ROBUST PRINCIPAL COMPONENT ANALYSIS BASED ON TRIMMING AROUND AFFINE SUBSPACES

C. Croux[1], L.A. García-Escudero[2], A. Gordaliza[2], C. Ruwet[3] and R. San Martín[2]

[1]*KU Leuven,* [2]*Universidad de Valladolid and* [3]*HE de la Province de Liège*

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. PCA suffers from a well-known lack of robustness. As a robust alternative, one can resort to an impartial trimming based approach and search 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 solutions to both the sample and population problems will be proven. Moreover, under mild conditions, the solutions of the sample problem are consistent toward the solutions of the population one. The robustness of the method is studied by proving qualitative robustness, computing the breakdown point, and deriving the influence functions. Furthermore, asymptotic efficiencies at the normal model are derived and finite sample efficiencies are studied by means of a simulation study.

Key words and phrases: Affine Subspaces, Dimension Reduction, Orthogonal Regression, Principal components, Multivariate statistics, Robustness, Trimming.

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 presence of outlying observations in the data set. There are

a lot of examples in the literature showing that 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 et al. (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 et al. (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 therein).

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 $n - \lfloor n\alpha \rfloor$, 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:

$$\min_{\mathcal{Y} \subset \mathcal{X},\, \#\mathcal{Y} \geq n - \lfloor n\alpha \rfloor} \; \min_{h \in \mathcal{A}_d(\mathbb{R}^p)} \frac{1}{\#\mathcal{Y}} \sum_{x_i \in \mathcal{Y}} \|x_i - \mathrm{Pr}_h(x_i)\|^2, \tag{1.1}$$

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 "best" subspace according to (1.1) 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 powerful tool to robustify statistical methods. The idea of discarding a symmetric proportion of extreme observations

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

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 tells us the best way of trimming a fixed proportion α of the data.

The problem defined in (1.1) 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, while asymptotic variances are obtained in Section 7. Section 8 provides finite-sample efficiencies. We also compare the robustness of different robust estimators for PCA by means of a simulation study. Section 9 contains a data example and we end with a concluding section. All the proofs are deferred to a supplementary file.

2. Notation and preliminary results

In this paper, X is a $\mathbb{R}^p$-valued random vector (r.v.) defined on a probability

space, β^p denotes the σ-algebra of all Borel sets in $\mathbb{R}^p$, P_X the probability measure induced by X on $(\mathbb{R}^p, \beta^p)$ and $\|\cdot\|$ 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 $0 \leq \beta \leq \alpha$,

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

A more general statement of the Robust Principal Component Analysis problem based on trimming can be given by using trimming functions instead of trimming subsets:

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

$$\inf_{\tau \in \mathcal{T}_{\alpha-}} \inf_{h \in \mathcal{A}_d(\mathbb{R}^p)} \frac{1}{\int \tau(x) dP_X(x)} \int \tau(x) \|x - \mathrm{Pr}_h(x)\|^2 dP_X(x). \tag{2.1}$$

The minimum value in (2.1) 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 (2.1) 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 (2.1). We recall that all proofs can be found in the supplementary file.

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

The next lemma shows that the solution in (2.1) is characterized by a *strip*. Given $h \in \mathcal{A}_d(\mathbb{R}^p)$ and $r \geq 0$, we define the strip 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

$$V_{d,\beta}(h) := \frac{1}{1-\beta}\int \tau_{h,\beta}(x)\|x - \mathrm{Pr}_h(x)\|^2 dP_X(x), \tag{2.2}$$

is the same for every $\tau_{h,\beta} \in \mathcal{T}_{h,\beta}$. We call (2.2) the *$\beta$-trimmed variation* of X around 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 (2.2). 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 (2.1).

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 above proposition allows us to simplify the original double minimization problem (2.1) to the single search of the optimal affine subspace. Once 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 (2.1), 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 (1.1) is equivalent to the problem (2.1) 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 (2.1). 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

$$V_{d,\alpha} = \inf_{h\in\mathcal{A}_d(\mathbb{R}^p)} V_{d,\alpha}(h) < \infty, \tag{3.1}$$

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 *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, h_0) is a solution of (2.1), then $I_{S(h_0,r_\alpha(h_0))} \leq \tau_0 \leq I_{\overline{S}(h_0,r_\alpha(h_0))}$, P_X-a.e. Moreover, if P_X is absolutely continuous w.r.t. the Lebesgue measure on $\mathbb{R}^p$, then $I_{S(h_0,r_\alpha(h_0))} = \tau_0$, P_X-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 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 (2.1) and 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 $P_X^{\tau_0}$ and then τ_0 and h_0 would not be a solution of (2.1).

4. Consistency

While Theorem 1 guarantees the existence of solutions for the population and the sample problems, 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.

From now on, $\{X_n\}_n$ is a sequence of $\mathbb{R}^p$-valued r.v. and $h_n \in \mathcal{A}_d(\mathbb{R}^p), n = 1, 2, ...$, is the d-dimensional trimmed principal component subspace 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 on 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. 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 *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 a.s. 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 r.v. 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 $\{Y_n\}_n$, 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 r.v. $\{X_n\}_n$, with probability distribution P_X and recall that problem stated in (1.1) is equivalent to the problem (2.1) taking P_n^ω instead of P_X. Furthermore, it is well-known that the set $\Omega_0 := \{\omega \in \Omega$ such that P_n^ω converges in distribution to $P_X\}$ satisfies $P(\Omega_0) = 1$. Thus, the desired consistency result follows as a simple consequence of Corollary 3:

Theorem 3 *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*

$$h_n^\omega \to h_0, P\text{-a.s.} \quad \text{and} \quad V_{d,\alpha}(P_n^\omega) \to V_{d,\alpha}(X), P\text{-a.s.}$$

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 r.v. X follows an elliptical symmetric distribution $X \sim E_p(\mu, \Sigma)$ if it admits a probability density function of the form

$$f_X(x) = |\Sigma|^{-\frac{1}{2}} h((x-\mu)'\Sigma^{-1}(x-\mu)) \text{ for } x \in \mathbb{R}^p \tag{5.1}$$

where h is a positive and non-increasing square integrable function called the *radial function*. The density f is called unimodal if the radial function h has a strictly positive derivative $\dot{h}$. The *location parameter* of the distribution is μ and 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. 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 *Let X be a random vector having an elliptically symmetric distribution as in* (5.1)*, 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 proof of the uniqueness result needs the application of a multivariate probability inequality in Davies (1987), which is given in the supplementary file. 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, from now on, 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

(5.1), 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

$$C(P) = \frac{1}{1-\alpha} \int_{S(P)} (x - m(P))(x - m(P))' dP(x). \tag{5.2}$$

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 *Let P be with density as in (5.1). 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 $cC(P)$ are equal to the first d eigenvalues and eigenvectors of the covariance matrix of P. At the multivariate normal distribution, one has $c = 1$.*

To get Fisher consistency, the matrix $C(P)$ needs to be multiplied by a constant c. In the sequel, the functional C will always 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 one.

6. Robustness

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.

He also defines 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.

Theorem 6 *The d-dimensional trimmed principal component subspace functional is weakly continuous and qualitatively robust at any absolutely continuous distribution P having a unique d-dimensional trimmed principal component subspace.*

Notice that we need again a uniqueness condition. This condition is similar to that needed for the population median in stating the qualitative robustness of the median estimator or the one 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 (IF) is the keystone of Hampel's infinitesimal approach to Robust Statistics (Hampel 1974 and Hampel et al. 1986). It quantifies the impact than an observation has on an estimator. The IF is a common measure of robustness, also in the context of principal components analysis. Furthermore, it a useful tool for computing 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, at 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 $IF(x_0; T, P) = \lim_{\varepsilon \downarrow 0}(T((1-\varepsilon)P + \varepsilon\delta_{\{x_0\}}) - T(P))/\varepsilon$, for those x_0 where this limit exists. Here $\delta_{\{x_0\}}$ denotes a Dirac distribution putting all 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 (5.2). 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.

Theorem 7 *At an elliptical distribution function P with probability density function given by* (5.1)*, with $\mu = 0$, and $\Sigma = diag(\lambda_1, \ldots, \lambda_p)$, we have that for any*

diagonal term of C:

$$IF(x_0; C, P)_{ii} = \frac{c}{1-\alpha} I_{S(P)}(x_0) \left(x_{0i}^2 - \frac{A_{ii}}{G} \right) - \Lambda_i(P) + \frac{cA_{ii}}{G} \qquad (6.1)$$

and for any 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 given in the supplementary file, section (S10).

We note that the influence functions are not bounded. This comes 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 their influence is 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 spanned by 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 $IF(x_0, \Lambda_i, P) = IF(x_0, C, P)_{ii}$ is obtained. 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

$$IF(x_0, V_i, P) = \sum_{j \neq i} \frac{\lambda_i\lambda_j}{\lambda_j - \lambda_i} \frac{I_{S(P)}(x_0) x_{0i} x_{0j}}{2H_{ij}} v_j, \qquad (6.2)$$

with v_j the jth eigenvector of Σ.

To conclude this section, Figures 6.1 and 6.2 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.

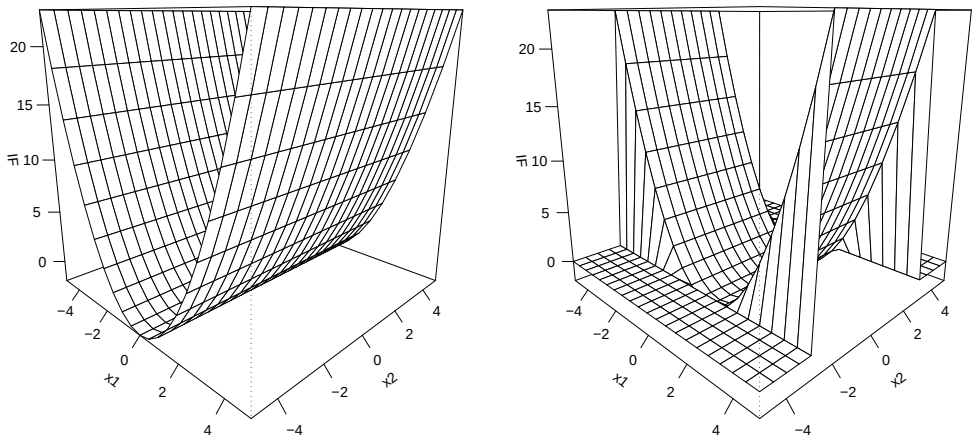

Figure 6.1: Influence function of the largest eigenvalue at $P = N(0, \text{diag}(2, 1))$ when $\alpha = 0$ (left panel) and $\alpha = 0.01$ (right panel).

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.

6.3. Breakdown Point

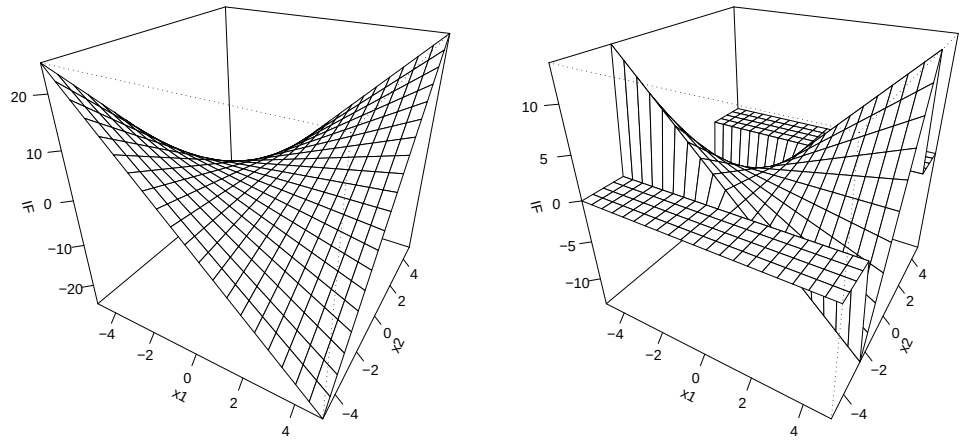

Figure 6.2: Influence function of the eigenvector associated to the largest eigenvalue at $P = N(0, \mathrm{diag}(2, 1))$ when $\alpha = 0$ (left panel) and $\alpha = 0.01$ (right panel).

The influence function 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 function of the estimator can be expected to extend. We 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 on 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. $\varepsilon_n^*(T, \mathcal{X}) = \min\left\{k/n;\ \sup_{\mathcal{X}'} \|T(\mathcal{X}) - T(\mathcal{X}')\| = \infty\right\}$, with $\mathcal{X}'$ ranging on the set of all possible samples obtained by replacing k original data points in the sample 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 PCA 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.

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 $\mathcal{X}$, satisfies*

$$\varepsilon_n^*(D,\mathcal{X}) = \min\left\{(\lfloor n\alpha\rfloor + d + 1)/n, (n - \lfloor n\alpha\rfloor)/n\right\} \to \alpha, \text{ as } n \to \infty.$$

Maronna (2005) also analyzed the breakdown point of this procedure. His result coincides with that in Theorem 8 but he focused on the breakdown of the "trimmed scale" target function (i.e., (1.1)) in terms of preventing it to become 0 or ∞ ("implosion" and "explosion"). In our result, we consider a different situation where the whole PCA subspace may be unbounded by taking an arbitrarily large "distance to the origin". He also introduced an alternative breakdown point concept based on "prediction bias" but he considered that the needed calculations seemed intractable even in the most simple case.

The breakdown point of the "distance to the origin" has its limitation. It considers breakdown due to shifts, but tells nothing about the orientation of the eigenvectors. It might be that the estimated eigenvectors go totally wrong, while the distance to the origin remains bounded. However, the latter type of breakdown is more difficult to formalize and to compute, and we refer to Tyler (2005) for further discussion of the definition of breakdown point for eigenvectors.

7. Asymptotic variances

Under the hypothesis that a functional T is Frechet differentiable, its asymptotic distribution is gaussian, and its asymptotic variance is given by

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

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

expressions of the influence function in section 6.2 allow us to compute the asymptotic variance in a rather easy way.

7.1. Asymptotic variances in the elliptical case

For an elliptical contoured distribution with $\mu = 0$ and $\Sigma = diag(\lambda_1, \ldots, \lambda_p)$, from (6.1) and (6.2), we can obtain the following expressions for the asymptotic variances for the associated eigenvalues and eigenvectors estimators:

$$\mathrm{ASV}(\Lambda_i, P) = \frac{c^2}{(1-\alpha)^2} \int_{S(P)} x_i^4 |\Sigma|^{-\frac{1}{2}} h(x'\Sigma^{-1}x)dx - \Lambda_i(P)^2$$

$$+\frac{\alpha}{1-\alpha}\left(\frac{cA_{ii}}{G}\right)^2 + 2\Lambda_i(P)\frac{cA_{ii}}{G}(\frac{-\alpha}{1-\alpha})$$

and

$$\mathrm{ASV}(V_i, P) = \sum_{j \neq i} \frac{\lambda_i^2 \lambda_j^2}{(\lambda_i - \lambda_j)^2} \frac{\int_{S(P)} x_i^2 x_j^2 dP(x)}{4H_{ij}^2} v_j {v_j}', \tag{7.1}$$

where the quantities G, A_{ii} and H_{ij} are again those given in the supplementary file.

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 $\Lambda_i(P) = \lambda_i$. These results allow for simpler expressions of the asymptotic variances of the eigenvalues with $1 \leq i \leq d$ as:

$$\mathrm{ASV}(\Lambda_i, P) = \frac{2}{1-\alpha}\lambda_i^2. \tag{7.2}$$

For the eigenvectors with $1 \leq i \leq d$, we obtain:

$$\mathrm{ASV}(V_i, P) = \frac{1}{1-\alpha} \sum_{j \neq i} \frac{\lambda_i \lambda_j c_j}{(\lambda_i - \lambda_j)^2} v_j {v_j}' \tag{7.3}$$

with c_j defined as

$$c_j^{-1} = \frac{\int_{S(P)} x_j^2 dP(x)}{(1-\alpha)\lambda_j}. \tag{7.4}$$

The availability of asymptotic variances under closed form expressions allows us to compute asymptotic relative efficiencies (ARE) with respect to maximum likelihood (ML) estimators at the gaussian model. 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 (7.2) that, for $1 \leq i \leq d$

$$\mathrm{ARE}(\Lambda_i, P) = \frac{\mathrm{ASV}(\Lambda_{ML;i}, P)}{\mathrm{ASV}(\Lambda_i, P)} = \frac{2}{2/(1-\alpha)} = 1-\alpha,$$

meaning that, for the first d eigenvalues, the efficiency is just the complementary of the trimming proportion. For instance, a trimming level of 10% yields a 90% efficiency for the eigenvalue estimators.

Regarding eigenvectors, we have from (7.3) that

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

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 (7.4), 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

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

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 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 performance 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), representing:

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

(b) 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 $d = 3,$ or 7 and $\alpha = 0.05, 0.1$ or 0.25.

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.

8.1 Finite-sample efficiencies

In this subsection we verify whether the asymptotic variances of the estimators, computed in Section 7, are confirmed by their finite sample counterparts.

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:

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

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 et al. (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.

From the MSE values, relative finite sample efficiencies are computed as

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

These finite sample efficiencies are reported in Table 8.1. 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 8.1, $p = 5, d = 3$, the efficiency decreases with an increasing trimming size. The finite sample efficiency of the eigenvalues tends to decrease towards the asymptotic value, while they increase for the eigenvectors towards the limit value with increasing sample size. The results for $p = 8$, where the trimming size is 0.25, show that if the model dimension d increases, everything else being fixed, a small increase in the efficiency of the eigenvectors is observed. This behavior has already been pointed out when studying the asymptotic efficiencies.

Under design (b), there is a large difference between the noise and non-noise levels. The convergence towards the asymptotic efficiencies is here slower than for simulation design (a). Note that some finite sample efficiencies are larger than one, which is possible since they are computed relative to the asymptotic variance of the ML estimator. The ML estimator itself also has finite sample efficiencies larger than one in these cases (see supplementary file). For $p = 8, d = 7$ the finite

Table 8.1: Finite sample efficiencies for the trimmed PCA w.r.t. the ML.

Design (a)										
p	d	α	n	Eigenvalues	Eigenvectors					
5	3	.05	50	.992	.754	.677	.590			
			100	.979	.918	.845	.710			
			500	.942	.927	.900	.852			
			∞	.950	.932	.922	.846			
5	3	.10	50	.985	.652	.608	.502			
			100	.912	.762	.782	.650			
			500	.905	.828	.809	.710			
			∞	.900	.869	.853	.736			
5	3	.25	50	.837	.458	.428	.356			
			100	.761	.586	.554	.436			
			500	.762	.662	.630	.476			
			∞	.750	.689	.659	.483			
8	3	.25	50	.806	.497	.447	.356			
			100	.762	.565	.513	.429			
			500	.722	.665	.654	.527			
			∞	.750	.692	.665	.502			
8	7	.25	50	.816	.532	.476	.444	.457	.427	.393	.353
			100	.791	.628	.629	.605	.593	.603	.554	.446
			500	.770	.755	.732	.714	.698	.689	.643	.517
			∞	.750	.746	.746	.742	.733	.712	.654	.435
Design (b)										
p	d	α	n	Eigenvalues	Eigenvectors					
5	3	.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	3	.25	100	1.026	.623	.564	.511			
			500	.813	.495	.485	.561			
			1000	.778	.651	.651	.678			
			∞	.750	.749	.749	.747			
8	3	.25	100	1.009	.614	.541	.485			
			500	.808	.541	.542	.618			
			1000	.752	.685	.669	.709			
			∞	.750	.748	.748	.745			
8	7	.25	100	1.523	1.253	1.306	1.094	.858	.669	.536	.479
			500	.969	.598	.521	.473	.485	.532	.572	.590
			1000	.866	.547	.505	.516	.552	.601	.641	.691
			∞	.750	.750	.750	.750	.750	.750	.749	.747

sample efficiencies first decrease, and then increase again with n. We do not have an explanation for this, but the same behavior is found for the untrimmed PCA.

8.2 Robustness at finite samples

In this subsection we generate samples containing outliers in order to study the robustness of the estimators at finite samples. Trimmed PCA is compared with 5 other approaches: (i) the ROPCA method of Hubert et al. (2005) (ii) the eigenvectors of the Minimum Covariance Determinant estimator (iii) the Projection Pursuit (PP) approach of Li and Chen (1985) (iv) the eigenvectors of the Sign Covariance Matrix (v) the eigenvectors of the sample covariance matrix. We use the implemented of the `rrcov` R-package, see Todorov and Filzmoser (2009). Similar simulation studies were carried out in Maronna (2005) and Engelen et al. (2005), among others.

We generate $M = 1000$ samples of size n, where $n - \lfloor n\epsilon \rfloor$ of the data are generated by the model distribution $N(0, \Sigma)$, with Σ as in the previous subsection. The $\lfloor n\epsilon \rfloor$ outliers follow a $N(10\mathbf{1}_p, 10\Sigma')$, where Σ' equals Σ with reversed diagonal elements, and $\mathbf{1}_p$ a vector of length p only ones. The outliers are at a large distance from the true principal component space, and also far away from the main data cloud. Hence they are bad leverage points. We performed similar experiments for good leverage points and vertical outliers, yielding comparable relative performance of the different methods. The percentage of outliers varies from 5% to 20%. For the trimmed PCA, we selected $\alpha = 0.25$ yielding a good compromise between robustness and efficiency. As performance criterion we take the expected squared distance between an observation from the model and the estimated subspace. We compute it as

$$D^2 = \text{Trace}\left(\Sigma \sum_{j=d+1}^{p} \hat{v}_j \hat{v}_j^t\right)$$

The lowed D^2, the better. Figure 8.3 presents the D^2, averaged over the M simulation runs, for the representative case $n = 50$, $p = 5$, $d = 3$, and design (a).

If no outliers are present, so $\epsilon = 0$, then the sample covariance matrix gives

For reasons of comparability between methods, we let the estimated subspace pass the true center of the distribution.

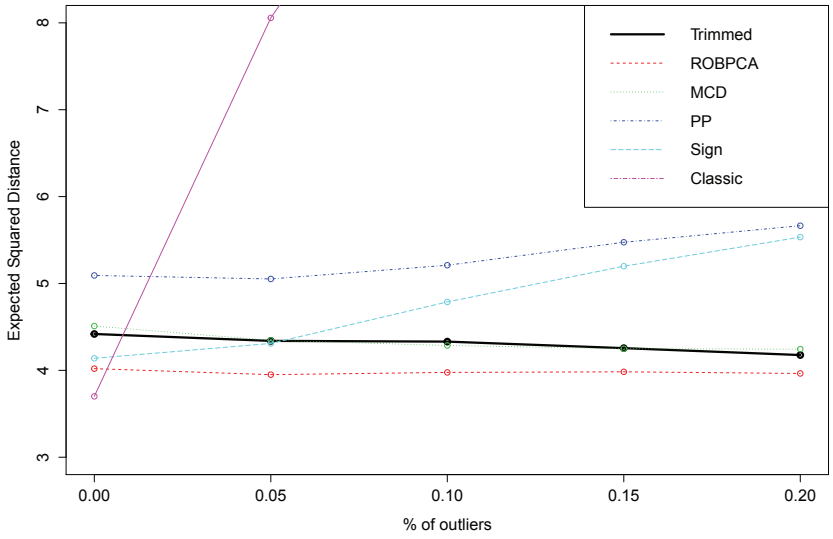

Figure 8.3: Simulated value of D^2 as a function of the percentage of outliers for 6 different estimators, for design (a) with $n = 50$, $p = 5$, and $d = 3$.

the best results, but its performance deteriorates quickly. The robust estimators are much more stable under contamination; the PP and the Sign covariance matrix start to perform worse in presence of outliers, but they do not explode. The Trimmed PCA, the MCD and ROBPCA yield the best results, where the D^2 does not increases further when outliers are added (the reason for this is that the more outliers there are, the less good observations are trimmed away). The ROBPCA method gives very good results, in line with previous simulation studies. ROBPCA is documented to work very well in practice, but no theoretical results are available for this approach. The MCD and the trimmed PCA method perform similar in this experiment, and are not too far from the ROBPCA. It is not surprising that MCD and trimmed PCA give similar results, since both yield eigenvectors from sample covariances matrices computed from trimmed samples. But trimmed PCA is the more natural approach in this setting, and it can also be computed for $n < p$ or when a majority of the data is lying exactly on a subspace.

In the supplementary file, we consider the worst case behavior of the esti-

mator over a larger range of outlier positions. We find that (i) the performance of the robust estimators is deteriorating if ϵ is getting larger (ii) intermediate outliers may be more dangerous than extreme outliers.

9. Data Example

In this section we illustrate the method using the Breast cancer data set, described in Chin et al. (2006) and available in the R-package `PMA`. We take the $p = 20$ comparative genomic hybridization (CGH) variables with largest standard deviation, measured for $n = 89$ patients for the first chromosome. Aim is to visualize the patients in a plane, and therefore we look for the optimal subspace of dimension $d = 2$. Outliers are to be expected in such datasets, and we take $\alpha = 0.25$ trimming level. In Figure 9.4 we plot the data projected on the trimmed principal subspace, together with a 95% tolerance ellipse. The tolerance ellipse uses the first $d = 2$ estimated eigenvalues. We add a plot of the squared distances of each observation to the α trimmed principal component subspace. We compare the outcomes of the trimmed case ($\alpha = 0.25$, top figure) and the non trimmed case ($\alpha = 0$, bottom figure). The different robust PCA methods give comparable results on this example.

We see from Figure 9.4 that the non trimmed approach gives a more spherical tolerance ellipsoid, and only one observation is detected as outlying in the subspace. The trimmed approach finds a subspace fitting well the large majority of the data; some observations have an unusual large distance (see top right panel) and may be atypical. The horizontal dashed line, that can be used as an heuristic device to diagnose observations with an unusual high distance, corresponds to the 95% critical value of a chi-squared distribution with the degrees of freedom estimated by the trimmed variation around the optimal subspace.

10. 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

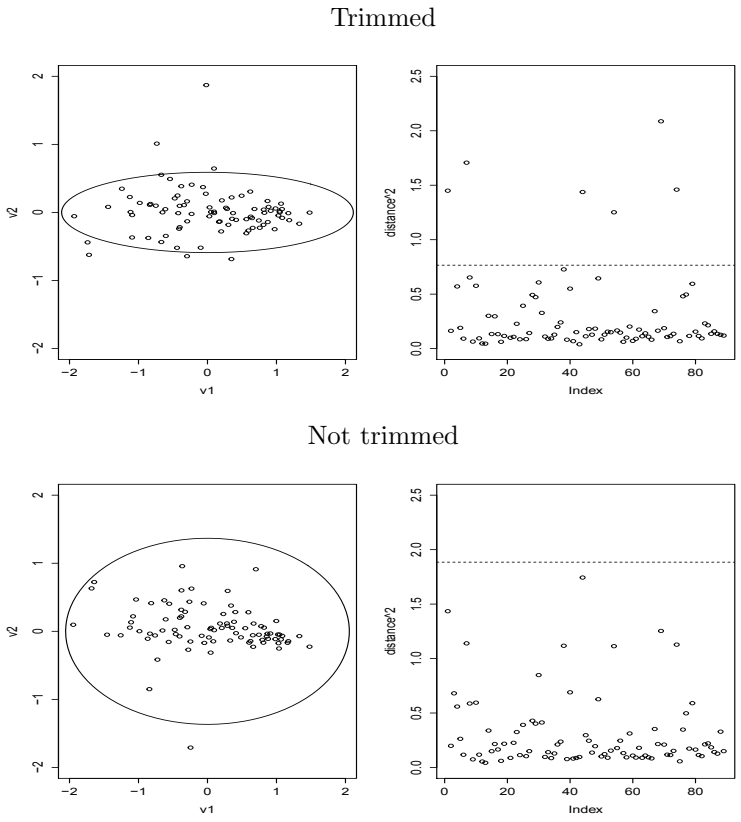

Figure 9.4: Projection on the α-trimmed optimal subspace (left) and squared distances (right) for 89 patients and $p = 20$. The top plot is for $\alpha = 0.25$, the bottom plot for $\alpha = 0$.

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, 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 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 prove, under mild conditions, consistency

of the sample trimmed PC space towards the population counterpart. The robustness of the method is studied by showing qualitative robustness, computing the breakdown point, and deriving the influence functions, which turn out to be bounded for bad leverage points. 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. A plot of the values of the trimmed variation as a function of d can be used to select the dimension of the subspace. If such a plot indicates that not much further reduction in trimmed variation can be gained by increasing d to $d+1$, the corresponding dimension can be selected.

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 et al. (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. As 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. Applications of the trimming approach in the multiple population case are in robust linear clustering (García-Escudero et al., 2009) and robust cluster analysis (García-Escudero et al., 2008).

Acknowledgment

We would first to thank the Editor, associate Editor, and referees for their comments and suggestions that helped improving the paper. The research of L.A. García-Escudero and A. Gordaliza was partially supported by the Spanish Ministerio de Ciencia e Innovación, grant MTM2014-56235-C2-1-P, and by Consejería de Educación y Cultura de la Junta de Castilla y León, grant VA212U13.

References

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

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

Chin, K., DeVries, S., Fridlyand, J., Spellman, P., Roydasgupta, R., Kuo, W., Lapuk, A., Neve, R., Qian, Z., Ryder, T., et al. (2006). "Genomic and transcriptional aberrations linked to breast cancer pathophysiologies", *Cancer Cell*, 10, 529-541.

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

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

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

Croux, C., 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.

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

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

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.

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

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.

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.

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

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

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.

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.

Engelen, S., Hubert, M. and Vanden Branden, K. (2005) "A comparison of three procedures for robust PCA in high dimensions", *Austrian Journal of Statistics,* 34, 117-126.

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

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.

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.

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

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

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

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.

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

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.

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

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

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.

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

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

Todorov, V., and Filzmoser, P. (2009) "An object-oriented framework for robust multivariate analysis", *Journal of Statistical Software,* 32, 1-47..

Tyler, D. (2005) "Breakdown and Groups-Discussion", *Annals of Statistics,* 33, 1009-1015.

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

KU Leuven, Naamsestraat 68, B3000 Leuven, Belgium.

E-mail: (christophe.croux@econ.kuleuven.ac.be)

IMUVA y Departamento de Estadística e Investigación Operativa, E.I.I., Universidad de Valladolid, Paseo del Cauce, 59, 47011 Valladolid, Spain.

E-mail: (lagarcia@eio.uva.es)

IMUVA y Departamento de Estadística e Investigación Operativa, E.I.I., Universidad de Valladolid, Paseo del Cauce, 59, 47011 Valladolid, Spain.

E-mail: (alfonsog@eio.uva.es)

Haute École de la Province de Liège (HEPL), Categorie technique, Quai Gloesener, 6, B4000 Liège, Belgium.

E-mail: (christel.ruwet@hepl.be)

Departamento de Estadística Investigación Operativa, E.T.S.I.A., Universidad de Valladolid, Avda. Madrid, 57, 34004 Palencia, Spain.

E-mail: (rsmartin@eio.uva.es)