Statistica Sinica Preprint No: SS-2015-0185R3	
Title	Robust Principal Component Analysis Based On Trimming Around Affine Subspaces
Manuscript ID	SS-2015-0185R3
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202015.0185
Complete List of Authors	Christophe Croux Luis Angel García-Escudero Alfonso Gordaliza Christel Ruwet and Roberto San Martín
Corresponding Author	Alfonso Gordaliza
E-mail	alfonsog@eio.uva.es

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0185

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 is given and existence of solutions to both the sample and population problems are proven. 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.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0185

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 familiar with this method due to its intuitive geometrical appealing and its implementation in most of statistical packages. One of the main drawbacks of PCA is the lack of robustness against the presence of outlying observations in the data set. There are examples in the literature showing that a single outlier, strategically placed, is enough to make classical PCA unreliable.

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 projection pursuit, where one looks for the direction maximizing a robust measure of scale of the data projected

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0185

on it (Li and Chen (1985); Croux and Ruiz-Gazen (2005)). A hybrid approach was taken by Hubert et al. (2005). Robust procedures have also been developed for kernel PCA (see, e.g., Debruyne and Verdonck (2010) and references therein) and in the learning machine literature (see, e.g., Xu, Caramanis and Sanghavi (2012) and references therein).

In this paper we aim at retrieving the lower-dimensional affine subspace best fitting the large majority of the data. More precisely, we look for the "best" subset of size $n - \lfloor n\alpha \rfloor$, with $0 \leq \alpha < 1$, 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. Thus, 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* that contains the observations surviving the trimming process.

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

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0185

To overcome the implicit hypothesis of symmetry and to extend the idea of trimming to such other frameworks 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 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. Gordaliza (1991) stated a functional or population version of some related trimming procedures in the multivariate setting and coined the term "impartial trimming" to mean that it is the data set itself which tells us the best way of trimming a fixed proportion α of the data.

The problem at (1.1) was also considered in Maronna (2005), who proposed a fast approximative algorithm to compute its solution. His paper mainly discussed computational aspects, while we present a theoretical study of the trimmed principal component subspace, including the 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

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0185

some preliminary results simplifying the problem and throwing light on the way 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 there is a conclusion section. Proofs are deferred to a supplementary file.

2. Notation and preliminary results

Here, 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.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0185

As introduced in Gordaliza (1991) and used in Cuesta-Albertos et al. (1997), trimming functions 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,

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

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 our trimming problem can be posed by using trimming functions instead of trimming subsets. 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) is denoted $V_{d,\alpha} \equiv V_{d,\alpha}(P_X) \equiv V_{d,\alpha}(X)$.

We need some technical results to simplify the problem at (2.1).

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

Given $h \in \mathcal{A}_d(\mathbb{R}^p)$ and $r \geq 0$, we take a strip around h and with radius r as $S(h,r) := \{x \in \mathbb{R}^p : \|x - \mathrm{Pr}_h(x)\| < r\}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0185

Lemma 2 *For any $h \in \mathcal{A}_d(\mathbb{R}^p)$ and $0 \leq \beta < 1$, let $r_\beta(h) = \inf\{r \geq 0 : P_X\big(S(h,r)\big) \leq 1-\beta \leq P_X\big(\overline{S}(h,r)\big)\}$ and $\mathcal{T}_{h,\beta} = \{\tau \in \mathcal{T}_\beta : I_{S(h,r_\beta(h))} \leq \tau \leq I_{\overline{S}(h,r_\beta(h))}, P_X\text{-a.e.}\}$,. 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 $\tau' \in \mathcal{T}_{h,\beta}$.*

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 it the *$\beta$-trimmed variation* of X around h,and $\tau_{h,\beta}$, 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 *In the notation of 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* $P_X(S(h, r_\alpha(h))) = 0$.

Thus, 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.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0185

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

Thus, one can to simplify the original double minimization problem (2.1) to the single search of the optimal affine subspace h. Once it 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 solving (2.1), is called a *d-dimensional* α*-trimmed principal component subspace of* X, the trimmed principal component subspace for short.

The results cover both the population and the sample problem. Thus, if we have a sample $\{X_i\}_{i=1}^n$ of size n from the probability distribution P_X, the associated empirical measure is

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

for ω in the sample space Ω. Given the outcome of a sample $X_1(\omega) = x_1, ..., X_n(\omega) = x_n$, the problem stated in (1.1) is equivalent to the problem (2.1) when taking P_n^ω instead of P_X.

3. Existence

Here we state the existence of solutions of (2.1), without moment conditions on the underlying distribution. This is important in terms of robustness, because outliers are often associated with the presence of heavy tails

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0185

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 $\tau_n = \tau_{h_n,\alpha}$, the radius $r_n = r_\alpha(h_n)$, and $S_n = S(h_n, r_n)$. We parameterize h_n through the distance to the origin, $d_n = \inf_{x \in h_n} \|x\|$, and the choice of d unitary vectors spanning the affine subspace.

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.

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.*

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0185

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 P_X^τ be the probability distribution with, 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 did 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

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 con-

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0185

verging unitary spanning vectors; the sequences of sample optimal radii and sample trimmed variations are then 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, and $V_n := V_{d,\alpha}(X_n), n = 0, 1, 2, ...,$ denotes the trimmed variation of X_n.

The proof main result on the consistency is similar to that used in Cuesta-Albertos et al. (1997) to establish consistency for trimmed k-means. Difficulties arise since the trimming functions have discontinuities on the boundaries of the corresponding strips, so the continuity of the probability distribution of the limit random vector is imposed.

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.

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 $n = 1, 2, \ldots$ and assume*

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

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0185

(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* $n \to \infty$*.*

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 $\{Y_n\}_n$ we get the following.

Corollary 3 *Theorem* 2 *holds if we replace condition* (a) *by* (a′) $X_n \to X_0$ *in distribution.*

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 (1.1) is equivalent to (2.1) taking P_n^ω instead of P_X. 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* $\omega \in \Omega$*. Assume* 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*

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0185

trimmed principal components of $\{P_n^\omega\}_n$*, then* $h_n^\omega \to h_0$*, P-a.s. and* $V_{d,\alpha}(P_n^\omega) \to V_{d,\alpha}(X)$*, P-a.s.*

The consistency result requires the uniqueness of the d-dimensional trimmed principal component subspace, which does not hold in general.

5. Uniqueness and Fisher consistency for elliptical distributions

Here, we consider elliptically contoured distributions. A $\mathbb{R}^p$-valued r.v. X has 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 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*; it is proportional to the covariance matrix if the distribution has a second moment. The ordered eigenvalues of Σ are denoted by $\lambda_1 \geq ... \geq \lambda_p > 0$ and the associated eigenvectors are $v_1, ..., v_p$, respectively. To have uniqueness we need $\lambda_d > \lambda_{d+1}$, where d is the dimension of the affine subspace we are looking for.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0185

Theorem 4 *Let X be a random vector with the elliptically symmetric density at* (5.1)*, and unimodal. If $\lambda_1 \geq ... \geq \lambda_p > 0$ are the eigenvalues of Σ with $\lambda_d > \lambda_{d+1}$, then*

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

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

The proof needs a multivariate probability inequality in Davies (1987); it is given in the supplementary file. If second moments exist, Σ is proportional to the covariance matrix and, therefore, the principal axes corresponding to the trimmed principal components are the same as those obtained by using the standard PCA.

From now on, we omit the reference to the random vector X in the notation P_X, writing P. For a given distribution P with density as in (5.1), denote by $S(P)$ the optimal strip associated with the trimmed principal component subspace. This strip is centered at μ and has the first d

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0185

eigenvectors of Σ as spanning vectors. We define the functional

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

and 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. Thus we can restrict our attention to elliptical distributions centered at the origin and with diagonal scatter matrix. In this case, $m(P) = 0$ and $C(P)$ is diagonal.

Theorem 5 *Let P be with density as in (5.1). Given finite second order moments, 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 the first d eigenvalues and eigenvectors of the covariance matrix of P. At the multivariate normal distribution, $c = 1$.*

In the sequel, the functional C is multiplied by this consistency factor c.

6. Robustness

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0185

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.*

6.2. Influence function

To investigate the infinitesimal 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 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.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0185

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 $C(P)$.

Theorem 7 *For an elliptical distribution function P with density at* (5.1)*, $\mu = 0$, and $\Sigma = diag(\lambda_1, \ldots, \lambda_p)$, 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).

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 outliers in the direction of the first d eigenvectors and still belonging to $S(P)$ can 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

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0185

is similar to that of the classical estimator for contaminations close to the subspace spanned by the first d eigenvectors.

One can now obtain the influence functions for eigenvectors and eigenvalues of C. 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} values 1 when $j = i$, 0 otherwise, 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 $1 \leq i \leq p$, 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

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

with v_j the jth eigenvector of Σ.

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 $\Sigma = \text{diag}(2, 1)$. We take $d = 1$. Only the non-zero component of the influence function of the eigenvector is represented.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0185

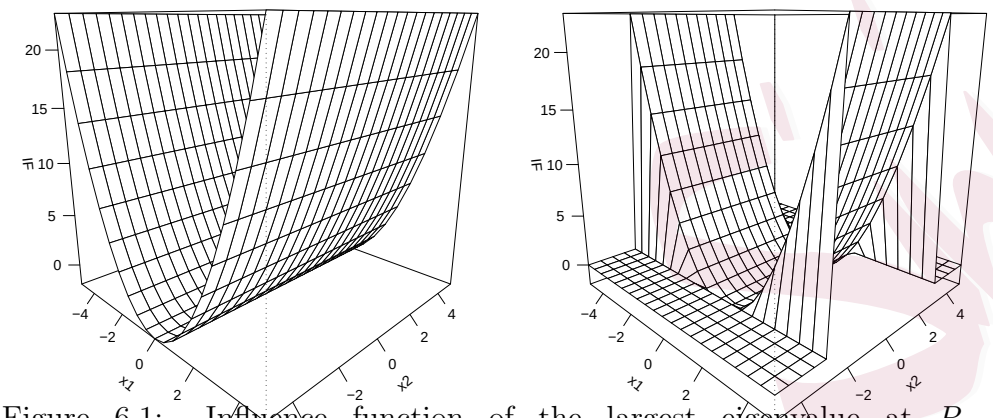

Figure 6.1: Influence function of the largest eigenvalue at $P = N(0, \mathrm{diag}(2, 1))$ when $\alpha = 0$ (left panel) and $\alpha = 0.01$ (right panel).

Inside the strip $S(P) = \{x_2 | x_2^2 \leq r^2(P)\}$, the influence function for the untrimmed and the trimmed influence functions have a similar behavior; outside the optimal strip the influence of the "trimmed" eigenvalue is zero, and is bounded for the "trimmed" eigenvectors. For the classical eigenvectors and eigenvalues, the influence functions is unbounded, 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

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0185

trimming remains limited, as is further explored in Section 7.

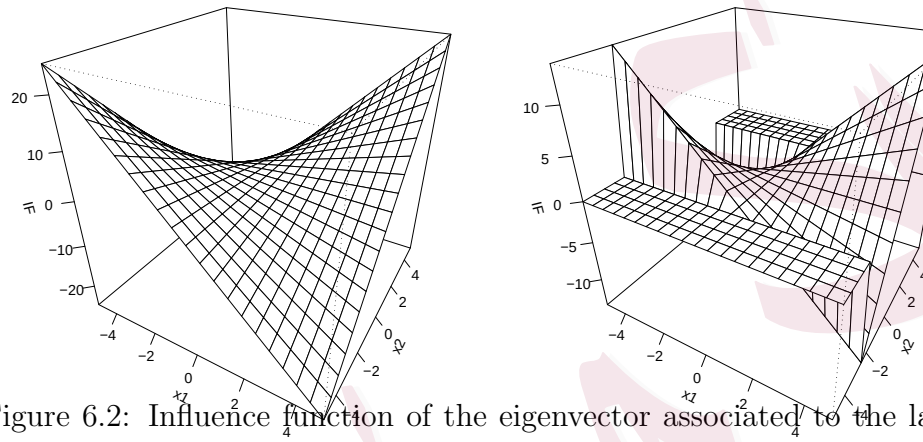

Figure 6.2: Influence function of the eigenvector associated to the largest eigenvalue at $P = N(0, \mathrm{diag}(2, 1))$ when $\alpha = 0$ (left panel) and $\alpha = 0.01$ (right panel).

6.3. Breakdown Point

The breakdown point 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 $\varepsilon_n^*(T, \mathcal{X}) = \min\left\{k/n;\ \sup_{\mathcal{X}'} \|T(\mathcal{X}) - T(\mathcal{X}')\| =\right.$

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0185

$\infty\}$, 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 the origin is $D(\mathcal{X}) := \inf_{x \in h_{\mathcal{X}}} \|x\|$, and we say that the procedure breaks down when $D(\mathcal{X}')$ can be made arbitrarily large.

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 $\varepsilon_n^*(T, \mathcal{X}) = (d+1)/n \to 0$ as $n \to \infty$, showing the lack of robustness of the classical estimator.

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 while focusing on the breakdown of the "trimmed scale" target function, (1.1), in terms of preventing it to

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0185

become 0 or ∞ ("implosion" and "explosion"). We consider a different situation where the whole PCA subspace may be unbounded by taking an arbitrarily large "distance to the origin".

The breakdown point of the "distance to the origin" has its limitation. It considers breakdown due to shifts, but says nothing about the orientation of the eigenvectors. We refer to Tyler (2005) for further discussion of the definition of breakdown point for eigenvectors.

7. Asymptotic variances

7.1. Asymptotic variances in the elliptical case

For an elliptical contoured distribution with $\mu = 0$ and $\Sigma = diag(\lambda_1, \ldots, \lambda_p)$, (6.1) and (6.2) allows us to obtain expressions for the asymptotic variances for the associated eigenvalues and eigenvectors estimators as

$$\begin{aligned} \mathrm{ASV}(\Lambda_i, P) &= \frac{c^2}{(1-\alpha)^2} \int_{S(P)} x_i^4 |\Sigma|^{-\frac{1}{2}} h(x'\Sigma^{-1}x)dx - \Lambda_i(P)^2 \\ &\quad + \frac{\alpha}{1-\alpha}\left(\frac{cA_{ii}}{G}\right)^2 + 2\Lambda_i(P)\frac{cA_{ii}}{G}\left(\frac{-\alpha}{1-\alpha}\right), \\ \mathrm{ASV}(V_i, P) &= \sum_{j\neq i} \frac{\lambda_i^2\lambda_j^2}{(\lambda_i-\lambda_j)^2} \frac{\int_{S(P)} x_i^2 x_j^2 dP(x)}{4H_{ij}^2} v_j v_j', \end{aligned} \tag{7.1}$$

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

7.2. Asymptotic relative efficiencies in the gaussian case

Using the preceding results, one can obtain information on the efficiency of the estimators of the eigenvectors and eigenvalues of C computed after

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0185

trimming. We restrict our attention here to gaussian distributions and we only consider the first d eigenvalues and eigenvectors.

From Section 5, the consistency factor c is 1 for the d first eigenvalues, and $\Lambda_i(P) = \lambda_i$. Thus the asymptotic variances of the eigenvalues with $1 \leq i \leq d$ are

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

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. As the ML estimator is the untrimmed PCA, its asymptotic variances are given by the above expressions for $\alpha = 0$. So it follows from (7.2) that, for $1 \leq i \leq d$,

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

and the efficiency is just the complementary of the trimming proportion. For instance, a trimming level of 10% yields a 90% efficiency for the eigen-

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0185

value 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), $c_j = 1$ 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

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

Globally, the efficiency is again determined by the trimming proportion, but 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 $(1-\alpha)/c$; (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) everything

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0185

else being fixed, if 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

Simulation experiments consisted 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 $\Sigma = \text{diag}(\lambda_1, \ldots, \lambda_p)$. Two sets of diagonal elements were considered, similar as in Maronna (2005): (a) a smooth decrease of the eigenvalues, $\lambda_j = 2^{p-j}$ for $1 \leq j \leq p$; (b) an abrupt decrease of the eigenvalues after λ_d, $\lambda_j = 20(1 + 0.5(d - j + 1))$ 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. Here, the approximative algorithm of Maronna (2005) was used. This algorithm follows the rationale behind the fast-MCD algorithm in Rousseeuw and van Driessen (1999) for computing the Minimum Covari-

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0185

ance Determinant (MCD) estimator, combining random starts and so-called "concentration" steps. We recommend taking the number of initial random starts equal to 500, and the number of concentration steps equal to 10.

8.1 Finite-sample efficiencies

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 was first applied and then the classical definition of MSE was 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 were computed as

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

These finite sample efficiencies are reported in Table 8.1. Since the efficiencies for the different eigenvalues of a particular setting are quite similar,

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0185

their average value is reported. In this table, the asymptotic relative efficiencies derived in the previous section appear in the rows referred as "$n = \infty$".

With smoothly decreasing eigenvalues, we can see from Table 8.1, $p = 5, d = 3$, that 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 slower than for design (a). 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 the supplementary file). For $p = 8, d = 7$ the finite sample efficiencies first decrease, and then increase again with n. We do not have an explanation for this, but the same behavior

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0185

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

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0185

is found for the untrimmed PCA.

8.2 Robustness at finite samples

We generated samples containing outliers in order to study the robustness of the estimators at finite samples. Trimmed PCA is compared with five 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 used 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 generated $M = 1000$ samples of size n, where $n - \lfloor n\epsilon \rfloor$ of the data were generated by the model distribution $N(0, \Sigma)$, with Σ as in the previous subsection. The $\lfloor n\epsilon \rfloor$ outliers followed a $N(10\mathbf{1}_p, 10\Sigma')$, where Σ' is Σ with reversed diagonal elements, and $\mathbf{1}_p$ a vector of ones of length p. 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. For reasons of comparability between methods, we let the estimated subspace

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0185

pass the true center of the distribution. The percentage of outliers varied 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 took the expected squared distance between an observation from the model and the estimated subspace. We computed 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, this for the representative case $n = 50$, $p = 5$, $d = 3$, and design (a).

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

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0185

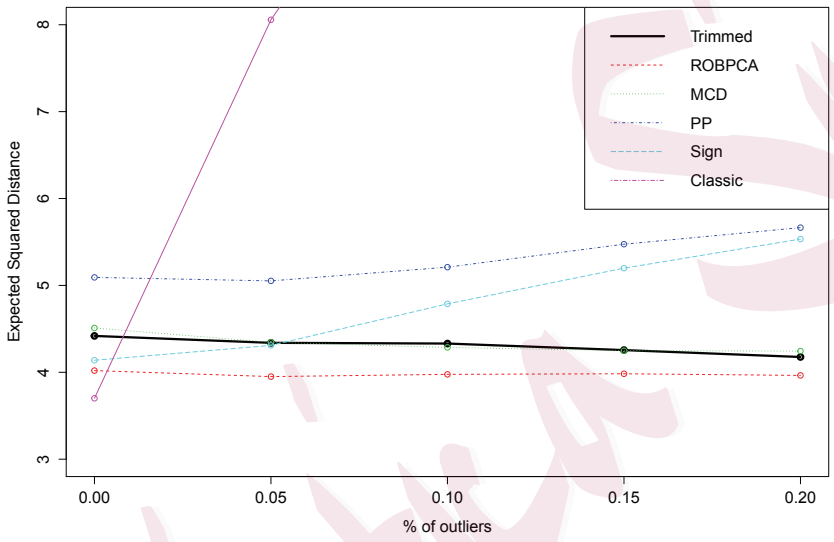

Figure 8.3: Simulated value of D^2 as a function of the percentage of outliers for 6 different estimators, for design (a) with $n = 50$, $p = 5$, and $d = 3$.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0185

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 estimator over a larger range of outlier positions. We find that the performance of the robust estimators is deteriorating if ϵ is getting larger, and 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 took the $p = 20$ comparative genomic hybridization (CGH) variables with largest standard deviation, measured for $n = 89$ patients for the first chromosome. The aim was to visualize the patients in a plane, and therefore we looked for the optimal subspace of dimension $d = 2$. Outliers were to be expected in such datasets, and we took the $\alpha = 0.25$ 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

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0185

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 that fits well to 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

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 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

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0185

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 found good performance of the method. He also applied it on several 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 ro-

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0185

bustification arises from the trimming, less by the different choices of the penalty function Φ. We could also adopt a "min-max" or L_∞ approach and search for the narrowest strip including a $1-\alpha$ proportion of the data points. Rousseeuw's LMS regression estimator 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)).

Supplementary Materials

Supplementary Materials Sections S1-S4 provide the proofs of Lemmas 1, 2, 3, and 4, respectively. The proofs of Theorem 1 and 2 are given in Sections S5 and S6. The proofs of Theorems 4, 5, 6, 7 and 8 are provided in Sections S7-S11. Section S12 provides details on obtaining asymptotic variances in the elliptical case, and Section S13 on obtaining the asymptotic relative efficiencies in the gaussian one. Section S14 presents additional simulation results.

Acknowledgment

We thank the Editor, an associate editor, and referees for their comments and suggestions that helped to improve the paper. C. Croux was supported by the Research Fund K.U. Leuven and the "Fonds voor Weten-

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0185

schappelijk Onderzoek". The work of C. Ruwet was partially supported by the IAP Research Network P7/06 of the Belgian State. 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.). Wiley, New York.

Campbell, N.A. (1980). Robust procedures in multivariate analysis I: Robust covariance estimation. *J. R. Stat. Soc. C.* **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

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0185

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, 257-269.

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.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0185

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

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0185

A Festschrift for Erich L. Lehmann (P. J. Bickel, K. Doksum, and J. L. Hodges, Jr., eds.) 157184. 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.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0185

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.* Wiley, New York.

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 theory 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*, (W. Grossmann, G. Pflug, I. Vincze, and W. Wertz, eds.) 283-297. Reidel Publishing

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0185

Company, Dordrecht.

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. *J. Stat. Softw.* **32**, 1-47.

Tyler, D. (2005). Discussion of "Breakdown and groups" by P. L. Davies and U. Gather. *Ann. Statist.* **33**, 1009-1015.

Xu, H., Caramanis, C., and Sanghavi, S. (2012). Robust PCA via outlier pursuit. *IEEE Trans. Inform. Theory* **58**, 3047-3064.

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.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0185

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)

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0185

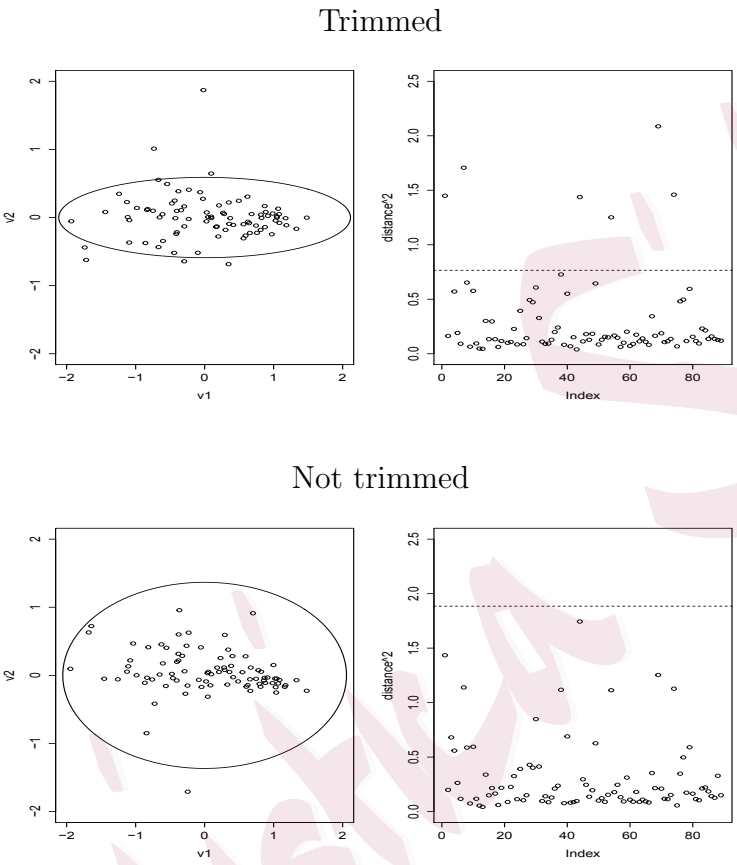

Figure 9.4: Projection on the α-trimmed optimal subspace (left) and squared distances (right) for 89 patients and $p = 20$. The top plot is for $\alpha = 0.25$, the bottom plot for $\alpha = 0$.