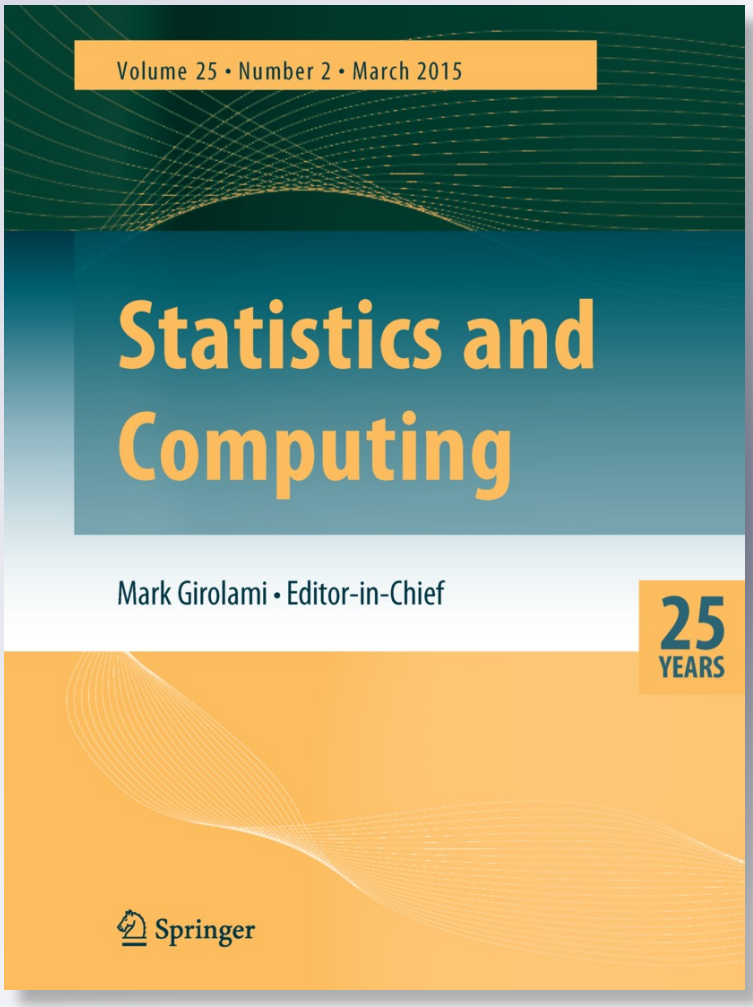
Robust nonlinear principal components

Ricardo A. Maronna, Fernanda Méndez & Víctor J. Yohai

Statistics and Computing

ISSN 0960-3174
Volume 25
Number 2

Stat Comput (2015) 25:439-448
DOI 10.1007/s11222-013-9442-0



Springer

Stat Comput (2015) 25:439–448
DOI 10.1007/s11222-013-9442-0

Robust nonlinear principal components

Ricardo A. Maronna · Fernanda Méndez · Víctor J. Yohai

Received: 29 July 2013 / Accepted: 28 November 2013 / Published online: 11 December 2013

Abstract All known approaches to nonlinear principal components are based on minimizing a quadratic loss, which makes them sensitive to data contamination. A predictive approach in which a spline curve is fit minimizing a residual M-scale is proposed for this problem. For a p-dimensional random sample $\mathbf{x}_i$ $(i = 1, \ldots, n)$ the method finds a function $\mathbf{h} : R \to R^p$ and a set $\{t_1, \ldots, t_n\} \subset R$ that minimize a joint M-scale of the residuals $\mathbf{x}_i - \mathbf{h}(t_i)$, where $\mathbf{h}$ ranges on the family of splines with a given number of knots. The computation of the curve then becomes the iterative computing of regression S-estimators. The starting values are obtained from a robust linear principal components estimator. A simulation study and the analysis of a real data set indicate that the proposed approach is almost as good as other proposals for row-wise contamination, and is better for element-wise contamination.

Keywords S-estimators · Splines · Principal curves

This research was partially supported by grants X–018 from University of Buenos Aires, PID 5505 from CONICET and PICTs 21407 and 00899 from ANPCyT.

R.A. Maronna (✉)
Faculty of Exact Sciences, University of La Plata, C.C. 172, 1900 La Plata, Argentina
e-mail: rmaronna@retina.ar

F. Méndez
Faculty of Economic Science and Statistics, University of Rosario, Bv. Oroño 1261, 2000 Rosario, Argentina

V.J. Yohai
Departamento de Matemática, Faculty of Natural and Exact Sciences, Universidad de Buenos Aires, Ciudad Universitaria, Pabellon 1, 1428 Buenos Aires, Argentina

1 Introduction

Principal components (henceforth PCs) are a well-established tool for data representation and compression. Let $\mathbf{x}$ be a random vector in R^p. The first PC can be defined by two linear functions $g : R^p \to R$ and $\mathbf{h} : R \to R^p$ such that $\mathrm{E}\|\mathbf{x} - \mathbf{h}(g(\mathbf{x}))\|^2 = \min$. It is well-known that the coefficients of g are those of the eigenvector corresponding to the largest eigenvalue of the covariance matrix of $\mathbf{x}$.

There have been several approaches to enlarge the family of functions considered. An early (and little known) proposal was made by Yohai et al. (1985). It was based on a predictive point of view, and it considers parametric families for g and $\mathbf{h}$. Later, Hastie and Stuetzle (1989) proposed principal curves, which became a popular tool. The subject of principal curves was later discussed by Tibshirani (1992) and Delicado (2001). A clever approach to principal curves, based on ideas similar to k-means clustering, was proposed by Verbeek et al. (2002). Gerber and Whitaker (2013) present an approach based on a new objective function. Several authors propose predictive approaches based on nonparametric or semiparametric fitting; see e.g. Bolton et al. (2003) and the references therein.

All the aforementioned approaches are based on second moments, and are therefore sensitive to atypical observations. The R function `principal.curve` that implements Hastie and Stuetzle's (1989) proposal, includes a robust option, in which a standard smoother is replaced by a robust one. The robust option of `principal.curve` has two drawbacks. The first one is that the starting values are given by classical linear PCs, and are therefore not robust.

In order to explain the second drawback, consider a data set $\mathbf{X} \in R^{n \times p}$ with n cases in p dimensions. We may consider two forms of outlier contamination. One is "row-wise", in which a proportion ε of the n rows is contaminated; the

other is "element-wise", in which a proportion ε of the np elements of $\mathbf{X}$ is contaminated. The second type of contamination presents serious problems in high-dimensional data, and has been considered by Maronna and Yohai (2008) and Alqallaf et al. (2009). The linear PC estimator proposed by Croux et al. (2003) may also be used for this type of contamination. The second drawback is that, even if the starting classical PCs are replaced by robust ones, the smoothing can be sensitive to element-wise contamination.

In this article we propose an approach basically similar to that of Yohai et al. (1985) with two changes. Firstly, the family of functions is enlarged to a broader family of smooth functions. Secondly, the quadratic criterion is replaced by another one based on robust scales, in a way that the resulting estimator is less sensitive to both row-wise and element-wise contamination, as will be explained in the next section.

Section 2 reviews Hastie and Stuetzle's (1989) principal curves and their robust versions. Section 3 presents the proposed estimator. Section 4 gives some theoretical results on both types of estimators. Section 5 shows the results of a small simulation study. Section 6 contains an example with a real data set. Section 7 shows the computing running times of the estimator. Finally Sect. 8 is an appendix containing proofs of theoretic results.

2 Principal curves

Given a curve $\mathbf{h}(t)$ where t ranges over an interval I, define the projection index of point $\mathbf{x}$ as

$$t_{\mathbf{x}} = t_{\mathbf{x}}(\mathbf{h}) = \arg\min_{t\in I} \left\|\mathbf{x} - \mathbf{h}(t)\right\|. \tag{1}$$

Then $\mathbf{h}$ defines a principal curve if the conditional expectation is

$$\mathrm{E}[\mathbf{x}|t_{\mathbf{x}}] = \mathbf{h}(t_{\mathbf{x}}).$$

Section 3 of Hastie and Stuetzle (1989) gives results on the relationships between principal curves and (linear) principal components.

Consider now sample principal curves. Let $\mathbf{X} = [\mathbf{x}_{ij}] \in R^{n\times p}$ be a p-dimensional data set, and put $\mathbf{x}_i = (x_{i1}, \ldots, x_{in})'$, $i = 1, \ldots, n$. Let $\mathbf{h}_0$ be an initial approximation to a principal curve, usually given by the first linear PC (i.e. $\mathbf{h}_0 = t\mathbf{a} + \mathbf{b}$ with $\mathbf{a}, \mathbf{b} \in R^p$ and $t \in R$). Then $\mathbf{h}$ is the result of an iterative procedure in which $\mathbf{h}_{k+1}$ is obtained by smoothing each of the columns of $\mathbf{X}$ on $t_{\mathbf{x}_i}(\mathbf{h}_k)$, $i = 1, \ldots, n$, and then $t_{\mathbf{x}_i}$ is updated through (1). Hastie and Stuetzle (1989) propose both lowess (Cleveland 1979) and smoothing splines as smoothing devices, and use the classical PC as the starting point. The robust version of this procedure as implemented in the R code `principal.curve` mentioned above employs the robust option for lowess, keeping the classical PC as the starting point. In order to improve the robustness of the procedure towards both row- and element-wise contamination, the classical PC must be replaced by a robust PC; see the beginning of Sect. 5.2. However, even using a robust starting point and robust smoothing, elementwise contamination may affect the updating of $t_{\mathbf{x}_i}$. In fact, if some of the coordinates of $\mathbf{x}_i$ are contaminated, then (1) may assign $\mathbf{x}_i$ to a wrong place on the curve.

3 Spline-based nonlinear principal components

We first describe the approach of Yohai et al. (1985). Given the functions $g : R^p \to R$ and $\mathbf{h} = (h_1, \ldots, h_p) : R \to R^p$, call the residuals $\mathbf{r}_i = \mathbf{r}_i(g, \mathbf{h}) = \mathbf{x}_i - \mathbf{h}(g(\mathbf{x}_i))$. Then the goal is to find $g \in \mathscr{G}$ and $h \in \mathscr{H}$—where $\mathscr{G}$ and $\mathscr{H}$ are some specified families of functions—such that the criterion

$$C_0(g, \mathbf{h}) = \sum_{j=1}^{p}\sum_{i=1}^{n} r_{ij}^2 = \min. \tag{2}$$

Yohai et al. (1985) consider nondecreasing quadratic functions for $\mathscr{G}$ and $\mathscr{H}$, but their approach can be employed for more general families. Note that given $\mathbf{h}$, g is determined by the minimization of the criterion C_0. We now describe our robust version of the former approach, also based on the minimization of a criterion C. Instead of expressing C as a function of $(g, \mathbf{h})$ as in (2), we shall define it as a function of $(\mathbf{t}, \mathbf{h})$ with $t_i = g(\mathbf{x}_i)$, which will yield more tractable expressions. Note that $\mathbf{t}$ is the nonlinear analogue of the first (linear) "principal component".

A scale M estimator (an M-scale for short) of the sample $\mathbf{z} = (z_1, \ldots, z_n)$ is the solution $S = S(\mathbf{z})$ of the equation

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{z_i}{S}\right) = \delta. \tag{3}$$

Here ρ is a "bounded ρ-function" in the sense of Maronna et al. (2006), namely, $\rho(t)$ is a nondecreasing function of $|t|$, $\rho(0) = 0$, $\rho(\infty) = 1$, and $\rho(t)$ is increasing for $t \geq 0$ such that $\rho(t) < 1$; and $\delta \in (0, 1)$ calibrates the estimator's breakdown point.

Given $\mathbf{h}$ and $\mathbf{t} = (t_1, \ldots, t_n)$ let $\mathbf{r}_i = \mathbf{x}_i - \mathbf{h}(t_i)$. Call $\mathbf{r}_{\cdot j} = (r_{1j}, \ldots, r_{nj})'$ the j-th column of the matrix $\mathbf{R}$ with rows $\mathbf{r}_1', \ldots, \mathbf{r}_n'$. Instead of quadratic loss we will employ the criterion

$$C(\mathbf{t}, \mathbf{h}) = \sum_{j=1}^{p} S(\mathbf{r}_{\cdot j})^2, \tag{4}$$

where S is an M-scale. Given $\mathbf{h}$, we choose $\mathbf{t}$ so that $C(\mathbf{t}, \mathbf{h})$ is minimum. Instead of a parametric family for $\mathbf{h}$, we adopt a family of "smooth" functions. We first give the definition of our proposed estimator, and then show its motivation. We choose for $\mathbf{h}$ the family of natural splines with a *fixed* number of knots N_{knots}, located at the quantiles s_k of order

$k/(n+1)$, $k = 1, \ldots, N_{\text{knots}}$. Let $(b_1(t), \ldots, b_{N_{\text{knots}}}(t))'$ be a functional basis of the space $NS^m(s_1, \ldots, s_{N_{\text{knots}}})$ of natural splines of order $2m$ on I with knots $s_1, \ldots, s_{N_{\text{knots}}}$, with $m = 2$ (cubic splines). Let

$$\mathbf{b}_i = \big(b_1(t_i), \ldots, b_{N_{\text{knots}}}(t_i)\big)', \quad i = 1, \ldots, n. \tag{5}$$

Then given $\mathbf{t}$, h_j is a linear combination of the b_ks, i.e.

$$h_j(t) = \sum_{k=1}^{N_{\text{knots}}} b_k(t)\beta_k^{(j)}, \tag{6}$$

where the coefficients $\beta_k^{(j)}$ have to be determined. Let $\boldsymbol{\beta}^{(j)} = (\beta_1^{(j)}, \ldots, \beta_{N_{\text{knots}}}^{(j)})$, and put for $\boldsymbol{\beta} \in R^{N_{\text{knots}}}$: $r_{ij} = r_{ij}(\boldsymbol{\beta}) = x_{ij} - \boldsymbol{\beta}'\mathbf{b}_i$. Then to minimize (4) define for $j = 1, \ldots, p$

$$\boldsymbol{\beta}^{(j)} = \arg\min_{\boldsymbol{\beta}} S\big(r_{\cdot j}(\boldsymbol{\beta})\big). \tag{7}$$

That is, $\boldsymbol{\beta}^{(j)}$ is a regression S estimator (Rousseeuw and Yohai 1984). Henceforth we choose for ρ in (3) the bisquare function:

$$\rho(t) = 1 - \big(1 - t^2\big)^3 \mathrm{I}\big(|t| \le 1\big)$$

where I(.) is the indicator function; and $\delta = 0.5(1 - N_{\text{knots}}/n)$. The reason for this choice of δ is that N_{knots} is the dimension of each $\boldsymbol{\beta}^{(j)}$, and therefore this δ maximizes the breakdown point of the regression S estimator in (7) (see Maronna et al. 2006).

Notice that the definition above involves some arbitrariness, since $\mathbf{h}(t)$ could be replaced by $\mathbf{h}(u(t))$ where u is a smooth strictly monotonic function. The justification of our choice of $\mathbf{h}$ is the following. For a general $\mathbf{h}$, an adequate way to obtain a smooth robust fit would be to penalize the "roughness" of $\mathbf{h}$ as follows:

$$\sum_{j=1}^{p} S\big(\mathbf{x}_{\cdot j} - h_j(\mathbf{t})\big)^2 + \lambda \sum_{j=1}^{p} \int_I \big(h_j''(s)\big)^2 ds = \min, \tag{8}$$

where λ is a penalty parameter and I is any interval that contains the elements of $\mathbf{t}$. When $\lambda = 0$ one gets a (rough) interpolating curve. It can be shown that for $\lambda > 0$ the resulting h_js are cubic splines with knots at each t_i. However, the selection of λ is very computationally expensive, for it requires re-computing the estimator for many values of λ, which makes the procedure impractically slow. Instead of (8) we could also have considered penalized splines as in Tharmaratnam et al. (2010); but with this alternative approach the problem of the choice of λ would remain. For this reason we prefer to employ a fixed set of knots. The fact that the resulting estimator has the form of a regression S estimator greatly simplifies the computation, as will be seen in the next section. In this work we deal only with the first PC.

3.1 Computing the spline-based PCs

The algorithm for the proposed PCs is iterative. Compute an initial approximation $(\mathbf{t}_0, \mathbf{h}_0)$. Given $(\mathbf{t}_m, \mathbf{h}_m)$ compute

$$\begin{aligned} &\mathbf{h}_{m+1} = \arg\min_{\mathbf{h}} C(\mathbf{t}_m, \mathbf{h}) \quad \text{and} \\ &\mathbf{t}_{m+1} = \arg\min_{\mathbf{t}} C(\mathbf{t}, \mathbf{h}_{m+1}). \end{aligned} \tag{9}$$

We cannot hope for absolute minima above, but we would at least want that each step decreases the criterion, i.e.,

$$\begin{aligned} &C(\mathbf{t}_m, \mathbf{h}_{m+1}) \le C(\mathbf{t}_m, \mathbf{h}_m) \quad \text{and} \\ &C(\mathbf{t}_{m+1}, \mathbf{h}_{m+1}) \le C(\mathbf{t}_m, \mathbf{h}_{m+1}). \end{aligned} \tag{10}$$

In similarity with principal curves, the initial $(\mathbf{t}_0, \mathbf{h}_0)$ are obtained from the first linear PC. In order to ensure robustness against both row- and element-wise contamination, we employ the PCs described in Maronna and Yohai (2008).

We now deal with the first half of (9) starting with $m = 0$. We have $(\mathbf{t}_m, \mathbf{h}_m)$ and want to compute $\mathbf{h}_{m+1}$ and therefore the respective $\boldsymbol{\beta}^{(j)}$s, ensuring at least (10). This part involves p robust linear regressions, each of which requires an adequate set of starting values for the respective iterative process. Choosing them does not seem an easy task. The standard approach to compute initial values in high breakdown point regression is subsampling (see Maronna et al. 2006, Sect. 5.7.2); but even if subsampling could ensure (10), it could become too expensive. We therefore employ another approach. Recall that S estimators can be computed by the "iterative reweighted least squares" (IRWLS) algorithm (Maronna et al. 2006, p. 136). Recall that $\mathbf{t}_m$ defines the $\mathbf{b}_i$s in (5). Compute the residuals $r_{ij} = x_{ij} - h_{m,j}(t_{m,j})$. Call $\boldsymbol{\gamma}^{(j)}$ the regression S estimator of $r_{\cdot j}$ on the $\mathbf{b}_i$s, computed through IRWLS and starting values of *zero*. Let $\mathbf{q}(t)$ have elements

$$q_j(t) = \sum_{k=1}^{N_{\text{knots}}} b_k(t)\gamma_k^{(j)}.$$

Finally define $\mathbf{h}_{m+1}(t) = \mathbf{h}_m(t) + \mathbf{q}(t)$. Then it can be shown that

$$C(\mathbf{t}_m, \mathbf{h}_{m+1}) \le C(\mathbf{t}_m, \mathbf{h}_m). \tag{11}$$

The proof is deferred to Sect. 8.1.

Now given $\mathbf{t}_m = (t_{m,1}, \ldots, t_{m,n})'$ and $\mathbf{h}_{m+1}$ we have to compute $\mathbf{t}_{m+1}$ as in the second half of (9). An exact result would involve simultaneous optimization over all elements of $\mathbf{t}$; therefore we shall employ a "greedy" approximation, optimizing one element at a time. For each $i = 1, \ldots, n$ we let all elements of $\mathbf{t}$ *fixed* except for the i-th, and optimize over this value. Instead of letting this value range over an interval, the candidates are just all elements of $\mathbf{t}$. More formally, define for $i = 1, \ldots, n$

$$i^* = \arg\min_{l=1,\ldots,n} C\big(\mathbf{t}^{(i,l)}, \mathbf{h}_{m+1}\big), \tag{12}$$

where $\mathbf{t}^{(i,l)} = \mathbf{t}_m$ except that its i-th element is $t_{m,l}$. Then put $t_{m+1,i}{=}t_{m,i^*}$.

This procedure requires recomputing C in (4)—and therefore the p scales $S(r_{.j})$—for each i. To save time, we employ just one iteration of IRWLS. The results are sufficiently similar to those of the "exact" procedure and represent considerable save in computing time.

As for the choice of N_{knots}, we prefer to think in terms of the ratio "observations/knots" $R_{\text{knots}} = n/N_{\text{knots}}$. Recall that the breakdown point of the S estimator is $0.5(1 - N_{\text{knots}}/n)$, and therefore R_{knots} should not be too small. Moreover, exploratory simulations indicate that if R_{knots} is small, a few outliers at the extremes of $\mathbf{t}$ may have a disastrous effect. To determine R_{knots} we tried both fixed values and cross-validation. Exploratory simulations indicate that in the trade-off between efficiency and robustness, the choice of R_{knots} between 10 and 15—i.e. one knot for each 10 to 15 observations—is better (and cheaper) than cross-validation.

4 Some theoretical results

Hastie and Stuetzle (1989) show in their Propositions 1 and 2 that if the functions involved are linear, then their principal curves coincide with the classical PCs. We would like to show the same for our proposed estimator. However, the fact that these estimators are nonlinear and do not have an explicit expression makes it very difficult to give general results. For these reasons we must restrict ourselves to distributions with some symmetry properties. We shall deal specifically with elliptically symmetric distributions. For a vector $\boldsymbol{\mu} \in R^p$, a symmetric positive definite matrix $\boldsymbol{\Sigma}$ and a nonnegative function f_0, we shall say that $\mathbf{x}$ has an *elliptic distribution* $\mathscr{E}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, f_0)$ if $\mathbf{x}$ has density

$$f(\mathbf{x}) = f_0\big((\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\big),$$

where in general $\mathbf{A}'$ denotes the transpose of $\mathbf{A}$. If f_0 is decreasing we call f *unimodal*. If $\mathbf{x}$ has second moments, then the covariance matrix of $\mathbf{x}$, $\mathrm{Var}(\mathbf{x})$, is a constant times $\boldsymbol{\Sigma}$, and therefore the principal directions are given by the eigenvectors of $\boldsymbol{\Sigma}$. It then makes sense to call these eigenvectors the "principal directions" of $\mathbf{x}$ even when $\mathrm{Var}(\mathbf{x})$ does not exist.

In order to obtain theoretical results, we have to deal with the population versions of the estimators, rather than with their original sample-based form. The following results correspond to Propositions 1 and 2 of Hastie and Stuetzle (1989). Define an M-scale of a random variable x as the solution $S = S(x)$ of

$$\mathrm{E}\rho\left(\frac{x}{S}\right) = \delta. \tag{13}$$

The S in (13) is the population version of the sample-based scale (3).

We have functions $g \in G$ and $h \in H$ where G and H are given families of functions that contain linear functions. Call $\mathbf{r} = \mathbf{x} - \mathbf{h}(g(\mathbf{x}))$ the residual vector. For a random vector $\mathbf{x}$ the population nonlinear PC's are defined by the criterion

$$C(g,h) = \sum_{j=1}^{p} S(r_j)^2 = \min. \tag{14}$$

Proposition 1 shows that when restricted to linear functions, the criterion (14) yields the classical PCs.

Proposition 1 *Let $\mathbf{x} \sim \mathscr{E}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, f_0)$ and unimodal, and assume that the largest eigenvalue of $\boldsymbol{\Sigma}$ is unique. If g and $\mathbf{h}$ are restricted to be linear, then the solution of (14) is given by $\mathbf{h}(t) = \boldsymbol{\mu} + ct\mathbf{b}$ and $g(\mathbf{x}) = \mathbf{b}'(\mathbf{x}-\boldsymbol{\mu})/c$, where $\mathbf{b}$ is the first eigenvector of $\boldsymbol{\Sigma}$. and c is any constant.*

Note that all values of c yield the same curve.

Proposition 2 shows that if the data lie around some straight line, when the criterion (14) is restricted to linear functions it has that same line as output.

Proposition 2 *Let $\mathbf{x} = \mathbf{b}_0 s + \mathbf{u}$ where $\mathbf{u} \sim \mathscr{E}(\mathbf{0}, \mathbf{I}, f_0)$ is unimodal and independent of the random variable s. If g and $\mathbf{h}$ are linear, then $\mathbf{h}(t) = \mathbf{b}_0 t$ and $g(\mathbf{x}) = \mathbf{b}_0'\mathbf{x}$.*

All proofs are given in Sect. 8.

5 Simulations

5.1 Scenario

Our simulation scenario is built on a basic smooth curve given by a smooth function $\mathbf{h}_0 : R \to R^p$, and a basic set of points along the curve: $\mathbf{x}_{0,i} = \mathbf{h}_0(s_i)$ where $(s_1, \ldots, s_n)$ are given values. In this study they are uniform random values: $s_i \sim \mathrm{U}(0,1)$ The "clean" observations $\mathbf{x}_{1,i}$ are obtained by adding normal noise to $\mathbf{x}_{0,i}$, namely $\mathbf{x}_{1,i} = \mathbf{x}_{0,i} + \sigma\mathbf{e}_i$, where $\mathbf{e}_i$ have a standard p-variate normal distribution $\mathrm{N}_p(\mathbf{0}, \mathbf{I})$ and σ is specified below. The final "data" x_{ij} $(i = 1, \ldots, n,$ $j = 1, \ldots, p)$ are obtained by contaminating $x_{1,ij}$ at a contamination rate ε. We have two types of contamination. The first type is element-wise contamination, in which each $x_{1,ij}$ is contaminated at random with probability ε, namely:

$$x_{ij} = x_{1,ij} + K\mathrm{I}(u_{ij} \le \varepsilon)v_{ij}, \tag{15}$$

where $u_{ij} \sim \mathrm{Un}(0,1)$ independent (where "Un" denotes the uniform distribution), v_{ij} are standard normal, and I(.) denotes the indicator. The second type is row-wise contamination. Let $m = [n\varepsilon]$ where [.] denotes the integer part. Choose m rows at random; then for the respective i's put

$$x_{ij} = x_{1,ij} + Kv_{ij}, \tag{16}$$

Table 1 Components of the function $\mathbf{h}_0(s)$

i	$h_{0i}(s)$	i	$h_{0i}(s)$
1	s	6	$\exp(-10s)$
2	s^3	7	$\exp(-(s-0.3)^2)$
3	s^5	8	$(1+10s)^{-1}$
4	$(s-0.4)^2$	9	$(1+10s^2)^{-1}$
5	$(s-0.6)^4$	10	$(1+10s)^{-2}$

with v_{ij} as above, and $x_{ij} = x_{1,ij}$ for the other rows.

There are infinite possible configurations for the basic curve. We want to include both monotonic and non–monotonic functions. We take $p = 10$. In Table 1 we give the components of $\mathbf{h}_0(s)$.

We take $\sigma = 0.1$, $\varepsilon = 0$, 0.05, 0.10, 0.15 and 0.20, and the values of K in (15) and (16) between 0 and $K_{\max} = 200$, namely $K = 0, 1, \ldots, 10, 20, 30, \ldots, 200$. We choose the sample size $n = 100$ and the number of simulation replicates as $N_{\text{rep}} = 200$.

5.2 Estimators

In order to obtain useful results, we have to compare our proposal to other *robust* nonlinear PCs.

The nonlinear estimators employed here require an initial linear approximation. There are several proposals for robust linear PCs; e.g. Maronna (2005) and Hubert et al. (2003), but they are not resistant to element-wise contamination. The proposal by Candès et al. (2011) is very fast and is resistant to element-wise contamination, but not to row-wise contamination. This fact can be inferred from the definition of their estimator, and we have verified it through simulations. As initial estimator we choose Maronna and Yohai's (2008) "Perturbed MM", which is resistant to both types of contamination. The estimator proposed by Croux et al. (2003) might also be used, but the results from Maronna and Yohai's (2008) show that it is outperformed by the "Perturbed MM" method.

The estimators considered are:

- the "classical" principal curves (henceforth "Pr.Cv.") computed with the standard version of code `principal.curve` (see Sect. 2) starting from classical PCs.
- the robust Pr.Cv. computed using the robust version of `principal.curve` and starting from Spherical Principal Components (SPC) (Locantore et al. 1999), which are fast to compute and are robust against row-wise but not against element-wise contamination.
- the same robust Pr.Cv., but starting from Maronna and Yohai's (2008) "Perturbed MM" PCs.
- and our spline-based PCs with $R_{\text{knots}} = 15$, starting also from the "Perturbed MM" PCs (henceforth "Spline-based" for short).

The parameters for the "Perturbed MM" estimator were the same as in the simulation in Maronna and Yohai (2008), i.e., $\gamma = 0.5$ and $m = 5$ (see Sect. 4 thereof). As shown on Table 8 thereof, the method is reasonably fast; its computation for one component requires $O(np)$ operations.

The R code for the estimators is available from the authors upon request.

5.3 Evaluation

For a given estimator and a given simulation sample let $(\mathbf{t}, \mathbf{h})$ be the final outcome, and let $\widehat{x}_{ij} = h_{ij}(t_i)$ be the "fitted values". We want to evaluate the estimator by comparing the fit with the data. However, we should not compare $\widehat{x}_{ij}$ with the observed x_{ij}, because with this criterion, the "best" curve would be one passing through all the observations. Therefore we define our "prediction errors" by comparing the fit with the "true" curve, namely

$$z_{ij} = x_{0,ij} - \widehat{x}_{ij}. \tag{17}$$

The classical way to measure the z's is the mean squared error (MSE) equal to the average of z_{ij}^2.

We have observed that for all the estimators considered the fitted curve is good overall, except for a small proportion of points where the fit can be very poor. For our procedure this happens at the extremes, while for principal curves it happens at the regions with highest curvature. Also, for some samples the `principal.curve` code did not converge, yielding a very poor fit. We therefore believe that a robust error measure of the set $\{z_{ij},\ i = 1, \ldots, n,\ j = 1, \ldots, p\}$ may be more informative than the MSE. The one we chose is a truncated MSE similar to the τ-scale defined by Yohai and Zamar (1988). Let

$$s = \frac{1}{0.675}\text{Median}\big(|z_{ij}|, i = 1, \ldots, n,\ j = 1, \ldots, p\big),$$

and define the scale τ by

$$\tau^2 = \frac{(cs)^2}{np}\sum_{j=1}^{p}\sum_{i=1}^{n}\rho(z_{iji}/cs)^2$$

where ρ is the truncation function $\rho(t) = \min\{|t|, 1\}$ and $c = 4$.

For replications $k = 1, \ldots, N_{\text{rep}}$ call respectively MSE_k and τ_k the MSE and the τ-scale corresponding to a given estimator. Then the overall error measures for the estimator are the root MSE and the mean τ, defined as

$$\text{RMSE} = \sqrt{\frac{1}{N_{\text{rep}}}\sum_{k=1}^{N_{\text{rep}}}\text{MSE}_k}, \qquad \overline{\tau} = \frac{1}{N_{\text{rep}}}\sum_{k=1}^{N_{\text{rep}}}\tau_k.$$

Table 2 Simulation: maximum mean τ-scales of prediction errors

Contamination	ε	Principal curves			Spline
		Class.	Rob-SPC	Rob-MM	
	0	0.037	0.057	0.058	0.085
Row	0.05	0.112	0.071	0.071	0.138
	0.10	0.660	0.086	0.085	0.155
	0.15	0.971	0.102	0.098	0.180
	0.20	1.187	0.119	0.104	0.226
Element	0.05	4.212	0.163	0.157	0.150
	0.10	4.290	0.408	0.404	0.200
	0.15	4.362	0.630	0.625	0.245
	0.20	4.408	0.759	0.764	0.327

Table 3 Simulation: maximum mean RMSEs of prediction errors

Contamination	ε	Principal curves			Spline
		Class.	Rob-SPC	Rob-MM	
	0	0.04	0.10	0.12	0.35
Row	0.05	4.46	2.01	1.99	6.65
	0.10	5.93	2.49	2,54	7.80
	0.15	6.79	2.99	2.97	8.09
	0.20	7.59	3.42	3.27	10.00
Element	0.05	6.73	1.38	1.08	1.03
	0.10	6.84	2.01	1.62	1.13
	0.15	6.90	2.79	2.41	1.70
	0.20	6.98	3.34	3.01	1.75

5.4 Results

For each contamination situation, Table 2 gives the maximum mean $\overline{\tau}$ of each estimator over all values of the outlier size K in (16) and (15). The headings correspond respectively to classical Pr.Cv., robust Pr.Cv. starting from SPC, robust Pr.Cv. starting from MM, and our spline-based estimator. In all cases, the values of the mean τ increased with the outlier size K, and therefore the maxima in the table correspond to $K = K_{\max}$.

It is seen that:

- For $\varepsilon = 0$, classical Pr.Cv. is the best estimator, as expected. The two robust Pr.Cv. perform better than the Spline-based; they all have low efficiency.
- Both robust Pr.Cv. have similar behaviors. For element-wise contamination, Rob-MM is only sightly better than Rob-SPC, which shows that the basic approach of Pr.Cv is not resistant to this type of contamination, whatever the starting values.
- For row-wise contamination, Spline-based is clearly more robust than classical Pr.Cv., but the robust Pr.Cv. performs better.
- The opposite happens with element-wise contamination, as is to be expected. The difference is more visible for large ε.

These results suggest that if we need an estimator to deal with just row-wise contamination, robust Pr.Cv. starting from SPC (which is much faster than "Perturbed MM") is to be preferred. But if we want an estimator able to deal with both types of contamination, then Spline-based is preferable.

For another point of view, Table 3 gives the estimators' maximum RMSEs.

Here we observe that

- for $\varepsilon = 0$ all estimators are rather inefficient as compared to classical Pr.Cv.
- in general, robust Pr.Cv. starting from MM is slightly better than starting from SPC
- For row-wise contamination, Spline-based shows the worst performance
- For element-wise contamination, Spline-based outperforms the other estimators.

The picture given by Table 3 does not exactly coincide with that from Table 2. However, as explained above, those differences depend on the behavior of the curves at just a small proportion of points, and therefore we consider the results from Table 2 as more representative.

To illustrate the results in Tables 2 and 3 we show the results from a realization of the simulation with element-wise contamination, with $n = 100$, $\varepsilon = 0.10$ and $K = 20$. For Spline-based and Pr.Cv.-MM we compute the norms of the prediction errors: $\sqrt{\sum_{j=1}^{p} z_{ij}^2}$ with z_{ij} defined in (17).

Figures 1 and 2 show the results for element- and row-wise contamination, respectively. The first one shows that the prediction errors of Spline-based are generally smaller than those of Pr.Cv.-MM, while the second one shows the opposite picture.

6 A real data set

We analyze a data set from Ein-Dor and Feldmesser (1987) in which $p = 8$ measures of the relative performance are given for $n = 209$ CPUs. The eight columns were normalized by their MADs. Contamination was then added to the data, with rates $\varepsilon = 0$, 0.1 and 0.2. For row-wise contamination, $m = [n\varepsilon]$ rows were chosen at random and the value K was added to each of their elements; for element-wise contamination, for each element x_{ij}, with probability ε, the quantity Kz_{ij} was added, with the z_{ij} i.i.d. standard normal. In both cases K ranged between 1 and 10. The criterion was the τ-scale of the errors $x_{ij} - \widehat{x}_{ij}$ where $\widehat{x}_{ij}$ is the

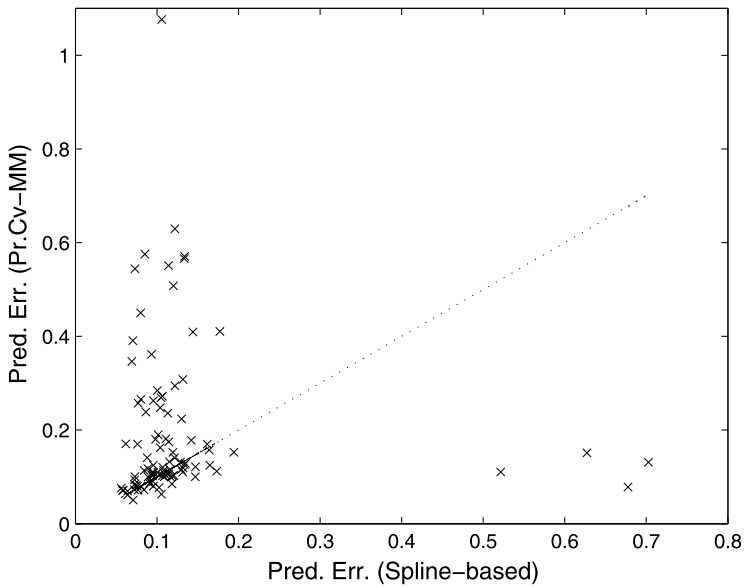


Fig. 1 Simulation: Prediction errors of Spline-based and Pr.Cv.-MM for elementwise contamination, with identity line for comparison

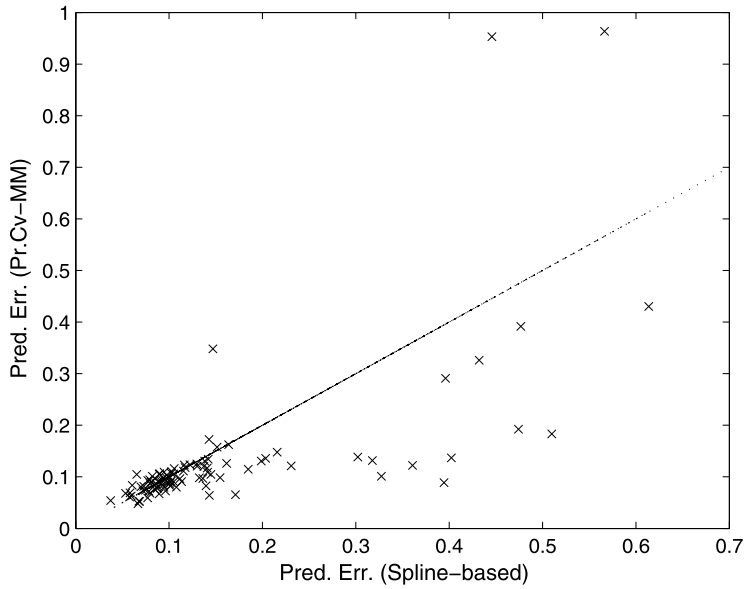


Fig. 2 Simulation: Prediction errors of Spline-based and Pr.Cv.-MM for row-wise contamination, with identity line for comparison

fitted value based on the contaminated data, and x_{ij} are the *uncontaminated* data. Besides the four estimators employed in the simulation, we added the linear classical and spherical PCs. Table 4 shows the results.

It is seen that Spline-based yields the lowest error scales in all cases. For another point of view, Table 5 shows the RMSEs.

Here for $\varepsilon = 0$ the classical Pr.Cv. clearly outperform the other estimators. For row-wise contamination, SPC is clearly the best. For element-wise contamination, SPC is the worst and the three nonlinear robust estimators have similar behaviors, with Spline-based as the best. Again, we recall that these results are heavily influenced by a small proportion of anomalous cases, and therefore we consider Table 4 as more representative.

Table 4 Computer data: maximum τ-scales of errors for row- and element-wise contamination

Cont.	ε	Class.	SPC	Principal curves			Spline
				Class	Rob-SPC	Rob-MM	
	0	1.20	1.15	0.67	0.64	0.66	0.56
Row	0.10	1.29	1.40	0.90	0.81	0.82	0.75
	0.20	1.86	1.41	1.36	1.08	1.09	1.05
Elem.	0.10	1.72	1.70	1.07	0.79	0.80	0.68
	0.20	2.82	3.59	2.64	1.60	1.35	0.87

Table 5 Computer data: maximum MSEs of errors for row- and element-wise contamination

Cont.	ε	Class.	SPC	Principal curves			Spline
				Class.	Rob-SPC	Rob-MM	
	0	2.38	2.40	1.17	1.74	1.75	2.02
Row	0.10	3.06	2.89	3.33	3.70	3.71	3.73
	0.20	4.27	3.47	4.65	4.96	4.95	4.84
Elem.	0.10	3.53	3.70	2.42	2.20	2.12	2.03
	0.20	4.59	4.67	2.98	2.44	2.46	2.24

Table 6 Running times in seconds of spline-based estimator

$n\backslash p$	10	20
50	28	30
100	50	55
200	117	133
400	316	413

7 Computing times

The running times of our proposal were computed for datasets of size $n \times 10$ generated as in Sect. 5. Then the squares of the data were added, to obtain datasets of size $n \times 20$. The times (in seconds) displayed in Table 6 are the averages of 10 runs, on a PC with an AMD Phenom II X2 560 processor with 3.30 GHz and 6 GB RAM, using an R code which is available from the authors.

It is seen that the estimator can be computed in a reasonable time for moderate datasets. An important proportion of the running time is consumed by the initial linear "Perturbed MM" estimator, which is $O(np)$, and by the search (12), which is $O(n^2 p)$. We have not been able to find an explanation for the slow increase of the running times with p.

8 Proofs of theoretical results

8.1 Proof of (18)

To prove (11), we want to show that each of the scales $S(\mathbf{r}_{\cdot j})$ in (4) decreases. Call $T(\mathbf{X}, \mathbf{y}; \boldsymbol{\beta}_0)$ a regression S estimator applied to the regression dataset $(\mathbf{X}, \mathbf{y})$ employing IRWLS starting from some $\boldsymbol{\beta}_0$. Let $\mathbf{X}_1$ and $\mathbf{X}_2$ be two predictor matrices. Let $\widehat{\boldsymbol{\beta}}_1$ be a regression estimator, and put $\widehat{\mathbf{y}}_1 = \mathbf{X}_1 \widehat{\boldsymbol{\beta}}_1$, $\boldsymbol{\gamma} = T(\mathbf{X}_2, \mathbf{y} - \widehat{\mathbf{y}}_1; \mathbf{0})$, and $\widehat{\mathbf{y}}_2 = \mathbf{X}_2 \boldsymbol{\gamma} + \widehat{\mathbf{y}}_1$. We want to show that

$$S(\mathbf{y} - \widehat{\mathbf{y}}_2) \leq S(\mathbf{y} - \widehat{\mathbf{y}}_1). \tag{18}$$

Since IRWLS ensures that the residual scale decreases at each iteration (Maronna et al. 2006, p. 328), we have $S(\mathbf{y} - \widehat{\mathbf{y}}_1 - \mathbf{X}_2 \boldsymbol{\gamma}) \leq S(\mathbf{y} - \widehat{\mathbf{y}}_1 - \mathbf{0})$, which is the same as (18).

8.2 Some auxiliary definitions and results for Sect. 4

Lemma 1 *Let $\mathbf{x} \sim \mathscr{E}(\mathbf{0}, \boldsymbol{\Sigma}, f)$ and let S be defined by (13). Then for $\mathbf{a} \in R^p$ we have $S(\mathbf{a}'\mathbf{x}) = K \mathbf{a}' \boldsymbol{\Sigma} \mathbf{a}$, where K is a constant depending on f.*

Proof Is straightforward. □

For the proofs in this section, we shall assume without loss of generality that $K = 1$.

Lemma 2 *If x has a symmetric unimodal distribution and S is an M-scale, then*

(a) *$S(x + t) > S(x)$ for all $t \neq 0$.*
(b) *If y is independent of x and $\mathrm{P}(y = 0) < 1$, then $S(x + y) > S(x)$.*

Proof Part (a) follows from Lemma 3.1 of Yohai (1987). Part (b) follows from the same lemma and conditioning on y. □

Now we deal with elliptical distributions.

For $\mathscr{E}(\mathbf{0}.\mathbf{I}, f_0)$ we characterize the one- and two dimensional marginals. Let

$$f_2(t) = \begin{cases} \int \cdots \int f_0(t + x_3^2 + \cdots + x_p^2) dx_3 \cdots dx_p \\ \quad \text{if } p > 2 \\ f_0(t) \quad \text{if } p = 2 \end{cases},$$

and

$$f_1(t) = \int f_2\big(t + x_2^2\big) dx_2.$$

If $\mathbf{x} \sim \mathscr{E}(\mathbf{0}.\mathbf{I}, f_0)$, then $(x_1, x_2)' \sim \mathscr{E}(\mathbf{0}.\mathbf{I}, f_2)$ and $x_1 \sim \mathscr{E}(0.1, f_1)$. If f_0 is decreasing, so are f_2 and f_1. For simplicity, we shall calibrate S so that $S(x) = 1$ if $x_1 \sim \mathscr{E}(0.1, f_1)$. As a consequence, if $\mathbf{x} \sim \mathscr{E}(\mathbf{0}, \boldsymbol{\Sigma}, f_0)$ and $\mathbf{a} \in R^p$, then by Lemma 1

$$S\big(\mathbf{a}'\mathbf{x}\big)^2 = \mathbf{a}' \boldsymbol{\Sigma} \mathbf{a}. \tag{19}$$

We now give a final auxiliary result:

Lemma 3 *Let $u = \mathbf{x}'\mathbf{a}$ and $v = \mathbf{x}'\mathbf{b}$ where $\mathbf{a}, \mathbf{b} \in R^p$, $\mathbf{x} \sim \mathscr{E}(\mathbf{0}, \boldsymbol{\Sigma}, f_0)$ unimodal. Call $f_{u|v}(u; v)$ the conditional density of u given v. Let*

$$u_v = v \frac{\mathbf{a}' \boldsymbol{\Sigma} \mathbf{b}}{\mathbf{b}' \boldsymbol{\Sigma} \mathbf{b}}.$$

Then $f_{u|v}(u - u_v; v)$, as a function of u, is symmetric and unimodal.

The proof is straightforward.

8.3 Proof of Proposition 1

We may assume $\boldsymbol{\mu} = 0$ without loss of generality. Let $\mathbf{r} = \mathbf{x} - \mathbf{h}(t)$, with $t = g(\mathbf{x})$. We want

$$\sum_{j=1}^{p} S(r_j)^2 = \min.$$

We assume

$$\mathbf{h}(t) = t\mathbf{b} + \mathbf{a}, \quad t = \mathbf{d}'\mathbf{x} + c, \tag{20}$$

where $\mathbf{a}, \mathbf{b}, \mathbf{c} \in R^p$ and we may assume $\|\mathbf{b}\| = 1$. Then

$$\mathbf{r} = \big(\mathbf{I} - \mathbf{b}\mathbf{d}'\big)\mathbf{x} + (\mathbf{a} - c\mathbf{b}). \tag{21}$$

For a given $j \in \{1, \ldots, p\}$ we have

$$r_j = \boldsymbol{\beta}'\mathbf{x} + \alpha, \quad \text{with } \alpha = a_j - cb_j;\ \boldsymbol{\beta} = \mathbf{e}_j - b_j\mathbf{d}. \tag{22}$$

Therefore $r_j \sim \mathscr{E}(\alpha, \boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta}, f_1)$, and hence by Lemma 2, $S(r_j)$ as a function of α is minimized for $\alpha = 0$, which obtains with $\mathbf{a} = \mathbf{0}$ and $c = 0$, and yields by Lemma 1

$$S^2(r_j) = \boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta} = \mathbf{e}_j'\boldsymbol{\Sigma}\mathbf{e}_j + b_j^2\mathbf{d}'\boldsymbol{\Sigma}\mathbf{d} - 2b_j\mathbf{e}_j'\boldsymbol{\Sigma}\mathbf{d}.$$

It follows that the criterion becomes

$$\sum_{j=1}^{p} S(r_j)^2 = \operatorname{tr}(\boldsymbol{\Sigma}) + \mathbf{d}'\boldsymbol{\Sigma}\mathbf{d} - 2\mathbf{b}'\boldsymbol{\Sigma}\mathbf{d}.$$

For each $\mathbf{d}$, this expression is minimized by $\mathbf{b} = \boldsymbol{\Sigma}\mathbf{d}/\|\boldsymbol{\Sigma}\mathbf{d}\|$, yielding $\mathbf{d}'\boldsymbol{\Sigma}\mathbf{d} - 2\|\boldsymbol{\Sigma}\mathbf{d}\|$.

Call $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_p > 0$ the eigenvalues of $\boldsymbol{\Sigma}$, and $\mathbf{v}_1, \ldots, \mathbf{v}_p$ the respective eigenvectors, so that $\boldsymbol{\Sigma} = \sum_{j=1}^{p} \lambda_j \mathbf{v}_j$, and let $q_j = \mathbf{v}_j'\mathbf{d}$. We must minimize

$$\mathbf{d}'\boldsymbol{\Sigma}\mathbf{d} - 2\|\boldsymbol{\Sigma}\mathbf{d}\| = \sum_{j=1}^{p} \lambda_j q_j^2 - 2\left(\sum_{j=1}^{p} \lambda_j^2 q_j^2\right)^{1/2}. \tag{23}$$

Differentiating with respect to q_j yields

$$\lambda_j q_j \left(1 - \frac{\lambda_j}{(\sum_{k=1}^{p} \lambda_k^2 q_k^2)^{1/2}}\right) = 0. \tag{24}$$

Let $J = \{j : q_j \neq 0\}$. Then it follows from (24) that all λ_j with $j \in J$ have the common value $\lambda = (\sum_{k=1}^{p} \lambda_k^2 q_k^2)^{1/2}$, which implies $\lambda = \lambda(\sum_{k=1}^{p} \lambda_k^2 q_k^2)^{1/2}$ and therefore $(\sum_{k=1}^{p} \lambda_k^2 q_k^2)^{1/2} = \|\mathbf{d}\| = 1$. It follows that (23) equals $-\lambda\|\mathbf{d}\|$, and this is minimized when $\lambda = \lambda_1$. □

8.4 Proof of Proposition 2

We may assume $\|\mathbf{b}_0\| = \|\mathbf{b}\| = 1$. Recall the notation (20). From (21) we get $\mathbf{r} = \mathbf{J}\mathbf{u} + \mathbf{J}\mathbf{b}_0 s - \boldsymbol{\alpha}$ with $\mathbf{J} = \mathbf{I} - \mathbf{b}\mathbf{d}'$ and $\boldsymbol{\alpha} = \mathbf{a} + c\mathbf{b}$. Put $\mathbf{K} = \mathbf{J}\mathbf{J}'$. Then we may write

$$\mathbf{r} = \boldsymbol{\mu}_s + \mathbf{v} \tag{25}$$

where $\boldsymbol{\mu}_s = \mathbf{J}\mathbf{b}_0 s - \boldsymbol{\alpha}$ and $\mathbf{v} \sim \mathscr{E}(0, \mathbf{e}_j'\mathbf{K}\mathbf{e}_j, f_1)$ is independent of s and hence of $\boldsymbol{\mu}_s$.

The criterion to be minimized is

$$C = C(\mathbf{a}, \boldsymbol{\mu}\mathbf{b}, c, \mathbf{d}) = \sum_{j=1}^{p} S\big(\mathbf{e}_j'\mathbf{r}\big)^2.$$

We will show that it is minimized at $(\mathbf{a}, \mathbf{b}, c, \mathbf{d}) = (\mathbf{0}, \mathbf{b}_0, 0, \mathbf{b}_0)$, which corresponds to $\boldsymbol{\mu}_s = \mathbf{0}$ and $\mathbf{K} = \mathbf{I} - \mathbf{b}_0\mathbf{b}_0'$.

It follows from (25) that $\mathbf{e}_j'\mathbf{r} = \mathbf{e}_j'\boldsymbol{\mu}_s' + \mathbf{e}_j'\mathbf{v}$ where $\mathbf{e}_j'\mathbf{v} \sim \mathscr{E}(0, \mathbf{e}_j'\mathbf{K}\mathbf{e}_j, f_1)$ and is independent of $\boldsymbol{\mu}_s$. Therefore Lemma 2(b) implies $S(\mathbf{e}_j'\mathbf{r}) > S(\mathbf{e}_j'\mathbf{v}) = \mathbf{e}_j'\mathbf{K}\mathbf{e}_j$, and then

$$\sum_{j=1}^{p} S\big(\mathbf{e}_j'\mathbf{r}\big)^2 > \sum_{j=1}^{p} \mathbf{e}_j'\mathbf{K}\mathbf{e}_j = \operatorname{tr}(\mathbf{K}).$$

Since $\|\mathbf{b}\| = \|\mathbf{b}_0\| = 1$, it is easy to verify that

$$\operatorname{tr}(\mathbf{K}) = p - 2\mathbf{b}'\mathbf{d} + \|\mathbf{d}\|^2 \geq p - 1 = \operatorname{tr}\big(\mathbf{I} - \mathbf{b}_0\mathbf{b}_0'\big),$$

which completes the proof. □

References

Alqallaf, F., Van Aelst, S., Yohai, V.J., Zamar, R.H.: Propagation of outliers in multivariate data. Ann. Stat. **37**, 311–331 (2009)

Bolton, R.J., Hand, D.J., Webb, A.R.: Projection techniques for non-linear principal components analysis. Stat. Comput. **13**, 267–276 (2003)

Candès, E., Li, X., Ma, Y., Wright, J.: Robust principal component analysis. J. ACM **58**(3), 11 (2011)

Cleveland, W.: Robust locally weighted regression and smoothing scatterplots. J. Am. Stat. Assoc. **74**, 829–836 (1979)

Croux, C., Filzmoser, P., Pison, G., Rousseeuw, P.J.: Fitting multiplicative models by robust alternating regressions. Stat. Comput. **13**, 23–36 (2003)

Delicado, P.: Another look at principal curves and surfaces. J. Multivar. Anal. **77**, 84–116 (2001)

Ein-Dor, P., Feldmesser, J.: Attributes of the performance of central processing units: a relative performance prediction model. Commun. ACM **30**, 308–317 (1987)

Gerber, S., Whitaker, R.: Regularization free principal curve estimation. J. Mach. Learn. Res. **14**, 1285–1302 (2013)

Hastie, T., Stuetzle, W.: Principal curves. J. Am. Stat. Assoc. **84**, 502–516 (1989)

Hubert, M., Rousseeuw, P.J., Verboven, S.: Robust PCA for high-dimensional data. In: Dutter, R., Filzmoser, P., Gather, U., Rousseeuw, P.J. (eds.) Developments in Robust Statistics, pp. 169–179. Physika Verlag, Heidelberg (2003)

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

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

Maronna, R.A., Martin, R.D., Yohai, V.J.: Robust Statistics: Theory and Methods. Wiley, New York (2006)

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

Rousseeuw, P.J., Yohai, V.J.: Robust regression by means of S estimators. In: Franke, J., Härdle, W., Martin, D. (eds.) Robust and Nonlinear Time Series Analysis. Lecture Notes in Statistics, vol. 26, pp. 256–272. Springer, New York (1984)

Tharmaratnam, K., Claeskens, G., Croux, C., Salibian-Barrera, M.: S-estimation for penalized regression splines. J. Comput. Graph. Stat. **19**, 609–625 (2010)

Tibshirani, R.: Principal curves revisited. Stat. Comput. **2**, 183–190 (1992)

Verbeek, J.J., Vlassis, N., Kröse, B.: A k-segments algorithm for finding principal curves. Pattern Recognit. Lett. **23**, 1009–1017 (2002)

Yohai, V.J.: High breakdown-point and high efficiency estimates for regression. Ann. Stat. **15**, 642–665 (1987)

Yohai, V.J., Ackerman, W., Haigh, C.: Nonlinear principal components. Qual. Quant. **19**, 53–71 (1985)

Yohai, V.J., Zamar, R.: High breakdown point estimates of regression by means of the minimization of an efficient scale. J. Am. Stat. Assoc. **86**, 403–413 (1988)