ROBUST PRINCIPAL COMPONENT ANALYSIS VIA PROJECTION PURSUIT

By

ZDENEK PATAK

B.Sc., The University of Brittish Columbia, 1987

A THESIS SUBMITTED IN PARTIAL FULLFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE

MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE STUDIES

(Department of Statistics)

We accept this thesis as conforming

to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

January 1990

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Statistics

The University of British Columbia
Vancouver, Canada

Date Sep. 28/1990

DE-6 (2/88)

Abstract

In principal component analysis (PCA), the principal components (PC) are linear combinations of the variables that minimize some objective function. In the classical setup the objective function is the variance of the PC's. The variance of the PC's can be easily upset by outlying observations; hence, Chen and Li (1985) proposed a robust alternative for the PC's obtained by replacing the variance with an M-estimate of scale. This approach cannot achieve a high breakdown point (BP) and efficiency at the same time. To obtain both high BP and efficiency, we propose to use MM- and τ-estimates in place of the M-estimate. Although outliers may cause bias in both the direction and the size of the PC's, Chen and Li looked at the scale bias only, whereas we consider both.

All proposed robust methods are based on the minimization of a non-convex objective function; hence, a good initial starting point is required. With this in mind, we propose an orthogonal version of the least median of squares (Rousseeuw and Leroy, 1987) and a new method that is orthogonal equivariant, robust and easy to compute. Extensive Monte Carlo study shows promising results for the proposed method. Orthogonal regression and detection of multivariate outliers are discussed as possible applications of PCA.

Contents

List of Tables

List of Figures

Acknowledgements

Dr. Ruben Zamar first introduced me to the subject of robustness. I thank him for his support, help and guidance during the writing of my thesis. I also like to thank Dr. Mohan Delampady for his helpful comments.

1 Introduction

This thesis discusses robust alternatives to principal component analysis (PCA) and orthogonal regression (OR). Classical methods and key robustness concepts are briefly discussed, existing robust procedures are described and new robust approaches are introduced. Several examples are included to illustrate the properties of the classical versus robust methods. Our goal is to extend the work of Chen and Li (1985) to MM- and τ-estimates of PC's and to discuss the breakdown properties not only of the scale of the PC's (Chen and Li only considered the properties of the scale of the PC's) but also of their direction.

The word *robust* is derived from the Latin word "robus" meaning strength. In different disciplines, it takes on different meanings. In statistics, robustness has usully been associated with methods and procedures that do not suffer greatly when a fraction of the data does not follow the model assumptions, i.e. outlying observations may be present. To expand on this, we say that estimator is robust whenever its value does not change appreciably after a number of aberrant observations has been introduced. This notion of robustness is very loose and can be made rigorous, yet it gives us a flavour of what is involved when we say that an estimator is robust.

Statistical inference is only in part based upon observations. Equally important are the explicit and implicit assumptions one makes about the underlying situation. In regression, whether classical or orthogonal, one generally assumes that observations are independent and errors are normally distributed with mean zero and some common variance. The presence of outliers in the data certainly violates this assumption and inference based on classical methods that are sensitive to even minor departures from these assumptions would be suspect. Robust alternatives have been proposed to deal with the inherent susceptibility of classical methods to 'disruptions'.

In 1964 Huber, in his milestone paper on robust location estimation, laid the foundation for modern robust statistical analysis. Since then robust methods have been developed for a variety of statistical procedures including orthogonal regression and the estimation of principal components, both of which constitute the focus of this paper.

Scatter matrices and their principal components are at the heart of multivariate data analysis. Unfortunately, the classical estimator - the sample covariance matrix and its eigenvalues and eigenvectors - are highly nonrobust. One outlying observation can distort or completely upset the classical estimator. An observation is considered an *outlier* if it does not follow the same model as the rest of the data. An observation is considered a *leverage* point if its relocation causes major changes in the parameters to be estimated.

A major problem with the detection of such cases in higher dimensional space is that an observation may not be extreme with respect to any of the original variables, but it can still be an outlier because it does not conform with the correlation structure of the remainder of the data. This type of outlier, called a structural outlier, will most likely distort the direction, an eigenvector, whereas a gross error outlier will most likely distort the scale, an eigenvalue. It may be possible to identify gross error outliers as they will stand out from the rest of the data but it is very difficult to detect structural outliers by looking solely at the original variables one at a time, or even two at a time.

Most classical diagnostic procedures can identify a single outlying or high leverage observation. However, they are rendered helpless in the presence of multiple contaminants especially when these outliers are grouped and have a high leverage. Often the physical process that generates these outliers causes them to cluster in a particular location. This phenomenon creates a 'masking effect', i.e., the removal of one observation from this cluster will not have a discernible effect on the estimated parameters. The

Table 1: Fire Claims in Belgium from 1976 to 1980.

Year	Number of Fires
x_i	y_i
76	16694
77	12271
78	12904
79	14036
80	13874

remaining contaminants will 'mask' any change that would be detected if the cluster contained only one outlying observation.

To deal with this problem, we employ robust procedures. They provide an objective analytic tool for identifying observations that do not conform with the structure of the rest of the data.

To illustrate the problems connected with the use of classical methods when model assumptions have been violated, we consider a simple data set (Rousseeuw, Leroy 1987) comprising the number of reported claims by Belgian fire-insurance companies in the five years from 1976 to 1980.

If one disregards the number of fire claims reported in 1976, it is clear that there is an annual upward trend. This is reflected in the estimates from the method of least median of absolute orthogonal deviations (ORLM), β_0 = -28872.7 and β_1 = 534.3. ORLM is one of the many robust alternatives to orthogonal regression which will be discussed later.

OR fits the data with a decreasing trend yielding β_0 = 244547.7 and β_1 = -2956.3. These results give one the false impression that the number of fire claims is going down

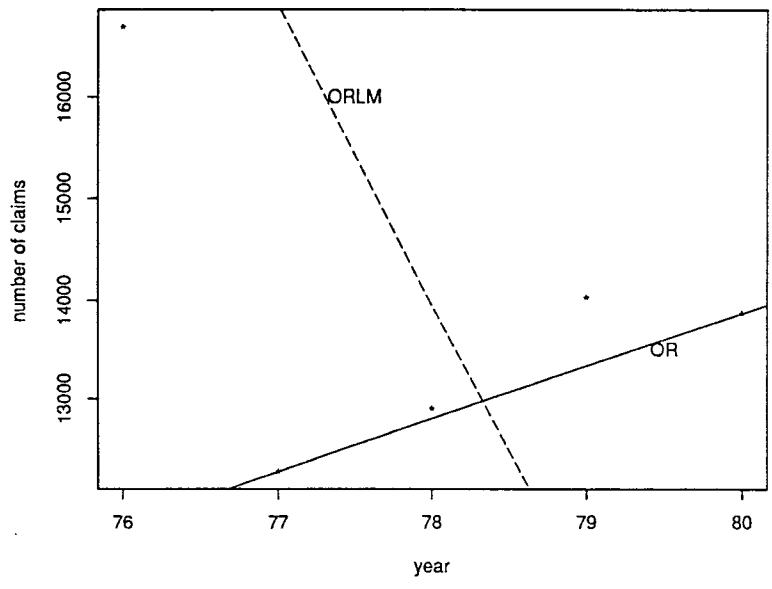

Figure 1: Plot of Fire Claims in Belgium from 1976 to 1980

while it is in fact going up. Although this example, with only one explanatory variable, is simple and the outlying observation can be easily identified because it is detached from the rest of the data, it illustrates how classical methods can lead to wrong conclusions in the presence of aberrant observations. Locating outliers in higher dimensions is usually more complicated and hence more reliable methods are needed.

The rest of the thesis is organized as follows. In section 2 we discuss some basic robustness concepts that are used to classify the performance of estimators. In section 3 we describe some robust estimates in the simple case of location and scale estimation. We focus our attention only on those estimates that are referred to in later sections. In section 4 we discuss classical PCA and the robust alternatives. We mainly focus on extending the work of Chen and Li (1985) beyond S-estimates of the scale of the PC's. We introduce MM- and τ-estimates of the direction and scale of the PC's. In section 5 we compute the lower bound for the maximum asymprotic direction bias and show that the BP of the scale is not inherited by the direction and that the BP of the direction

depends on the ratio of the adjacent eigenvalues. In section 6 we propose a robust estimate of the direction and the scale of the PC's and discuss its properties. We show that the proposed estimate is orthogonal equivariant, Fisher consistent and robust. In the last two sections we include Monte Carlo results to show the bias characterictics of different estimates and mention possible applications of robust PCA.

2 Basic Robustness Concepts

Here, we introduce several concepts and definitions aimed at assessing the performance of estimators. Some stem from robustness while others are universally applicable.

2.1 Influence Function

The influence function (IF) measures the sensivity of an estimator to infinitesimal perturbations. The IF of an estimator T at a point x and a distribution F is given by

$$IF(x;T,F) = \lim_{\epsilon \to 0} \frac{T((1-\epsilon)F + \epsilon\delta_x) - T(F)}{\epsilon} \tag{1}$$

for those points x of the sample space where the limit exists. (Here, δ_x is the point-mass distribution at x.). We define the *gross error sensitivity* (GES) as

$$GES = \sup_x \|IF(T,x)\| \ . \tag{2}$$

Observe that for ϵ near zero,

$$\|T((1-\epsilon)F + \epsilon\delta_x) - T(F)\| \approx \epsilon \|IF(x;T,F)\| \ .$$

This implies that

$$\sup_x \|T((1-\epsilon)F + \epsilon\delta_x) - T(F)\| \approx \epsilon \ GES \ ,$$

where $\sup_x \|T((1-\epsilon)F + \epsilon\delta_x) - T(F)\|$ is the maximum bias induced in an estimator by a fraction ϵ of contamination. The restriction to point mass contamination implies no loss of generality (see Martin, Yohai and Zamar, 1990).

2.2 Maximum Bias Curve and Breakdown Point

An important notion of robustness of an estimator is the *breakdown point* (BP). It measures the extent to which an estimator is able to cope with contamination. It is

often helpful in understanding the robustness properties of the estimator and can also be used to classify its performance. There are several definitions of the breakdown point of an estimator. For simplicity, only the finite sample version due to Donoho and Huber (1983) will be introduced here. To define the breakdown point, let us suppose we have a data set

$$X = \{\mathbf{x}_1, ..., \mathbf{x}_n\} = \{(x_{11}, x_{12}, ..., x_{1p}), ..., (x_{n1}, x_{n2}, ..., x_{np})\} .$$

T(X) is the value of an estimator at the sample X. Consider all corrupted samples X' obtained by replacing any fraction $\epsilon \in (0,1)$ of the original data points by arbitrary values. $T(X')$ is the value of an estimator at the contaminated sample X'. First, we define the *maximum bias* as

$$B(\epsilon; T, X) = \sup \|T(X') - T(X)\| \quad , \tag{3}$$

where the supremum is taken over all ϵ-contaminated samples. Plotting the function B versus the fraction of contamination ϵ produces the *maximum bias curve* which is a carrier of both the local and global robustness properties of the estimate. The breakdown point is the value of ϵ where the asymptote to the maximum bias curve crosses the x-axis. It is defined as

$$\epsilon^*(T, X) = \inf\{\epsilon : B(\epsilon; T, X) = \infty\} , \tag{4}$$

i.e., $\epsilon^*(T, X)$ is the smallest fraction of contamination that can cause $T(X')$ to take values arbitrarily far from T(X). Asymptotic counterparts of ϵ^* *and* B have been defined (Hampel, 1986, for example). Under certain regularity conditions, the GES is the value of the derivative of the maximum asymptotic bias curve at zero, that is,

$$GES = B'(0)$$

and therefore can be used to give a linear approximation for $B(\epsilon)$ for ϵ near zero.

2.3 Equivariance

Equivariance is a concept that reaches beyond robustness as it pertains to a property that in one form or another is desired of all estimators. We shall distinguish between four types of equivariance: location, scale, orthogonal and affine. In the context of orthogonal regression and principal component analysis, the first three are a natural requirement for any estimator.

2.3.1 Definitions

Suppose we have a collection of vectors $\mathbf{x}_1, ..., \mathbf{x}_n$ in R^p. An estimator T is said to be

1. *location equivariant* if $T(\mathbf{x}_1 + \mathbf{v}, ..., \mathbf{x}_n + \mathbf{v}) = T(\mathbf{x}_1, ..., \mathbf{x}_n) + \mathbf{v}$, where $\mathbf{v}$ is any vector in R^p.

2. *scale equivariant* if $T(c\mathbf{x}_1, ..., c\mathbf{x}_n) = |c|T(\mathbf{x}_1, ..., \mathbf{x}_n)$, where $c \in R$.

3. *affine equivariant* if $T(\mathbf{A}\mathbf{x}_1 + \mathbf{v}, ..., \mathbf{A}\mathbf{x}_n + \mathbf{v}) = \mathbf{A}T(\mathbf{x}_1, ..., \mathbf{x}_n) + \mathbf{v}$, where $\mathbf{A}$ is any nonsingular matrix and $\mathbf{v}$ is any vector in R^p.

4. *orthogonal equivariant* if $T(\mathbf{\Gamma}\mathbf{x}_1 + \mathbf{v}, ..., \mathbf{\Gamma}\mathbf{x}_n + \mathbf{v}) = \mathbf{\Gamma}T(\mathbf{x}_1, ..., \mathbf{x}_n) + \mathbf{v}$, where $\mathbf{\Gamma}$ is an orthogonal matrix and $\mathbf{v}$ is any vector in R^p.

For a scatter matrix $\mathbf{C}$, *affine equivariance* is defined as

$$\mathbf{C}(\mathbf{A}\mathbf{x}_1 + \mathbf{v}, ..., \mathbf{A}\mathbf{x}_n + \mathbf{v}) = \mathbf{A}\mathbf{C}(\mathbf{x}_1, ..., \mathbf{x}_n)\mathbf{A}^T \quad ,$$

where $\mathbf{A}$ and $\mathbf{v}$ are as above. This means that if a point cloud is rotated or rescaled, then any measurement of its orientation will rotate and any measurement of its size will scale correspondingly. Orthogonal equivariance of a scatter matrix is defined similarly with $\mathbf{A}$ replaced by $\mathbf{\Gamma}$.

3 Robust Estimators - An overview

There exists several different families of robust estimators. However here we will focus our attention only on those estimators that are referred to in later sections. They are the M-, S- and τ - estimators. Each is briefly described for the simple case of location and scale estimation.

3.1 M - estimates

Suppose we have a set $\{x_i\}_{i=1,..,n}$ of independent identically distributed observations from $F_{\theta,\sigma}$. The maximum likelihood estimators for location and scale at the Gaussian model are defined as the solutions to

$$\frac{1}{n}\sum_{i=1}^{n}\Psi\left(\frac{x_i-\theta}{\sigma}\right)=0$$

$$\frac{1}{n}\sum_{i=1}^{n}\chi\left(\frac{x_i-\theta}{\sigma}\right)=b\ ,$$

where $\Psi(x)=x$, $\chi(x)=x^2$ and $b=1$. Both Ψ and χ favour "large" observations. To reduce the influence of these observations on the estimated location parameter, we can use a function Ψ satisfying

C1. Ψ is odd, bounded, with at most a finite number of discontinuities.

Examples of such Ψ's are the Huber's function

$$\Psi_c^H(x) \;=\; \begin{cases} x & \text{if } |x|\le c \\ c\ sgn(x) & \text{if } |x|>c,\ \ c\in(0,\infty) \end{cases}$$

or the Tukey's biweight function

$$\Psi_c^T \;=\; \begin{cases} x\left(1-\left(\frac{x}{c}\right)^2\right)^2 & \text{for } |x|\le c \\ 0 & \text{for } |x|>c,\ \ c\in(0,\infty)\ . \end{cases}$$

To robustify the scale estimate, we often choose a χ that meets the following conditions:

C2. χ is symmetric, differentiable almost everywhere and $\chi(0) = 0$.

C3. χ is strictly increasing on [0,c) and constant on [c,∞) .

An example of a function that satisfies C2 and C3 is the Huber's χ-function defined as

$$\chi_c^H = \begin{cases} \frac{x^2}{c^2} & \text{for } |x| \leq c \\ 1 & \text{for } |x| > c, \ c \in (0, \infty) \end{cases} .$$

Using general Ψ and χ functions, we define the generalized maximum likelihood estimates (M-estimates) $\hat{t}$ and $\hat{s}$ of location and scale as the solutions to

$$\frac{1}{n}\sum_{i=1}^{n} \Psi\left(\frac{x_i - t}{s}\right) = 0 \tag{5}$$

$$\frac{1}{n}\sum_{i=1}^{n} \chi\left(\frac{x_i - t}{s}\right) = b \, , \tag{6}$$

where b is usually taken to be the $E\chi(Z)$ and Z is the standard normal random variable.

Huber (1964) defined M-estimates of location and described some of their asymptotic properties. These include $\sqrt{n}$ convergence rate to a normal distribution and a fairly high efficiency.

3.2 S - estimates

Let t be a tentative location of the center of a set of numbers, $\{x_i\}_{i=1,..,n}$. Consider the residuals, $r_i(t) = x_i - t$. The corresponding M-estimate of scale, $s(t)$, is implicitly defined by

$$\frac{1}{n}\sum_{i=1}^{n} \chi\left(\frac{r_i(t)}{s(t)}\right) = b \, , \tag{7}$$

where χ and b are as above. The S-estimate of location is then defined as

$$\hat{T} = argmin_t \, s(t) \, .$$

Notice that $s(\hat{T}) = \hat{s}$ is a robust estimate of scale. In fact, it can be shown that for χ satisfying C2 and C3, the BP of $\hat{s}$ is $\min\{\frac{b}{\chi(\infty)}, 1-\frac{b}{\chi(\infty)}\}$. (see Huber, 1981, for instance).

It is clear from their definition in (5), that M-estimates of location are sensitive to the choice of $\hat{s}$. To obtain good robustness properties for the M-estimate, we need to use a measure of dispersion of the residuals $r_i(t)$ that has the most bias resistance, i.e. a BP of 1/2. This criterion is met by $\hat{s}$. This approach produces a new type of estimator called the MM estimator. It combines the efficiency of M-estimates with the robustness of S-estimates.

3.3 τ - estimates

Introduced in 1988 by Yohai and Zamar, the τ-estimates combine efficiency with good breakdown properties. They are defined by

$$\hat{T} \;=\; argmin_t\, \tau^2(t)\ ,$$

where

$$\tau^2(t) \;=\; \frac{1}{n}\, s^2(t) \sum_{i=1}^{n} \chi_2\left(\frac{x_i - t}{s(t)}\right)\ ,$$

$s(t)$ is implicitly defined by

$$\frac{1}{n}\sum_{i=1}^{n} \chi_1\left(\frac{x_i - t}{s(t)}\right) = b$$

and χ_1 and χ_2 satisfy conditions C2 and C3. The corresponding "tuning constants" (explained below) for χ_1 and χ_2 are $c_1 = 1.548$ to yield a high BP for the scale $s(t)$ and $c_2 = 6.08$ to yield 95% efficiency for the location $\hat{t}$.

Suppose that $\left|\frac{x_i - t}{s(t)}\right|$ is small and χ_2 is quadratic near zero. Then

$$s^2(t)\chi_2\left(\frac{x_i - t}{s(t)}\right) \;=\; s^2(t)\left(\frac{x_i - t}{s(t)}\right)^2 \;=\; (x_i - t)^2\ .$$

Hence for non-contaminated samples, the τ estimators of location and scale reduce to the sample mean and variance respectively. This property gives the τ estimator its high

efficiency. If on the other hand, the absolute residual, $\left|\frac{x_i-t}{s(t)}\right|$, is large, i.e, greater than the tuning constant, c_2, its influence is diminished because

$$\chi_2\left(\frac{x_i-t}{s(t)}\right) \;=\; 1\ .$$

This property gives the estimator its high BP.

4 Principal Component Analysis

The objective of Principal Component Analysis (PCA) is to reduce the dimensionality of a data set containing a large number of correlated variables while retaining as much as possible of the variability present in the data. This is done by transforming to a new set of variables, the *principal components*, which are uncorrelated.

Suppose $\mathbf{x}$ is a vector of p random variables with mean μ and covariance Σ. Unless p is small or the covariance structure is very simple, not much insight can be obtained from looking at the p variances and $\frac{1}{2}p(p-1)$ covariances. An alternative approach is to look for a few "principal components" that retain most of the information contained in the variance-covariance structure.

The first step is to look for a linear combination of the components of $\mathbf{x}$, $\alpha^T\mathbf{x}$, that has maximum variance; i.e, $max_\alpha\ var(\alpha^T\mathbf{x}) = \alpha^T\Sigma\alpha$. It is clear that without a suitable constraint the maximum will not be achieved for finite α. The conventional constraint here is $\alpha^T\alpha = 1$. The problem then becomes

$$\text{maximize } \alpha^T\Sigma\alpha \quad \text{subject to } \quad \alpha^T\alpha = 1$$

To obtain a solution, we can use the method of Lagrange multipliers (see for example Jolliffe, 1986) and maximize

$$J_1(\alpha, \lambda) = \alpha^T\Sigma\alpha - \lambda(\alpha^T\alpha - 1) \ ,$$

where λ is a Lagrange multiplier. Differentiating $J_1(\alpha, \lambda)$ with respect to α_1 and setting the derivative to zero yields

$$\frac{\partial}{\partial\alpha} J_1(\alpha, \lambda) = \Sigma\alpha - \lambda\alpha = 0 \ .$$

By premultiplying both sides of the equation by α^T and using the constraint $\alpha^T\alpha = 1$, we get

$$\alpha^T\Sigma\alpha = \lambda \ .$$

Therefore, the solution to the constrained maximization problem is the eigenvector α_1 associated with the largest eigenvalue λ_1 of Σ. The linear function $\alpha_1^T \mathbf{x}$ is the first principal component.

Next, we look for a linear combination of the elements of $\mathbf{x}$, $\alpha^T \mathbf{x}$, that has maximum variance and satisfies the constraints (i) $\alpha^T \alpha = 1$ and (ii) $\alpha^T \alpha_1 = 0$. The solution can be obtained, again by using the method of Lagrange multipliers, by maximizing

$$J_2(\alpha, \lambda, \phi) = \alpha^T \Sigma \alpha - \lambda(\alpha^T \alpha - 1) - \phi \alpha^T \alpha_1 \ ,$$

where λ and ϕ are Lagrange multipliers. Differentiating $J_2(\alpha, \lambda, \phi)$ with respect to α and setting the derivative to zero yields

$$\frac{\partial}{\partial \alpha} J_2(\alpha, \lambda, \phi) = \alpha^T \Sigma \alpha - \lambda \alpha - \phi \alpha = 0 \ . \tag{8}$$

Premultiplying both sides of the equation by α_1^T will result in

$$\alpha_1^T \Sigma \alpha - \lambda \alpha_1^T \alpha - \phi \alpha_1^T \alpha_1 = 0 \ .$$

By noticing that (i) $\alpha_1^T \alpha_1 = 1$ and (ii) $\alpha_1^T \Sigma = \lambda \alpha_1^T$ we have $\phi = 0$. Substituting for ϕ in (8) and premultiplying by α results in

$$\alpha^T \Sigma \alpha = \lambda \ .$$

Therefore, the solution to the doubly constrained maximization problem is the eigenvector α_2 associated with the second largest eigenvalue λ_2 of Σ. The linear function $\alpha_2^T \mathbf{x}$ is the second principal component.

This process is repeated until all principal components are computed. It can be shown that $\alpha_1, \alpha_2, ..., \alpha_p$ are the eigenvectors of Σ corresponding to $\lambda_1, \lambda_2, ..., \lambda p$, respectively, where the λ_i's are in decreasing order. It can also be shown that

$$var(\alpha_i^T \mathbf{x}) \ = \ \lambda_i \ \text{ for } i = 1, 2, ..., p.$$

Note that in the classical PCA, proceeding from either minimizing or maximizing the variance of a linear combination of the elements of $\mathbf{x}$ yields the same principal components. As we will show later, this is not the case in the the robust setting.

4.1 Robust PCA

PCA is an important tool used in many fields where there is a need for reducing the dimensionality of a data set. It has been a popular technique with psychologists who routinely collect a multitude of information on their patients and then try to construct a few "indices" to explain their behavior. It is important that such indices should be as reliable as possible to prevent misdiagnoses; however, because the classical approach is unable to cope with aberrant observations, its reliability can be questioned.

As we have seen, principal components are obtained via successive constrained maximizations of the variance of a linear function of elements of $\mathbf{x}$. One extreme observation may inflate the variance enough to upset the order of the principal components. To illustrate the weaknesses of the classical PCA, we have generated a hundred point, two variable sample. The first variable y is distributed as N(0,1), the second variable x is 90% distributed as N(0,0.1) and 10% as

1. N(0,0.1)
2. N(3,0.1)
3. N(4,0.1)

We have computed the first PC for each of the sampling situations. It is clear from the accompanying plot that contamination can adversely affect the direction of a PC. In this example the size and location of the outliers are not severe yet the direction of the first PC is completely upset. In the two-dimensional case we can surely identify the

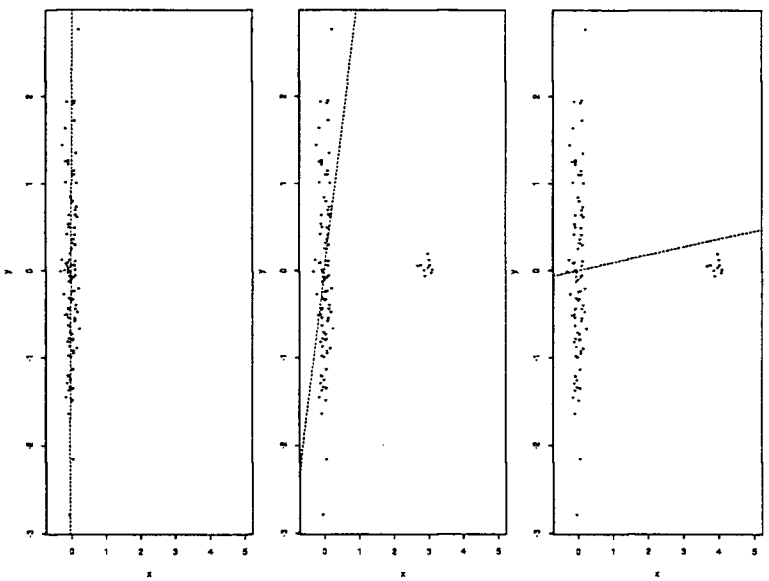

Figure 2: Plot of the First PC's

outliers from looking at an $x - y$ plot. In higher dimensions we may not be able to do so.

To make the PCA contamination resistant, statisticians have adopted one of two possible approaches. One is to obtain a robust estimate of the scatter matrix and then proceed with an eigenvalue-eigenvector decomposition (Boente, 1987) to compute the principal components. The second approach is to replace the variance in the maximization by some robust estimate of scale, s, and then proceed as above (Chen and Li, 1985). Note that, classically, we obtain the same principal components whether we begin by maximizing the variance or by minimizing it. This is generally not the case when $var(\alpha^T \mathbf{x})$ is replaced by $s(\alpha^T \mathbf{x})$. In the robust setup, the minimization approach is more attractive in view of the many minimization algorithms developed for classical and orthogonal regression that can be used as building blocks.

4.1.1 Robust PCA via the Robustification of Scatter Matrices

A natural way to obtain robust estimates of location and scatter is to extend the definition of one-dimensional M-estimators of location and scale to the multivariate setup. Maronna (1976) proposed such M-estimators, defined as solutions of a system of equations of the form

$$\frac{1}{n}\sum_{i=1}^{n} u_1[\{(\mathbf{x}_i - \mathbf{t})'\mathbf{V}^{-1}(\mathbf{x}_i - \mathbf{t})\}^{1/2}](\mathbf{x}_i - \mathbf{t}) = \mathbf{0}$$

$$\frac{1}{n}\sum_{i=1}^{n} u_2[(\mathbf{x}_i - \mathbf{t})'\mathbf{V}^{-1}\mathbf{x}_i - \mathbf{t})](\mathbf{x}_i - \mathbf{t})(\mathbf{x}_i - \mathbf{t})' = \mathbf{V},$$

where u_1 and u_2 are functions satisfying a set of general assumptions (see Maronna, 1976). Boente (1987) studied the asymptotic distribution of the eigenvalues and eigenvectors of the above M-estimator of scatter. She has shown that they are consistent and asymptotically normal at the usual rate, $\sqrt{n}$. However, it can be shown that that M-estimates of scatter attain a BP of at most $1/p$ (Maronna, 1976), where p is the number of variables. This makes them of limited use when p is large. Also, it is not clear whether the BP of the covariance matrix estimate is inherited by the direction of the PC's.

Some other techniques for computing robust scatter matrices include convex peeling (Barnett, 1976, Bebbington, 1978), ellipsoidal peeling (Titterington , 1978, Helbling, 1983), iterative trimming (Gnanadesikan and Kettenring, 1972, Devlin et al., 1975) and depth trimming based on the concept of depth (Tukey, 1974). Unfortunately, they all possess a breakdown point of at most 1/(p+1) (Donoho, 1982).

The first affine equivariant estimator with high BP was constructed independently by Stahel (1981) and Donoho (1982). It measures the "outlyingness" of a point $\mathbf{x}$ relative to some center location. For each observation $\mathbf{x}_i$, one looks for the one-dimensional

projection leaving it most exposed:

$$\zeta_i = \sup_{\|\mathbf{v}\|=1} \frac{|\mathbf{v}^T\mathbf{x}_i - MED_j(\mathbf{v}^T\mathbf{x}_j)|}{MED_k|\mathbf{v}^T\mathbf{x}_k - MED_j(\mathbf{v}^T\mathbf{x}_j)|}\ ,\ \mathbf{v} \in R^p\ ,$$

where MED is the median. Now consider a weight function $w_i = w(\zeta_i)$, where $w : [0,\infty) \rightarrow [0,\infty)$ is decreasing with $\sup \|\zeta\ w(\zeta)\| < \infty$. A robust covariance estimator based on the w_i's can be computed then as

$$\mathbf{C} = \frac{\sum_{i=1}^n w_i^2(\mathbf{x}_i - \mathbf{t})(\mathbf{x}_i - \mathbf{t})^T}{\sum_{i=1}^n w_i^2}\ ,$$

where the multivariate estimate of location, $\mathbf{t}$, is defined to be

$$\mathbf{t} = \frac{\sum_{i=1}^n w_i\mathbf{x}_i}{\sum_{i=1}^n w_i}\ .$$

The estimators obtained this way combine high BP with affine equivariance (Donoho, 1982). However, for each random vector $\mathbf{x}_i$, we have to solve a nontrivial maximization problem. This would be computationally prohibitive even for the computers aboard the Enterprise.

Rousseeuw (1983, 1987) proposed another estimator, the minimum volume ellipsoid (MVE), that combines the properties of affine equivariance and high BP. Define an ellipsoid $E_{\mathbf{C},\mu}$ by

$$E_{\mathbf{C},\mu} = \{\mathbf{x} :\ (\mathbf{x} - \mu)^T\mathbf{C}^{-1}(\mathbf{x} - \mu) \leq 1\}$$

and the set $\mathcal{C}$ by

$$\mathcal{C} = \{(\mathbf{C},\mu) :\ \#(E_{\mathbf{C},\mu} \cap \text{data})\ \geq [n/2] + 1\}\ .$$

The MVE is $(\mathbf{C},\mu) = \text{argmin}|\mathbf{C}|$, where $|\mathbf{C}|$ is the determinant of $\mathbf{C}$. In most cases it is not feasible to consider all "halves" of the data and to compute the volume of the smallest ellipsoid that surrounds them. Hence to compute the MVE, we use a method similar to the bootstrap.

Given a set of random vectors $\{\mathbf{x}\}_n$ in R^p, we draw repeatedly a subsample of $p+1$ different observations. The number of subsamples, m, drawn must be large enough so that the probability of a subsample containing only "good" data points is high. For large data sets with many variables, we limit the number of subsamples to whatever is computationally feasible (this is usually in the range of 100 to 3000 depending on p).

We find the mean $\bar{\mathbf{x}}_k$ and the covariance matrix $\mathbf{C}_k$ for the k^{th} subsample. Denote by

$$E_k = \{\mathbf{x} : (\mathbf{x} - \bar{\mathbf{x}}_k)^T \mathbf{C}_k^{-1} (\mathbf{x} - \bar{\mathbf{x}}_k) \leq 1\}$$

the ellipsoid corresponding to $\mathbf{C}_k$ and $\bar{\mathbf{x}}_k$. It contains the observations $\mathbf{x}_i$ that are within a $\mathbf{C}_k$ unit distance from $\bar{\mathbf{x}}_k$. The volume of this ellipsoid is related to $|\mathbf{C}_k|$, that is,

$$\text{vol}\ \{\mathbf{x} : (\mathbf{x} - \bar{\mathbf{x}}_k)^T \mathbf{C}_k^{-1} (\mathbf{x} - \bar{\mathbf{x}}_k) \leq 1\} = k_p |\mathbf{C}_k|^{1/2} \ ,$$

where

$$k_p = \frac{2\pi^{p/2}}{p\Gamma(\frac{p}{2})}$$

and $\Gamma(x)$ is the gamma function evaluated at x (see for example Johnson and Wichern, 1988). To envelop $[n/2]+1$ points, the ellipsoid E_k has to be inflated or deflated by being multiplied by some correction factor, the median Mahalanobis distance (MMD)

$$MMD_k = \text{median}_{i=1,\ldots,n} (\mathbf{x}_i - \bar{\mathbf{x}}_k)' \mathbf{C}_k^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}_k) \ .$$

Observe that the resulting ellipsoid

$$E_k' = \{\mathbf{x} : (\mathbf{x} - \bar{\mathbf{x}}_k)^T \mathbf{C}_k^{-1} (\mathbf{x} - \bar{\mathbf{x}}_k) \leq MMD_k\}$$

contains exactly 50% of the observations. The volume of E_k' is

$$\begin{aligned}
&\text{vol}(E_k') = \text{vol}\ \{\mathbf{x} : (\mathbf{x} - \bar{\mathbf{x}}_k)^T \mathbf{C}_k^{-1} (\mathbf{x} - \bar{\mathbf{x}}_k) \leq MMD_k\} \\
&= \text{vol}\{\mathbf{x} : (\mathbf{x} - \bar{\mathbf{x}}_k)^T (\mathbf{C}_k MMD_k)^{-1} (\mathbf{x} - \bar{\mathbf{x}}_k) \leq 1\} \\
&= k_p |MMD_k \mathbf{C}|^{1/2} \\
&= k_p MMD_k^{p/2} |\mathbf{C}|^{1/2} \ .
\end{aligned}$$

Hence the volume of the ellipsoid E_k' is proportional to

$$MMD_k^{p/2}|\mathbf{C}_k|^{1/2} \ . \tag{9}$$

Suppose that $\mathrm{MMD}_{k^*}^{p/2}|\mathbf{C}_{k^*}|^{1/2}$ minimizes (9) over all subsamples $k = 1, .., m$. Then the MVE covariance estimator is expressed as

$$\mathbf{C}_{k^*} MMD_{k^*} (\chi^2_{p;0.5})^{-1} \ ,$$

where $\chi^2_{p;0.5}$ is the median of a chi distribution with p degrees of freedom; it is used as a correction factor to obtain consistency at the multivariate normal model. In the univariate case the MVE reduces to the SHORTH. Given a set of numbers $\{x_i\}_{i=1,..,n}$, it is the length of the shortest line segment that contains at least $[n/2]+1$ such numbers. It can be shown (Martin and Zamar, 1989) that the SHORTH is the most resistant measure of dispersion with respect to minimizing the maximum bias among all M-estimates of scale.

Davies (1989) showed that the MVE converges weakly at rate of $\sqrt[3]{n}$ to a limiting distribution that is nonnormal. To improve the rate of convergence, Rousseeuw (1983) also proposed the minimum covariance determinant (MCD). The MCD covariance matrix estimate is the sample variance of the observations contained in the minimum volume elipsoid, the MCD multivariate location estimate is their sample mean. Butler and Juhn (1988) showed that the MCD is asymptotically normal at the rate $\sqrt{n}$.

Estimators analogous to S-estimates, MM-estimates and τ-estimates of location and scale in the univariate have been extended to the multivariate setup by Lopuhaä (1990) who discusses their properties at length in his Ph.D. thesis.

The above estimators have one characteristic in common; they are affine equivariant. To attain affine equivariance and a high BP, we must sacrifice computational efficiency. In the PCA, the principal components of a covariance matrix define an orthogonal

basis in the factor space. Hence, we need only consider estimators that are orthogonal equivariant, i.e., preserve the basis. Weakening the assumption of affine equivariance allows us to develop a new PCA estimator that is robust, yet easy to compute. A brief description of the estimator follows while its properties are discussed in section 6.

Suppose we have a sample $\{\mathbf{x}\}_{i=1,..,n}$ with some initial robust estimate of covariance $\hat{\mathbf{S}}_0(\mathbf{x})$. We propose to use an iterative procedure for computing weighted estimates of multivariate location and scatter with weights based on the principal components of $\hat{\mathbf{S}}_k$, the estimate of the covariance matrix at the kth iteration. Note that Maronna's M-estimate is also a reweighted covariance matrix, but the weights are based on the Mahalanobis distance (this is the reason for the low BP of the Maronna's estimate).

Let $\mathbf{a}_j$ be the eigenvector of $\hat{\mathbf{S}}_k$ associated with the jth largest eigenvalue. Then the jth principal component of $\mathbf{x}_i$ is $PC_{ij} = \mathbf{a}_j^T \mathbf{x}_i$. Next, we consider a weight function, $w_{ij} = w(PC_{ij})$, where w: [0,∞) $\rightarrow$ [0,∞) is decreasing with $\sup \|PC_{ij}\ w(PC_{ij})\| < \infty$. The weighted estimators of multivariate location and scatter are

$$\mathbf{t}_{k+1} = \frac{\sum_{i=1}^n W_i^k \mathbf{x}_i}{\sum_{i=1}^n W_i^k}$$

and

$$\mathbf{S}_{k+1} = \frac{\sum_{i=1}^n (\mathbf{x}_i - \mathbf{t}_{k+1})(\mathbf{x}_i - \mathbf{t}_{k+1})^T (W_i^k)^2}{\sum_{i=1}^n (W_i^k)^2},$$

where W_i^k is the product of w_{ij}'s computed at step k. The W_i^k satisfy the following condition

$$W_i^k = \begin{cases} W_i^k & \text{if } W_i^k < W_i^{k-1} \\ W_i^{k-1} & \text{otherwise} \end{cases}.$$

The weights W^k are forced to decrease at each iteration. The lower bound for the weights will be zero by the assumptions on the weight generating function $w(\bullet)$. This ensures convergence of the method. We will show in section 6 that at least $p+1$ weights will be larger than zero for observations that do not lie in a lower dimensional hyperplane.

This will prevent the scatter estimate from being singular. Aside from being easy to compute, the estimators are orthogonal equivariant, consistent and have a BP of 1/2 as will be shown in section 6.

4.1.2 Robust PCA via Projection Pursuit

Classical PCA is a type of projection pursuit method. Consider a set of random vectors $\{\mathbf{x}\}_{i=1,\ldots,n}$. In this method, one searches R^p for a direction in which the variance of a linear function, $\alpha^T\mathbf{x}$, of the elements of $\mathbf{x}$ attains a critical value. The BP of variance, as classically defined, tends to zero with increasing sample size and even a tiny fraction of contamination may cause it explode (blow up to ∞). To make the PCA more resistant, a robust scale estimate S of $\alpha^T\mathbf{x}$ is used in place of the variance. The unmodified problem is as follows:

$$\min_{\|\mathbf{a}_0\|=1} S(\mathbf{a}_0) \quad \text{subject to} \quad \frac{1}{n}\sum_{i=1}^{n}\chi\left(\frac{\mathbf{a}_0^T\mathbf{x}_i}{S(\mathbf{a}_0)}\right) = b\,,$$

where χ is a nondecreasing even function that limits the influence of outlying or influential observations. b is usually taken to be $E\{\chi(Z)\}$, where Z is the standard normal random variable. To make the minimization feasible, one can employ a method first introduced by Chen and Li (1985). It consists of three steps: reparametrization, minimization and projection.

First, we note that a (p+1)-dimensional unit vector $\mathbf{a}_0$ can be reparametrized as

$$\mathbf{a}_0 = \frac{1}{\sqrt{1+\beta^T\beta}}(1,-\beta)^T\,, \text{ where } \beta \in R^p\,.$$

The purpose of the reparametrization is to eliminate the constraint $\|\mathbf{a}_0\| = 1$ and make the computations simpler by doing so. The problem thus becomes

$$\min_{\beta} S(\beta) \quad \text{subject to} \quad \frac{1}{n}\sum_{i=1}^{n}\chi\left(\frac{x_{0i}-\beta_1 x_{1i}-\ldots-\beta_p x_{pi}}{\sqrt{1+\beta^T\beta}\;S(\beta)}\right) = b\,, \qquad (10)$$

where $\beta \in R^p$. Denote by $\hat{\beta} \in R^p$ the minimizer of $S(\beta)$. Next, we notice that the solution to

$$\min_{\|\mathbf{a}\|=1, \mathbf{a}^T\hat{\mathbf{a}}_0=0} S(\mathbf{a})$$

lies in the nullspace, $\mathcal{N}$, of $\hat{\mathbf{a}}_0$, where $\hat{\mathbf{a}}_0 = \frac{1}{\sqrt{1+\hat{\beta}^T\hat{\beta}}}(1, -\hat{\beta})^T$. We project our data onto this nullspace and proceed as above, i.e obtain a solution to an unconstrained problem of one less dimension by means of reparametrization. Let $\tilde{\mathbf{a}}_1, .., \tilde{\mathbf{a}}_p$ be the orthonormal basis of $\mathcal{N}$. Then any $\mathbf{a}_1 \in \mathcal{N}$ can be expressed by

$$\mathbf{a}_1 = \sum_{k=1}^{p} \alpha_k \tilde{\mathbf{a}}_k \ ,$$

where $\alpha \in R^p$ and $\alpha^T\alpha = 1$. Let $\mathbf{y}_j = (\tilde{\mathbf{a}}_1^T\mathbf{x}_j, ..., \tilde{\mathbf{a}}_p^T\mathbf{x}_j)^T$ be the projection of the $(p+1)$-dimensional vector of observations $\mathbf{x}_j$ onto $\mathcal{N}$. Observe that the $\mathbf{y}_i$'s have one less dimension than the $\mathbf{x}_i$'s. Now we minimize the scale $S(\mathbf{a}_1)$ over all unit vectors in $\mathcal{N}$, that is,

$$\min_{\|\mathbf{a}_1\|=1} S(\mathbf{a}_1) \quad \text{subject to} \quad \frac{1}{n}\sum_{i=1}^{n} \chi\left(\frac{\mathbf{a}_1^T\mathbf{x}_i}{S(\mathbf{a}_1)}\right) = \frac{1}{n}\sum_{i=1}^{n} \chi\left(\frac{\alpha^T\mathbf{y}_i}{S(\alpha)}\right) = b \ .$$

By reparameterizing α as

$$\alpha = \frac{1}{\sqrt{1+\beta^T\beta}}(1, -\beta)^T, \quad \text{where} \quad \beta \in R^{p-1} \ ,$$

the above minimization problem becomes

$$\frac{1}{n}\sum_{i=1}^{n} \chi\left(\frac{y_{1i} - \beta_1 y_{2i} - \ \ldots \ - \beta_{p-1} y_{pi}}{\sqrt{1+\beta^T\beta}\ S(\beta)}\right) \ = \ b \ .$$

By solving the above equation we obtain

$$\hat{\mathbf{a}}_1 = \sum_{k=1}^{p} \hat{\alpha}_i \tilde{\mathbf{a}}_i \ ,$$

where $\hat{\alpha} = (\hat{\alpha}_1, .., \hat{\alpha}_p)^T$ is given by

$$\alpha = \frac{1}{\sqrt{1+\hat{\beta}^T\hat{\beta}}}(1, -\hat{\beta})^T \ .$$

$\hat{\beta} \in R^{p-1}$ is the minimizer of $S(\beta)$. We continue this process until the dimension of the nullspace of the existing vectors is reduced to 0. Vectors obtained in this fashion will be the robust eigenvectors and the square of the scale estimates will be the robust eigenvalues. The robust scatter matrix can then be reconstructed as

$$\hat{\Sigma} = \sum_{i=0}^{p} S_i^2 \hat{\mathbf{a}}_i \hat{\mathbf{a}}_i^T .$$

The beauty of this method lies in the fact that the three steps described above reduce a problem in (p+1) dimensions to the same problem in one less dimension. The robust scale estimate S used in equation (10) is a nonlinear function with several minima. To minimize it, one requires a good initial starting point.

In the linear regression setup, one uses a particular S-estimate of regression, the least median of squares (LMS). Essentially, one computes the regression estimates by minimizing the median of the absolute residuals. Analytically, the solution to this problem can be written as

$$\min_{\beta} \; S(\beta) \quad \text{subject to} \quad \frac{1}{n} \sum_{i=1}^{n} \chi_a \left(\frac{y_i - \beta^T \mathbf{x}_i}{S(\beta)} \right) = 1/2 , \tag{11}$$

where χ is defined as

$$\chi_a(x) = \begin{cases} 0 & \text{if } |x| \leq a \\ 1 & \text{if } |x| > a \end{cases} , \; a \in R .$$

χ is referred to as the *jump function*. Minimizing (11) is computationally very difficult given the discontinuous nature of χ_a.

In practice, the computation of the LMS estimate involves a technique similar to the bootstrap. Given a set of observations $\{\mathbf{x}_i\}$: i=1,..,n in R^{p+1}, we begin by drawing subsamples of (p+1) points. Again, note that a large enough number of subsamples must be drawn to increase the probability that a particular subsample contains only "good" data. For very large data sets with multiple variables, the optimal number of

subsamples considered is limited to whatever is computationally feasible (usually 100 to 3000 depending on p).

We apply the method of least squares (LS) to each subsample k to obtain the regression coefficients $\hat{\beta}_k$. Then we compute the LS residuals corresponding to $\hat{\beta}_k$ for all observations. The LMS estimate is the set of coefficients that minimizes the median of the squared residuals.

The method of LMS is CPU intensive and its rate of convergence has been shown to be only $\sqrt[3]{n}$. This is because for each sample k we find the median absolute residual, a quantity that is not uniquely defined for even number of observations. The median absolute residual is an M-estimate of scale based on a jump function χ. The nature of the function causes slower than usual convergence rate.

To improve the speed of convergence to the usual rate, we can use a smoothed version of the function χ (Tukey's χ, for example). The M-scale equation will now have a unique solution; however, the computation of the minimum M-scale remains nontrivial.

To compute it, we can use the resampling scheme described above. However, this requires us to solve an equation of the robust residual scale similar to (7) m times. This can be extremely time consuming especially for large p. To improve computational efficiency, we make use of Yohai's suggestion (Yohai and Zamar, 1990). We solve (7) only when it becomes necessary.

Observe that

$$\frac{1}{n}\sum_{i=1}^{n}\chi\left(\frac{y_i - \mathbf{x}_i^T\beta}{s(\beta)}\right) < \frac{1}{2} \tag{12}$$

if $s(\beta)$ is overestimated. To initialize the procedure, we compute an M-estimate of scale for the first set of residuals to obtain $s_1(\beta)$. Subsequently to minimize $s(\beta)$, we only solve for $s(\beta)$ if (12) is satisfied (the scale used in (12) is the minimum M-scale computed thus far). Implementing this suggestion will reduce the number of equations that we

have to solve from m, the number of subsamples drawn, to $\ln(m)$.

In the context of projection pursuit, we propose an analog of LMS, the least median of absolute orthogonal errors (ORLM). This method differs from LMS only in the computation of errors. In LMS, an error is defined as the vertical distance between the observed response and the fitted response. In ORLM, it is defined as the Euclidean distance of the observed response from the fitted regression line.

Both methods, LMS and ORLM, although bias robust, have a rate of convergence of $\sqrt[3]{n}$. This results in low efficiency even if the errors are really normally distributed. To improve the speed of convergence, one uses a smoothed-out version of the ORLM obtained by replacing the jump function χ with a continuous χ, for instance Huber's

$$\chi_c(x) = \min\{1, \frac{x^2}{c^2}\}$$

or Tukey's

$$\chi_c(x) = \min\{1, \frac{3}{c^2}(x^2 - \frac{x^4}{c^2} + \frac{x^6}{3c^4})\} .$$

c is called the *tuning constant*. The S-estimate of scale in a regression context based on either χ function is asymptotically normal at the usual rate. Tukey's χ is usually preferred because it has continuous first and second derivatives and hence equation (11) can be more easily minimized by Newton-Raphson type of methods. A disadvantage of S-estimates is their inability to achieve efficiency with high BP at the same time. There has always been a trade-off between the two.

If an S-estimate is to achieve maximal BP, a low tuning constant c is used. For Tukey's χ, c = 1.548. This will cause some non-outlying observations to be viewed as outliers and penalized accordingly; any observation with the standardized robust residual larger than 1.548 will be considered an outlier since from that point on all observations will have the same value at the function χ). This will in turn increase the asymptotic variance of the estimated parameters caused by the loss of information.

This results in a reduction in efficiency (it can be shown that efficiency is an increasing function of c). It turns out, from extensive Monte Carlo simulation runs, that minimizing the robust scale estimate of the orthogonal errors gives rise to regression estimates that are about six times less efficient than the MLE at the Gaussian model.

Estimators, MM and τ, that combine maximal BP with efficiency have been proposed in the linear regression setup. Both can be made 95% efficient at the Gaussian model and their maximum asymptotic bias characteristics are comparable to those of the maximal BP S-estimates.

In the orthogonal regression setup we propose estimators analogous to the MM-estimates and τ-estimates of classical regression. In the computation of the MM-estimates, we use a fixed S-estimate of scale, say S_n, of the orthogonal residuals that has a BP of 1/2 and solve for β, the regression parameter. This allows us to increase the tuning constant c to gain efficiency (c = 4.7 for 95% efficiency at the normal model) by drawing closer to the quadratic χ of the Gaussian model.

The first estimator we define is the orthogonal regression MM estimator (ORMM). It is the solution of the minimization problem

$$\min_{\beta \in R^p} \sum_{i=1}^{n} \chi \left(\frac{y_i - \beta^T \mathbf{x}_i}{\sqrt{1 + \beta^T \beta}\ S(\beta)} \right) ,$$

where χ is Tukey's χ with c = 4.7 and S_n is as above.

The second estimator is the orthogonal regression τ estimator (OR-τ) which is defined as the solution to

$$\min_{\beta \in R^p} \tau^2(\beta) = S^2(\beta) \sum_{i=1}^{n} \chi_2 \left(\frac{y_i - \beta^T \mathbf{x}_i}{\sqrt{1 + \beta^T \beta}\ S(\beta)} \right) ,$$

where $S(\beta)$ is implicitly defined by

$$\frac{1}{n} \sum_{i=1}^{n} \chi_1 \left(\frac{y_i - \beta^T \mathbf{x}_i}{\sqrt{1 + \beta^T \beta}\ S(\beta)} \right) = 1/2 .$$

The tuning constant $c_1 = 1.548$ for χ_1 is chosen so that the maximal BP is achieved, the tuning constant $c_2 = 6.08$ for χ_2 is chosen so that 95% efficiency is achieved at the Gaussian model. The τ estimator is an adaptive combination of a high efficiency M-estimate with a maximal BP M-estimate. If data are contaminated with a large fraction of outliers, the robust M-estimate dominates. If there are no outliers in the data or only a small fraction, the efficient M-estimate dominates. Hence the τ estimator combines both bias robustness and efficiency.

We have seen that robust OR can be used as a building block in robust PCA. This approach was pioneered by Chen and Li (1985). In their paper they considered the bias properties of an S-estimate of the scale of the principal components. However, they ignored the breakdown properties of the direction in which this scale is minimized.

We have extended Chen and Li's method to efficient robust MM and τ estimators that attain maximal BP with respect to the size of the PC's. We have also considered the direction bias and we will show that the breakdown properties of the scale are not inherited by the direction. We further show that the S-estimate of the direction of the PC's cannot be made robust and efficient at the same time. On the other hand, the MM and τ estimates can be made 95% efficient while retaining a high level of robustness which, as we will show in the next section, depends on the ratio of adjacent eigenvalues. The larger the ratio the higher the BP. As this ratio approaches ∞ the BP tends toward 1/2.

5 Direction Bias Computation

The claims made in this section without formal proof are intuitively clear and can be rigorously established along the lines of Zamar (1989).

Classical methods are usually derived under Gaussian assumptions and are optimal when the data follow these assumptions. Gross-error-models of the form

$$F = (1-\epsilon)F_0 + \epsilon H \ ,$$

where F_0 is a multivariate normal with mean μ (we take $\mu = \mathbf{0}$, for simplicity and without loss of generality) and some covariance matrix $\mathbf{\Sigma}$, are used to describe situations in which a certain fraction ϵ of the data do not follow the central Gaussian model.

Let $\mathbf{\Sigma}_0 = \mathbf{\Sigma}(F_0)$ be the covariance matrix at F_0. We denote by $\lambda_0 < \lambda_1 < .. < \lambda_p$ the eigenvalues of $\mathbf{\Sigma}_0$ and by $\mathbf{a}_0, \mathbf{a}_1, .., \mathbf{a}_p$ the associated eigenvectors.

The treatment of the direction bias is asymptotic. We denote by $\hat{\mathbf{a}}_i(F)$ the estimating direction functional at F. It can be shown that the direction functionals are Fisher consistent, that is, $\hat{\mathbf{a}}_i(F_0) = \mathbf{a}_i$. The bias of the direction $\hat{\mathbf{a}}_i$ at F is then defined (Zamar ,1989) as

$$B_i(F) = 1 - |\hat{\mathbf{a}}_i^T(F)\hat{\mathbf{a}}_i(F_0)| = 1 - |\hat{\mathbf{a}}_i^T(F)\mathbf{a}_i| \ .$$

Note that $B_i(F)$ lies between zero, no bias, and one, complete breakdown of the direction $\hat{\mathbf{a}}_i$. The maximum direction biases $\bar{B}_i(\epsilon)$ are then

$$\bar{B}_i(\epsilon) = \sup_F B_i(F) \ .$$

The breakdown point of $\hat{\mathbf{a}}_i(F)$ is then (Zamar, 1989)

$$BP_i = \sup_{\epsilon \in (0,1)} \{\epsilon : \bar{B}_i(\epsilon) < 1\} \ .$$

The breakdown point of $\hat{\mathbf{a}}_i(F)$ is achieved when $\hat{\mathbf{a}}_i(F)$ becomes orthogonal to $\mathbf{a}_i$. Given a pair of adjacent PC's, this occurs when an outlying observation inflates the scale

of the smaller PC enough to make it larger than the scale of the larger PC. Intuitively, this can be done most easily when the variances of the adjacent PC's are similar in size.

We can find a lower bound $b_i(\epsilon)$ for the $\bar{B}_i(\epsilon)$. The lower bound $b_i(\epsilon)$ depends on the fraction of contamination ϵ and on the ratio of two adjacent eigenvalues λ_{i+1} and λ_i. We denote the square root of the ratio λ_{i+1}/λ_i by r_i. The larger the r_i, the smaller the lower bound $b_i(\epsilon)$. It can be shown that the lower bound $b_0(\epsilon)$ is sharp, that is,

$$b_0(\epsilon) = \bar{B}_0(\epsilon) \ .$$

We conjecture that $b_i(\epsilon) = \bar{B}_i(\epsilon)$ for $i = 1, .., p$, as well; however, we are unable to prove it.

To find the lower bound $b_i(\epsilon)$, we should, in principle, consider all directions $\mathbf{a}(F)$ that are generated by introducing outliers at different locations in the (p+1)-dimensional space. However, it can be shown, as in Zamar (1989), that the largest deviation from the true direction $\mathbf{a}_i = \hat{\mathbf{a}}_i(F_0)$ is obtained by moving toward $\hat{\mathbf{a}}_{i+1}(F_0)$ which is the path of the least resistance.

Hence it is enough to consider a biased direction $\mathbf{a}(\gamma)$ of the form

$$\mathbf{a}(\gamma) = (1-\gamma)\mathbf{a}_i + \sqrt{1-(1-\gamma)^2}\ \mathbf{a}_{i+1} \ ,$$

where γ denotes the lower bound for the asymptotic direction bias. We also define the direction $\tilde{\mathbf{a}}(\gamma)$ that is orthogonal to $\mathbf{a}(\gamma)$ by

$$\tilde{\mathbf{a}}(\gamma) = \sqrt{1-(1-\gamma)^2}\ \mathbf{a}_i - (1-\gamma)\mathbf{a}_{i+1} \ .$$

It turns out that the point mass contamination at

$$y = \kappa\tilde{\mathbf{a}}(\gamma) \ , \ \kappa \to \infty$$

results in the most pessimistic scenario. The gross error model associated with a point mass at y, δ_y, is denoted by

$$F^{\infty} = (1-\epsilon)F_0 + \epsilon\delta_y \ .$$

It is enough to consider F^∞ alone because being the most pessimistic contamination model it induces the largest lower bound.

5.1 τ - estimate

The asymptotic version $\hat{\mathbf{a}}(F_0)$ of the τ-estimate is

$$\hat{\mathbf{a}}(F) = \arg\min_{\|\mathbf{a}\|=1} \tau^2(F, \mathbf{a}) \ ,$$

where

$$\tau^2(F, \mathbf{a}) = s^2(F, \mathbf{a}) E_F \chi \left(\frac{\mathbf{a}^T \mathbf{x}}{s(F, \mathbf{a})} \right) \ .$$

The estimate of scale $s(F, \mathbf{a})$ is implicitly defined by

$$E_F \chi \left(\frac{\mathbf{a}^T \mathbf{x}}{s(F, \mathbf{a})} \right) = 1/2 \ .$$

We assume without loss of generality that $\chi(\infty) = 1$. We also define the function $g(t)$ as

$$g(t) = E\chi(\sqrt{t}Z) \ , \text{where } Z \sim N(0,1) \ .$$

To compute $b_i(\epsilon)$, we consider two situations (to simplify notation we assume $i = 0$). First we calculate the value of the τ-scale when the contamination is ignored, that is, when $\mathbf{a} = \hat{\mathbf{a}}_0(F_0) = \mathbf{a}_0$. In this case the distribution of $\mathbf{a}_0^T \mathbf{x}$ under F^∞ is

$$(1-\epsilon) N(0, \lambda_0) + \epsilon \delta_\infty \ .$$

Second we compute the value of the τ-scale when the contamination is fitted exactly, that is, when $\mathbf{a} = \hat{\mathbf{a}}_0(F) = \mathbf{a}(\gamma)$. In this case the distribution of $\mathbf{a}_0^T(\gamma)\mathbf{x}$ is

$$(1-\epsilon) N(0, (1-\gamma)^2 \lambda_0 + \Delta^2 \lambda_1) + \epsilon \delta_0 \ ,$$

where $\Delta^2 = [1 - (1-\gamma)^2]$. We can show that $\tau(F^\infty, \mathbf{a}(\gamma)) > \tau(F^\infty, \mathbf{a}_0)$ implies $\hat{\mathbf{a}}(F^\infty) = \mathbf{a}_0$. On the other hand, $\tau(F^\infty, \mathbf{a}(\gamma)) < \tau(F^\infty, \mathbf{a}_0)$ implies that $\hat{\mathbf{a}}(F^\infty) = \mathbf{a}(\gamma)$. The lower

bound for the maximum bias is thus obtained by equating

$$\tau^2(F^\infty, \mathbf{a}_0) = \tau^2(F^\infty, \mathbf{a}(\gamma)) \ .$$

We accomodate the contamination as long as the resulting scale is smaller than the scale we obtain by ignoring the contamination.

Consider the two modelling situations:

1. Contamination is ignored $(F^\infty, \mathbf{a}_0)$.

   The defining equation for the τ-scale can be written as

   $$\tau^2(F^\infty, \mathbf{a}_0) = s^2(F^\infty, \mathbf{a}_0)[(1-\epsilon)E_{F_0}\chi\left(\frac{\mathbf{a}_0^T\mathbf{x}}{s(F^\infty, \mathbf{a}_0)}\right) + \epsilon] \ ,$$

   or in terms of $g(t)$

   $$\tau^2(F^\infty, \mathbf{a}_0) = s^2(F^\infty, \mathbf{a}_0)[(1-\epsilon)g_2\left(\frac{\lambda_0}{s^2(F^\infty, \mathbf{a}_0)}\right) + \epsilon] \ . \tag{13}$$

   $s(F^\infty, \mathbf{a}_0)$ satisfies the following equation

   $$(1-\epsilon)g_1\left(\frac{\lambda_0}{s^2(F^\infty, \mathbf{a}_0)}\right) + \epsilon = 1/2 \ ,$$

   hence

   $$s^2(F^\infty, \mathbf{a}_0) = \frac{\lambda_0}{g_1^{-1}\left(\frac{0.5-\epsilon}{1-\epsilon}\right)} \ .$$

   Substituting for $s^2(F^\infty, \mathbf{a}_0)$ into (13) yields

   $$\tau^2(F^\infty, \mathbf{a}_0) = (1-\epsilon)\left[\frac{\lambda_0}{g_1^{-1}\left(\frac{0.5-\epsilon}{1-\epsilon}\right)}\right] g_2(g_1^{-1}\left(\frac{0.5-\epsilon}{1-\epsilon}\right)) + \epsilon\frac{\lambda_0}{g_1^{-1}\left(\frac{0.5-\epsilon}{1-\epsilon}\right)} \ .$$

2. Contamination is fitted exactly $(F^\infty, \mathbf{a}(\gamma))$.

   The equation for the τ-scale expressed in terms of g(t) is

   $$\tau^2(F^\infty, \mathbf{a}(\gamma)) = s^2(F^\infty, \mathbf{a}(\gamma))(1-\epsilon)g_2\left(\frac{(1-\gamma)^2\lambda_0 + \Delta^2\lambda_1}{s^2(F^\infty, \mathbf{a}(\gamma))}\right) \ , \tag{14}$$

where $s(F^\infty, \mathbf{a}(\gamma))$ satisfies the following equation

$$(1-\epsilon)g_1\left(\frac{(1-\gamma)^2\lambda_0 + \Delta^2\lambda_1}{s^2(F^\infty, \mathbf{a}(\gamma))}\right) = 1/2 \ .$$

Solving for $s(F^\infty, \mathbf{a}(\gamma))$ we obtain

$$s^2(F^\infty, \mathbf{a}(\gamma)) = \frac{(1-\gamma)^2\lambda_0 + \Delta^2\lambda_1}{g_1^{-1}\left(\frac{0.5}{1-\epsilon}\right)} \ .$$

We then substitute for $s(F^\infty, \mathbf{a}(\gamma))$ into (14) to obtain

$$\tau^2(F^\infty, \mathbf{a}(\gamma)) = (1-\epsilon)\frac{(1-\gamma)^2\lambda_0 + \Delta^2\lambda_1}{g_1^{-1}\left(\frac{0.5}{1-\epsilon}\right)} g_2(g_1^{-1}\left(\frac{0.5}{1-\epsilon}\right)) \ .$$

Equating $\tau^2(F^\infty, \mathbf{a}(\gamma))$ and $\tau^2(F^\infty, \mathbf{a}(\gamma))$ and solving for γ, we obtain the equation for the lower bound for the maximum direction bias as a function of the fraction of contamination ϵ and the eigenvalue ratio λ_1/λ_0, that is,

$$b_\tau(\epsilon) = 1 - \sqrt{\frac{\lambda_1/\lambda_0 - f(\epsilon)}{\lambda_1/\lambda_0 - 1}} \ , \tag{15}$$

where

$$f(\epsilon) = \frac{\left[g_2(g_1^{-1}\left(\frac{0.5-\epsilon}{1-\epsilon}\right)) + \frac{\epsilon}{1-\epsilon}\right]}{g_2(g_1^{-1}\left(\frac{0.5}{1-\epsilon}\right))} \frac{g_1^{-1}\left(\frac{0.5}{1-\epsilon}\right)}{g_1^{-1}\left(\frac{0.5-\epsilon}{1-\epsilon}\right)} \ . \tag{16}$$

5.2 S - estimate

The derivation of the lower bound for the maximum asymptotic direction bias for the S-estimate is performed in a way similar to that of the τ-estimate. We again consider two cases, first where the outlier is fitted exactly and second where the outlier is ignored. The lower bound is then obtained by equating the scale estimates of the smallest PC computed in the two cases.

The defining equation for $s(F^\infty, \mathbf{a}_0)$ is the following

$$(1-\epsilon)g\left(\frac{\lambda_0}{s^2(F^\infty, \mathbf{a}_0)}\right) + \epsilon = b \ .$$

Solving it for $s(F^\infty, \mathbf{a}_0)$ yields

$$s^2(F^\infty, \mathbf{a}_0) = \frac{\lambda_0}{g^{-1}\left(\frac{b-\epsilon}{1-\epsilon}\right)} .$$

Following the same steps as above we find the expression for $s(F^\infty, \mathbf{a}(\gamma))$

$$s^2(F^\infty, \mathbf{a}(\gamma)) = \frac{(1-\gamma)^2\lambda_0 + \Delta^2\lambda_1}{g^{-1}\left(\frac{b}{1-\epsilon}\right)} .$$

Equating $s^2(F^\infty, \mathbf{a}(\gamma))$ and $s^2(F^\infty, \mathbf{a}(\gamma))$ and solving for γ that denotes the lower bound for the maximum asymptotic direction bias, we get

$$b_S(\epsilon) = 1 - \sqrt{\frac{\lambda_1/\lambda_0 - f(\epsilon)}{\lambda_1/\lambda_0 - 1}} ,$$

where

$$f(\epsilon) = \frac{g^{-1}\left(\frac{b}{1-\epsilon}\right)}{g^{-1}\left(\frac{b-\epsilon}{1-\epsilon}\right)} . \tag{17}$$

Notice that if the function $g_2 = g_1$, then $E_F\chi\left(\frac{\mathbf{a}^T\mathbf{x}}{s(F,\mathbf{a})}\right) = 1/2$ and minimizing $\tau^2(F, \mathbf{a})$ becomes equivalent to minimizing $s^2(F, \mathbf{a})$ and the τ-estimate of direction will be equal to the S-estimate of direction. Equation (16) reduces to equation (17) resulting in $b_\tau(\epsilon) = b_S(\epsilon)$.

5.3 MM - estimate

To compute the MM-estimate of direction, we minimize the criterion

$$J(F, \mathbf{a}; \hat{s}(F)) = E_F\chi\left(\frac{\mathbf{a}^T\mathbf{x}}{\hat{s}(F)}\right) ,$$

where $\hat{s}(F)$ is an S-estimate of scale with the maximal BP, that is, $\hat{s}(F) = s(F, \mathbf{a}(F))$. We assume that the direction bias γ_0 of this robust S-estimate is smaller than the direction bias γ of the 95% efficient MM-estimate. Notice that for $\gamma > \gamma_0$,

$$s(F^\infty, \mathbf{a}(\gamma)) > s(F^\infty, \mathbf{a}_0)$$

and hence the robust S-estimate of direction will be equal to $\mathbf{a}_0$. Hence, for all $\gamma > \gamma_0$,

$$\hat{s} = \hat{s}(F^\infty) = s(F^\infty, \mathbf{a}_0) = \frac{\lambda_0}{g_1^{-1}\left(\frac{0.5-\epsilon}{1-\epsilon}\right)} .$$

Again consider the two modelling situations:

1. Contamination is ignored $(F^\infty, \mathbf{a}_0)$.

   In terms of the function g, the criterion $J(F^\infty, \mathbf{a}_0; \hat{s})$ can be written as

   $$J(F^\infty, \mathbf{a}_0; \hat{s}) = (1-\epsilon) g_2\left(\frac{\lambda_0}{\hat{s}}\right) + \epsilon . \tag{18}$$

   Substituting for $\hat{s}$ in (18) we get

   $$J(F^\infty, \mathbf{a}_0; \hat{s}) = (1-\epsilon) g_2(g_1^{-1}\left(\frac{0.5-\epsilon}{1-\epsilon}\right)) + \epsilon .$$

2. Contamination is fitted exactly $(F^\infty, \mathbf{a}(\gamma))$.

   The criterion $J(F^\infty, \mathbf{a}(\gamma); \hat{s})$ can be writen as

   $$J(F^\infty, \mathbf{a}(\gamma); \hat{s}) = (1-\epsilon) g_2\left(\frac{(1-\gamma)^2\lambda_0 + \Delta^2\lambda_1}{\hat{s}}\right) . \tag{19}$$

   We then substitute for $\hat{s}$ into (19) to obtain

   $$J(F^\infty, \mathbf{a}(\gamma); \hat{s}) = (1-\epsilon) g_2\left\{\frac{[(1-\gamma)^2\lambda_0 + \Delta^2\lambda_1] g_1^{-1}\left(\frac{0.5-\epsilon}{1-\epsilon}\right)}{\lambda_0}\right\} .$$

Again, the achievable bias occurs when fitting the outliers is better (from the criterion point of view) than ignoring them, that is, when

$$J(F^\infty, \mathbf{a}_0; \hat{s}) > J(F^\infty, \mathbf{a}(\gamma); \hat{s}) .$$

The lower bound for the maximum asymptotic direction bias, $b_{MM}(\epsilon)$, is attained when we equate $J(F^\infty, \mathbf{a}_0; \hat{s})$ to $J(F^\infty, \mathbf{a}(\gamma); \hat{s})$ and solve for γ. Doing so we obtain

$$b_{MM}(\epsilon) = 1 - \sqrt{\frac{\lambda_1/\lambda_0 - f(\epsilon)}{\lambda_1/\lambda_0 - 1}} , \tag{20}$$

where

$$f(\epsilon) = \frac{g_2^{-1}\left[g_2(g_1^{-1}\left(\frac{0.5-\epsilon}{1-\epsilon}\right) + \frac{\epsilon}{1-\epsilon}\right]}{g_1^{-1}\left(\frac{0.5-\epsilon}{1-\epsilon}\right)} . \tag{21}$$

In Figure 3, we plot the lower bound for the asymptotic direction bias as a function of the fraction of contamination ϵ for the 95% efficient S-estimate, the maximum BP S-estimate, the MM-estimate and the τ-estimate when the ratio r_i is equal to ten. The plot gives a global picture of the performance of the four estimates. First we notice that the S-estimate cannot achieve robustness and efficiency at the same time. The S-estimate of direction based on the 50% BP scale estimate gives the greatest bias protection but it is very inefficient (eff=28%) at the Gaussian model. On the other hand, the 95% efficient S-estimate is not bias robust and breaks down at about 11% of contamination when the ratio $r_i = \sqrt{\lambda_1/\lambda_0} = 10$ (see Figure 3a) and around 7% of contamination when $r_i = 2$ (see Figure 3b).

Second we observe that the BP of the S-estimate of direction depends on the r_i and falls short of the BP of the corresponding scale estimate which is 50% for the robust and 12% for the efficient scales, respectively. The BP of the robust S-estimate of direction is about 40% when $r_i = 10$ and about 23% when $r_i = 2$. The corresponding breakdown points of the efficient S-estimates are 11% and 8%, respectively. From Figure 4, we notice that as $r_i \to \infty$, the BP of the direction approaches that of the corresponding scale estimate.

Also notice that the biggest increase in the BP occurs when $1 \leq r \leq 5$; the BP curves remain fairly flat (although increasing toward the scale BP) for $r > 5$. Finally, we notice (see Figure 4) that as r approaches one, the BP tends to zero. However, when $r \approx 1$, switching the order of adjacent principal components will not seriously harm the analysis of the data.

Figure 3b is a plot of the lower bound of the direction bias as a function of the

fraction of contamination ϵ for an r_i of two. This ratio has been used in the Monte Carlo study that was carried out to assess the finite sample performance of the four estimates.

As we expected, the MM- and τ-estimates of direction can combine efficiency and robustness. From the accompanying pictures we notice that the bias performance of the 95% efficient MM- and τ-estimates is much better than that of the efficient S-estimate. Although the bias behaviour of the MM- and τ-estimates is in general worse than that of the robust S-estimate, the performance gap quickly narrows for increasing r_i. From (15) and (20) we notice that the BP of the MM- and τ-estimates occurs when $f(\epsilon)$ (defined by (16) and (21)) is equal to λ_1/λ_0, that is,

$$BP = f^{-1}(\lambda_1/\lambda_0) \ .$$

Hence, $f^{-1}(\lambda_1/\lambda_0) \to 1/2$ as $\lambda_1/\lambda_0 \to \infty$.

Finally, notice that the τ-estimate has visibly better bias characteristics than the MM-estimate for fractions of contamination larger than 0.27. This is to be expected because the τ-estimate, being of an adaptive nature, should accomodate better a large percentage of outliers than the fixed scale MM-estimate. For smaller ϵ, the MM-estimate is marginally better than the τ-estimate.

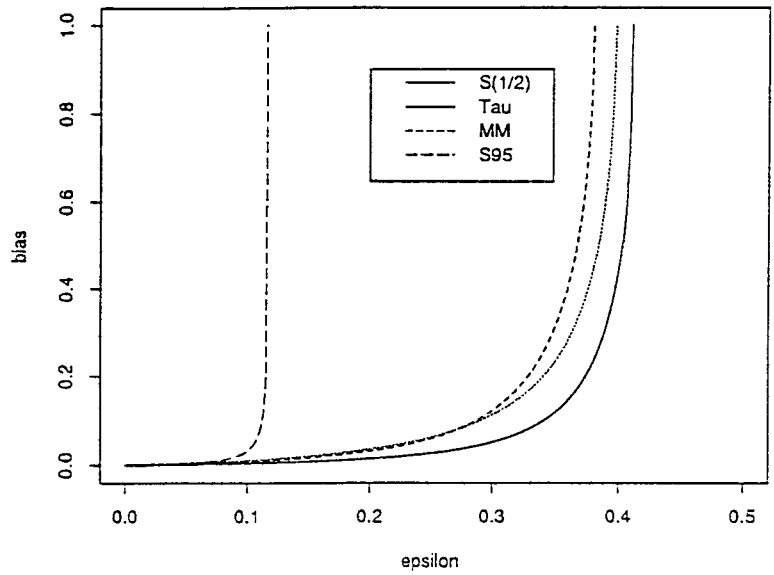

Figure 3a: Plot of Direction Bias vs Fraction of Contamination ($\sqrt{\lambda_1/\lambda_0} = 10$).

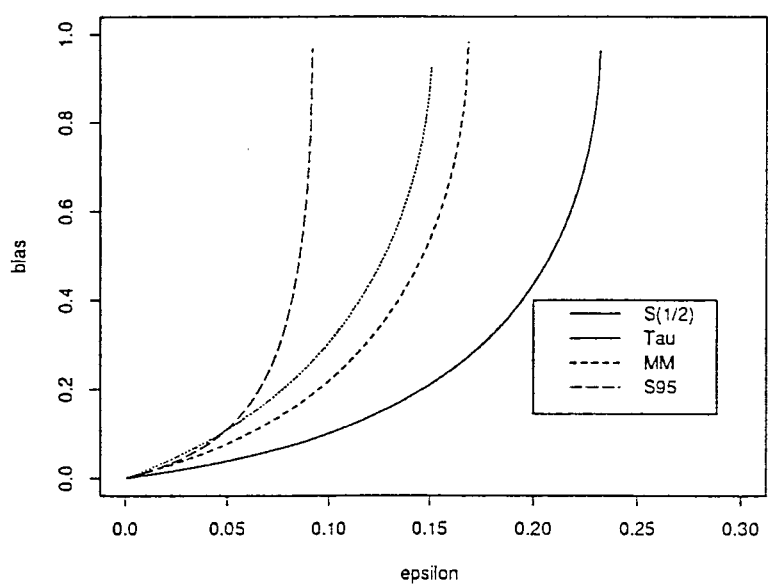

Figure 3b: Plot of Direction Bias vs Fraction of Contamination ($\sqrt{\lambda_1/\lambda_0} = 2$).

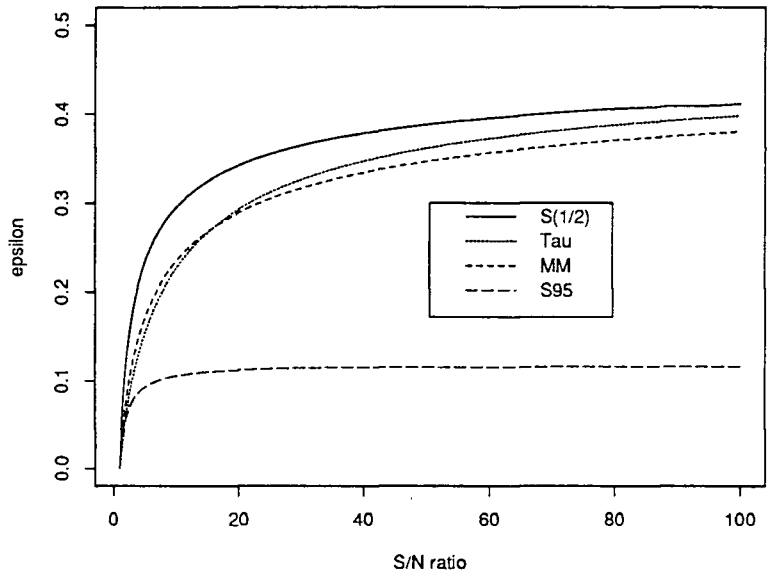

Figure 4: Plot of Fraction of Contamination versus λ_1/λ_0.

6 Proposed Estimator

Although there are a number of robust scatter matrix estimators in existence, they are computed via the minimization of a nonconvex function with multiple minima. This is both time consuming and difficult to achieve. Also, to increase the probability of convergence to a global minimum, a "good" initial starting value for the function must be provided. The method proposed here is quick and easy to compute. A Monte Carlo study will show that it compares favourably with other robust methods. It can also serve as an initial estimate for other, more complicated procedures.

Among the properties of the proposed estimator are high breakdown point and orthogonal equivariance. Usually, one requires an estimator to have a somewhat stronger type of equivariance, affine equivariance. However, PCA is based upon successive minimizations of the variance of a linear combination of the elements of $\mathbf{x}$. This combination is given by a particular eigenvector of the scatter matrix. The set of such eigenvectors forms an orthogonal basis in the factor space that is to be preserved under transforma-

tion. An orthogonal equivariant estimator satisfies such a requirement.

Here we describe in detail the procedure used to compute the proposed estimate. The algorithm has three steps:

1. Center the data.

   First we robustly center the data. In order to achieve orthogonal equivariance and a high BP for the estimate of scatter, we want to center the data by a multivariate location estimate that is orthogonal equivariant, robust, independent of the estimate of the scatter matrix and easy to compute. These requirements are satisfied by the L_1 estimate of location defined (see for example Lopuhaä, 1990) as the solution to the minimization problem

   $$\min_{\mathbf{T}_n} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{T}_n\| \ .$$

   The asymptotic BP of this location estimate is $1/2$. Once $\mathbf{T}_n$ is computed, it is subtracted from the data so that we obtain a new centered data set $\{\tilde{\mathbf{x}}_1, .., \tilde{\mathbf{x}}_n\}$, $\tilde{\mathbf{x}}_i = \mathbf{x}_i - \mathbf{T}_n$. For simplicity, the tilde will be dropped from the notation.

2. Compute the initial estimate of scatter, $\hat{\mathbf{S}}_0$.

   To initialize the procedure, we compute a robust weighted estimate of scatter. The weights used are generated by the function $u(r)$ that is deacreasing on $[0,\infty)$ and such that $u(r)$ is equal to zero when its argument exceeds some previously specified constant. Let $\hat{s}$ be the robust scale imlicitly defined by

   $$\frac{1}{n}\sum_{i=1}^{n} \chi_a \left(\frac{\|\mathbf{x}_i\| - \text{median}_j(\|\mathbf{x}_j\|)}{s(\|\mathbf{x}\|)} \right) = 1/2 \ .$$

   χ_a is defined as before and $a = \Phi(3/4)$. The above maximal BP M-estimate of scale is known as the median absolute deviation from the median, MAD. The

initial estimate of scatter $\hat{\mathbf{S}}_0$ is

$$\hat{\mathbf{S}}_0(\mathbf{x}_1,..,\mathbf{x}_n) = \frac{\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T u^2\left(\frac{\|\mathbf{x}_i\|}{\hat{s}}\right)}{\sum_{i=1}^n u^2\left(\frac{\|\mathbf{x}_i\|}{\hat{s}}\right)} .$$

3. Compute $\hat{\mathbf{S}}_{k+1}$ given $\hat{\mathbf{S}}_k$

Let $\hat{\mathbf{a}}_1,..,\hat{\mathbf{a}}_p$ be the eigenvalues of $\hat{\mathbf{S}}_k$, the current estimate. Then the $(k+1)^{th}$ estimate of the scatter matrix is of the form

$$\hat{\mathbf{S}}_{k+1}(\mathbf{x}_1,..,\mathbf{x}_n) = \frac{1}{\sum_{i=1}^n W_{ik}^2}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T W_{ik}^2$$

The W_{ik}'s are defined as the product of weights that are based on the standardized absolute principal components, that is,

$$W_{ik} = \prod_{j=1}^p w\left(\frac{|\hat{\mathbf{a}}^T\mathbf{x}_i|}{\hat{s}_j}\right) .$$

The weight generating function $w(r)$ is decreasing on [0,∞) and such that $w(r)$ is zero whenever its argument exceeds a previously specified constant. The scales $\hat{s}_j$, j=1,..,p are the MAD, that is,

$$\frac{1}{n}\sum_{i=1}^n \chi_a\left(\frac{\hat{\mathbf{a}}_j^T\mathbf{x}_i}{s_j}\right) = 1/2 .$$

Termination criterion:

We repeat the third step until the maximum difference between the current weights and the weights obtained at the previous step is less than some δ (in the Monte Carlo study δ was taken to be 0.001) or until the maximum number of iterations (user specified) is reached.

6.1 Some Properties of Proposed Estimate

In this section, we will discuss the properties of the proposed estimates of the direction and size of the principal components (PC). These estimates are based on a reweighted

scatter matrix which, at convergence, is of the form

$$\hat{\mathbf{S}}(\mathbf{x}_1, .., \mathbf{x}_n) = \frac{1}{\sum_{i=1}^{n} W_i^2(\mathbf{x}_i)} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T W_i^2(\mathbf{x}_i) \tag{22}$$

where W_i is the weight assigned to observation $\mathbf{x}_i$. The weight generating function, w, is even, decreasing on [0,∞) and such that $|w(r)r|$ is bounded. We will show that $\hat{\mathbf{S}}(\mathbf{x}_1, .., \mathbf{x}_n)$ is orthogonal equivariant and Fisher consistent. This will in turn imply that the estimates of the directions of the PC's are orthogonal equivariant and Fisher consistent and that the estimates of the size of the PC's are Fisher consistent. We will also show that the BP of $\hat{\mathbf{S}}(\mathbf{x}_1, .., \mathbf{x}_n)$ tends to 1/2 as the sample size approaches ∞. The BP is inherited by the estimate of the size of the PC's; however, the BP of the direction may be smaller as this was indeed the case for the S-, MM- and τ-estimates. However, as we will see in the next section, simulation results suggest that the BP of the direction of the PC's based on the proposed method is higher than that of the robust S-scale. Further study is required to investigate finite sample properties of the proposed estimate.

Theorem 1 *Let* $\{\mathbf{x}_1, .., \mathbf{x}_n\}$ *be a p-dimensional sample of size n. Suppose the weight functions* $u(t)$ *and* $w(t)$ *are even, decreasing on [0,*∞*) and such that* $|w(t)t|$ *and* $|u(t)t|$ *are bounded. Then the estimate* $\hat{\mathbf{S}}_n$ *given by (22) is orthogonal equivariant, that is,*

$$\hat{\mathbf{S}}_n(\mathbf{QX}) = \mathbf{Q}\hat{\mathbf{S}}_n(\mathbf{X})\mathbf{Q}^T ,$$

where $\mathbf{Q}$ *is any orthogonal matrix.*

Proof: First of all we notice that the data can be robustly centered by using, for example, the L_1 estimate of multivariate location (Lopuhaä, 1990) which is clearly orthogonal equivariant and has breakdown point 1/2. Hence we can assume without loss of generality that the location estimate $T(\mathbf{x}_1, .., \mathbf{x}_n)$ is equal to zero. Let $\mathbf{y}_i =$

$\mathbf{Qx}_i$ (i=1,..,n). Recall that the initial weights $u(\|\mathbf{x}_i\|)$ are based on the Euclidean norms. Since $\|\mathbf{x}_i\| = \|\mathbf{Qx}_i\| = \|\mathbf{y}_i\|$, the initial estimate of scatter $\hat{\mathbf{S}}_1(\mathbf{x}_1, .., \mathbf{x}_n)$ is orthogonal equivariant, that is,

$$\begin{aligned}
\hat{\mathbf{S}}_1(\mathbf{y}_1, .., \mathbf{y}_n) &= \frac{\sum_{i=1}^n \mathbf{y}_i \mathbf{y}_i^T u^2(\|\mathbf{y}_i\|)}{\sum_{i=1}^n u^2(\|\mathbf{y}_i\|)} \\
&= \frac{\sum_{i=1}^n (\mathbf{Qx}_i)(\mathbf{Qx}_i)^T u^2(\|\mathbf{x}_i\|)}{\sum_{i=1}^n u^2(\|\mathbf{x}_i\|)} \\
&= \mathbf{Q}\left\{\frac{\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T u^2(\|\mathbf{x}_i\|)}{\sum_{i=1}^n u^2(\|\mathbf{x}_i\|)}\right\}\mathbf{Q}^T \\
&= \mathbf{Q}\hat{\mathbf{S}}_1(\mathbf{x}_1, .., \mathbf{x}_n)\mathbf{Q}^T .
\end{aligned}$$

The proof is now completed by induction. Suppose that $\hat{\mathbf{S}}_m(\mathbf{x}_1, .., \mathbf{x}_n)$ is orthogonal equivariant. Let $\lambda_1 < \lambda_2 < ... < \lambda_p$ and $\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_p$ be the eigenvalues and eigenvectors of $\hat{\mathbf{S}}_m(\mathbf{x}_1, .., \mathbf{x}_n)$. Let $\mathbf{b}_1, \mathbf{b}_2, .., \mathbf{b}_p$ be the eigenvectors of $\hat{\mathbf{S}}_m(\mathbf{y}_1, .., \mathbf{y}_n)$. Then

$$\hat{\mathbf{S}}_m(\mathbf{x}_1, .., \mathbf{x}_n)\mathbf{a} = \lambda\mathbf{a} .$$

Premultiplying the above by $\mathbf{Q}$, we get

$$\mathbf{Q}\hat{\mathbf{S}}_m(\mathbf{x}_1, .., \mathbf{x}_n)\mathbf{a} = \lambda\mathbf{Qa} .$$

By using the fact $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$, we have

$$[\mathbf{Q}\hat{\mathbf{S}}_m(\mathbf{x}_1, .., \mathbf{x}_n)\mathbf{Q}^T][\mathbf{Qa}] = [\hat{\mathbf{S}}_m(\mathbf{y}_1, .., \mathbf{y}_n)][\mathbf{Qa}] = \lambda[\mathbf{Qa}] .$$

Hence $\mathbf{Qa}$ is an eigenvector of $\hat{\mathbf{S}}_m(\mathbf{y}_1, .., \mathbf{y}_n)$ corresponding to λ. Therefore $\mathbf{b}_i = \mathbf{Qa}_i$. Let

$$s_{xj} = \text{Median}_i(|\mathbf{a}_j^T\mathbf{x}_i|) = \text{Median}_i(|\mathbf{b}_j^T\mathbf{y}_i|) = s_{yj} \tag{23}$$

and

$$W_{xi} = \prod_{j=1}^{p} w\left(\frac{\mathbf{a}_j^T\mathbf{x}_i}{s_{xj}}\right) = \prod_{i=1}^{p} w\left(\frac{\mathbf{b}_j^T\mathbf{y}_i}{s_{yj}}\right) = W_{yi} . \tag{24}$$

Then, by (23) and (24), $\hat{\mathbf{S}}_{m+1}(\mathbf{x}_1, .., \mathbf{x}_n)$ is orthogonal equivariant, that is,

$$\begin{aligned}
\hat{\mathbf{S}}_{m+1}(\mathbf{y}_1, .., \mathbf{y}_n) &= \frac{\sum_{i=1}^n \mathbf{y}_i \mathbf{y}_i^T W_{yi}^2}{\sum_{i=1}^n W_{yi}^2} \\
&= \frac{\sum_{i=1}^n (\mathbf{Q}\mathbf{x}_i)(\mathbf{Q}\mathbf{x}_i)^T W_{xi}^2}{\sum_{i=1}^n W_{xi}^2} \\
&= \mathbf{Q}\left\{\frac{\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T W_{xi}^2}{\sum_{i=1}^n W_{xi}^2}\right\}\mathbf{Q}^T \\
&= \mathbf{Q}\hat{\mathbf{S}}_{m+1}(\mathbf{x}_1, .., \mathbf{x}_n)\mathbf{Q}^T \ .
\end{aligned}$$

This concludes the proof.

Theorem 2 *Let $F_{\mathbf{\Sigma}}(\mathbf{x})$ be an elliptical distribution with location parameter μ and scatter parameter $\mathbf{\Sigma}$. The corresponding density is $f_{\mathbf{\Sigma}}(\mathbf{x}) = |\mathbf{\Sigma}|^{-1/2} g(\|(\mathbf{x}-\mu)\mathbf{\Sigma}^{-1/2}\|)$. Suppose the weight functions $u(t)$ and $w(t)$ satisfy the assumptions of Theorem 1. Then the estimating functional $\hat{\mathbf{S}}(F_{\mathbf{\Sigma}})$ is Fisher consistent, that is,*

$$\hat{\mathbf{S}}(F_{\mathbf{\Sigma}}) = \mathbf{\Sigma} \ .$$

Proof: Let $\lambda_1 < \lambda_2 < ... < \lambda_p$ be the eigenvalues of $\mathbf{\Sigma}$ and $\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_p$ be the corresponding eigenvectors. By Theorem 1, we can assume without loss of generality that

$$\{\mathbf{a}_0, \mathbf{a}_1, ..., \mathbf{a}_p\} = \{\mathbf{e}_0, \mathbf{e}_1, ..., \mathbf{e}_p\} \ ,$$

where $\mathbf{e}_i = (0_1, 0_2, .., 0_{i-1}, 1, 0_{i+1}, .., 0)^T$. Therefore,

$$\mathbf{\Sigma} = diag(\lambda_1, ..., \lambda_p) \ .$$

and the corresponding density function is

$$f_{\mathbf{\Sigma}}(\mathbf{x}) = \frac{1}{\sqrt{\prod_{i=1}^p \lambda_i}} g[\sum_{i=1}^p \left(\frac{\mathbf{x}_i^2}{\lambda_i}\right)] \ .$$

Asymptotically, convergence occurs in one step; however, the proof is done in two steps to allow the reader to follow the natural course of our argument. First we show that

the initial estimating functional $\hat{\mathbf{S}}_0(F_\Sigma)$ (using the weight function $u(\|\mathbf{x}\|)$) is diagonal. To see that $\hat{\mathbf{S}}_0(F_\Sigma)$ is diagonal, notice that by symmetry we obtain

$$E[x_i x_k u(\|\mathbf{x}\|)] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} x_i x_k f_\Sigma(\mathbf{x}) u(\sqrt{x_1^2 + \cdots x_p^2}) dx_1 \cdots dx_p = 0 \quad \text{for i}\neq\text{k} .$$

Recall that the weight function used after the initial step is based on the principal components. Hence the weights we use in the second step are based on the natural coordinates of $\mathbf{x}$, that is, the projections $\mathbf{e}_j\mathbf{x}$ (j=1,..p). Again we have to show that the cross-product vanishes to ensure that $\hat{\mathbf{S}}(F_\Sigma)$ remains diagonal. To see that this is the case, notice that by symmetry

$$E[x_i x_k \prod_{j=1}^{p} w\left(\frac{\mathbf{x}_j}{\hat{s}_j}\right)] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} x_i x_k f_\Sigma(\mathbf{x}) \prod_{j=1}^{p} w\left(\frac{\mathbf{x}_j}{\hat{s}_j}\right) dx_1 \cdots dx_p = 0 \quad \text{for i}\neq\text{k} .$$

Next we show that the diagonal elements of $\hat{\mathbf{S}}_1(F_\Sigma)$, $\hat{\lambda}_1, .., \hat{\lambda}_p$, satisfy

$$\hat{\lambda}_i = K(F_\mathbf{I})\lambda_i \quad \text{i=1,..,p} ,$$

where $K(F_\mathbf{I})$ is a known constant. Notice that in such a case $\hat{\mathbf{S}}(F_\Sigma) = \hat{\mathbf{S}}_1(F_\Sigma)$, that is, convergence occurs in one step. To show that $\hat{\lambda}_i = \lambda_i K(F_\mathbf{I})$, we have to evaluate

$$\hat{\lambda}_i = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} x_i^2 f_\Sigma(\mathbf{x}) \prod_{j=1}^{p} w\left(\frac{\mathbf{x}}{\hat{s}_j}\right) dx_1 \cdots dx_p , \tag{25}$$

By noticing that $\hat{s}_j$ (the median absolute deviation, MAD, was used) is Fisher consistent, i.e. $\hat{s}_j = \sqrt{\lambda_j}$, we can write (25) as

$$\hat{\lambda}_i = \frac{1}{\sqrt{\prod_{i=1}^{p} \lambda_i}} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} x_i^2 f_\mathbf{I}(\frac{x_1^2}{\lambda_1} + \cdots \frac{x_p^2}{\lambda_p}) \prod_{j=1}^{p} w\left(\frac{x_j}{\sqrt{\lambda_j}}\right) dx_1 \cdots dx_p .$$

Finally, we make a change of variables $\tilde{x}_i = x_i/\sqrt{\lambda_i}$ to obtain

$$\hat{\lambda}_i = \lambda_i \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \tilde{x}_i^2 f_\mathbf{I}(\tilde{x}_1^2 + \cdots \tilde{x}_p^2) \prod_{j=1}^{p} w(\tilde{x}_j) d\tilde{x}_1 \cdots d\tilde{x}_p = \lambda_i K(F_\mathbf{I}) .$$

This results in Fisher consistency for the direction of the principal components. To obtain Fisher consistency for the size of the principal components, the scale estimate $\hat{\lambda}_i$ must be divided by the known constant $K(F_\mathbf{I})$.

Theorem 3 *At any p-dimensional sample* $\{\mathbf{x}_1, .., \mathbf{x}_n\}$ *in general position, that is, there are no more than p points in any* $(p-1)$*-dimensional hyperplane, the breakdown point of the proposed estimate equals*

$$\epsilon_n^*(T, X) = ([\frac{n}{2}] - p)/n$$

which converges to 1/2 as $n \to \infty$.

Proof: Notice that the weight function w is positive for some constant c_1 greater than one and zero for some finite constant c_2 that is larger than c_1. Consider any sample $\{\mathbf{y}_1, .., \mathbf{y}_n\}$ obtained by replacing at most $[n/2]-(p+1)$ points in $\{\mathbf{x}_1, .., \mathbf{x}_n\}$ by arbitrary values. For any unit vector $\mathbf{a}$ in the p-dimensional space we have

$$s(\mathbf{a}^T\mathbf{y}_1, \mathbf{a}^T\mathbf{y}_2, ..., \mathbf{a}^T\mathbf{y}_n) \leq s(\|\mathbf{y}_1\|, \|\mathbf{y}_2\|, ..., \|\mathbf{y}_n\|) \leq \|\mathbf{y}\|_{([n/2]+p)} = d < \infty \ .$$

Notice that we can write any observation $\mathbf{y}_i$ as a linear combination of orthonormal unit vectors, $\{\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_p\}$, that span the p-dimensional space, that is,

$$\mathbf{y}_i = \sum_{j=1}^{p} \alpha_{ji}\mathbf{a}_j \ ,$$

where the α_{ij}'s are non-negative. Hence $\|\mathbf{y}_i\|^2 = \sum_{j=1}^{p} \alpha_{ji}^2$. Suppose that a large outlier, say $\mathbf{y}_k$, has a norm $\|\mathbf{y}_k\|^2 > p(c_2 d)^2$. Then at least one $\alpha_{jk} > c_2 d$, say α_{1k}. The weight w_{1k} for the observation $\mathbf{y}_k$ is

$$w_{1k} = w\left(\frac{\mathbf{a}_1^T\mathbf{y}_k}{s(\mathbf{a}_1)}\right) \leq w\left(\frac{\alpha_{1k}}{d}\right) \leq w\left(\frac{c_2 d}{d}\right) = w(c_2) = 0 \ .$$

We define the weight W_k assigned to the k^{th} observation as the product of the individual weights, w_{jk}, obtained for the projections $\mathbf{a}_j^T\mathbf{y}_k$ for i=1,..p associated with the k^{th} observation. Hence, $W_k = \prod_{j=1}^{p} w_{jk} = 0$; an observation that is large with respect to the "good" data is assigned a weight of zero. This will bound the largest scale away from ∞.

To show that the smallest scale does collapse to zero, consider the most pessimistic direction, say $\mathbf{a}_0$, in which all outliers have a null projection and p observations lie in a $(p-1)$-dimensional hyperplane. Hence, we will have a sequence of absolute ordered projections

$$\{0_{(1)}, 0_{(2)}, ..., 0_{([n/2]-(p+1))}, 0_{([n/2]-p)}, ..., 0_{([n/2]-1)}, k_{([n/2])}, ..., k_{([n/2]+p)}, ..., k_{(n)}\} \ ,$$

where $k_l > 0$ for l=([n/2]),..,(n). The corresponding scale of is the $([n/2]+p)^{th}$ absolute projection, that is,

$$s(\mathbf{a}_0) = k_{([n/2]+p)} \ .$$

Hence there will be at least $p+1$ "good" points of $\{\mathbf{y}_1, .., \mathbf{y}_n\}$ that will have standardized absolute projections less than or equal to one. From the definition of the weight function, these points will have nonzero weights. Since these points are in a general position, they determine a convex hull with nonzero volume and the corresponding matrix is nonsingular.

7 Tables - Simulation Results

The core model (contamination-free) used in the Monte Carlo study carried out to assess the finite sample performance of the estimates was the following

$$y_i = \beta \mathbf{X}_i + v_i$$

$$\mathbf{x}_i = \mathbf{X}_i + \mathbf{u}_i \ .$$

v_i and $\mathbf{u}_i$ are normal random errors. The errors are assumed to be uncorrelated with mean zero and variance σ_v^2 and $\sigma_{\mathbf{u}}^2\mathbf{I}$. We considered two situations, $p = 1$ (x is one-dimensional) and $p = 4$ ($\mathbf{x}$ is four-dimensional).

In the one-dimensional case we considered sample sizes of twenty and sixty observations. The parameter β was set to 0, 1, 5 and 10. The X_i's were distributed as N(0,1), v_i ~N(0,0.25) and u_i ~N(0,0.25) ($\sqrt{\lambda_1/\lambda_0} = 2$). The fraction of contamination ϵ was set to 0 (Gaussian model), 0.05, 0.10, 0.15, 0.20 and 0.25 with contamination-generating distributions N(3,0.25), N(5,0.25), N(10,0.25), N(15,0.25), N(20,0.25) and N(25,0.25). The outliers were generated in either y or x (more complicated contamination models were not considered due to the high cost of simulation and time constraint). Each sampling situation was replicated two hundred times. The following was used to assess the performance of the estimates

$$B_T = \sum_{i=1}^{200} B_i \ ,$$

where B_i is the direction bias of the estimate at the i^{th} replicate. The estimate of the direction $\hat{\mathbf{a}}$ is defined as

$$\hat{\mathbf{a}} = \frac{1}{\sqrt{1 + \hat{\beta}^T\hat{\beta}}}(1, -\hat{\beta})^T \ .$$

B_T was computed for each sampling situation.

Notice that B_T ranges from 0 (no direction bias has been induced) to 200 (maximum lower bound for the direction bias was reached at every replicate). The lower the value of B_T the better the bias behaviour of the estimate.

In the four-dimensional case we considered sample sizes of forty and seventy five. The parameter β is now a vector of parameters and in the Monte Carlo study it was taken to be one of $\mathbf{0}, \mathbf{1}, \mathbf{5}$ and $\mathbf{10}$. The vector $\mathbf{X}_i$ was distributed as N($\mathbf{0}$,1$\mathbf{I}$). The error structure was as above with $\mathbf{u}_i$ being N($\mathbf{0}$, 0.25$\mathbf{I}$). The fraction of contamination was as above with contamination-generating distributions N(5,0.25), N(10,0.25), N(15,0.25), N(20,0.25) and N(30,0.25). The outliers were generated in either y or x_1. Each sampling situation was replicated a hundred times. B_T was again used to compare the bias behaviour of the estimates.

The random number generator used in the Monte carlo study was adapted from an article by Schrage (1979). Eigenvalue-eigenvector decomposition was done using the QR algorithm.

The estimates considered in the Monte Carlo were the following:

1. Orthogonal regression.

2. Orthogonal regression analog of LMS based on Tukey's χ_c (c=1.548).

3. 95% efficient S-estimate using Tukey's χ_c (c=4.70).

4. S-estimate (BP=1/2) using Tukey's χ_c (c=1.548).

5. MM-estimate using Tukey's χ_c (c=4.70).

6. τ-estimate using Tukey's χ_c (c=6.08).

7. Proposed estimate with $u(r)$ equal to one for r less than 2.50. The weight function $w(r)$ is defined as follows

$$w(r) = \begin{cases} 1 & \text{if } |r| \leq 1 \\ \frac{1}{|r|} & \text{if } 1 < |r| \leq 2.5 \\ 0 & \text{otherwise} \end{cases}$$

To improve the performance of the proposed estimate, we considered a slightly different initial estimate $\hat{\mathbf{S}}_0$. To compute it, we have adopted an approach that is conceptually closely related to the Donoho-Stahel estimate. However, instead of looking at all one-dimensional projections that leave an observation $\mathbf{x}_i$ most exposed, we limit our search to a set of randomly generated orthonormal bases (including the canonical basis). The $\mathbf{x}_i$'s are projected onto the basis vectors. We define a projection index γ_i as

$$\gamma_i = \sup_{\mathbf{v} \in \ basis} \frac{|\mathbf{v}^T\mathbf{x}_i - \text{median}_j(\mathbf{v}^T\mathbf{x}_j)|}{\text{median}_k|\mathbf{v}^T\mathbf{x}_k - \text{median}_j(\mathbf{v}^T\mathbf{x}_j)|} .$$

The weights $w(\gamma_i)$ are assigned to each observation, $\mathbf{x}_i$, according to the weight function $w(\bullet)$ defined as above. The resulting multivariate estimates of location and scatter are

$$\hat{\mathbf{t}}_0 = \frac{\sum_{i=1}^n w_i\mathbf{x}_i}{\sum_{i=1}^n w_i} ,$$

$$\hat{\mathbf{S}}_0 = \frac{\sum_{i=1}^n (\mathbf{x}_i - \hat{\mathbf{t}}_0)(\mathbf{x}_i - \hat{\mathbf{t}}_0)^T w_i^2}{\sum_{i=1}^n w_i^2} .$$

This estimate is relatively easy to compute and has a breakdown point of 1/2.

8. One-step reweighted estimate with zero-one type of weights based on the proposed estimate. An observation is assigned a weight of zero when the corresponding proposed method weight is zero, otherwise the weight is equal to one.

The following tables summarize simulation results for the seven robust estimates considered, ORLM, 95% efficient S-estimate, S-estimate with BP of 1/2, MM-estimate, τ-estimate, one-step reweighted OR with weights based on the proposed method (WOR) and the proposed method estimate (MPP) together with classical orthogonal regression for comparison. We only include tables that show clearly the merits of using the robust estimates in place of classical estimates when the data are contaminated by outlying observations.

Table 2: Contamination in Y, N(3,0.25), β=0, n=20, m=200.

ϵ	OR	ORLM	S(95%)	S(BP=0.5)	MM	τ	WOR	MPP
0.00	1.729	6.849	2.893	7.065	3.058	3.027	2.160	7.997
0.05	16.128	7.343	3.850	7.368	3.568	5.844	5.125	5.780
0.10	56.041	8.518	29.662	9.530	10.789	12.254	13.938	8.836
0.15	93.977	18.839	86.486	17.807	26.621	33.595	24.401	9.535
0.20	113.490	20.039	119.962	18.549	58.561	55.115	39.843	15.937
0.25	137.259	31.845	146.006	35.817	110.012	99.240	48.995	18.023

Table 3: Contamination in Y, N(3,0.25), β=0, n=60, m=200.

ϵ	OR	ORLM	S(95%)	S(BP=0.5)	MM	τ	WOR	MPP
0.00	0.520	1.985	0.540	1.960	0.546	0.571	0.520	1.418
0.05	5.489	2.142	0.916	1.757	0.863	1.203	1.130	1.486
0.10	43.451	1.916	7.440	2.137	1.191	2.462	1.588	1.315
0.15	105.231	3.451	90.313	2.659	6.754	11.586	4.707	1.791
0.20	138.850	2.585	145.655	2.960	65.327	51.431	10.920	1.567
0.25	157.160	5.416	167.097	5.958	155.097	131.435	23.000	3.232

Table 4: Contamination in Y, N(5,0.25), β=0, n=20, m=200.

ϵ	OR	ORLM	S(95%)	S(BP=0.5)	MM	τ	WOR	MPP
0.00	1.729	6.849	2.893	7.065	3.058	3.027	2.160	7.997
0.05	99.114	7.070	6.737	8.468	5.548	7.403	5.488	9.534
0.10	156.263	11.718	63.213	11.867	9.269	10.800	7.134	9.049
0.15	171.012	15.320	178.990	17.221	14.268	19.381	9.104	10.145
0.20	175.600	17.266	182.988	20.762	32.522	35.111	5.428	6.867
0.25	176.412	29.579	182.560	37.020	94.562	92.341	10.439	11.378

Table 5: Contamination in Y, N(5,0.25), β=0, n=60, m=200.

ϵ	OR	ORLM	S(95%)	S(BP=0.5)	MM	τ	WOR	MPP
0.00	0.520	1.985	0.540	1.960	0.546	0.571	0.520	1.418
0.05	106.260	1.879	0.500	1.838	0.523	0.843	0.483	1.291
0.10	174.419	1.616	49.046	1.672	0.673	1.180	0.641	1.384
0.15	181.852	1.570	187.940	1.817	0.845	2.285	0.729	1.273
0.20	184.547	1.356	189.764	2.831	11.795	16.112	0.778	1.501
0.25	186.338	3.379	192.827	4.844	81.398	79.776	1.554	1.786

Table 6: Contamination in X, N(20,0.25), β=5, n=20, m=200.

ϵ	OR	ORLM	S(95%)	S(BP=0.5)	MM	τ	WOR	MPP
0.00	0.0648	0.2405	0.0669	0.2208	0.0724	0.1150	0.0648	0.2764
0.05	35.9934	0.2109	0.0837	0.2317	0.0985	0.1550	0.0718	0.2356
0.10	105.4132	0.2694	0.0753	0.2834	0.0821	0.1848	0.0967	0.2536
0.15	130.9318	0.1580	111.8115	0.2381	0.0934	0.1866	0.0694	0.1480
0.20	143.1987	0.1580	131.8617	0.2966	0.1183	0.2603	0.0840	0.3040
0.25	144.1752	0.1986	141.6511	0.4552	0.1949	0.4162	0.1064	0.2136

Table 7: Contamination in X, N(20,0.25), β=5, n=60, m=200.

ϵ	OR	ORLM	S(95%)	S(BP=0.5)	MM	τ	WOR	MPP
0.00	0.0153	0.0653	0.0164	0.0627	0.0168	0.0169	0.0153	0.0595
0.05	18.8693	0.0610	0.0193	0.0565	0.0201	0.0284	0.0185	0.0608
0.10	105.5620	0.0607	0.0224	0.0615	0.0238	0.0415	0.0202	0.0587
0.15	131.1035	0.0523	131.7244	0.0573	0.0247	0.0449	0.0213	0.0411
0.20	144.2054	0.0445	147.7639	0.1046	0.0368	0.0873	0.0223	0.0489
0.25	143.9072	0.0528	150.3505	0.1888	0.0682	0.1701	0.0261	0.0409

Table 8: Contamination in Y (dim=5), N(10,0.25), β=**0**, n=40, m=100.

ϵ	OR	ORLM	S(95%)	S(BP=0.5)	MM	τ	WOR	MPP
0.00	1.794	8.972	1.895	8.183	2.181	2.425	1.800	7.076
0.05	92.399	10.130	4.316	9.940	4.713	7.021	3.280	7.191
0.10	94.399	16.350	16.796	17.030	13.129	15.472	6.825	11.372
0.15	96.311	18.532	97.659	22.511	16.425	25.028	1.918	8.824
0.20	96.393	37.183	97.190	45.574	45.832	51.641	6.040	11.085
0.25	96.985	56.955	97.529	73.062	89.264	91.446	6.806	12.250

Table 9: Contamination in X1 (dim=5), N(10,0.25), β=**5**, n=40, m=100.

ϵ	OR	ORLM	S(95%)	S(BP=0.5)	MM	τ	WOR	MPP
0.00	1.393	5.328	1.422	4.944	1.636	2.461	1.446	9.401
0.05	12.663	9.130	5.544	8.795	4.724	7.204	3.167	6.726
0.10	15.156	11.331	16.162	11.099	8.948	11.027	3.144	7.648
0.15	15.607	15.550	15.781	15.953	14.629	15.708	4.772	9.468
0.20	15.591	17.365	16.977	19.181	17.749	18.091	5.070	6.022
0.25	15.278	19.201	16.556	20.420	16.711	17.622	6.016	8.205

An extensive simulation was done to assess the breakdown properties of different robust orthogonal regression estimates. Given that the upper bound for the maximum direction bias is one, we assume here that an estimate has broken down if the 95^{th} quantile of the direction bias is larger than 0.90. Using this criterion, the empirical BP's for the direction are

Table 10: Empirical BP's of the estimates.

	OR	ORLM	S(95%)	S(BP=0.5)	MM	τ	WOR	MPP
ϵ^*	0.00	> 0.25	≈ 0.10	> 0.25	≈ 0.20	≈ 0.20	> 0.25	> 0.25

The empirical BP's are slightly higher than what is expected theoretically (note that the ratio $\sqrt{\lambda_1/\lambda_0}$ used in the simulations was two). This is probably because the simulations do not reflect the most damaging type of contamination.

The Monte Carlo study only confirms what has been theorized. The efficient S-estimate performs well at the Gaussian model and for small fractions of contamination. However, it breaks down early. The maximal BP S-estimate does not break down until about 25% of contamination but its performance at the Gaussian model is unsatisfactory. The same applies to the ORLM. The MM- and τ- estimates perform well at the Gaussian model and, for larger fractions of contamination, attain nearly the same level of bias robustness as the maximal BP S-estimate.

The proposed estimate appears to have better bias characteristics than the robust S-estimate throughout the ϵ-range; however, it is quite inefficient at the Gaussian model. To improve efficiency while retaining a high BP, we considered a one-step reweighted estimate with weights based on the proposed estimate. It is evident from the tables above that this approach yields superior results. The one-step estimate is more efficient than the MM- and τ-estimates at the Gaussian model and its bias performance over the ϵ-range is exemplary. Although the proposed estimate and the one-step estimate give

good results, further study of their properties is required.

8 Applications of PCA

8.1 Orthogonal Regression

Orthogonal regression (OR) is the maximum likelihood procedure at the Gaussian error-in-variables (EV) model. In classical regression, the response variable consists of a deterministic part (assumed known) $\beta^T\mathbf{X}$ and a random part e where e is assumed to be normally distributed with mean zero and covariance equal to some multiple of the identity matrix. The $\mathbf{X}_i$'s can also be random but they are assumed to be observed without error. In OR, the $\mathbf{X}_i$'s are either random independent identically distributed vectors (*structural* EV model) or they are non-random but unknown (*functional* EV model).

Let

$$\begin{aligned} y_i &= \alpha + \beta^T\mathbf{X}_i + v_i \\ \mathbf{x}_i &= \mathbf{X}_i + \mathbf{u}_i \ , \end{aligned} \tag{26}$$

where α is the intercept, β is the vector of regression parameters and $\mathbf{u}$ and v are errors with zero mean and some variance, possibly different. $\mathbf{u}$ and v are uncorrelated. Then the OR estimates are the solution of the following minimization problem

$$\min_{\alpha;\beta\in R^p} \sum_{i=1}^{n} \chi\left(\frac{y_i - \alpha - \beta^T\mathbf{x}_i}{\sqrt{1+\beta^T\beta}}\right) \ .$$

In the classical setup $\chi(x) = x^2$. To make the OR method robust, the function χ is chosen to reduce the influence of outlying observations.

The above orthogonal regression problem can be restated as follows. Let $\mathbf{x}i = \mathbf{X}_i + \mathbf{u}_i$ be a set of random vectors satisfying the condition $\mathbf{a}_0\mathbf{X}_i = b_0$, where $\mathbf{a}_0^T\mathbf{a}_0 = 1$ and b_0 is some constant. Then the vector $\hat{\mathbf{a}}$ and the number $\hat{b}$ are found by minimizing

$$\frac{1}{n}\sum_{i=1}^{n}(\mathbf{a}^T\mathbf{x}_i - b)^2 \ .$$

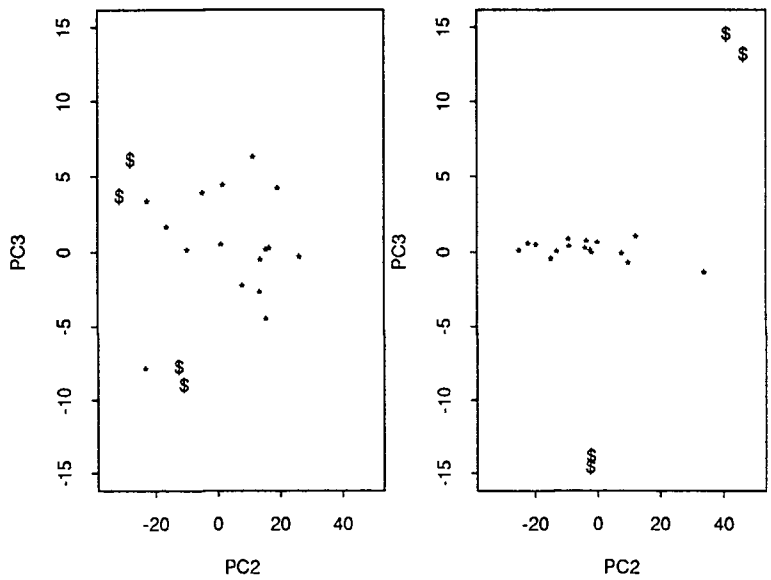

Figure 5: Plot of Classical versus Robust PCA

$(\mathbf{a}^T\mathbf{x}_i - b)^2$ is the square of the orthogonal distance from $\mathbf{x}_i$ to the hyperplane $H(\mathbf{a}, b) = \{\mathbf{x} : \mathbf{a}^T\mathbf{x} = b\}$ (Zamar, 1989). It can be shown that $\hat{\mathbf{a}}$ is the principal component of the sample covariance matrix corresponding to the smallest eigenvalue. Thus orthogonal regression reduces to finding the the direction of the smallest principal component (PC). To obtain robust orthogonal regression estimates we find the direction of the smallest robust PC.

8.2 Outlier Detection using PCA

Robust PCA can also be used for detecting outliers in higher dimensional spaces. This can be done by identifying observations that give rise to unusually large principal components. To emphasize why classical PCA is not well suited to this purpose, let us consider the artificial data set in Zamar (1989). It consists of twenty three-dimensional vectors four of which have been made outlying, observations 1, 2, 19, 20 (marked with a \$ sign). We have plotted all pairwise combinations of the PC's for the classical and

robust PCA. In the classical PCA, we do not have a clear indication that some of the observations may be outliers. It is possible that an experienced data analyst could identify the outliers from the PC1-PC3 plot (not shown). However, his conclusions would be subjective, based on his or her experience.

To make outlier detection objective, we use robust methods. In this example we have used the proposed estimator. From the robust PC2-PC3 plot it appears that observations 1, 2, 19 and 20 are unusually large in the third PC. This indicates that they may be possible candidates for outliers as these observations do not conform with the structure of the rest of the data. These points have been purposely designed to be outliers with the intention of upsetting the classical estimates. The classical PCA failed to identify these observations because they inflated the scale of the third PC so that outliers would not be detectable; in fact, the classical scale was almost ten times larger than its robust counterpart. On the other hand, the proposed estimator clearly distinguished between "good" data and "bad" data.

The example illustrates the dangers of relying on classical methods for outlier detection and how robust methods can provide a better picture of the situation by identifying aberrant observations and remaining stable in their presence.

9 References

BARNETT, V. (1976). The ordering of multivariate data. *Journal of the Royal Statistical Society A* **138** 318-344.

BEBBINGTON, A.C. (1978). A method of bivariate trimming for estimation of the correlation coefficient. *Applied Statistics* **38** 221-226.

BOENTE, G. (1987). Asymptotic theory for robust principal components. *Journal of Multivariate Analysis* **21** 67-78.

BUTLER, R.W. and JUHN, M. (1987). Asymptotic for trimmed multivariate data. Revised Preprint November 1987, University of Michigan and University of Florida.

CHEN, Z. and LI, G. (1985). Projection-pursuit approach to robust dispersion matrices and principal components: Primary theory and Monte Carlo. *JASA* **80** 759-766.

DAVIES, P.L. (1989). Improving S-estimators by means of k-step M-estimators. Technical report, GHS - Essen.

DEVLIN, S.J. et. al. (1975). Robust estimation and outlier detection with correlation coefficients. *Biometrika* **62** 531-545

DONOHO, D.L. (1982). Breakdown properties of multivariate location estimators. Ph.D. qualifying paper. Harvard University.

DONOHO, D.L. and HUBER, P.J. (1983). The notion of breakdown point. In *A Festschrift for Erich L. Lehmann* (P.J.Bickel, K.A.Doksum, J.L.Hodges Jr., eds.) 157-184. Wadsworth, Belmont, California.

GNANADESIKAN, R. and KETTENRING. J.R. (1972). Robust estimates, residuals and outlier detection with multiresponse data. *Biometrics* **28** 81-124.

HAMPEL, F.R., RONCHETTI, E.M., ROUSSEEUW, P.J. and STAHEL, W.A. (1986). *Robust statistics : The approach based on influence functions.* Wiley, New York.

HELBLING, J.M. (1983). Ellipsoïdes minimaux de couverture en statistique multivariée, Ph.D. thesis, École Polytechnique Fédérale de Lausanne, Switzerland.

HUBER, P.J. (1964). Robust estimation of a location paramater. *Annals of Statistics* **35** 73-101.

HUBER, P.J. (1981). *Robust statistics*. Wiley, New York.

JOHNSON, R.A. and WICHERN, D.W. (1988). *Applied multivariate statistical analysis*. Prentice-Hall, 2nd ed.

JOLLIFFE, I.T. (1986). *Principal component analysis*. Springer -Verlag, New York.

LOPUHAÄ, R. (1990). *Estimation of location and covariance with high breakdown point*. Ph.D. thesis. Delft University of Technology, Netherlands.

MARONNA, R.A. (1976). Robust M-estimates of multivariate location and scatter. *Annals of Statistics* **4** 51-67.

MARTIN, R.D., YOHAI, V.J. and ZAMAR, R.H. (1989). Min-max bias robust regression. *Annals of Statistics* **17** 1608-1630.

MARTIN, R.D. and ZAMAR, R.H. (1989). Bias robust estimation of scale when location is unknown. Technical report.

ROUSSEEUW, P.J. and LEROY, A.M. (1987). *Robust regression and outlier detection*. Wiley, New York.

SCHRAGE, L. (1979). A more portable Fortran random number generator. *ACM Trans. Math. Softw.* **5** 132-138.

STAHEL, W.A. (1981). Robust Estimation: Infinitesimal Optimality and Covariance Matrix Estimators. Ph.D. thesis (in German), ETH, Zurich.

TITTERINGTON, D.M. (1978). Estimation of correlation coefficients by ellipsoidal trimming. *Applied Statistics* **27** 227-234.

TUKEY, J.W. (1974). T6: Order statistics, in mimeographed notes for Statistics 411,

Princeton University.

YOHAI, V.J. and ZAMAR, R. (1988). High breakdown-point of estimates of regression by means of the minimization of an efficient scale. *JASA* **83** 406-413.

YOHAI, V.J. and ZAMAR, R.H. (1990). Discussion of Paper No. 90 SM 493-7 PWRS.

ZAMAR, R.H. (1989). Robust estimation in the errors-in-variables model. *Biometrika* **76** 149-160.