Robust Principal Component Analysis Based On Trimming Around Affine Subspaces

Christophe Croux[a], Luis A. García-Escudero[b], Alfonso Gordaliza[b,*], Christel Ruwet[c], Roberto San Martín[b]

[a] *KU Leuven, Naamsestraat 68, B3000 Leuven, Belgium.*
[b] *Departamento de Estadística Investigación Operativa, Universidad de Valladolid, Prado de la Magdalena, S/N Valladolid 47005, Spain.*
[c] *University of Liege (ULg), Grande Traverse 7, B4000 Liege, Belgium.*

Abstract

Principal Component Analysis (PCA) is a widely used technique for reducing dimensionality of multivariate data. The principal component subspace is defined as the affine subspace of a given dimension d giving the best fit to the data. However, PCA suffers from a well-known lack of robustness. As a robust alternative, one can resort to an impartial trimming based approach. Here one searches for the best subsample containing a proportion $1-\alpha$ of the observations, with $0 < \alpha < 1$, and the best d-dimensional affine subspace fitting this subsample, yielding the trimmed principal component subspace.

A population version will be given and existence of a solution to both the sample and population problem will be proven. Moreover, under mild conditions, the solutions of the sample problem are consistent toward the solutions of the population problem. The robustness of the method is studied

*Corresponding author
Email addresses: christophe.croux@econ.kuleuven.ac.be (Christophe Croux), lagarcia@eio.uva.es (Luis A. García-Escudero), alfonsog@eio.uva.es (Alfonso Gordaliza), cruwet@ulg.ac.be (Christel Ruwet), rsmartin@eio.uva.es (Roberto San Martín)

by proving quantitative robustness, computing the breakdown point, and deriving the influence functions. Furthermore, asymptotic efficiencies at the normal model are derived, and finite sample efficiencies of the estimators are studied by means of a simulation study.
Keywords: Affine Subspaces, Dimension Reduction, Orthogonal Regression, Principal components, Multivariate statistics, Robustness, Trimming
2000 MSC: 62F35, 62H25

---

1. Introduction

When analyzing multivariate data sets, one of the primary goals is to reduce the dimension of the data set at hand with a minimal loss of information. This is often a preliminary step to carry out other statistical analysis such as classification, regression fits and so on. Principal Component Analysis (PCA) is the most commonly used technique for doing this task and most practitioners of statistics are familiarized with this method due to its intuitive geometrical appealing and its implementation in most of statistical packages. As it happens with many classical statistical methods, one of the main drawbacks of PCA is the lack of robustness against the possible presence of outlying observations in the data set. There are a lot of examples in the literature showing that the presence of one single outlier, strategically placed, is enough to make classical PCA providing unreliable results.

During the past years, there have been several proposals to robustify classical PCA. Most of them use robust estimates of the covariance matrix and compute eigenvectors and eigenvalues from it. As such, Campbell (1980)

and Devlin et al. (1981) use M estimates, Croux and Haesbroeck (2000) take high breakdown point covariance matrix estimators such as the Minimum Covariance Determinant estimator and Croux, Ollila and Oja (2002) use sign and rank covariance matrices. Another approach is based on the "projection pursuit" idea, where one looks for the direction maximizing a robust measure of scale of the data projected on it (Li and Chen, 1985; Croux and Ruiz-Gazen 2005). A hybrid approach combining projection pursuit and robust covariance matrices was followed by Hubert, Rousseeuw, and Vanden Branden (2005). Robust procedures have also been developed for kernel PCA (see, e.g., Debruyne and Verdonck, 2010 and references therein) or in the learning machine literature (see, e.g., Xu, Caramanis and Sanghavi, 2012 and references there in).

In this paper one aims at retrieving directly the lower dimensional affine subspace best fitting the large majority of the data. More precisely, we are looking for the "best" subset of size $n - \lfloor n\alpha \rfloor$[1], with $0 \leq \alpha < 1$, hereby trimming a portion α of the data, and the corresponding best fitting affine subspace of a given dimension, where the goodness of fit is measured by the sum of squared Euclidean distances between the subspace and the selected observations. More formally, given a sample $\mathcal{X} = \{x_1, ..., x_n\}$ of observations in $\mathbb{R}^p$ and $0 \leq \alpha < 1$, one looks for the solution of the problem:

$$\min_{\mathcal{Y} \subset \mathcal{X},\, \#\mathcal{Y} \geq\, n - \lfloor n\alpha \rfloor} \; \min_{h \in \mathcal{A}_d(\mathbb{R}^p)} \frac{1}{\#\mathcal{Y}} \sum_{x_i \in \mathcal{Y}} \|x_i - \mathrm{Pr}_h(x_i)\|^2, \tag{1}$$

where $\mathcal{A}_d(\mathbb{R}^p)$ denotes the set of d-dimensional ($1 \leq d < p$) affine subspaces in $\mathbb{R}^p$ and $\mathrm{Pr}_h(\cdot)$ denotes the orthogonal projection on $h \in \mathcal{A}_d(\mathbb{R}^p)$. The

[1] $\lfloor x \rfloor$ represents the largest integer not greater than x.

"best" subspace according to (1) is called the *trimmed principal component subspace*. The "best" $\mathcal{Y}$ with $n - \lfloor n\alpha \rfloor$ observations is the *optimal set* which contains the observations surviving the trimming process.

Trimming procedures have revealed as a very powerful tool to robustify statistical methods. The idea of discarding a symmetric proportion of extreme observations in both sides of the sample is a very old and appealing proposal for robustifying the classical univariate sample mean. In order to overcome the implicit hypothesis of symmetry and to extend the idea of trimming to other frameworks such as multivariate estimation and regression, trimming procedures based on the idea of searching for the "best" subsample containing a fixed proportion of the data were introduced by Rousseeuw (1984, 1985). That gave raise to the well known Least Median of Squares (LMS) and Least Trimmed Squares (LTS) procedures in the robust regression context and the Minimum Volume Ellipsoid (MVE) and the Minimum Covariance Determinant (MCD) in the robust multivariate estimation context. Later on, Gordaliza (1991) stated a functional or population version of some related trimming procedures in the multivariate setting and coined the term "impartial trimming" which means that it is the data set itself which tell us the best way of trimming a fixed proportion α of the data.

The problem defined in (1) is also considered in Maronna (2005), who proposed a fast approximative algorithm to compute its solution. His paper mainly discussed computational aspects, while this paper presents a theoretical study of the trimmed principal component subspace, including existence, consistency, influence function and asymptotic variance of the estimators.

The outline of the paper is as follows. In Section 2 we state the functional

version of the problem by using trimming functions and we prove some preliminary results simplifying the problem and throwing light on the way how impartial trimming proceeds in this case. Section 3 is devoted to a general existence result, not requiring any conditions on the distribution. Consistency is proven in Section 4 for absolutely continuous random variables. Special attention is paid to the case of elliptical distributions in Section 5. Robustness aspects are considered in Section 6 including qualitative robustness, influence functions and breakdown point. We take advantage of the influence functions previously derived to obtain asymptotic variances in Section 7. Section 8 provides finite-sample efficiencies obtained by means of a simulation study. The last section contains the conclusions, while the Appendix contains all the proofs.

2. Notation and preliminary results

In this paper X is a $\mathbb{R}^p$-valued random vector defined on a probability space, β^p denotes the σ-algebra of all Borel sets in $\mathbb{R}^p$, P_X denotes the probability measure induced by X on $(\mathbb{R}^p, \beta^p)$ and $\|\cdot\|$ denotes the usual norm on $\mathbb{R}^p$. For a set $S \subset \mathbb{R}^p$, $\overline{S}$ denotes its closure, S^c its complementary set and $I_S(\cdot)$ its associated indicator function. For $1 \leq d < p$, $\mathcal{A}_d(\mathbb{R}^p)$ denotes the set of d-dimensional affine subspaces in $\mathbb{R}^p$ and for $h \in \mathcal{A}_d(\mathbb{R}^p)$, $\mathrm{Pr}_h(\cdot)$ denotes the orthogonal projection on h.

We recall the notion of "trimming function" introduced in Gordaliza (1991) and used in Cuesta-Albertos et al. (1997). Trimming functions are introduced in order to allow impartial trimming of observations and play an important technical role. For $0 \leq \alpha < 1$, $\mathcal{T}_\alpha = \mathcal{T}_\alpha(X)$ denotes the nonempty

set of trimming functions for X at level α, i.e.,

$$\mathcal{T}_\alpha = \{\tau : \mathbb{R}^p \to [0,1] \text{ measurable}, \int \tau(x) dP_X(x) = 1 - \alpha\},$$

and $\mathcal{T}_{\alpha-} = \mathcal{T}_{\alpha-}(X)$ denotes the set of trimming functions for level $0 \leq \beta \leq \alpha$,

$$\mathcal{T}_{\alpha-} = \{\tau : \mathbb{R}^p \to [0,1] \text{ measurable}, \int \tau(x) dP_X(x) \geq 1 - \alpha\} = \bigcup_{\beta \leq \alpha} \mathcal{T}_\beta.$$

Now we state a generalized version of the sample problem (1) using trimming functions instead of trimming subsets:

PROBLEM STATEMENT: For $\alpha \in (0,1)$ and $1 \leq d < p$, search for an affine subspace $h_0 \in \mathcal{A}_d(\mathbb{R}^p)$ and a trimming function $\tau_0 \in \mathcal{T}_{\alpha-}$ solution of the double minimization problem:

$$\inf_{\tau \in \mathcal{T}_{\alpha-}} \inf_{h \in \mathcal{A}_d(\mathbb{R}^p)} \frac{1}{\int \tau(x) dP_X(x)} \int \tau(x) \|x - \mathrm{Pr}_h(x)\|^2 dP_X(x). \tag{2}$$

The minimum value in (2) will be denoted $V_{d,\alpha} \equiv V_{d,\alpha}(P_X) \equiv V_{d,\alpha}(X)$.

We first state some technical results devoted to simplify the problem (2) and to make the proofs of the existence and consistency results easier. The next result guarantees the boundedness of the optimal value of the objective function in (2). We recall that all proofs can be found in the Appendix.

Lemma 1. *For any $1 \leq d < n$ and any $0 \leq \alpha < 1$, we have $V_{d,\alpha}(X) < \infty$.*

The next lemma shows that the optimal solution in (2) is characterized by a *strip*. Given $h \in \mathcal{A}_d(\mathbb{R}^p)$ and $r \geq 0$, we define the strip $S(h,r)$ around h and with radius r as

$$S(h,r) = \{x \in \mathbb{R}^p : \|x - \mathrm{Pr}_h(x)\| < r\}.$$

Lemma 2. *For any $h \in \mathcal{A}_d(\mathbb{R}^p)$ and $0 \leq \beta < 1$, let us denote*

$$r_\beta(h) = \inf\{r \geq 0 : P_X\big(S(h,r)\big) \leq 1-\beta \leq P_X\big(\overline{S}(h,r)\big)\}$$

and

$$\mathcal{T}_{h,\beta} = \{\tau \in \mathcal{T}_\beta : I_{S(h,r_\beta(h))} \leq \tau \leq I_{\overline{S}(h,r_\beta(h))}, P_X\text{-a.e.}\},$$

then, for all $\tau \in \mathcal{T}_{h,\beta}$ we have:

(a) $\int \tau(x)\|x - \mathrm{Pr}_h(x)\|^2 dP_X(x) \leq \int \tau'(x)\|x - \mathrm{Pr}_h(x)\|^2 dP_X(x)$ *for all the trimming functions $\tau' \in \mathcal{T}_\beta$*

(b) *The equality in* (a) *holds if and only if $\tau' \in \mathcal{T}_{h,\beta}$.*

Take $\tau_{h,\beta}$ any trimming function in $\mathcal{T}_{h,\beta}$. From Lemma 2 (b) it follows that

$$V_{d,\beta}(h) := \frac{1}{1-\beta}\int \tau_{h,\beta}(x)\|x - \mathrm{Pr}_h(x)\|^2 dP_X(x), \tag{3}$$

is the same for every $\tau_{h,\beta} \in \mathcal{T}_{h,\beta}$. We call (3) the *$\beta$-trimmed variation* of X around the affine subspace h. Unless necessary, no explicit reference to any particular choice in $\mathcal{T}_{h,\beta}$ will be made and the notation $\tau_{h,\beta}$ will be used for any trimming function in $\mathcal{T}_{h,\beta}$. Lemma 2 (a) says that taking another trimming function τ cannot decrease the value of (3). Hence, $\tau_{h,\beta}$, which is essentially an indicator function of the strip $S(h, r_\beta(h))$ around h, is the optimal trimming function for the problem (2).

Lemma 3. *With the same notation as in Lemma 2, if $\beta \leq \alpha$, we have:*

(a) $V_{d,\alpha}(h) \leq V_{d,\beta}(h)$;

(b) *The equality in* (a) *holds if and only if* $r_\alpha(h) = r_\beta(h)$ *and* $P_X(S(h, r_\alpha(h))) = 0$.

It follows from Lemma 3 that, in order to minimize the α-trimmed variation around h, it is strictly better to trim the exact proportion α, except in the case that all the probability mass of $\overline{S}(h, r_\alpha(h))$ is supported on its boundary. Lemma 2 and Lemma 3 together result in

Proposition 1. *For any $h \in \mathcal{A}_d(\mathbb{R}^p)$ and $0 \leq \alpha < 1$, it holds that*

$$V_{d,\alpha} = \inf_{h \in \mathcal{A}_d(\mathbb{R}^p)} V_{d,\alpha}(h).$$

The previous proposition allows us to simplify the original double minimization problem (2) to the single search of the optimal affine subspace. Once that the optimal affine subspace h is determined, the optimal trimming function is essentially the indicator function of the associated strip $S(h, r_\alpha(h))$. Any affine subspace h_0 satisfying $V_{d,\alpha}(h_0) = V_{d,\alpha}$, i.e. being a solution of the problem stated in (2), will be called a d-dimensional α-trimmed principal component subspace of X. The shorter name trimmed principal component subspace will be also used.

Note that the previous problem statement covers both the population and the sample problem. In the sample case P_X is replaced by the empirical measure P_n^ω. That is, if we have a sample $\{X_i\}_{i=1}^n$ of size n from the probability distribution P_X, the associated empirical measure is defined as

$$P_n^\omega(A) = \frac{1}{n} \sum_{i=1}^{n} I_A(X_i(\omega))$$

for ω in the sample space Ω. Now, given the outcome of a sample $X_1(\omega) = x_1, ..., X_n(\omega) = x_n$, we can see that the problem stated in (1) is equivalent to the problem (2) when taking P_n^ω instead of P_X.

3. Existence

The main goal of this section is to state the existence of solutions of problem (2). The result would guarantee the existence of solutions of both the population and the sample problem. We do not assume any moment condition on the underlying distribution. This is important in terms of robustness, because outliers are often associated with the presence of heavy tails for the underlying distribution, where moment conditions are not realistic.

From Lemma 1 and Proposition 1, we have that

$$V_{d,\alpha} = \inf_{h \in \mathcal{A}_d(\mathbb{R}^p)} V_{d,\alpha}(h) < \infty, \tag{4}$$

so we can take a sequence of subspaces $\{h_n\}_n \subset \mathcal{A}_d(\mathbb{R}^p)$ such that $V_{d,\alpha}(h_n) \downarrow V_{d,\alpha}$ as $n \to \infty$. For any affine subspace h_n in that sequence, let us denote $\tau_n = \tau_{h_n,\alpha}$, the radius $r_n = r_\alpha(h_n)$ and $S_n = S(h_n, r_n)$. Moreover, we parameterize h_n through the distance to the origin, denoted by $d_n = \inf_{x \in h_n} \|x\|$, and the choice of d unitary vectors spanning the affine subspace. The boundedness of the sequences $\{d_n\}_n$ and $\{r_n\}_n$ follows from the following lemma:

Lemma 4. *If $\{h_n\}_n$ is a sequence of affine subspaces in $\mathcal{A}_d(\mathbb{R}^p)$ satisfying $V_{d,\alpha}(h_n) \downarrow V_{d,\alpha}$ as $n \to \infty$, then $\{d_n\}_n$ and $\{r_n\}_n$ are bounded sequences.*

Furthermore, as all d sequences of unitary vectors are bounded and $\mathbb{R}^p$ is a complete space, $\{h_n\}_n$ contains a convergent subsequence in the sense that the corresponding subsequences of unitary spanning vectors, distances to the origin $\{d_n\}_n$, and the radii $\{r_n\}_n$, are all convergent. We pass to this convergent subsequence without changing notation. We now state the existence result:

Theorem 1 (Existence). *Let X be a random vector, $\alpha \in (0,1)$ and $1 \leq d < p$. Then there exists a d-dimensional α-trimmed principal component of X.*

Now that existence of the trimmed principal component subspace is established, we can formulate two important corollaries. The first one says that the optimal trimming function is essentially the indicator function of a strip whose axis is the optimal affine subspace. The second one establishes that the trimmed principal component subspace is spanned by the eigenvectors associated with the largest eigenvalues of the covariance matrix obtained with respect to the probability distribution P_X "restricted" through the optimal trimming function.

Corollary 1. *Under the hypotheses of Theorem 1, if τ_0 and h_0 are a solution of (2), then*

$$I_{S(h_0,r_\alpha(h_0))} \leq \tau_0 \leq I_{\overline{S}(h_0,r_\alpha(h_0))},\ P_X\text{-}a.e.$$

Moreover, if P_X is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^p$, then

$$I_{S(h_0,r_\alpha(h_0))} = \tau_0,\ P_X\text{-}a.e.$$

For every $\tau \in \mathcal{T}_\alpha$, let us denote P_X^τ the probability distribution induced on $\mathbb{R}^p$ by the restriction of X through the trimming function τ, i.e. for every Borel set A,

$$P_X^\tau(A) = \frac{1}{1-\alpha} \int_A \tau(x) dP_X(x).$$

Corollary 2. *Under the hypotheses of Theorem 1, if τ_0 and h_0 are a solution of (2) and the random variable X has finite second order moments, then h_0*

is the affine subspace spanned by the ordinary principal components of the probability distribution $P_X^{\tau_0}$.

If Corollary 2 would not hold, the α-trimmed variation could be strictly diminished by replacing h_0 by the affine subspace spanned by the ordinary principal components of the probability distribution $P_X^{\tau_0}$ and then τ_0 and h_0 would not be a solution of (2).

4. Consistency

While Theorem 1 guarantees the existence of solutions for the population and the sample problem, we now prove the convergence of the sample solutions to the population ones. The convergence between affine subspaces is stated as the convergence of the distances to the origin and the possible choice of a sequence of converging unitary spanning vectors. Obviously, the sequences of sample optimal radii and sample trimmed variations will then also be consistent.

In what follows, $\{X_n\}_n$ is a sequence of $\mathbb{R}^p$-valued random vectors and $h_n \in \mathcal{A}_d(\mathbb{R}^p), n = 1, 2, ...,$ are the d-dimensional trimmed principal component subspaces for X_n with associated optimal trimming function $\tau_n = \tau_{h_n,\alpha}(X_n)$ and optimal radius r_n. Moreover, $V_n := V_{d,\alpha}(X_n), n = 0, 1, 2, ...,$ denotes the trimmed variation of X_n.

The main result related to the consistency of the trimmed principal component subspace is based on a continuity result as well as on the Skorohod representation theorem. This scheme of the proof is similar to that used in Cuesta-Albertos et al. (1997) to establish consistency for trimmed k-means.

Similar as in Cuesta-Albertos et al. (1997) difficulties arise since the trimming functions have discontinuities on the boundaries of the corresponding strips. To overcome this, the continuity of the probability distribution of the limit random vector will be imposed.

As in the existence proof, the first step is to show that $\{h_n\}_n$ contains a convergent subsequence by showing that their unitary vectors, the distances to the origin $\{d_n\}_n$ and the radii sequences $\{r_n\}_n$ are bounded.

Lemma 5. *Let $\{X_n\}_n$ be a sequence of $\mathbb{R}^p$-valued random vectors such that $X_n \to X_0$, P-a.e. Then $\{d_n\}_n$ and $\{r_n\}_n$ are bounded sequences.*

The proof of this lemma is essentially the same as that of Lemma 4. One only needs to take into account that the sequence $\{X_n\}_n$ is tight. Now we are ready to formulate the "continuity" result.

Theorem 2 (Continuity). *Let $\{X_n\}_n$ be a sequence of $\mathbb{R}^p$-valued random vectors, $\alpha \in (0,1)$ and $1 \leq d < p$. Let $\{h_n\}_n \subset \mathcal{A}_d(\mathbb{R}^p)$ be the sequence of d-dimensional trimmed principal component of X_n, for $n = 1, 2, \ldots$ Assume that:*

(a) *$X_n \to X_0$, P-a.e.;*

(b) *P_{X_0} is an absolutely continuous distribution;*

(c) *h_0 is the unique d-dimensional trimmed principal component of X_0.*

Then $h_n \to h_0$ and $V_n \to V_0$ as $n \to \infty$.

We can replace the almost sure convergence condition in Theorem 2 by a convergence in distribution. By applying the a.s. Skorohod representation

theorem, there exists a sequence $\{Y_n\}_n$ of $\mathbb{R}^p$-valued random vectors such that $P_{X_0} \equiv P_{Y_0}$, $P_{X_n} \equiv P_{Y_n}$ and $Y_n \to Y_0$ $P-$a.s. Hence, by applying Theorem 2 to the sequence $\{Y_n\}_n$, it follows that

Corollary 3. *Theorem* 2 *holds if we replace condition* (a) *by*

(a′) $X_n \to X_0$ *in distribution.*

Finally, to obtain the desired consistency result, consider a sequence of independent, identically distributed random vectors $\{X_n\}_n$, with probability distribution P_X and recall that problem stated in (1) is equivalent to the problem (2) taking P_n^ω instead of P_X. Furthermore, it is well-known that the set

$$\Omega_0 := \{\omega \in \Omega \text{ such that } P_n^\omega \text{ converges in distribution to } P_X\}$$

has probability equal to 1. Thus, the desired consistency result follows as a simple consequence of Corollary 3:

Theorem 3 (Consistency). *Let* $\{X_n\}_n$ *be a sequence of independent, identically distributed* $\mathbb{R}^p$*-valued random vectors with distribution* P_X *and let* $\{P_n^\omega\}$ *be the sequence of empirical probability measures, for any* $\omega \in \Omega$*. Let us assume that* P_X *is absolutely continuous having a unique* d*-dimensional trimmed principal component subspace* $h_0 \in \mathcal{A}_d$*. If* $\{h_n^\omega\}_n$ *is a sequence of empirical* d*-dimensional trimmed principal components of* $\{P_n^\omega\}_n$*, then*

(a) $h_n^\omega \to h_0$*, P-a.s.*

(b) $V_{d,\alpha}(P_n^\omega) \to V_{d,\alpha}(X)$*, P-a.s.*

The consistency result requires the uniqueness of the d-dimensional trimmed principal component subspace, which does not hold in general. The uniqueness property may be guaranteed resorting to certain "geometrical" conditions on the probability distribution P_X. In the next section, a uniqueness result is obtained for elliptically contoured distributions.

5. Uniqueness and Fisher consistency for Elliptical distributions

In this section we focus on the interesting case of the elliptically contoured distributions. We say that a $\mathbb{R}^p$-valued random variable X follows an elliptical symmetric distribution $X \sim E_p(\mu, \Sigma)$ if it admits a probability density function of the form

$$f_X(x) = |\Sigma|^{-\frac{1}{2}} h((x-\mu)'\Sigma^{-1}(x-\mu)) \text{ for } x \in \mathbb{R}^p \tag{5}$$

where h is a positive and non-increasing square integrable function called the *radial function*. The symmetric positive definite matrix Σ is called the *scatter matrix*, and is proportional to the covariance matrix if the distribution has a second moment. The ordered eigenvalues of Σ will be denoted by $\lambda_1 \geq ... \geq \lambda_p > 0$ and the associated eigenvectors will be $v_1, ..., v_p$, respectively. The *location parameter* of the distribution is μ. The density f is called unimodal if the radial function h has a strictly positive derivative $\dot{h}$.

The proof of the uniqueness result for elliptically contoured distributions needs the application of the following multivariate probability inequality, whose proof can be found in Davies (1987):

Lemma 6. *Let $\mu \in \mathbb{R}^p$ and Σ be a symmetrix positive definite matrix. Let*

ξ and $g : \mathbb{R}^+ \to \mathbb{R}^+$ be nonincreasing functions with $\int g(x'x)dx < \infty$. Then

$$\int \xi((x-\mu)'\Sigma^{-1}(x-\mu))g(x'x)dx \leq \int \xi(x'\Sigma^{-1}x)g(x'x)dx.$$

To have uniqueness we need an additional restriction on the eigenvalues. There needs to be a difference between λ_d and λ_{d+1}, where d is the dimension of the affine subspace we are looking for. The other eigenvalues may coincide. This condition guarantees that the space spanned by the first d eigenvectors of Σ is uniquely determined.

Theorem 4 (Uniqueness). *Let X be a random vector having an elliptically symmetric distribution as in* (5)*, with unimodal density. Let $\lambda_1 \geq ... \geq \lambda_p > 0$ be the eigenvalues of Σ satisfying $\lambda_d > \lambda_{d+1}$. Then,*

(a) *For every $\alpha > 0$ and every $d < p$, the d-dimensional trimmed principal component subspace of X is unique. That subspace passes through μ and is spanned by the d largest eigenvectors of the matrix Σ.*

(b) *If X has finite second order moments, then the trimmed d-dimensional principal component subspace coincides with the ordinary principal component subspace of dimension d.*

The theorem above tells us that, at any elliptically symmetric distribution, the trimmed principal component subspace passes through the location parameter μ and it is spanned by the largest d eigenvectors of the scatter matrix Σ. If the second moment exist, then Σ is proportional to the covariance matrix and, therefore, the principal axis corresponding to the trimmed principal components are the same as those obtained by using the standard PCA.

We also give a Fisher consistency result for elliptical contoured distributions. At this point, some functional notations are needed. To avoid notational complexity, we omit the reference to the random vector X in the notation P_X by just writing P. For a given distribution P with density as in (5), let us denote by $S(P)$ the optimal strip associated with the trimmed principal component subspace. By Theorem 4 and the hypothesis on the eigenvalues of Σ, this strip is centered at μ and has the first d eigenvectors of Σ as spanning vectors. We define the functional giving us the average over this space

$$m(P) = \frac{1}{1-\alpha} \int_{S(P)} x dP(x).$$

Analogously, we introduce the (restricted) covariance matrix

$$C(P) = \frac{1}{1-\alpha} \int_{S(P)} (x - m(P))(x - m(P))' dP(x). \tag{6}$$

Due to orthogonal and translation equivariance of the loss function defining the optimal strip, these functionals are orthogonal and translation equivariant. Based on this property, we restrict our attention to elliptical distributions centered at the origin and with diagonal scatter matrix, i.e. $\mu = 0$ and Σ is a diagonal matrix. In this case, it is easy to see that $m(P) = 0$ and $C(P)$ is diagonal.

Theorem 5 (Fisher consistency). *Let P be with density as in (5). If we assume finite second order moments, then there exists a real constant c depending only on the distribution P via the radial function h and the trimming constant α, such that the first d eigenvalues and eigenvectors of*

$$c\, C(P)$$

are equal to the first d eigenvalues and eigenvectors of the covariance matrix of P. At the multivariate normal distribution, one has $c = 1$

The above property is called Fisher consistency of the the first d eigenvalues and eigenvectors of $C(P)$. To get this Fisher consistency, the matrix C needs to be multiplied by a constant c. In the sequel, the functional C will always be be multiplied by this consistency factor c. At the multivariate normal distribution, no such correction is needed, but at other types of elliptical distributions c may be different from zero.

6. Robustness

In this section, to avoid notational complexity, we keep on omitting the reference to the random vector X in the notation P_X by just writing P.

6.1. Qualitative Robustness

Hampel (1971) introduces the qualitative robustness of a sequence of estimators $\{T_n\}_{n=1}^{\infty}$ as the equicontinuity of the mappings $\{P \to \mathcal{L}_P(T_n)\}_{n=1}^{\infty}$, where $\mathcal{L}_P(T_n)$ denotes the distribution of the estimator T_n under the distribution P.

Hampel (1971) also defined a "continuity" condition for a sequence of estimators at a distribution F. If T_n is such that $T_n = T(P_n^{\omega})$ with P_n^{ω} the empirical distribution, the continuity condition is analogous to that of T being a weak continuous functional. If we have a sequence of distributions Q_n, $n = 1, 2, ...$, converging weakly to P, we can obtain through the Skorohod Representation theorem some random vectors Z_n, $n = 1, 2, ...$, and Z_0 with distributions Q_n, $n = 1, 2, ...$, and P, respectively, and converging almost

surely. Thus, we can apply Theorem 2 to the sequence $\{Z_n\}_{n=0}^{\infty}$ to obtain the weak continuity.

Lemma 7. *The d-dimensional trimmed principal component subspace functional is weakly continuous at any absolutely continuous distribution P admitting an unique d-dimensional trimmed principal component subspace.*

If P^n denotes the product measure on $\mathbb{R}^{n\times p}$, the weak continuity together with the continuity of T_n as a point function on $\mathbb{R}^n$, except for a set of P^n-measure 0, would imply the qualitative robustness of T_n (Theorem 1.a in Hampel 1971). In our case, the weak continuity follows from Lemma 7 and the point continuity is achieved, except perhaps in those points where we have (at least) two optimal subsets of the sample $\mathcal{X}$ reaching the same minimum value in expression (2). However, for absolutely continuous distributions with respect to the Lebesgue measure, those points are a finite union of P^n-measure 0 zones, so those points have null P^n-measure.

Theorem 6. *The d-dimensional trimmed principal component subspace functional is qualitatively robust under the assumptions of Lemma 7.*

Notice that we need an uniqueness condition. This condition may be seen as being similar to that of the uniqueness of the population median in stating the qualitative robustness of the median estimator. A similar uniqueness condition was needed to state the qualitative robustness of the trimmed k-means estimator in García-Escudero and Gordaliza (1999).

6.2. Influence function

The influence function is the keystone of Hampel's infinitesimal approach to Robust Statistics (Hampel 1974 and Hampel et al. 1986), providing a very

rich information about the robustness of an estimator. It is also a useful tool for exploring asymptotic variances. Thus, to further investigate the robustness and asymptotic properties of the trimmed principal component subspace estimator, we compute its influence function, for the eigenvalues and eigenvectors, for elliptical contoured distributions. The main ideas will follow Croux and Haesbroeck (1999). The IF of a functional T at a distribution P is given by

$$IF(x_0; T, P) = \lim_{\varepsilon \downarrow 0} \frac{T((1-\varepsilon)P + \varepsilon\delta_{\{x_0\}}) - T(P)}{\varepsilon},$$

for those x_0 where this limit exists. Here $\delta_{\{x_0\}}$ denotes a Dirac distribution putting al its mass at x_0.

For deriving the influence function of the eigenvectors and eigenvalues at elliptical distributions, we first need the influence function for the functional C, defined in (6). For $j = 1, \ldots, p$, we denote by $\Lambda_j(P)$ and $V_j(P)$ the jth eigenvalue and eigenvector of $C(P)$. Thanks to the orthogonal and translation equivariance of the functional, we may assume that $\mu = 0$ and take Σ diagonal. The following result is proven in the Appendix.

Theorem 7. *At an elliptical distribution function P with probability density function given by* (5)*, with $\mu = 0$, and $\Sigma = diag(\lambda_1, \ldots, \lambda_p)$, we have that for a diagonal term of C:*

$$IF(x_0; C, P)_{ii} = \frac{c}{1-\alpha} I_{S(P)}(x_0) \left(x_{0i}^2 - \frac{A_{ii}}{G} \right) - \Lambda_i(P) + \frac{cA_{ii}}{G}$$

and for an off-diagonal term ($i \neq j$)

$$IF(x_0; C, P)_{ij} = -\frac{(\Lambda_j(P) - \Lambda_i(P))\lambda_i\lambda_j}{2(\lambda_j - \lambda_i)} \frac{I_{S(P)}(x_0)x_{0i}x_{0j}}{H_{ij}}.$$

The quantities G, A_{ii} *and* H_{ij} *are defined in the Appendix, see formulas* (A.16),(A.14), *and* (A.20).

We note that the influence functions are not bounded. This come from the unboundedness of the strip $S(P)$ along the first d eigenvectors of $C(P)$. However, the influence function reveals that only good leverage points, i.e. outliers in the direction of the first d eigenvectors and still belonging to $S(P)$, may have huge influence. On the other hand, bad outliers have bounded influence, and are even redescending to zero for the non diagonal elements. The influence function is alike the one of the classical estimator for contaminations close to the subspace span of the first d eigenvectors.

Using the above theorem, one readily obtains the influence functions for eigenvectors and eigenvalues of C. Indeed, for Σ diagonal, Lemma 3 of Croux and Haesbroeck (2000) yields

$$IF(x_0, V_{ji}, P) = \frac{IF(x_0, C, P)_{ji}}{\Lambda_i(P) - \Lambda_j(P)}(1 - \delta_{ij})$$

where δ_{ij} is a boolean that takes value 1 when $j = i$, and the corresponding result for eigenvalues

$$IF(x_0, \Lambda_i, P) = IF(x_0, C, P)_{ii}.$$

The case of eigenvalues is therefore immediate

$$IF(x_0, \Lambda_i, P) = \frac{c}{1-\alpha} I_{S(P)}(x_0) \left(x_{0i}^2 - \frac{A_{ii}}{G} \right) - \Lambda_i(P) + \frac{cA_{ii}}{G}, \qquad (7)$$

for $1 \leq i \leq p$. For an eigenvector V_i, with $1 \leq i \leq p$, we have that the influence function of its ith component is zero, while for component $j \neq i$

$$IF(x_0, V_i, P)_j = \frac{\lambda_j \lambda_i}{\lambda_j - \lambda_i} \frac{I_{S(P)}(x_0) x_{0i} x_{0j}}{2H_{ij}}.$$

In another form

$$IF(x_0, V_i, P) = \sum_{j \neq i} \frac{\lambda_i \lambda_j}{\lambda_j - \lambda_i} \frac{I_{S(P)}(x_0) x_{0i} x_{0j}}{2H_{ij}} v_j, \tag{8}$$

with v_j the jth eigenvector of Σ.

To conclude this section, Figures 1 and 2 picture the influence functions of the largest eigenvalue and its associated eigenvector for a bivariate normal distribution with zero mean and covariance matrix $\Sigma = \text{diag}(2, 1)$. Furthermore, we take $d = 1$. Only the non-zero component of the influence function of the eigenvector, i.e. only the second component, is represented. We make plots of the IF for the functionals with $\alpha = 0$ (left panel - no trimming) and $\alpha = 0.01$ (right panel - 1% trimming).

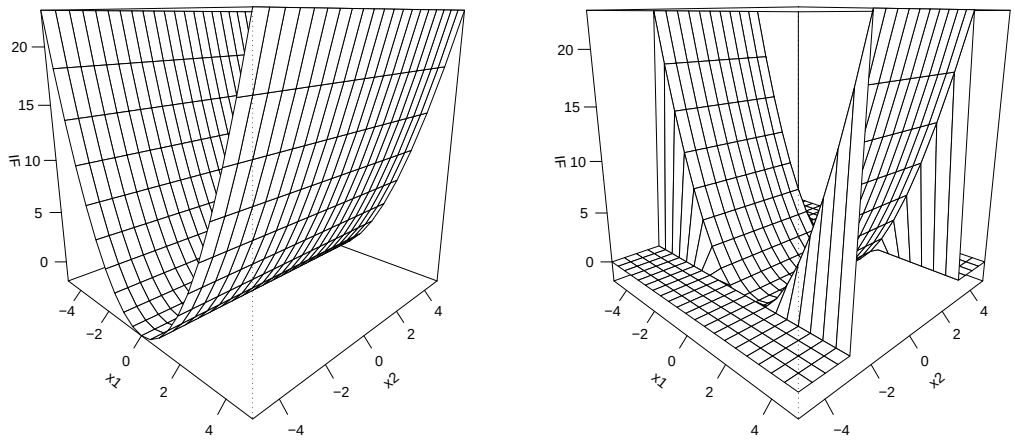

Figure 1: Influence function of the largest eigenvalue at $P = N(0, \text{diag}(2, 1))$ when $\alpha = 0$ (left panel) and $\alpha = 0.01$ (right panel).

Inside the strip $S(P)$, which is here given by $S(P) = \{x_2 | x_2^2 \leq r^2(P)\}$,

the influence function for the untrimmed and the trimmed influence functions have a similar behavior. But outside the optimal strip the influence of the "trimmed" eigenvalue becomes zero, and bounded for the "trimmed" eigenvectors. For the untrimmed or classical eigenvectors and eigenvalues, the influence functions goes beyond all bounds, also outside the optimal strip. The plots illustrate that the trimmed principal components bound the influence of bad leverage points (outside the optimal strip), while they still give unbounded influence to good leverage points. The latter property ensures that the loss in statistical efficiency due to the trimming remains limited, as will be further explored in Section 7.

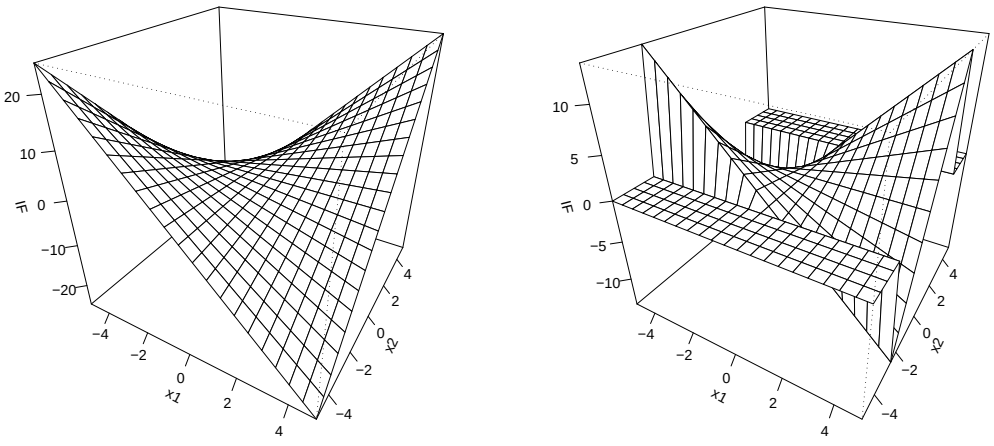

Figure 2: Influence function of the eigenvector associated to the largest eigenvalue at $P = N(0, \mathrm{diag}(2, 1))$ when $\alpha = 0$ (left panel) and $\alpha = 0.01$ (right panel).

6.3. Breakdown Point

As it is well known, the influence functions provides just a local description of the behavior of a functional at a probability model and we always need to complement this description with a measure of global reliability. This complementary measure is the breakdown point, that provides a measure of how far from the model the good properties derived from the influence functions of the estimator can be expected to extend. We will consider Donoho and Huber's (1983) sample version. Given $\mathcal{X} = \{x_1, ..., x_n\}$ a sample of n points and T an estimator based in that sample, let us denote by $\varepsilon_n^*(T, \mathcal{X})$ the smallest fraction of corrupted observations needed to breakdown the estimator T, i.e.

$$\varepsilon_n^*(T, \mathcal{X}) = \min \left\{ \frac{k}{n};\ \sup_{\mathcal{X}'} \|T(\mathcal{X}) - T(\mathcal{X}')\| = \infty \right\},$$

with $\mathcal{X}'$ ranging on the set of all possible samples obtained by replacing k original data points in the sample $\mathcal{X}$ by arbitrary ones.

We consider the "distance to the origin" of the empirical optimal trimmed principal component subspace based on the sample $\mathcal{X}$. If $h_{\mathcal{X}}$ denotes the empirical optimal subspace for the sample, the distance to origin is $D(\mathcal{X}) := \inf_{x \in h_{\mathcal{X}}} \|x\|$, and, we would say that the procedure breaks down when $D(\mathcal{X}')$ can be made arbitrarily large.

It is not difficult to see that for the "distance to the origin" estimator associated with classical Principal Components Analysis it suffices to replace $d + 1$ data points strategically placed in order to obtain an affine subspace whose distance to the origin is arbitrarily large. Hence $\varepsilon_n^*(T, \mathcal{X}) = (d+1)/n$, which asymptotically reaches the worst possible value 0, showing the lack of robustness of the classical estimator.

For the trimming based method, the next result shows that the breakdown point of the "distance to the origin" estimator is asymptotically equal to α. Maronna (2005) also analyzed the breakdown point for an alternative estimator. His results are coincident with Theorem 8 but his proof mainly applies to estimators based on minimizing M-scales. We give in the Appendix a detailed proof for the "distance to the origin" estimator resulting from trimmed principal components.

Theorem 8. *Let $\alpha \in (0, 1/2]$ and $1 \leq d < p$. The breakdown point of the "distance to the origin" estimator D, at any p-dimensional sample $\mathcal{X}$ is*

$$\varepsilon_n^*(D, \mathcal{X}) = \min\left\{\frac{\lfloor n\alpha \rfloor + d + 1}{n}, \frac{n - \lfloor n\alpha \rfloor}{n}\right\}. \tag{9}$$

One has $\varepsilon_n^*(D, \mathcal{X}) \to \alpha$ as $n \to \infty$.

7. Asymptotic variances

In this section, we will use similar notation as in Subsection 6.2. Under the hypothesis that a functional T is Frechet differentiable, its asymptotic distribution is gaussian, and its asymptotic variance is given by

$$\mathrm{ASV}(T, P) = \int_{\mathbb{R}^d} IF(x, T, P) IF(x, T, P)' dP(x)$$

The question of Frechet differentiability of the functionals is not addressed in this paper.

7.1. Asymptotic variances in the elliptical case

For an elliptical contoured distribution with $\mu = 0$ and $\Sigma = diag(\lambda_1, \ldots, \lambda_p)$, expression (7) gives

$$\mathrm{ASV}(\Lambda_i, P) = \int_{\mathbb{R}^d} \left\{\frac{c}{1-\alpha} I_{S(P)}(x)\left(x_i^2 - \frac{A_{ii}}{G}\right) - \Lambda_i(P) + \frac{cA_{ii}}{G}\right\}^2 |\Sigma|^{-\frac{1}{2}} h(x'\Sigma^{-1}x) dx$$

$$= \frac{c^2}{(1-\alpha)^2}\int_{S(P)} x_i^4 |\Sigma|^{-\frac{1}{2}} h(x'\Sigma^{-1}x)dx + \left[\frac{-c\alpha A_{ii}}{(1-\alpha)G} - \Lambda_i(P)\right]$$

$$\cdot \frac{2c}{1-\alpha}\int_{S(P)} x_i^2 |\Sigma|^{-\frac{1}{2}} h(x'\Sigma^{-1}x)dx + \frac{\alpha}{1-\alpha}\left(\frac{cA_{ii}}{G}\right)^2 + \Lambda_i(P)^2$$

$$= \frac{c^2}{(1-\alpha)^2}\int_{S(P)} x_i^4 |\Sigma|^{-\frac{1}{2}} h(x'\Sigma^{-1}x)dx - \Lambda_i(P)^2 + \frac{\alpha}{1-\alpha}\left(\frac{cA_{ii}}{G}\right)^2 + 2\Lambda_i(P)\frac{cA_{ii}}{G}\left(\frac{-\alpha}{1-\alpha}\right).$$

For the eigenvectors, using (8) results in

$$\begin{aligned}\mathrm{ASV}(V_i, P) &= \int_{\mathbb{R}^d}\left(\sum_{j\neq i}\frac{\lambda_i\lambda_j}{\lambda_j-\lambda_i}\frac{I_{S(P)}(x)x_ix_j}{2H_{ij}}v_j \cdot \right. \\ &\qquad \left. \sum_{k\neq i}\frac{\lambda_i\lambda_k}{\lambda_k-\lambda_i}\frac{I_{S(P)}(x)x_ix_k}{2H_{ik}}{v_k}'\right)dP(x).\end{aligned}$$

By symmetry of $S(P)$ and P, the terms for $j \neq k$ integrate to zero. Hence there remains

$$\mathrm{ASV}(V_i, P) = \sum_{j\neq i}\frac{\lambda_i^2\lambda_j^2}{(\lambda_i-\lambda_j)^2}\frac{\int_{S(P)} x_i^2x_j^2dP(x)}{4H_{ij}^2}v_j{v_j}'. \tag{10}$$

7.2. Asymptotic relative efficiencies in the gaussian case

Using the preceding results, one may obtain information on the efficiency of the estimators of the eigenvectors and eigenvalues of C computed after trimming. We restrict our attention here to gaussian distributions, where further simplifications in the expressions derived for the asymptotic variances can be made. Furthermore, we only consider the first d eigenvalues and eigenvectors (which are also the only once retained in practical data analysis).

In Section 5 we showed that the consistency factor c is equal to 1 for the d first eigenvalues, and that $\Lambda_i(P) = \lambda_i$. By definition of G and A, see (A.14) and (A.16), and by using the property that the marginals of a multivariate

normal are independent, we have that

$$\frac{A_{ii}}{G} = \frac{\int_{\mathbb{R}} y^2 \phi(y/\sqrt{\lambda_i}) dy}{\int_{\mathbb{R}} \phi(y/\sqrt{\lambda_i}) dy} = \lambda_i,$$

where $\phi(\cdot)$ denotes the probability density function of the standard normal. These results allow for simpler expressions of the asymptotic variance of the eigenvalues with $1 \leq i \leq d$:

$$\begin{aligned}
\mathrm{ASV}(\Lambda_i, P) = & \frac{1}{(1-\alpha)^2} \int_{S(P)} x_i^4 |\Sigma|^{-\frac{1}{2}} h(x'\Sigma^{-1}x) dx \\
& - \lambda_i^2 + \frac{\alpha}{1-\alpha} \left[\left(\frac{A_{ii}}{G} \right)^2 - 2\lambda_i \frac{A_{ii}}{G} \right] \\
= & \frac{1}{(1-\alpha)^2} \int_{S(P)} x_i^4 dP(x) - \lambda_i^2 \frac{1}{1-\alpha} = \frac{1}{1-\alpha} \lambda_i^2 (3-1) = \frac{2}{1-\alpha} \lambda_i^2.
\end{aligned} \tag{11}$$

For the eigenvectors with $1 \leq i \leq d$, the definition of H_{ij} in (A.20) together with the fact that, under the gaussian assumption, we have $\dot{h}\left(y'\Sigma^{-1}y\right) = -\frac{1}{2} h\left(y'\Sigma^{-1}y\right)$ gives

$$H_{ij} = -\frac{1}{2} \int_{S(P)} x_i^2 x_j^2 f(x) dx. \tag{12}$$

Inserting (12) in the expression for the asymptotic variance (10) gives

$$\mathrm{ASV}(V_i, P) = \sum_{j \neq i} \frac{\lambda_i^2 \lambda_j^2}{(\lambda_i - \lambda_j)^2} \frac{1}{\int_{S(P)} x_i^2 x_j^2 dP(x)} v_j {v_j}'.$$

Now, since $i \leq d$, we have

$$\int_{S(P)} x_i^2 x_j^2 dP(x) = \lambda_i \lambda_j \frac{1-\alpha}{c_j},$$

with

$$c_j^{-1} = \frac{\int_{S(P)} x_j^2 dP(x)}{(1-\alpha)\lambda_j} \tag{13}$$

for $1 \leq j \leq p$. We finally obtain

$$\mathrm{ASV}(V_i, P) = \frac{1}{1-\alpha} \sum_{j \neq i} \frac{\lambda_i \lambda_j c_j}{(\lambda_i - \lambda_j)^2} v_j {v_j}'. \tag{14}$$

The availability of asymptotic variances under closed form expression allows us to compute asymptotic relative efficiencies (ARE) with respect to maximum likelihood (ML) estimators at the gaussian model. Those are defined by

$$\mathrm{ARE}(\Lambda_i, P) = \frac{\mathrm{ASV}(\Lambda_{ML;i}, P)}{\mathrm{ASV}(\Lambda_i, P)} \text{ and } \mathrm{ARE}(V_i, P) = \frac{\mathrm{trace}(\mathrm{ASV}(V_{ML;i}, P))}{\mathrm{trace}(\mathrm{ASV}(V_i, P))},$$

for $1 \leq i \leq d$. Note that the ML estimator is the untrimmed PCA, and its asymptotic variances are given by the above expressions for $\alpha = 0$. So it follows from (11) that

$$\mathrm{ARE}(\Lambda_i, P) = \frac{2}{2/(1-\alpha)} = 1 - \alpha,$$

meaning that the efficiency is just given by the trimming proportion for the first d eigenvalues. A trimming level of 10% yields a 90% efficiency for the eigenvalue estimators.

Regarding eigenvectors, we have from (14)

$$\mathrm{ARE}(V_i, P) = \frac{\sum_{j \neq i} \frac{\lambda_j}{(\lambda_i - \lambda_j)^2}}{\frac{1}{1-\alpha} \sum_{j \neq i} \frac{\lambda_j c_j}{(\lambda_i - \lambda_j)^2}}.$$

We evaluate the above expression for the spherical noise situation, where the $p - d$ last eigenvalues are assumed to be equal, say, to λ. Observations generated by a spherical noise model are lying in the same subspace, with some spherical noise added. Using (13), one can readily see that $c_j = 1$ for $j \leq d$, and $c_j = \tilde{c}$ for $j > d$, with $\tilde{c}^{-1} = E[Z_1^2 I(\|Z\| \leq \tilde{r})]$ and $\tilde{r}^2$

the $1-\alpha$ quantile of a chi-square distribution with $p-d$ degrees of freedom. The constant $\tilde{c}$ is the same as the consistency factor needed for the Minimum Covariance determinant estimator computed in Croux and Haesbroeck (1999, p.165). We get

$$\mathrm{ARE}(V_i, P) = (1-\alpha)\frac{\sum_{j\neq i, j\leq d}\frac{\lambda_j}{(\lambda_i-\lambda_j)^2} + (p-d)\frac{\lambda}{(\lambda_i-\lambda)^2}}{\sum_{j\neq i, j\leq d}\frac{\lambda_j}{(\lambda_i-\lambda_j)^2} + (p-d)\tilde{c}\frac{\lambda}{(\lambda_i-\lambda)^2}} \tag{15}$$

This result calls for a few remarks. Globally, the efficiency is again determined by the trimming proportion. But here, other effects appear. For instance (i) If the the noise level tends to zero, or $\lambda \downarrow 0$, the efficiency tends to $1-\alpha$; (ii) If the eigenvalue λ_i gets closer to the noise level λ, the efficiency decreases to $(1-\alpha)/c$. Adding noise tends to decrease the efficiency of the trimmed principal components; (iii) If the space dimension p rises for fixed model dimension d, the efficiency reaches $1-\alpha$ for very high space dimensions, since $\tilde{c}$ tends to 1 with p going to infinity; (iv) If, everything else being fixed, the model dimension d rises, numerical computations show that the efficiency increases in almost all scenarios (except for high trimming levels and low initial noise dimension).

8. Simulations

This section studies the finite sample efficiency of the trimmed PCA. The simulation experiment consists of $m = 1000$ replications of p-dimensional samples of size n with $p = 5$ or $p = 8$ and $n = 50, 100, 500$ or 1000. The samples were generated according to a normal distribution with a zero mean and a diagonal covariance matrix $\Sigma = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$. Two sets of diagonal elements were considered, similar as in Maronna (2005):

(a) one representing a smooth decrease of the eigenvalues, i.e. $\lambda_j = 2^{p-j}$ for $1 \leq j \leq p$;

(b) one representing an abrupt decrease of the eigenvalues after λ_d, i.e. $\lambda_j = 20(1 + 0.5(d - j + 1))$ for $1 \leq j \leq d$ and $\lambda_j = 1 + 0.1(p - j + 1)$ for $d + 1 \leq j \leq p$.

For each dataset, the d-dimensional α-trimmed PCA method was applied with $d = 3, 4$ or 7 and $\alpha = 0.05$ or 0.1.

The computation of the empirical d-dimensional α-trimmed PC has a high computational complexity, since one needs to optimize over the space of all subsets of a given size. Exact algorithms are, in general, no longer possible. In the simulation study that follows, the approximative algorithm of Maronna (2005) is used. This algorithm follows the rationale behind the fast-MCD algorithm in Rousseeuw and van Driessen (1999) for computing the Minimum Covariance Determinant (MCD) estimator, combining random starts and so-called "concentration" steps. We recommend to take the number of initial random starts equal to 500, and the number of concentration steps equal to 10.

To assess the performance of the estimators of the eigenvalues and eigenvectors, mean squared error (MSE) were computed. For the eigenvalues, a correction for bias is first applied and then the classical definition of MSE is used :

$$\text{MSE}(\Lambda_j) = \frac{1}{m} \sum_{i=1}^{m} (\hat{\hat{\lambda}}_j^{(i)} - \lambda_j)^2$$

where $\hat{\hat{\lambda}}_j^{(i)} = \hat{\lambda}_j^{(i)} \times \left(\frac{1}{m} \sum_{k=1}^{m} \hat{\lambda}_j^{(k)} / \lambda_j \right)^{-1}$ and $\hat{\lambda}_j^{(i)}$ is the estimate of λ_j computed from the ith generated sample. For the eigenvectors, following Croux,

Ollila and Oja (2002), the MSE is defined as

$$\mathrm{MSE}(V_j) = \frac{1}{m}\sum_{i=1}^{m}\left(\cos^{-1}|v_j^t\hat{v}_j^{(i)}|\right)^2$$

where $\hat{v}_j^{(i)}$ is the estimate of v_j computed from the ith generated sample.

Based on these MSE values, relative finite sample efficiencies were computed as

$$\mathrm{Eff}_n(\Lambda_j) = \frac{\mathrm{ASV}(\Lambda_{ML;j}, P)}{n\,\mathrm{MSE}(\Lambda_j)} \text{ and } \mathrm{Eff}_n(V_j) = \frac{\mathrm{ASV}(V_{ML;j}, P)}{n\,\mathrm{MSE}(V_j)}.$$

These finite sample efficiencies are reported in Tables 1 and 2. Since the efficiencies for the different eigenvalues of a particular setting are quite similar, their average value is reported. In this table, the asymptotic relative efficiencies derived in the previous section appear in the rows referred as "$n = \infty$".

We first discuss the results for the model with smoothly decreasing eigenvalues. As we can see from Table 1, the efficiency decreases with an increasing trimming size. The finite sample efficiency of the eigenvalues seems to decrease towards the asymptotic value, while they increase for the eigenvectors towards the limit value with increasing sample size. If the model dimension d increases, everything else being fixed, one observes a small increase in the efficiency of the eigenvectors. When it is the space dimension p that increases, except for small sample sizes, the efficiency of the eigenvectors increases. These last two behaviors have already been pointed out when studying the asymptotic efficiencies.

Under scenario (b), there is a large difference between the noise and non-noise levels. The results in Table 2 show that some finite sample efficiencies

Table 1: Finite sample efficiencies of the eigenvalues and eigenvectors of the trimmed PCA method w.r.t. the ML method under design (a).

p	d	α	n	Eigen values	Eigen vectors						
5	3	.05	50	.992	.754	.677	.590				
			100	.979	.918	.845	.710				
			500	.942	.927	.900	.852				
			∞	.950	.935	.927	.869				
5	3	.10	50	.985	.652	.608	.502				
			100	.912	.762	.782	.650				
			500	.905	.828	.809	.710				
			∞	.900	.874	.861	.768				
5	4	.10	50	.961	.668	.622	.552	.472			
			100	.948	.747	.738	.718	.614			
			500	.892	.877	.894	.858	.693			
			∞	.900	.886	.881	.853	.713			
8	3	.10	50	.977	.620	.596	.521				
			100	.947	.776	.725	.669				
			500	.908	.871	.876	.751				
			∞	.900	.883	.876	.823				
8	7	.10	50	.972	.642	.595	.587	.556	.523	.516	.462
			100	.914	.735	.765	.774	.737	.726	.701	.614
			500	.897	.856	.890	.901	.860	.832	.814	.745
			∞	.900	.898	.898	.897	.893	.884	.855	.717

Table 2: Finite sample efficiencies of the eigenvalues and eigenvectors of the trimmed PCA method w.r.t. the ML method under design (b).

p	d	α	n	Eigenvalues	Eigenvectors						
5	2	.05	100	1.275	.697	.642	.642				
			500	1.040	.688	.702	.851				
			1000	.951	.899	.927	.967				
			∞	.950	.950	.950	.949				
5	3	.10	100	1.196	.686	.593	.546				
			500	.919	.689	.665	.713				
			1000	.923	.831	.829	.882				
			∞	.900	.899	.899	.898				
5	4	.10	100	1.179	.667	.628	.592	.640			
			500	.955	.729	.718	.789	.846			
			1000	.894	.833	.809	.823	.858			
			∞	.900	.899	.900	.899	.896			
8	3	.10	100	1.426	1.168	1.168	.974				
			500	1.167	.597	.560	.607				
			1000	.995	.695	.674	.737				
			∞	.900	.900	.900	.900				
8	7	.10	100	1.594	1.196	1.106	.914	.734	.612	.562	.599
			500	1.087	.644	.590	.572	.608	.676	.760	.782
			1000	.978	.716	.668	.689	.745	.774	.848	.896
			∞	.900	.900	.900	.900	.900	.900	.899	.897

even become larger than one, showing the good performance of the trimming approach to find the best affine subspace. The convergence towards the asymptotic efficiencies is here slower than for simulation design (a).

9. Conclusions

Principal Component Analysis (PCA) is a technique for reducing dimensionality in multivariate data analysis. For p-dimensional observations, and a given dimension d, with d typically much lower than p, classical PCA yields the best fitting affine subspace of dimension d, in the sense of minimizing the sum of squared Euclidean distances between the subspace and the observations. The robust alternative studied in this paper relies on an impartial trimming based approach, where a proportion α of the observations is discarded, and the best fitting d-dimensional affine subspace is determined from the non-discarded observations. The difficulty is to find this "best" subsample of observations yielding the "best" affine subspace, which we called the trimmed PC subspace. While an algorithm for computing the trimmed PC subspace was already proposed by Maronna (2005), its theoretical properties were not studied yet.

As a first result we could prove existence of the trimmed PC subspace without making any moment restrictions. While standard PCA requires existence of second moments, this is not required for its trimmed version. Hence, the trimmed PC subspace exists at a multivariate Cauchy distribution, for example, where standard PCA is not feasible. We also proved, under mild conditions, consistency of the sample trimmed PC space towards the population counterpart. The robustness of the method is studied by showing

quantitative robustness, computing the breakdown point, and deriving the influence functions. The influence function turns out to be bounded in the region where the outliers are, but good leverage points still may have an unbounded influence. Furthermore, asymptotic efficiencies at the normal model are derived, while finite sample efficiencies of the estimators are obtained by means of a simulation study. It is shown that, by selecting an appropriate trimming proportion α, both a high breakdown point and a high efficiency are attainable.

A distinct feature of the proposed method compared to other approaches for robust PCA is that it directly aims at finding the best fitting affine subspace. The population version, which we presented in Section 2 and of which we showed existence in Section 3, has a clear geometric interpretation, also at non-elliptical distributions. If one would use, for example, the space spanned by the first d eigenvectors of a robust estimate of the covariance matrix as best fitting subspace, then it is not clear whether the corresponding population quantity has any optimality property, unless at elliptically symmetric distributions. When the aim of the robust principal component analysis is to perform dimension reduction, and to find an optimal subspace of a certain dimension, then trimmed PCA is a natural candidate. Of course, the aim of the principal component analysis might be different. Sometimes one looks for the optimal linear combination of the variables having maximal dispersion, and then then robust prospection pursuit approach of Li and Chen (1985) becomes more appealing.

Maronna (2005) conducted a simulation study and did found good performance of the method. He also applied it on several real data sets. An

application in robust multivariate error-in-variables modeling was studied in Croux, Fekri and Ruiz-Gazen (2009). Serneels and Verdonck (2009) showed its good performance when applied to principal component regression for data containing outliers.

There are several extensions possible of the trimmed principal components method we studied. One could consider general penalty functions $\Phi(\cdot)$ for quantifying the discrepancy between the point x and the affine subspace h through $\Phi(\|x - \mathrm{Pr}_h(x)\|)$, instead of merely considering the squared loss. However, similar as in in García-Escudero and Gordaliza (1999), we expect that the main robustification arises from the trimming and less by the different choices of the penalty function Φ. We can also adopt a "min-max" or L_∞ approach. In other words, we would search for the narrowest strip (i.e., having the smallest radius as possible) including a $1 - \alpha$ proportion of the data points. Notice that Rousseeuw's LMS regression estimator also shares that idea. A second possible avenue for future research is the extension of the theoretical results to a multiple population setting. Applications of the trimming approach in the multiple population case are in robust linear clustering (Garcia-Escudero et al, 2009) and robust cluster analysis (Garcia-Escudero et al, 2008).

Appendix A. Proofs

Proof of Lemma 1: Let us consider $h^d \in \mathcal{A}_d(\mathbb{R}^p)$, the affine subspace spanned by the origin and the first d vectors of the canonical basis in $\mathbb{R}^p$. Take $r > 0$ such that $P_X(S(h^d, r)) \geq 1 - \alpha$ and consider the trimming

function $\tau_d = I_{S(h^d,r)} \in \mathcal{T}_{\alpha-}$. We have

$$V_{d,\alpha}(X) \leq \frac{1}{P_X(S(h^d,r))} \int_{S(h^d,r)} \|x - \Pr_{h^d}(x)\|^2 dP_X(x) < r^2. \ \square$$

Proof of Lemma 2: For every $\tau \in \mathcal{T}_{h,\beta}$ and $\tau' \in \mathcal{T}_\beta$, we have that

$$\begin{aligned} \tau(x)(1-\tau'(x)) &= 0 \text{ for all } x \notin \overline{S}(h, r_\beta(h)), \\ \int \tau(x)(1-\tau'(x))dP_X(x) &= \int \tau'(x)(1-\tau(x))dP_X(x), \text{ and,} \\ \tau'(x)(1-\tau(x)) &= 0 \text{ for all } x \in S(h, r_\beta(h)). \end{aligned}$$

Hence, by applying the above equalities, we have

$$\int \tau(x)(1-\tau'(x))\|x - \Pr_h(x)\|^2 dP_X(x) \leq r_\beta^2(h) \int \tau(x)(1-\tau'(x))dP_X(x) \tag{A.1}$$

$$= r_\beta^2(h) \int \tau'(x)(1-\tau(x))dP_X(x) \leq \int \tau'(x)(1-\tau(x))\|x - \Pr_h(x)\|^2 dP_X(x). \tag{A.2}$$

So, we have

$$\begin{aligned} &\int \tau(x)\|x - \Pr_h(x)\|^2 dP_X(x) \\ &= \int \tau(x)\tau'(x)\|x - \Pr_h(x)\|^2 dP_X(x) + \int \tau(x)(1-\tau'(x))\|x - \Pr_h(x)\|^2 dP_X(x) \\ &\leq \int \tau(x)\tau'(x)\|x - \Pr_h(x)\|^2 dP + \int \tau'(x)(1-\tau(x))\|x - \Pr_h(x)\|^2 dP_X(x) \\ &= \int \tau'(x)\|x - \Pr_h(x)\|^2 dP_X(x). \end{aligned}$$

Moreover, the equality holds if and only if (A.1) and (A.2) are equalities. However, (A.1) is an equality if and only if

$$\int_{S(h,r_\beta(h))} \tau(x)(1-\tau'(x))dP_X(x) = 0,$$

which implies that

$$\int_{S(h,r_\beta(h))} (1 - \tau'(x)) dP_X(x) = 0,$$

and, thus, we conclude that $I_{S(h,r_\beta(h))} \leq \tau'$, P_X-a.e. . The equality in (A.2) would analogously imply $\tau' \leq I_{\overline{S}(h,r_\beta(h))}$, P_X-a.e. Therefore, assertion (b) in this Lemma is also proven. □

Proof of Lemma 3: Without loss of generality, we can assume that $\tau_{h,\beta} \geq \tau_{h,\alpha}$, P_X-a.e., for $\beta \leq \alpha$ (in fact, we can always choose $\tau_{h,\beta}$ and $\tau_{h,\alpha}$ such that $\tau_{h,\beta} \geq \tau_{h,\alpha}$ pointwise).

Now, we can see that

$$\begin{aligned}
&\int \tau_{h,\alpha}(x) dP_X(x) \int (\tau_{h,\beta}(x) - \tau_{h,\alpha}(x)) \|x - \mathrm{Pr}_h(x)\|^2 dP_X(x) \\
&\quad \geq \int \tau_{h,\alpha}(x) dP_X(x) \cdot r_\alpha^2(h) \int (\tau_{h,\beta}(x) - \tau_{h,\alpha}(x)) dP_X(x) \qquad \text{(A.3)} \\
&\quad \geq \int \tau_{h,\alpha}(x) \|x - \mathrm{Pr}_h(x)\|^2 dP_X(x) \cdot \int (\tau_{h,\beta}(x) - \tau_{h,\alpha}(x)) dP_X(x), \qquad \text{(A.4)}
\end{aligned}$$

and then we have

$$\begin{aligned}
&\int \tau_{h,\alpha}(x) dP_X(x) \int \tau_{h,\beta}(x) \|x - \mathrm{Pr}_h(x)\|^2 dP_X(x) \\
&\quad = \int \tau_{h,\alpha}(x) dP_X(x) \int \tau_{h,\alpha}(x) \|x - \mathrm{Pr}_h(x)\|^2 dP_X(x) \\
&\qquad\quad + \int \tau_{h,\alpha}(x) dP_X(x) \int (\tau_{h,\beta}(x) - \tau_{h,\alpha}(x)) \|x - \mathrm{Pr}_h(x)\|^2 dP_X(x) \\
&\quad \geq \int \tau_{h,\alpha}(x) dP_X(x) \int \tau_{h,\alpha}(x) \|x - \mathrm{Pr}_h(x)\|^2 dP_X(x) \\
&\qquad\quad + \int \tau_{h,\alpha}(x) \|x - \mathrm{Pr}_h(x)\|^2 dP_X(x) \int (\tau_{h,\beta}(x) - \tau_{h,\alpha}(x)) dP_X(x) \\
&\quad = \int \tau_{h,\beta}(x) dP_X(x) \int \tau_{h,\alpha}(x) \|x - \mathrm{Pr}_h(x)\|^2 dP_X(x).
\end{aligned}$$

Now, by using $\int \tau_{h,\alpha}(x)dP_X(x) = 1-\alpha$ and $\int \tau_{h,\beta}(x)dP_X(x) = 1-\beta$, we have

$$\frac{1}{1-\beta}\int \tau_{h,\beta}(x)\|x - \mathrm{Pr}_h(x)\|^2 dP_X(x) \;\geq\; \frac{1}{1-\alpha}\int \tau_{h,\alpha}(x)\|x - \mathrm{Pr}_h(x)\|^2 dP_X(x).$$

and result (a) is derived.

Moreover, the equality in (a) holds if and only if (A.3) and (A.4) are equalities. Now, the equality (A.3) holds if and only if

$$\int_{\overline{S}(h,r_\alpha(h))^c} (\tau_{h,\beta}(x) - \tau_{h,\alpha}(x))dP_X(x) = 0,$$

which holds if and only if $r_\alpha(h) = r_\beta(h)$. Analogously, (A.4) is an equality if and only if

$$\int_{S(h,r_\alpha(h))} \tau_{h,\alpha}(x)dP_X(x) = 0,$$

which implies $P_X(S(h, r_\alpha(h))) = 0$. In other words, all the probability mass is concentrated on the boundary of $S(h, r_\beta(h))$. □

Proof of Lemma 4: Let us consider a ball B centered at the origin and with radius $R > 0$, such that $P_X(B) > \max\{1-\alpha, \alpha\}$. As $P_X(S_n) \leq 1-\alpha \leq P_X(\overline{S}_n)$, it can be easily seen that $\overline{S}_n \cap B \neq \emptyset$ and $B \not\subseteq S_n$. Therefore, $d_n - R \leq r_n \leq d_n + R$ for every $n \in \mathbb{N}$, and $\{r_n\}_n$ will be bounded if and only if $\{d_n\}_n$ is bounded. We will prove that $\{d_n\}_n$ is a bounded sequence.

Let $\{\varepsilon_n\}_n$ and $\{\gamma_n\}_n$ be two sequences of positive numbers such that $\varepsilon_n \downarrow 0$, $\gamma_n \uparrow \infty$ and $P_X(B(0,\gamma_n)) > 1-\varepsilon_n$. If $\{d_n\}_n$ were not bounded, we could find a subsequence (denoted as the original one) such that $d_n > 2\gamma_n$

for every $n \in \mathbb{N}$. Then, we would have

$$\begin{aligned} V_{d,\alpha}(h_n) &= \frac{1}{1-\alpha}\int \tau_n(x)\|x - \mathrm{Pr}_{h_n}(x)\|^2 dP_X(x) \\ &\geq \frac{1}{1-\alpha}\int_{B(0,\gamma_n)} \tau_n(x)\|x - \mathrm{Pr}_{h_n}(x)\|^2 dP_X(x) \\ &\geq \frac{1}{1-\alpha}\int_{B(0,\gamma_n)} \tau_n(x)\,\gamma_n^2\, dP_X(x) \\ &\geq \gamma_n^2 \frac{1-\alpha-\varepsilon_n}{1-\alpha} \uparrow \infty \text{ as } n \to \infty, \end{aligned}$$

contradicting (4). Thus, $\{d_n\}_n$ and $\{r_n\}_n$ are bounded. □

Proof of Theorem 1: Taking into account the comments and results at the beginning of Section 3, we can take a sequence $\{h_n\}_n \subset \mathcal{A}_d(\mathbb{R}^p)$ satisfying $V_{d,\alpha}(h_n) \downarrow V_{d,\alpha}$ as $n \to \infty$, and such that the corresponding sequences of unitary director vectors, distances to the origin and radius are convergent. Let us denote $h_0 \in \mathcal{A}_d(\mathbb{R}^p)$ the limit subspace, r_0 the limit of the radius sequence and $S_0 = S(h_0, r_0)$ the corresponding limit strip.

We have that

$$I_{S_0}(X) \leq \liminf_n \tau_n(X) \leq \limsup_n \tau_n(X) \leq I_{\overline{S}_0}(X),$$

and then, Fatou's Lemma implies

$$\begin{aligned} \int I_{S_0}(x) dP_X(x) &\leq \int \liminf_n \tau_n(x) dP_X(x) \leq 1-\alpha \\ &\leq \int \limsup_n \tau_n(x) dP_X(x) \leq \int I_{\overline{S}_0}(x) dP_X(x), \end{aligned}$$

which means that $r_0 = r_\alpha(h_0)$ and $S_0 = S(h_0, r_\alpha(h_0))$.

We can consider a trimming function $\tau_0 := \tau_{h_0,\alpha} \in \mathcal{T}_\alpha$ associated to the limit strip S_0. If we prove that h_0 satisfies $\lim_{n\to\infty} V_{d,\alpha}(h_n) = V_{d,\alpha}(h_0)$, then

$$V_{d,\alpha}(h_0) = V_{d,\alpha} = \inf_{h \in \mathcal{A}_d(\mathbb{R}^p)} V_{d,\alpha}(h),$$

and the proof would be finished. To do this task, we need to prove that

$$\left| \int \tau_n(x)\|x - \mathrm{Pr}_{h_n}(x)\|^2 dP_X(x) - \int \tau_0(x)\|x - \mathrm{Pr}_{h_0}(x)\|^2 dP_X(x) \right| \to 0.$$

Let us denote $E_n = S_0^c \cap \overline{S}_n$, $F_n = \overline{S}_0 \cap S_n^c$ and $G_n = S_0 \cap S_n$. Note that the convergence of the sequence of strips S_n toward the strip S_0 implies that $P_X(E_n) \to 0$ and $P_X(F_n) \to 0$ as $n \to \infty$. Thus, taking into account that $\tau_n(x) = \tau_0(x) = 0$ for $x \in (E_n \cup F_n \cup G_n)^c$, we can decompose

$$\begin{aligned}
&\left| \int \tau_n(x)\|x - \mathrm{Pr}_{h_n}(x)\|^2 dP_X(x) - \int \tau_0(x)\|x - \mathrm{Pr}_{h_0}(x)\|^2 dP_X(x) \right| \\
&\leq \left| \int_{E_n} \tau_n(x)\|x - \mathrm{Pr}_{h_n}(x)\|^2 dP_X(x) - \int_{E_n} \tau_0(x)\|x - \mathrm{Pr}_{h_0}(x)\|^2 dP_X(x) \right| \\
&+ \left| \int_{F_n} \tau_n(x)\|x - \mathrm{Pr}_{h_n}(x)\|^2 dP_X(x) - \int_{F_n} \tau_0(x)\|x - \mathrm{Pr}_{h_0}(x)\|^2 dP_X(x) \right| \\
&+ \left| \int_{G_n} \tau_n(x)\|x - \mathrm{Pr}_{h_n}(x)\|^2 dP_X(x) - \int_{G_n} \tau_0(x)\|x - \mathrm{Pr}_{h_0}(x)\|^2 dP_X(x) \right| \\
&:= A_n^{(1)} + A_n^{(2)} + A_n^{(3)}.
\end{aligned}$$

We need to prove that $A_n^{(1)}$, $A_n^{(2)}$ and $A_n^{(3)}$ converge to 0. For $A_n^{(1)}$, recalling the bounded character of the sequence $\{r_n\}_n$ from Lemma 4, we have:

$$\begin{aligned}
A_n^{(1)} &\leq \left| \int_{E_n} \tau_n(x)\|x - \mathrm{Pr}_{h_n}(x)\|^2 dP_X(x) \right| \\
&\qquad + \left| \int_{E_n} \tau_0(x)\|x - \mathrm{Pr}_{h_0}(x)\|^2 dP_X(x) \right| \\
&\leq r_n^2 \int_{E_n} \tau_n(x) dP_X(x) + r_0^2 \int_{E_n} \tau_0(x) dP_X(x) \\
&\leq r_n^2 P_X(E_n) + r_0^2 P_X(E_n) = (r_n^2 + r_0^2) P_X(E_n) \to 0.
\end{aligned}$$

In a similar way we can prove that $A_n^{(2)}$ converges to 0. To study the

convergence of $A_n^{(3)}$ we can obtain the following decomposition:

$$A_n^{(3)} \leq \left| \int_{G_n} \tau_n(x)(\|x - \mathrm{Pr}_{h_n}(x)\|^2 - \|x - \mathrm{Pr}_{h_0}(x)\|^2) dP_X(x) \right| + \left| \int_{G_n} (\tau_n(x) - \tau_0(x)) \|x - \mathrm{Pr}_{h_0}(x)\|^2 dP_X(x) \right| := A_n^{(3,a)} + A_n^{(3,b)}.$$

As for $x \in G_n$ it holds $\tau_n(x) = \tau_0(x) = 1$ and then $\tau_n(x) - \tau_0(x) = 0$, we have $A_n^{(3,b)} = 0$ and it only remains the convergence of $A_n^{(3,a)}$. Now, taking into account the uniform continuity of the real valued quadratic function $g(x) = x^2$ on the compact set $[0, \sup_n r_n]$, we have

$$A_n^{(3,a)} \leq \sup_{x \in G_n} \left\{ \|x - \mathrm{Pr}_{h_n}(x)\|^2 - \|x - \mathrm{Pr}_{h_0}(x)\|^2 \right\} (1 - \alpha) \to 0,$$

and the proof is complete. □

Proof of Theorem 2: It suffices to prove that every subsequence of $\{h_n\}_n$ (resp. $\{V_n\}_n$) admits a new subsequence which converges to h_0 (resp. V_0). Along the proof, all subsequences will be denoted as the original sequences.

For every $n = 1, 2, ...$, let us denote by $\tau'_n = \tau'_n(X_n)$ a trimming function in $\mathcal{T}_{h_0,\alpha}$. So, with $r'_n, n = 1, 2, ...$, the radius associated to τ'_n, that is,

$$r'_n = \inf\{r \geq 0 : P_{X_n}(S(h_0, r)) \leq 1 - \alpha \leq P_{X_n}(\overline{S}(h_0, r))\},$$

we have $I_{S(h_0,r'_n)} \leq \tau'_n \leq I_{\overline{S}(h_0,r'_n)}$. Moreover, denote

$$V'_n = \frac{1}{1 - \alpha} \int \tau'_n(x) \|x - \mathrm{Pr}_{h_0}(x)\|^2 dP_{X_n}(x).$$

Obviously, $\{r'_n\}_n$ is a bounded sequence, so we can assume, without loss of generality, that $r'_n \to r'_0$ for some $r'_0 \in \mathbb{R}$. Then, because of the continuity of P_{X_0}, we have $\tau'_n(X_n) \to I_{S(h_0,r'_0)}(X_0)$ P-a.e., and, then, taking into account that $|\tau'_n| \leq 1$, we may write

$$1 - \alpha = \int \tau'_n(x) dP_{X_n}(x) \to \int I_{S(h_0,r'_0)}(x) dP_{X_0}(x), \text{ as } n \to \infty.$$

Therefore, we have $I_{S(h_0,r_0')}(X_0) = \tau_0(X_0), P$-a.e. .

The sequence $\{\tau_n'(X_n)\|X_n - \mathrm{Pr}_{h_0}(X_n)^2\|\}_n$ is uniformly bounded and satisfies

$$\tau_n'(X_n)\|X_n - \mathrm{Pr}_{h_0}(X_n)\|^2 \to \tau_0(X_0)\|X_0 - \mathrm{Pr}_{h_0}(X_0)\|^2, P\text{-a.e. .}$$

Hence we have

$$\begin{aligned} V_n \leq V_n' &= \frac{1}{1-\alpha}\int \tau_n'(x)\|x - \mathrm{Pr}_{h_0}(x)\|^2 dP_{X_n}(x) \\ &\to \frac{1}{1-\alpha}\int \tau_0(x)\|x - \mathrm{Pr}_{h_0}(x)\|^2 dP_{X_0}(x) = V_0 \end{aligned}$$

and, consequently, recalling the optimal character of V_n for X_n, we have

$$\limsup_n V_n \leq \limsup_n V_n' \leq V_0. \tag{A.5}$$

Taking into account Lemma 5 and the boundedness of the sequences of unitary spanning vectors, we can take a subsequence of $\{h_n\}_n \subset \mathcal{A}_d(\mathbb{R}^p)$ such that the corresponding sequences of unitary spanning vectors, distances to the origin and radius are convergent. Let us denote $h^0 \in \mathcal{A}_d(\mathbb{R}^p)$ the limit subspace, r^0 the limit of the radius sequence and $S^0 = S(h^0, r^0)$ the corresponding limit strip.

In order to prove that $S^0 = S(h^0, r^0)$ provides trimming function of level α for X_0, we note that $\lim_n \tau_n(X_n) = I_{S^0}(X_0), P$-a.e.. Now, by taking into account that $|\tau_n| \leq 1$ for every $n = 1, 2, ...$, we have

$$1 - \alpha = \int \tau_n(x) dP_{X_n}(x) \to \int I_{S^0}(x) dP_{X_0}(x),$$

so that I_{S^0} is a trimming function of level α for X_0. Let us denote V^0 the associated trimmed variation around h^0, i.e.

$$V^0 = \frac{1}{1-\alpha}\int I_{S^0}(x)\|x - \mathrm{Pr}_{h^0}(x)\|^2 dP_{X_0}(x).$$

Moreover, the sequence $\{\tau_n(X_n)\|X_n - \mathrm{Pr}_{h_n}(X_n)^2\|\}_n$ is uniformly bounded and satisfies

$$\tau_n(X_n)\|X_n - \mathrm{Pr}_{h_n}(X_n)\|^2 \to I_{S^0}(X_0)\|X_0 - \mathrm{Pr}_{h^0}(X_0)\|^2, P\text{-a.e. .}$$

Then, we have

$$\begin{aligned} V_n &= \frac{1}{1-\alpha}\int \tau_n(x)\|x - \mathrm{Pr}_{h_n}(x)\|^2 dP_{X_n}(x) \\ &\to \frac{1}{1-\alpha}\int I_{S^0}(x)\|x - \mathrm{Pr}_{h^0}(x)\|^2 dP_{X_0}(x) \end{aligned}$$

and, consequently, recalling the optimal character of V_0 for X_0, we have

$$\liminf_n V_n = V^0 \geq V_0. \tag{A.6}$$

Finally, from (A.5) and (A.6) we obtain

$$\limsup_n V_n \leq \limsup_n V_n' \leq V_0 \leq V^0 \leq \liminf_n V_n, \tag{A.7}$$

i.e., $\lim_n V_n = V_0, P$-a.e. and the convergence of the variations holds.

Moreover, from (A.7) we also have $V_0 = V^0$ and then h^0 is optimal for X_0, but taking into account the uniqueness of the d-dimensional trimmed principal component subspace of X_0 we must have $h_0 = h^0$, P_{X_0}-a.e., and then it also holds the convergence of the optimal affine subspaces. □

Proof of Theorem 4: Without loss of generality, let us assume that $\mu = 0$. The proof will be arranged in two steps:

1. Any optimal affine subspace pass trough $\mu = 0$. In the first step we will prove that given any $h \in \mathcal{A}_d(\mathbb{R}^p)$, the value of the target function $V_{d,\alpha}(h)$ is strictly decreased when choosing the affine subspace $h_0 \in \mathcal{A}_d(\mathbb{R}^p)$ parallel to h and passing through the origin. I.e., let us consider h_0 the affine subspace

passing through the origin and spanned by the columns of a matrix U, where the d columns of U are unitary and orthogonal vectors spanning the original affine subspace h. Consider an orthonormal basis spanning $h_0^{\perp}$, the $p-d$ affine subspace orthogonal to h_0, and let us denote V the $p \times (p-d)$ matrix having these vectors as columns. Notice that

$$d(x, h_0)^2 = \|x - \mathrm{Pr}_h(x)\|^2 = \|\mathrm{Pr}_{h_0^{\perp}}(x)\|^2 = \|V'x\|^2.$$

As h is an affine subspace parallel to h_0, then there exists $x_1 \in \mathbb{R}^p$ such that $h \equiv h_0 + x_1$ and

$$d(x, h)^2 = \|x - \mathrm{Pr}_h(x)\|^2 = \|V'(x - x_1)\|^2.$$

Without loss of generality, we have assumed $\mu = 0$ and, then, we have $X \sim E_p(0, \Sigma)$ and $Y = V'X \sim E_{p-d}(0, V'\Sigma V)$ with p.d.f. equal to

$$f_Y(y) = |V'\Sigma V|^{-1/2} h(y'\big(V'\Sigma V\big)^{-1}y\big),$$

with h a decreasing radial density function.

If we denote $r_0 = r_\alpha(h_0)$, the trimmed variation around h_0 can be written as

$$\begin{aligned} V_{d,\alpha}(h_0) &= \frac{1}{1-\alpha}\int_{S(h_0,r_0)} \|x - \mathrm{Pr}_{h_0}(x)\|^2 f_X(x)dx \\ &= \frac{1}{1-\alpha}\int_{S(h_0,r_0)} x'VV'x f_X(x)dx = \frac{1}{1-\alpha}\int_{B(0,r_0)} \|y\|^2 f_Y(y)dy \end{aligned}$$

where $B(0, r_0) \subset \mathbb{R}^{p-d}$ denotes the ball with radius r_0 around $0 \in \mathbb{R}^{p-d}$.

In a similar fashion, we get

$$\begin{aligned} V_{d,\alpha}(h) &= \frac{1}{1-\alpha}\int_{S(h,r_\alpha(h))} \|x - \mathrm{Pr}_h(x)\|^2 f_X(x)dx \\ &= \frac{1}{1-\alpha}\int_{B(y_1,r_\alpha(h))} \|y - y_1\|^2 f_Y(y)dy, \end{aligned}$$

with $B(y_1, r_\alpha(h)) \subset \mathbb{R}^{p-d}$ and $y_1 = V'x_1$.

Now, we take $\xi(\cdot) = |V'\Sigma V|^{-1/2}h(\cdot)$ and $g(\cdot) = (r_0^2 - \cdot)I_{[0,r_0]}(\cdot)$, for appliyng Lemma 6 with $\theta = -y_1$ and the positively defined matrix $V'\Sigma V$, we obtain that (I1) $\leq$ (I2) with

$$\begin{aligned}
\text{(I1)} &= \int \xi((y+y_1)'(V'\Sigma V)^{-1}(y+y_1))g(y'y)dy \\
&= \int f_Y(y+y_1)g(y'y)dy \\
&= r_0^2 \int_{B(0,r_0)} f_Y(y+y_1)dy - \int_{B(0,r_0)} \|y\|^2 f_Y(y+y_1)dy \\
&= r_0^2 \int_{B(y_1,r_0)} f_Y(y)dy - \int_{B(y_1,r_0)} \|y-y_1\|^2 f_Y(y)dy \\
&= r_0^2 \int_{S(h,r_0)} f_X(x)dx - \int_{S(h,r_0)} \|x - \mathrm{Pr}_h(x)\|^2 f_X(x)dx
\end{aligned}$$

and

$$\begin{aligned}
\text{(I2)} &= \int \xi(y'(V'\Sigma V)^{-1}y)g(y'y)dy \\
&= \int f_Y(y)g(y'y)dy \\
&= r_0^2 \int_{B(0,r_0)} f_Y(y)dy - \int_{B(0,r_0)} \|y\|^2 f_Y(y)dy \\
&= r_0^2 \int_{S(h_0,r_0)} f_X(x)dx - \int_{S(h_0,r_0)} \|x - \mathrm{Pr}_{h_0}(x)\|^2 f_X(x)dx.
\end{aligned}$$

Then, we have

$$\begin{aligned}
& r_0^2 \int_{S(h_0,r_0)} f_X(x)dx - (1-\alpha)V_{d,\alpha}(h_0) \\
\geq\ & r_0^2 \int_{S(h,r_0)} f_X(x)dx - \int_{S(h,r_0)} \|x - \mathrm{Pr}_h(x)\|^2 f_X(x)dx.
\end{aligned}$$

Now, adding and subtracting $(1-\alpha)V_{d,\alpha}(h)$ and rearranging terms in the previous expression, we obtain the inequality

$$\begin{aligned}
&(1-\alpha)[V_{d,\alpha}(h) - V_{d,\alpha}(h_0)] \\
&\quad \geq \quad r_0^2 \left[\int_{S(h,r_0)} f_X(x)dx - \int_{S(h_0,r_0)} f_X(x)dx \right] \\
&\qquad - \left[\int_{S(h,r_0)} \|x - \mathrm{Pr}_h(x)\|^2 f_X(x)dx - \int_{S(h,r_\alpha(h))} \|x - \mathrm{Pr}_h(x)\|^2 f_X(x)dx \right] \\
&\quad = \quad r_0^2 \left[\int_{S(h,r_0)} f_X(x)dx - \int_{S(h,r_\alpha(h))} f_X(x)dx \right] \\
&\qquad - \left[\int_{S(h,r_0)} \|x - \mathrm{Pr}_h(x)\|^2 f_X(x)dx - \int_{S(h,r_\alpha(h))} \|x - \mathrm{Pr}_h(x)\|^2 f_X(x)dx \right],
\end{aligned}$$

where we have used $P_X(S(h_0,r_0)) = P_X(S(h,r_\alpha(h)) = 1-\alpha$. Now, taking into account that $r_\alpha(h) > r_0$ (that is a trivial consequence of the Anderson lemma for strictly unimodal distributions (Anderson 1955)), $f_X(x) > 0$ and $\|x - \mathrm{Pr}_h(x)\|^2 > r_0^2$ for all $x \in \overline{S}(h,r_0)^c \cap S(h,r_\alpha(h))$, we have

$$\begin{aligned}
&(1-\alpha)[V_{d,\alpha}(h) - V_{d,\alpha}(h_0)] \\
&\quad \geq \quad \left[\int_{S(h,r_0)^c \cap S(h,r_\alpha(h))} \|x - \mathrm{Pr}_h(x)\|^2 f_X(x)dx \right] \\
&\qquad\qquad\qquad - r_0^2 \left[\int_{S(h,r_0)^c \cap S(h,r_\alpha(h))} f_X(x)dx \right] \\
&\quad > \quad r_0^2 \left[\int_{S(h,r_0)^c \cap S(h,r_\alpha(h))} f_X(x)dx \right] - r_0^2 \left[\int_{S(h,r_0)^c \cap S(h,r_\alpha(h))} f_X(x)dx \right] = 0.
\end{aligned}$$

Thus, it holds the desired inequality $V_{d,\alpha}(h) - V_{d,\alpha}(h_0) > 0$.

2. The optimal affine subspace is spanned by the d largest eigenvectors of the scatter matrix Σ. Once proved that the d-dimensional trimmed

principal component pass through the origin, we will search for the directions of the optimal subspace. Without loss of generality, we continue assuming that $\mu = 0$ and, thus, that $X \sim E_p(0, \Sigma)$. Let us consider again $Y = V'X \sim E_{p-d}(0, V'\Sigma V)$. When trying to minimize $V_{d,\alpha}(h)$ on $h \in \mathcal{A}_d(\mathbb{R}^p)$, but restricted to affine subspaces passing through the origin, we need to minimize

$$\min_V \int_{B(0,r_{\alpha,Y})} \|y\|^2 f_Y(y)dy,$$

(0 now stands for the zero vector in $\mathbb{R}^{p-d}$) where $r_{\alpha,Y}$ is defined as

$$r_{\alpha,Y} := \inf\{r : P_Y(B(0,r)) \geq 1-\alpha\}.$$

Take $Z = (V'\Sigma^{1/2})^{-1}Y$ (in such a way that $Z \sim E_{p-d}(0, I_{p-d})$). We have that

$$\begin{aligned}
\int_{B(0,r_{\alpha,Y})} \|y\|^2 f_Y(y)dy &= |\Sigma|^{1/2} \int_{B(0,z_\alpha)} \|V'\Sigma^{1/2}z\|^2 f_Z(z)dz \\
&= |\Sigma|^{1/2} \int_{B(0,z_\alpha)} \text{trace}[V'\Sigma^{1/2}zz'\Sigma^{1/2}V] f_Z(z)dz \\
&= |\Sigma|^{1/2} \text{trace}\left[V'\Sigma^{1/2}\left(\int_{B(0,z_\alpha)} zz' f_Z(z)dz\right)\Sigma^{1/2}V\right],
\end{aligned} \tag{A.8}$$

with $z_\alpha := \inf\{r : P_Z(B(0,r)) \geq 1-\alpha\}$.

It can be seen (see, e.g., Theorem 8.1 of Liu et al. 1999) that there exists a positive constant ζ_α depending only on α, the dimension $p-d$, and, the elliptical family considered, such that

$$\int_{B(0,z_\alpha)} zz' f_Z(z)dz = \zeta_\alpha I_{p-d}.$$

Therefore, from (A.8), the problem reduces to the minimization of

$$\min_V \Big[\text{trace}[V'\Sigma V]\Big],$$

where V is a $p\times(p-d)$ matrix with unitary orthogonal vectors in its columns. This problem admits a unique solution if the eigenvalues of Σ, $\lambda_1 \geq ... \geq \lambda_p > 0$, satisfies $\lambda_d > \lambda_{d+1}$. Moreover, the solution is obtained from the matrix with columns equal to the eigenvectors associated to these d largest eigenvalues, see for example Jolliffe (2002). □

Proof of Theorem 5: Without loss of generality, we assume that $\mu = 0$ and that Σ is diagonal with decreasing diagonal elements. Theorem 4 and Corollary 2 yields that the d largest eigenvectors of $C(P)$ are the same as those of Σ, showing Fisher consistency for the eigenvectors. So we restrict attention to the eigenvalues. The first d eigenvectors are the first d canonical basis vectors, and they span the axis of the strip $S(P)$. Hence

$$S(P) = \{x = (x_1, \ldots, x_p)' \in \mathbb{R}^p |\, x_{d+1}^2 + \ldots + x_p^2 \leq r^2(P)\},$$

with $r(P)$ the radius of the strip. The strip is thus unbounded in the first d coordinates, and we get that the first d eigenvalues of $c\,C(P)$ are given by

$$\Lambda_j(P) = c\,C(P) = \frac{c}{1-\alpha} E[X_j^2 I(X_{d+1}^2 + \ldots + X_p^2 \leq r^2(P))]$$

for $1 \leq j \leq d$ and with $X = (X_1, \ldots, X_p)'$). Denoting λ_j the eigenvectors of the covariance matrix, which is a multiple of Σ, we have that the distribution of $X_j/\sqrt{\lambda_j}$ is the same for every j. Hence

$$\Lambda_j(P) = \frac{c\lambda_j}{1-\alpha} E[(X_1/\lambda_1)^2 I(X_{d+1}^2 + \ldots + X_p^2 \leq r^2(P))].$$

We get $\Lambda_j(P) = \lambda_j$, hence Fisher consistency, for $1 \leq j \leq d$, if we set

$$c = \frac{1-\alpha}{E[(X_1^2/\lambda_1)I(X_{d+1}^2 + \ldots + X_p^2 \leq r(P))]}. \tag{A.9}$$

Note that the expression above is not depending on j. For the normal distributions we can use independency of the marginals, resulting in

$$c = \frac{1-\alpha}{E[(X_1^2/\lambda_1)]P(X_{d+1}^2 + \ldots + X_p^2 \leq r(P))} = 1.$$

Hence at the normal distribution no correction for Fisher consistency is needed. □

PROOF OF THEOREM 7: Let P be an absolutely continuous distribution satisfying (5) and let $T(P) = (d_0, r_0, V_0)$ denotes its unique d-dimensional trimmed principal component. Let us consider the point-mass contaminated distribution $P_{\varepsilon,x_0} = (1-\varepsilon)P + \varepsilon\delta_{\{x_0\}}$, and $T(P_{\varepsilon,x_0})$ the corresponding trimmed principal components. As $P_{\varepsilon,x_0} \to P$ when $\varepsilon \downarrow 0$, we can use Corollary 3 to obtain the convergence $T(P_{\varepsilon,x_0}) \to T(P_0)$.

We now start with the derivation of the influence function. Recall that we assumed $\mu = 0$, and Σ a diagonal matrix with decreasing diagonal elements. Denote

$$m_\varepsilon = m(P_{\varepsilon,x_0}) = \frac{1-\varepsilon}{1-\alpha}\int_{S(P_{\varepsilon,x_0})} x dP(x) + \frac{\varepsilon}{1-\alpha} I_{S(P_{\varepsilon,x_0})}(x_0)x_0$$

and

$$C_\varepsilon = C(P_{\varepsilon,x_0}) = c\left\{\frac{1-\varepsilon}{1-\alpha}\int_{S(P_{\varepsilon,x_0})} xx' dP(x) + \frac{\varepsilon}{1-\alpha} I_{S(P_{\varepsilon,x_0})}(x_0)x_0x'_0 - m_\varepsilon m'_\varepsilon\right\}. \tag{A.10}$$

We have by definition

$$IF(x_0, C, P) = \left(\frac{\partial C_\varepsilon}{\partial \varepsilon}\right)_{|_{\varepsilon=0}}.$$

Differentiating (A.10) gives

$$\frac{\partial C_\varepsilon}{\partial \varepsilon}_{|\varepsilon=0} = c\Big\{ -\frac{1}{1-\alpha}\int_{S(P)} xx'dP(x) + \frac{1}{1-\alpha}\frac{\partial}{\partial\varepsilon}\int_{S(P_{\varepsilon,x_0})} xx'dP(x)_{|\varepsilon=0} + \frac{1}{1-\alpha} I_{S(P)}(x_0)x_0x'_0\Big\}. \quad (A.11)$$

By definition, the first term is just $-C(P)$. The third term is easily handled. We now turn to the differentiation of the second term. We introduce the notation

$$I(\varepsilon) = \int_{S(P_{\varepsilon,x_0})} xx'dP(x)$$

To obtain a tractable integration domain, we apply the change of variables $y = V_\varepsilon^{-1}(x - m_\varepsilon)$, where V_ε is the matrix of eigenvectors of C_ε. To obtain an admissible change of variable, it must be that all eigenvalues are distinct. However, it is easily seen that, making an arbitrary choice where needed, the results for the first d eigenvalues and eigenvectors. To avoid further notational complications, we develop the argument assuming all eigenvalues to be distinct. The domain of integration then becomes a strip of the same radius r_ε but with axis equal to the span of the first d coordinates of the space.

$$I(\varepsilon) = |V_\varepsilon||\Sigma|^{-\frac{1}{2}} \int_{y_{d+1}^2+\ldots+y_p^2\le r_\varepsilon^2} (V_\varepsilon y + m_\varepsilon)(V_\varepsilon y + m_\varepsilon)' \, h\left((V_\varepsilon y + m_\varepsilon)'\Sigma^{-1}(V_\varepsilon y + m_\varepsilon)\right) dy.$$

Now to make differentiation easier, we rewrite the last $p-d$ coordinates in polar form : $(y_{d+1},\ldots,y_p)' = r\,e(\theta)$ with $r \in [0, r_\varepsilon]$, $\theta \in \Theta = [0,\pi[\times\ldots[0,\pi[\times[0,2\pi[$, and $e(\theta) \in S^{p-d-1}$, the unit hypersphere in $p-d$ dimensions. Denote by $J(r,\theta)$ the Jacobian of this transformation. We get,

with $y_{1:d} = (y_1, \ldots, y_d)'$,

$$I(\varepsilon) = |V_\varepsilon||\Sigma|^{-\frac{1}{2}} \int_{\mathbb{R}^d} \int_{[0,r_\varepsilon]} \int_\Theta J(r,\theta) \left[V_\varepsilon(y_{1:d}, r\, e(\theta)) + m_\varepsilon\right] \left[V_\varepsilon(y_{1:d}, r\, e(\theta)) + m_\varepsilon\right]'$$
$$h\left(\left[V_\varepsilon(y_{1:d}, re(\theta)) + m_\varepsilon\right]' \Sigma^{-1} \left[V_\varepsilon(y_{1:d}, r\, e(\theta)) + m_\varepsilon\right]\right) d\theta\, dr\, dy_{1:d}.$$

By matrix differentiation, and since $V_0 = I$ we have

$$\frac{\partial \det(V_\varepsilon)}{\partial \varepsilon}_{|\varepsilon=0} = \text{trace}(IF(x_0, V, P)). \tag{A.12}$$

Applying the Leibniz formula, the derivative of the integral in the expression of $I(\varepsilon)$ is given by

$$\frac{\partial r_\varepsilon}{\partial \varepsilon}_{|\varepsilon=0} A(r_0, h, \Sigma) + \int_{S(P)} \frac{\partial}{\partial \varepsilon} B(\varepsilon, h, \Sigma, y)_{|\varepsilon=0} dy \tag{A.13}$$

where

$$A \equiv A(r_0, h, \Sigma) = |\Sigma|^{-\frac{1}{2}} \int_\Theta J(r_0, \theta) \int_{\mathbb{R}^d} [y_{1:d}, r_0 e(\theta)] \, [y_{1:d}, r_0 e(\theta)]'$$
$$\cdot h\left([y_{1:d}, r_0 e(\theta)]' \Sigma^{-1} [y_{1:d}, r_0 e(\theta)]\right) dy_{1:d}\, d\theta \tag{A.14}$$

and

$$B(\varepsilon, h, \Sigma, y) = \left[V_\varepsilon y + m_\varepsilon\right] \left[V_\varepsilon y + m_\varepsilon\right]' |\Sigma|^{-\frac{1}{2}} h\left(\left[V_\varepsilon y + m_\varepsilon\right]' \Sigma^{-1} \left[V_\varepsilon y + m_\varepsilon\right]\right).$$

Using symmetry arguments, it is clearly seen that A is a diagonal matrix. An exact expression is available for some specific distributions, but in the general case, there seem to be no further simplification.

Differentiating B one gets

$$\frac{\partial}{\partial \varepsilon} B(\varepsilon, h, \Sigma, y)_{|\varepsilon=0} = \Big\{ IF(x_0, V, P) yy' + yy' IF(x_0, V, P)'$$
$$+ IF(x_0, m, P) y' + y IF(x_0, m, P)' \Big\} |\Sigma|^{-\frac{1}{2}} h(y'\Sigma^{-1} y)$$

$$+yy'|\Sigma|^{-\frac{1}{2}}\dot{h}\left(y'\Sigma^{-1}y\right)\left\{2y'\Sigma^{-1}IF(x_0,m,P)+2y'\Sigma^{-1}IF(x_0,V,P)y\right\}. \quad \text{(A.15)}$$

Due to the symmetry of integration domain and distribution, the quantities with an odd number of y's integrate to zero. This implies that terms including $IF(x_0, m, P)$ give a zero contribution to the integral.

Now let us take care of

$$\frac{\partial r_\varepsilon}{\partial \varepsilon}_{|\varepsilon=0}.$$

By definition of a solution strip, one has

$$1-\alpha=(1-\varepsilon)\int_{S(P_{\varepsilon,x_0})} dP(x)_{|\varepsilon=0}+\varepsilon I_{S(P_{\varepsilon,x_0})}(x_0).$$

Differentiating both sides w.r.t. ε yields

$$0=-\int_{S(P)} dP(x)+\frac{\partial}{\partial\varepsilon}\int_{S(P_{\varepsilon,x_0})} dP(x)_{|\varepsilon=0}+I_{S(P)}(x_0).$$

In a similar fashion as was already done, one easily verifies that

$$\begin{aligned}\frac{\partial}{\partial\varepsilon}\int_{S(P_{\varepsilon,x_0})} dP(x)_{|\varepsilon=0} &= (1-\alpha)\text{trace}(IF(x_0,V,P))+\frac{\partial r_\varepsilon}{\partial\varepsilon}_{|\varepsilon=0} G(r_0,h,\Sigma)\\ &\quad +2|\Sigma|^{-\frac{1}{2}}\int_{S(P)}\dot{h}\left(y'\Sigma^{-1}y\right)y'\Sigma^{-1}IF(x_0,V,P)ydy\end{aligned}$$

where

$$G\equiv G(r_0,h,\Sigma)=|\Sigma|^{-\frac{1}{2}}\int_\Theta J(r_0,\theta)\int_{\mathbb{R}^d} h\left(\left[y_{1:d},r_0e(\theta)\right]'\Sigma^{-1}\left[y_{1:d},r_0e(\theta)\right]\right)dy_{1:d}\,d\theta. \quad \text{(A.16)}$$

By symmetry of the integration domain, the integral in the last term reduce to the diagonal terms, hence:

$$\begin{aligned}\frac{\partial r_\varepsilon}{\partial\varepsilon}_{|\varepsilon=0} &= \frac{1}{G}\Bigg((1-\alpha)(1-\text{trace}(IF(x_0,V,P))) \quad \text{(A.17)}\\ &\quad -2|\Sigma|^{-\frac{1}{2}}\int_{S(P)}\dot{h}\left(y'\Sigma^{-1}y\right)y'\Sigma^{-1}\text{diag}(IF(x_0,V,P))ydy-I_{S(P)}(x_0)\Bigg)\end{aligned}$$

At this point, we have an expression of $IF(x_0, C, P)$ as a function of $IF(x_0, V, P)$ where V is the matrix of eigenvectors of C. By Lemma 3 in Croux and Haesbroeck (2000), the last influence function elements may be expressed in term of the firsts. So we'll end up with an expression involving only $IF(x_0, C, P)$ and some constants.

Combining (A.11), (A.12), (A.13), (A.15), and (A.17), $IF(x_0, C, P)$ becomes

$$
\begin{aligned}
= \quad & \frac{c}{1-\alpha} I_{S(P)}(x_0) \left(x_0 x'_0 - \frac{1}{G} A \right) \\
& + (\text{trace}(IF(x_0, V, P)) - 1) \left(C(P) - \frac{cA}{G} \right) \\
& - \frac{cA}{(1-\alpha)G} 2|\Sigma|^{-\frac{1}{2}} \int_{S(P)} \dot{h}\left(y'\Sigma^{-1}y\right) y'\Sigma^{-1}\text{diag}(IF(x_0, V, P))y dy \\
& + \frac{c}{(1-\alpha)} |\Sigma|^{-\frac{1}{2}} \int_{S(P)} \{IF(x_0, V, P)yy' + yy' IF(x_0, V, P)'\} h(y'\Sigma^{-1}y) dy \\
& + \frac{c}{1-\alpha} |\Sigma|^{-\frac{1}{2}} \int_{S(P)} yy' \dot{h}\left(y'\Sigma^{-1}y\right) 2y'\Sigma^{-1} IF(x_0, V, P) y dy
\end{aligned}
$$

Using Lemma 3 of Croux & Haesbroeck (2000) and the diagonality of $C(P)$, the diagonal elements of the influence function of V, the matrix of eigenvector functionals, is zero, hence is also the trace, and that the non diagonal elements are given by

$$
IF(x_0, V, P)_{jk} = \frac{IF(x_0, C, P)_{jk}}{\Lambda_k(P) - \Lambda_j(P)} \tag{A.18}
$$

So that we have to assume that all eigenvalues are distinct. As was mentioned above, a more refined change of variable, with the identity on the last $p - d$ coordinates, would avoid this assumption. We end up with the simplified

form for $IF(x_0, C, P)$:

$$\begin{aligned} = & \frac{c}{1-\alpha} I_{S(P)}(x_0) \left(x_0 x'_0 - \frac{1}{G} A \right) - C(P) + \frac{cA}{G} \\ & + \frac{c}{1-\alpha} |\Sigma|^{-\frac{1}{2}} \int_{S(P)} \left\{ IF(x_0, V, P) yy' + yy' IF(x_0, V, P)' \right\} h(y'\Sigma^{-1} y) dy \\ & + \frac{c}{1-\alpha} |\Sigma|^{-\frac{1}{2}} \int_{S(P)} yy' \dot{h} \left(y'\Sigma^{-1} y \right) 2y'\Sigma^{-1} IF(x_0, V, P) y dy, \end{aligned}$$

where the two terms in the integral of the second line above involve the matrix $C(P)$, so that we can further simplify into

$$\begin{aligned} IF(x_0, C, P) = & \frac{c}{1-\alpha} I_{S(P)}(x_0) \left(x_0 x'_0 - \frac{1}{G} A \right) - C(P) + \frac{cA}{G} \\ & + IF(x_0, V, P) C(P) + C(P) IF(x_0, V, P)' \\ & + \frac{c}{1-\alpha} |\Sigma|^{-\frac{1}{2}} \int_{S(P)} yy' \dot{h} \left(y'\Sigma^{-1} y \right) 2y'\Sigma^{-1} IF(x_0, V, P) y dy. \end{aligned} \tag{A.19}$$

Regarding the second line in the above formula, we are using (A.18) for the element in position i, j:

$$\begin{aligned} & \left[IF(x_0, V, P) C(P) + C(P) IF(x_0, V, P)' \right]_{ij} \\ & \qquad = IF(x_0, V, P)_{ij} \Lambda_j(P) + IF(x_0, V, P)_{ji} \Lambda_i(P) \\ & \qquad = (1 - \delta_{ij}) \left(\frac{IF(x_0, C, P)_{ij}}{\Lambda_j(P) - \Lambda_i(P)} \Lambda_j(P) + \frac{IF(x_0, C, P)_{ji}}{\Lambda_i(P) - \Lambda_j(P)} \Lambda_i(P) \right) \\ & \qquad = (1 - \delta_{ij}) IF(x_0, C, P)_{ij} \end{aligned}$$

since $IF(x_0, C, P)$ is a symmetric matrix, and $C(P)$ has its eigenvalues $\Lambda_j(P)$ on its diagonal.

Let us now treat the last integral in (A.19):

$$J = \frac{c}{1-\alpha} |\Sigma|^{-\frac{1}{2}} \int_{S(P)} yy' \dot{h} \left(y'\Sigma^{-1} y \right) 2y'\Sigma^{-1} IF(x_0, V, P) y dy.$$

We consider a typical element of the resulting matrix

$$J_{ij} = \frac{c}{1-\alpha}|\Sigma|^{-\frac{1}{2}} \int_{S(P)} y_i y_j \dot{h}\left(y'\Sigma^{-1}y\right) 2y'\Sigma^{-1} IF(x_0, V, P) y dy.$$

The integrand is given by

$$2\dot{h}\left(y'\Sigma^{-1}y\right) \sum_{k,l=1}^{p} y_i y_j y_k y_l \frac{IF(x_0, V, P)_{kl}}{\lambda_k},$$

where we recall that $\lambda_1, \ldots, \lambda_p$ are the eigenvalues and also diagonal elements of the matrix Σ. By symmetry of the integration domain, only those terms where the indices i, j, k, l are such that only even powers of y are present will contribute to the integral. Moreover, for $k = l$ the influence function is zero. That is, non-zero contributions come from $k \neq l$ and $i = k, j = l$ or $i = l, j = k$. So that for the element i, j of J, the contribution comes from

$$2\dot{h}\left(y'\Sigma^{-1}y\right) y_i^2 y_j^2 \left(\frac{IF(x_0, V, P)_{ij}}{\lambda_i} + \frac{IF(x_0, V, P)_{ji}}{\lambda_j}\right).$$

Using (A.18) and the symmetry of $IF(x_0, C, P)$, we get

$$\begin{aligned} J_{ij} &= (1-\delta_{ij}) IF(x_0, C, P)_{ij} \frac{\lambda_j - \lambda_i}{(\Lambda_j(P) - \Lambda_i(P))\lambda_i\lambda_j} \frac{2c}{1-\alpha} |\Sigma|^{-\frac{1}{2}} \int_{S(P)} y_i^2 y_j^2 \dot{h}\left(y'\Sigma^{-1}y\right) dy \\ &= (1-\delta_{ij}) IF(x_0, C, P)_{ij} \frac{\lambda_j - \lambda_i}{(\Lambda_j(P) - \Lambda_i(P))\lambda_i\lambda_j} \frac{2c}{1-\alpha} H_{ij}, \end{aligned}$$

with

$$H_{ij} = |\Sigma|^{-\frac{1}{2}} \int_{S(P)} y_i^2 y_j^2 \dot{h}\left(y'\Sigma^{-1}y\right) dy. \tag{A.20}$$

In the end, following up on (A.19), we can write an element i, j of the influence function for C:

$$\begin{aligned} IF(x_0, C, P)_{ij} &= \frac{c}{1-\alpha} I_{S(P)}(x_0) \left(x_{0i}x_{0j} - \frac{A_{ij}}{G}\right) - C(P)_{ij} + \frac{cA_{ij}}{G} \\ &\quad + (1-\delta_{ij}) IF(x_0, C, P)_{ij} \left(1 + \frac{2c}{(1-\alpha)} \frac{\lambda_j - \lambda_i}{(\Lambda_j(P) - \Lambda_i(P))\lambda_i\lambda_j} H_{ij}\right). \end{aligned}$$

Given the expression above, it is profitable to give separate expression for diagonal and off-diagonal terms. We have

$$IF(x_0, C, P)_{ii} = \frac{c}{1-\alpha} I_{S(P)}(x_0) \left(x_{0i}^2 - \frac{A_{ii}}{G} \right) - \Lambda_i(P) + \frac{cA_{ii}}{G}$$

and for an off-diagonal term ($i \neq j$), we use that A in (A.14) is diagonal,

$$\begin{aligned} IF(x_0, C, P)_{ij} = \frac{c}{1-\alpha} I_{S(P)}(x_0) x_{0i} x_{0j} \\ + IF(x_0, C, P)_{ij} \left(1 + \frac{2c}{1-\alpha} \frac{\lambda_j - \lambda_i}{(\Lambda_j(P) - \Lambda_i(P))\lambda_i \lambda_j} H_{ij} \right) \end{aligned}$$

which gives

$$IF(x_0, C, P)_{ij} = -\frac{(\Lambda_j(P) - \Lambda_i(P))\lambda_i \lambda_j}{2(\lambda_j - \lambda_i)} \frac{I_{S(P)}(x_0) x_{0i} x_{0j}}{H_{ij}}. \quad \square$$

PROOF OF THEOREM 8: Let $\mathcal{X} = \{x_1, ..., x_n\} \subset \mathbb{R}^p$ be the original sample and $h_{\mathcal{X}} \in \mathcal{A}_d(\mathbb{R}^p)$ the empirical d-dimensional trimmed principal component of $\mathcal{X}$. Assume, without loss of generality, that $D(\mathcal{X}) = d(h_{\mathcal{X}}, 0) = 0$. Let us denote $R = \max_{i=1,...,n} d(x_i, 0)$. Then the original sample satisfies $\mathcal{X} \subset B(0, R)$.

We will develop the proof for the case $(\lfloor n\alpha \rfloor + d + 1)/n \leq (n - \lfloor n\alpha \rfloor)/n$. In the other case the proof is easier. Note that this inequality may be equivalently rewritten as $n - 2\lfloor n\alpha \rfloor - d \geq 1$.

The proof will be arranged in two steps:

1. We first prove that $\varepsilon_n^*(D, \mathcal{X}) \geq (\lfloor n\alpha \rfloor + d + 1)/n$. If we replace at most $\lfloor n\alpha \rfloor + d$ points of $\mathcal{X}$ in order to obtain a corrupted sample $\mathcal{X}'$, then at least $n - \lfloor n\alpha \rfloor - d$ original points remain in $\mathcal{X}'$. Let $h_{\mathcal{X}'} \in \mathcal{A}_d(\mathbb{R}^p)$ be the empirical d-dimensional trimmed principal component of $\mathcal{X}'$, which is based on a subsample $\mathcal{Y}' \subset \mathcal{X}'$ containing $n - \lfloor n\alpha \rfloor$ data points. Note that

$n-\lfloor n\alpha\rfloor-(\lfloor n\alpha\rfloor+d)=n-2\lfloor n\alpha\rfloor-d\geq 1$, therefore any subsample $\mathcal{Y}'\subset\mathcal{X}'$ contains at least 1 data point from the original sample $\mathcal{X}$.

Assume that for any arbitrarily large constant $C>(\sqrt{n}+1)R$, we could get a contaminated sample satisfying $D(\mathcal{X}')=d(h_{\mathcal{X}'},0)\geq C$. Then, we would have

$$(n-\lfloor n\alpha\rfloor)V_{d,\alpha}(h_{\mathcal{X}'})=\sum_{y\in\mathcal{Y}'}d(y,h_{\mathcal{X}'})^2\geq\sum_{y\in\mathcal{Y}'\cap\mathcal{X}}d(y,h_{\mathcal{X}'})^2\geq(C-R)^2>nR^2.$$

On the other hand, if we considered a subsample $\mathcal{Y}^*\subset\mathcal{X}'$ made of $n-\lfloor n\alpha\rfloor-d$ points belonging to $\mathcal{X}$ together with d arbitrary points belonging to $\mathcal{X}'-\mathcal{X}$ and the affine subspace $h_{\mathcal{Y}^*}\in\mathcal{A}_d(\mathbb{R}^p)$ containing the origin 0 and those d arbitrary points, then we would have

$$(n-\lfloor n\alpha\rfloor)V_{d,\alpha}(h_{\mathcal{Y}^*})=\sum_{y\in\mathcal{Y}^*}d(y,h_{\mathcal{Y}^*})^2=\sum_{y\in\mathcal{Y}^*\cap\mathcal{X}}d(y,h_{\mathcal{Y}^*})^2\leq nR^2.$$

Then, we would get $V_{d,\alpha}(h_{\mathcal{Y}^*})<V_{d,\alpha}(h_{\mathcal{X}'})$, contradicting the fact that $h_{\mathcal{X}'}$ is a d-dimensional trimmed principal component of $\mathcal{X}'$. Therefore, $\sup_{\mathcal{X}'}d(h_{\mathcal{X}'},0)<\infty$ and the first inequality is proven.

2. We now prove that $\varepsilon_n^*(D,\mathcal{X})\leq(\lfloor n\alpha\rfloor+d+1)/n$. The goal is now to built a corrupted sample $\mathcal{X}'$ replacing at least $\lfloor n\alpha\rfloor+d+1$ points of $\mathcal{X}$ in such a way that the optimum $h_{\mathcal{X}'}$ satisfies that $D(\mathcal{X}')$ is arbitrarily large. Firstly, note that $h_{\mathcal{X}'}$ would be based on a subsample $\mathcal{Y}'\subset\mathcal{X}'$ of size $n-\lfloor n\alpha\rfloor$ containing at least $d+1$ corrupted observations belonging to $\mathcal{X}'-\mathcal{X}$ and at most $n-\lfloor n\alpha\rfloor-d-1$ points from the original sample $\mathcal{X}$. Given $M>0$, let us consider a d-dimensional subspace h_0 parallel to $h_{\mathcal{X}}$ and satisfying $d(h_0,h_{\mathcal{X}})=M$. Take a corrupted sample $\mathcal{X}'$ satisfying:

(i) $\mathcal{X}'-\mathcal{X}\subset h_0$,

(ii) $d(y, y') \geq M$ for every $y, y' \in \mathcal{X}' - \mathcal{X}$ for $y \neq y'$, and,

(iii) every subset of $d+1$ points in $\mathcal{X}' - \mathcal{X}$ are in general position.

Some technicalities that will be here omitted (see San Martín (2008) for details) lead us to $\lim_{M\to\infty} d(h_{\mathcal{X}'}, 0) = \infty$ and the result is proven. □

Acknowledgements:

C. Croux was supported by the Research Fund K.U. Leuven and the "Fonds voor Wetenschappelijk Onderzoek". The work of C. Ruwet was partially supported by the IAP Research Network P7/06 of the Belgian State. The research of the other authors was partially supported by Ministerio de Ciencia y Tecnología and FEDER grant MTM2005-08519-C02-01 and by Consejería de Educación y Cultura de la Junta de Castilla y León grant PAPIJCL VA102A06.

References

[1] Anderson, T. W. (1955), "The integral of a symmetric unimodal function over a symmetric convex set and some probability inequalities", *Proc. Amer. Math. Soc.,* 6, 170-176.

[2] Billingsley, P. (1986), *Probability and Measure* (2nd Ed.), New York: Wiley.

[3] Campbell, N.A. (1980), "Robust Procedures in Multivariate Analysis I: Robust Covariance Estimation", *J. R. Stat. Soc. Ser. C. Appl. Stat.,* 29, 231-237.

[4] Croux, C., Filzmoser, P. and Fritz, H. (2013), "Robust Sparse Principal Component Analysis", *Technometrics,* 55, 202-214

[5] Croux, C. and Haesbroeck, G. (1999), "Influence function and efficiency of the minimum covariance determinant scatter matrix estimator", *J. Multivariate. Anal.,* 71, 161-190.

[6] 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.

[7] Croux, C., Ollila, E. and Oja, H. (2002). "Sign and Rank covariance matrices: Statistical properties and applications to principal components analysis", *Statistical data analysis based on the L1-norm and related methods (Neuchtel, 2002)*, Statistics for Industry and Technology, 257269.

[8] Croux, C., and Ruiz-Gazen, A. (2005), "High Breakdown Estimators for Principal Components: the Projection-Pursuit Approach Revisited", *J. Multivariate. Anal.,* 95, 206-226

[9] Croux, C., Fekri, M. and Ruiz-Gazen, A. (2009), "Fast and Robust estimation of the Multivariate Errors in Variables Model," *Test*, 19, 286-303.

[10] Cuesta-Albertos, J.A. and Matrán, C. (1988), "The Strong Law of Large Numbers for k-Means and Best Possible Nets of Banach Valued Random Variables", *Probab. Theory Relat. Fields,* 78, 523-534.

[11] Cuesta-Albertos, J.A., Gordaliza, A. and Matrán, C. (1997), "Trimmed k-means: An attempt to robustify quantizers", *Ann. Statist.,* 25, 553-576.

[12] Cuesta-Albertos, J.A., Gordaliza, A. and Matrán, C. (1998), "Trimmed best k-nets: A robustified version of an L_∞-based clustering method", *Statist. Probab. Lett.,* 36, 401-413.

[13] Cuesta-Albertos, J.A., García-Escudero, L. A. and Gordaliza, A. (2002), "On the Asymptotics of Trimmed Best k-Nets", *J. Multivariate. Anal.,* 82, 486-516.

[14] Davies, P.L. (1987). "Asymptotic behaviour of S-estimates of multivariate location parameters and dispersion matrices", *Ann. Statist.,* 15, 1269-1292.

[15] Debruyne, M. and Verdonck, T. (2010), "Robust kernel principal component analysis and classification", *Adv. Data Anal. Classif.,* 4, 151-167.

[16] Devlin, S.J., Gnanadesikan, R. and Kettering, J.R. (1981), "Robust Estimation of Dispersion Matrices and Principal Components", *J. Amer. Statist. Assoc.,* 76, 354-362.

[17] Donoho, D.L. and Huber, P.J. (1983), The notion of breakdown point. In: Bickel, P., Doksum, K. and Hodges Jr. J.L., eds., *A Festschrift for Erich L. Lehmann.* Wadsworth, Belmont, CA.

[18] García-Escudero, L.A. and Gordaliza, A. (1999), "Robustness Properties of k-Means and Trimmed k-Means'," *J. Amer. Statist. Assoc.,* 94, 956-969.

[19] García-Escudero, L. A., Gordaliza, A., Matrán, C. and Mayo-Iscar, A. (2008), "A general trimming approach to robust cluster analysis", *Ann. Statist.,* 36, 1324-1345.

[20] García-Escudero, L. A., Gordaliza, A., San Martín, R., Van Aelst, S. and Zamar, R. (2009), "Robust linear clustering", *J. R. Statist. Soc. B,* 71, 301-318.

[21] Gordaliza, A. (1991), "Best approximations to random variables based on trimming procedures", *J. Approx. Theory,* 64, 162-180.

[22] Hampel, F.R. (1971), "A General Qualitative Definition of Robustness", *Ann. Math. Statist.,* 42, 1887-1896.

[23] Hampel, F.R. (1974), "The Influence Function and its Role in Robust Estimation", *J. Amer. Statist. Assoc.,* 69, 383-393.

[24] Hampel, F.R., Rousseeuw, P.J., Ronchetti, E. and Stahel, W.A. (1986), *Robust Statistics, The Approach Based on The Influence Function,* New York: Wiley.

[25] Hubert, M., Rousseeuw, P.J., and Vanden Branden, K. (2005), "ROBPCA: A new approach to robust principal component analysis", *Technometrics,* 47, 64-79.

[26] Jolliffe, I.T.(2002), *Principal Component Analysis, 2nd ed,* Berlin: Springer-Verlag.

[27] Li, G., and Chen, Z. (1985), "Projection Pursuit Approach to Robust Dispersion Matrices and Principal Components: Primary Thery and Monte Carlo", *J. Amer. Statist. Assoc.*, 80, 759-766.

[28] Liu, R.Y., Parelius, J.M and Singh, K. (1999). Multivariate Analysis

by Data Depth: Descriptive Statistics, Graphics and Inference. *Ann. Statist.,* 27, 783-840.

[29] Maronna, R. (2005). "Principal Components and Orthogonal Regression Based on Robust Scales", *Technometrics,* 47, 264-273.

[30] Rousseeuw, P.J. (1984). "Least median of squares regression", *J. Amer. Statist. Assoc.,* 79, 871-880.

[31] Rousseeuw, P.J. (1985). "Multivariate Estimation with High Breakdown Point", in *Mathematical Statistics and Applications*, edited by W. Grossmann, G. Pflug, I. Vincze, and W. Wertz, Reidel Publishing Company, Dordrecht, 283-297.

[32] Rousseeuw, P.J. and Van Driessen, K. (1999), "A Fast Algorithm for the Minimum Covariance Determinant Estimator", *Technometrics*, 41, 212-223.

[33] San Martín, R. (2008), Recortes Imparciales en torno a Subespacios Afines. Aplicaciones al Análisis Multivariante Robusto, *Unpublished Ph.D. Thesis.*

[34] Serneels, S., and Verdonck, T. (2009) "Principal component regression for data containing outliers and missing elements", *Comput. Statist. Data Anal.,* 53, 3855-3863.

[35] Xu, H., Caramanis, C., and Sanghavi, S. (2012), "Robust PCA via Outlier Pursuit", *IEEE Trans. Inform. Theory,* 58, 30473064.