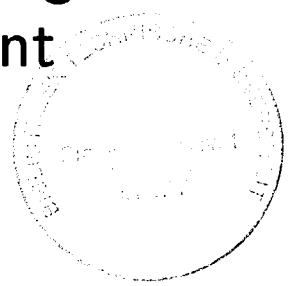
Estimation of Location and Covariance with high Breakdown Point

Estimation of Location and Covariance with high Breakdown Point



PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Technische Universiteit Delft, op gezag van de Rector Magnificus, Prof. Drs. P. A. Schenck, in het openbaar te verdedigen ten overstaan van een commissie aangewezen door het College van Dekanen op donderdag 26 april 1990 te 16.00 uur

door

Hendrik Paul Lopuhaä

geboren te Amsterdam,

Doctorandus in de Wiskunde.

Dit proefschrift is goedgekeurd door de promotor

Prof.Dr. P.J. Rousseeuw

Voor Irma

Acknowledgments

Peter Rousseeuw introduced me to the subject of this thesis. I thank him for his suggestions and remarks, and for the several discussions about robustness. I also like to express my gratitude to him and his wife Lieve for their hospitality during my visits to Fribourg, Switserland and to Edegem, Belgium. The research for this thesis has been performed at the Delft University of Technology. Gerard Hooghiemstra, Rudolf Grübel, Piet Groeneboom and Bert van Zomeren were kind enough to spend some of their time discussing probabilistic, statistical and TEXnical problems. It is a pleasure to thank them and all other members of the "vakgroep SSOR" in Delft, not the least my roommates Marco Martens, Ronald Meester and Jacques Resing, for a most enjoyable four years. I thank the Delft University of Technology and Shell Nederland for their financial support which made it possible to visit the robustness conference held at the Institute of Mathematics and Applications at the University of Minnesota in Minneapolis, where I have had many stimulating discussions with David Tyler and several other people. Finally, I thank the Dutch Organisation for Scientific Research (NWO) who financed my stay in Delft, as well as the several visits to Fribourg and Edegem.

Copyright

CONTENTS

INTRODUCTION

COPIES OF THE FOUR PAPERS

[1] Tentatively accepted by the Annals of Statistics

[2] Appeared in the Annals of Statistics 1989, Vol. 17, No. 4, 1662-1683

1. Summary

Consider a collection of n points in a p-dimensional Euclidean space. We are interested in estimating a point around which the collection is located and how the collection is scattered around this point. The sample mean and the sample covariance are no doubt the most widely known estimators to do this. These estimators are in some sense the most accurate ones, but they are also notorious for their sensitivity to outliers, i.e. points whose position is 'unusually' far from the majority of the collection. A single aberrant point has a tremendous influence on the value of these estimators. Alternatively, one may use robust estimators for multivariate location and scatter. Such estimators will be less sensitive to outliers, but on the other hand they will be less accurate than the sample mean and the sample covariance in case no outliers are present.

Whether outliers are present is generally unknown and it is difficult, if not impossible, to determine this for collections in higher dimensions. When the collection is just a set of real numbers, a simple plot will be enough to reveal possible outliers. This may still be possible in two dimensions, and possibly even in three dimensions using some sophisticated statistical software package. However, plots or graphs will no longer be of help when p is larger than three. In these situations the need arises for robust multivariate estimators of location and scatter that automatically pay attention to possible outliers. Such methods are already well known in univariate situations; the trimmed mean and the median are only a few examples.

Covariance matrices and the associated ellipsoids are often used for describing the overall shape of distributions of points in a p-dimensional Euclidean space. Important examples occur in principal component and factor analysis, and in discrimant analysis. Because of their high sensitivity to outliers, the sample mean and the sample covariance are not particularly well suited for this purpose. This drawback may be overcome using outlier resistant alternatives.

Another important application of robust estimators of location and scatter, especially in high dimensional situations, is to use them as a diagnostic tool to detect possible outliers. When the outliers are detected, one may assign a smaller weight to them, or even delete them and use classical methods on the remaining points.

In this thesis we will investigate the robustness and the asymptotic properties of multivariate estimators of location and scatter. Robustness of the estimators

will be measured in two different ways. The global sensitivity of an estimator is measured by means of its breakdown point, which is roughly the smallest fraction of outliers in the collection that can take the estimator over all bounds. It describes the global behaviour of an estimator under large pertubations. The local robustness is measured by the influence function, which describes the alteration of the estimator under infinitesimal pertubations at some point $\mathbf{x}$. A robust estimator will typically have a high breakdown point and a bounded influence function. To investigate the asymptotic properties, the collection is assumed to be a sample generated by a distribution P on $\mathbf{R}^p$, and the behaviour of the estimators is studied as the sample size n tends to infinity. Of main interest will be the rate of convergence, the limiting distribution and the asymptotic efficiency. As a special case we will consider the usual location-scale model, where the distribution P is assumed to be elliptically contoured with an unkown location and scale parameter.

We will mainly be interested in estimators that commute with affine transformations of the points. This means that the location estimator is translation and scale equivariant, and that the covariance estimator is translation invariant and scale equivariant. This seems a natural property for multivariate estimators, especially if one wants to estimate the location and scale parameters of an elliptically contoured distribution. To construct univariate estimators of location and scale with this property which can also resist large amounts of outliers is no problem. The univariate sample median is an example of this. However, in the multivariate setting the situation turns out to be completely different. Affine equivariant multivariate M-estimators for location and scatter, probably the most well known robust alternatives to the sample mean and the sample covariance, can only resist at most a fraction $1/(p+1)$ of outliers, which means that these estimators become more sensitive when p increases. This result was rather disappointing, since these estimators were shown to be optimally robust from different other points of view. The poor breakdown properties of M-estimators, as well as of several other affine equivariant estimators that were thought to be robust, has tempted people to think that it was impossible to combine affine equivariance with a high breakdown point. Nevertheless, the first affine equivariant estimators for multivariate location and scatter with a high breakdown point were constructed in the early 1980's.

The earliest affine equivariant estimators for location and scatter with a high breakdown point exibit relatively poor asymptotic properties. The rate of convergence is generally slower than the usual $\sqrt{n}$ rate, the limiting distribution may not be normal, or the asymptotic efficiency is disappointingly low. The main purpose of this thesis is to construct affine equivariant estimators of multivariate location and scatter that combine good global and local robustness, i.e. a high breakdown point and a bounded influence function, together with good asymptotic properties, i.e. $\sqrt{n}$ rate of convergence towards a normal distribution with a reasonable efficiency relative to the sample mean and the sample covariance.

The findings have been written down in four different papers. These papers have been reproduced at the end of the thesis and are preceded by an introduction. After introducing some notation and basic definitions in Chapter 2, we discuss

the concepts of breakdown point and influence function in Chapter 3. Chapter 4 briefly mentions two other approaches to robustness. The first one is the minimax variance approach. It corresponds with the class of M-estimators, a generalization of maximum likelihood estimators. The second one is the minimax bias approach, which has only recently been investigated. In Chapter 5 we discuss some of the first affine equivariant estimators of multivariate location and scatter with a high breakdown point. These chapters of the introduction are meant to put the four papers in perspective. The remaining three chapters discuss three different new proposals for estimating multivariate location and scatter. This is done to a much larger extent than in the actual papers, since in that case space was limited because of the constraints set by the journals where the papers were submitted to.

The first paper, together with Peter Rousseeuw, studies the breakdown point of several estimators of multivariate location and scatter and illustrates the role of various equivariance properties. Furthermore, a striking relation between the breakdown point of univariate location estimators and a measure of large deviations is extended to multivariate location estimators. Most of the contents of this paper are briefly summarized in Chapter 3 and Section 7.2 of the introduction.

The second paper studies multivariate S-estimators, which are smoothed versions of Rousseeuw's minimum volume ellipsoid estimator. They can be seen as a first step towards combining good robustness with good asymptotic properties. Chapter 6 summarizes the results concerning these estimators.

The third paper investigates an affinely scaled location M-estimator. This estimator is basically a location M-estimator based on the sample that arises after scaling with an affine equivariant covariance estimator with a high breakdown point. The resulting location estimator is affine equivariant, has a high breakdown point and a bounded influence function, and converges at rate $\sqrt{n}$ to a normal distribution, with good efficiency. This estimator is one example of combining high breakdown estimators with multivariate M-estimators in a suitable way. This example and a few others are briefly discussed in Chapter 7.

The last paper studies the class of multivariate τ-estimators, which is a further extension of the class of multivariate S-estimators. The resulting estimators of location and scatter are affine equivariant, and combine a high breakdown point and a bounded influence function with a $\sqrt{n}$ rate of convergence to a normal distribution and good efficiency relative to the sample mean and sample covariance. Chapter 8 discusses this class of estimators.

2. Notation and Basic Definitions

2.1 Notation

We first fix the notation that will be used and give the definitions of a few basic concepts that will be needed. Let $\mathbf{R}^p$ denote the p-dimensional Euclidean space. To distinguish vectors and matrices from ordinary real numbers in the case that p is larger than 2 we use boldface lowercase letters, such as $\mathbf{v} = (v_1 \cdots v_p)^T$ or $\boldsymbol{\mu} = (\mu_1 \cdots \mu_p)^T$ to denote elements of $\mathbf{R}^p$, where T stands for transpose, and we use boldface uppercase letters, such as $\mathbf{A} = (a_{ij})_{i,j=1}^p$ or $\boldsymbol{\Sigma} = (\sigma_{ij})_{i,j=1}^p$, to denote $p \times p$-matrices.

For real numbers y we have to make a distinction between $\lfloor y \rfloor$, defined as the largest integer less than or equal to y, and $\lceil y \rceil$, defined as the smallest integer greater than or equal to y. When y is not integer valued, then $\lceil y \rceil = \lfloor y \rfloor + 1$, however, $\lceil y \rceil = \lfloor y \rfloor = y$ when $y \in \mathbf{N}$. Both quantities will be needed to describe the different breakdown points later on.

Denote by $\|\cdot\|$ the Euclidean distance, which will either be between two vectors in $\mathbf{R}^p$ or between two matrices $\mathbf{A}$ and $\mathbf{B}$. The determinant of a $p \times p$-matrix $\mathbf{A}$ is denoted by $|\mathbf{A}|$ and the eigenvalues of such $\mathbf{A}$ are denoted by $\lambda_p(\mathbf{A}) \leq \cdots \leq \lambda_1(\mathbf{A})$. We will mainly be concerned with $p \times p$-matrices that are positive definite and symmetric and we will denote by PDS(p) the class of all such matrices. Recall that every element $\mathbf{A}$ of this class has a root $\mathbf{R}$, i.e. a matrix $\mathbf{R}$ such that $\mathbf{A} = \mathbf{R}\mathbf{R}^T$.

An ellipsoid in $\mathbf{R}^p$ with center $\mathbf{m}$ and covariance structure $\mathbf{M}$ is denoted by

$$E(\mathbf{m}, \mathbf{M}, r) = \{\mathbf{x} \in \mathbf{R}^p : (\mathbf{x} - \mathbf{m})^T \mathbf{M}^{-1} (\mathbf{x} - \mathbf{m}) \leq r^2\} \tag{2.1}$$

where r is a positive real number which together with $\mathbf{M}$ determines the magnitude of the ellipsoid. When $\mathbf{M} = \mathbf{I}$, (2.1) reduces to the ball with center $\mathbf{m}$ and radius r. Recall that the volume of an ellipsoid $E(\mathbf{m}, \mathbf{M}, r)$ is a multiple α_p of $\sqrt{|\mathbf{M}|}$, where $\alpha_p = (\pi r^2)^{p/2} / \Gamma(\frac{p}{2} + 1)$, and that the p different axes of $E(\mathbf{m}, \mathbf{M}, r)$ have lengths $2r\sqrt{\lambda_j(\mathbf{M})}$ for $j = 1, 2, \ldots, p$.

The *Mahalanobis distance* between a vector $\mathbf{v}$ and a vector $\mathbf{m}$ with respect to a

matrix $\mathbf{M}$ in PDS(p) is defined as

$$d(\mathbf{v},\mathbf{m},\mathbf{M}) = \sqrt{(\mathbf{v}-\mathbf{m})^T\mathbf{M}^{-1}(\mathbf{v}-\mathbf{m})}. \tag{2.2}$$

It may be interpreted either as the smallest factor that is needed to inflate the ellipsoid $E(\mathbf{m},\mathbf{M},1)$ such that it will cover $\mathbf{v}$, or as the largest factor allowed to deflate $E(\mathbf{m},\mathbf{M},1)$ such that it still covers $\mathbf{v}$. When $\mathbf{M}=\mathbf{I}$, (2.2) is simply the ordinairy Euclidean distance between $\mathbf{v}$ and $\mathbf{m}$.

A collection of n points in $\mathbf{R}^p$ will be typically denoted by $\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n$, where $\mathbf{x}_i = (x_{i1}\cdots x_{ip})^T$ for $i=1,2,\ldots,n$. Such a collection may contain multiple copies of one point. We say that a collection with at least $p+1$ points is in *general position* if no $p+1$ points are contained in some lower dimensional hyperplane. In $\mathbf{R}^2$ this simply means that no three points may be situated on the same line. Obviously, if a collection is in general position, all points must be different. When a collection $\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n$ is assumed to be a random sample from some probability distribution on $\mathbf{R}^p$, we will write $X_1,X_2,\ldots,X_n$ instead, where $X_i = (X_{i1}\cdots X_{ip})^T$ is a random vector for $i=1,2,\ldots,n$. Note that if the sample distribution is absolutely continuous with respect to the Lebesgue measure, the sample $X_1,X_2,\ldots,X_n$ is in general position with probability one.

Distributions on $\mathbf{R}^p$ will be denoted either by $P,Q,\ldots$, or by means of their corresponding distribution functions $F,G,\ldots$. When the distributions have densities, we will denote these by $f,g,\ldots$. The empirical distribution on $\mathbf{R}^p$, that puts mass $1/n$ at each $\mathbf{x}_i$ for $i=1,2,\ldots,n$, will be denoted by P_n. By $\delta_{\mathbf{x}}$ is meant the Dirac measure on $\mathbf{R}^p$, which has all its probability mass concentrated in the point $\mathbf{x}$. Expectation and variance with respect to a distribution P will denoted by $\mathbf{E}_P$ and $\mathbf{V}_P$ respectively, or simply by $\mathbf{E}$ and $\mathbf{V}$ if it does not cause any confusion.

A *location estimator* based on a collection of n points is a vector valued function of $\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n$, which is typically denoted by $\mathbf{t}_n = \mathbf{t}(\mathbf{x}_1,\ldots,\mathbf{x}_n)$. Similarly, a *covariance estimator*, or estimator of scatter, is a function of $\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n$, which takes on values in PDS(p), and which is typically denoted by $\mathbf{C}_n = \mathbf{C}(\mathbf{x}_1,\ldots,\mathbf{x}_n)$. When $p=1$, we will use a more standard notation and write σ_n^2 instead of c_n.

We will only be interested in estimators that are *permutation invariant*, i.e.

$$\begin{aligned}\mathbf{t}(\mathbf{x}_{\pi(1)},\ldots,\mathbf{x}_{\pi(n)}) &= \mathbf{t}(\mathbf{x}_1,\ldots,\mathbf{x}_n)\\ \mathbf{C}(\mathbf{x}_{\pi(1)},\ldots,\mathbf{x}_{\pi(n)}) &= \mathbf{C}(\mathbf{x}_1,\ldots,\mathbf{x}_n)\end{aligned} \tag{2.3}$$

for every permutation $\pi(1),\pi(2),\ldots,\pi(n)$ of $1,2,\ldots,n$. It means that the estimators will be independent of the numbering of the points $\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n$. Hence, we can also write the estimators as a function of a whole collection $\mathbf{X} = (\mathbf{x}_1,\ldots,\mathbf{x}_n)$, which will be more convenient in some cases. We will then write $\mathbf{t}_n(\mathbf{X})$ or $\mathbf{C}_n(\mathbf{X})$. Equivalently, we will sometimes write the estimators as a function of the empirical distribution, i.e. $\mathbf{t}(P_n)$ or $\mathbf{C}(P_n)$.

2.2 Equivariance

A location estimator $\mathbf{t}_n$ is called *translation equivariant*, if

$$\mathbf{t}(\mathbf{x}_1 + \mathbf{v}, \ldots, \mathbf{x}_n + \mathbf{v}) = \mathbf{t}(\mathbf{x}_1, \ldots, \mathbf{x}_n) + \mathbf{v} \tag{2.4}$$

for every vector $\mathbf{v}$ in $\mathbf{R}^p$. This means that if we translate a collection $\mathbf{X}$ over a vector $\mathbf{v}$, the location estimate based on the translated collection is equal to the translated location estimate based on $\mathbf{X}$. Intuitively, this is how one expects a reasonable location estimator to behave. Most of the location estimators that we will discuss will satisfy (2.4).

We say that a location estimator $\mathbf{t}_n$ is *affine equivariant*, if

$$\mathbf{t}(\mathbf{A}\mathbf{x}_1 + \mathbf{v}, \ldots, \mathbf{A}\mathbf{x}_n + \mathbf{v}) = \mathbf{A}\mathbf{t}(\mathbf{x}_1, \ldots, \mathbf{x}_n) + \mathbf{v} \tag{2.5}$$

for every nonsingular matrix $\mathbf{A}$ and every vector $\mathbf{v}$ in $\mathbf{R}^p$. This means that a location estimator commutes with affine transformations in the same way as the expectation operator does : $\mathbf{E}(\mathbf{A}X + \mathbf{v}) = \mathbf{A}\mathbf{E}(X) + \mathbf{v}$. Affine equivariance may seem a natural condition for location estimators. However, several multivariate estimators fail to satisfy this condition. Nevertheless, affine equivariance remains a desirable property, and we will mainly be interested in multivariate location estimators that satisfy it.

One can weaken affine equivariance by requiring (2.5) only for orthogonal matrices $\mathbf{A}$ and vectors $\mathbf{v}$. Property (2.5) is then referred to as *orthogonal equivariance*. Sometimes this is called *rigid motion equivariance*, because in this case the group of transformations $\mathbf{x} \mapsto \mathbf{A}\mathbf{x} + \mathbf{v}$ are the so-called rigid motions, such as translations, rotations and reflexions.

We say that a covariance estimator $\mathbf{C}_n$ is affine equivariant if

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

for every nonsingular matrix $\mathbf{A}$ and every vector $\mathbf{v}$ in $\mathbf{R}^p$. It means that the covariance estimator is invariant under translations, and that it commutes with multiplications in the same way as the variance operator does : $\mathbf{V}(\mathbf{A}X + \mathbf{v}) = \mathbf{A}\mathbf{V}(X)\mathbf{A}^T$. The covariance estimators that we will discuss, will all satisfy this property.

When we study asymptotic properties of $\mathbf{t}(X_1, \ldots, X_n)$ and $\mathbf{C}(X_1, \ldots, X_n)$, such as consistency or asymptotic efficiency, we will consider as a special case an underlying distribution which is a member of a family of elliptically contoured distributions, or briefly *elliptical distributions* (see for instance Kelker 1970, Cambanis, Huang and Simons 1981). By this we mean a family of distributions $P_{\mu,\Sigma}$, where $\mu \in \mathbf{R}^p$ and $\Sigma \in \mathrm{PDS}(p)$, of which each member has a corresponding density

$$|\mathbf{B}|^{-1} f(\|\mathbf{B}^{-1}(\mathbf{x} - \mu)\|)$$

where $\mathbf{B}\mathbf{B}^T = \mathbf{\Sigma}$ and f is a known function. A special member of the family is the *spherically symmetric* distribution with density

$$f(\|\mathbf{x}\|).$$

This distribution generates the whole family, since each member of the family can be obtained from the spherically symmetric distribution by means of the affine transformation $\mathbf{x} \mapsto \mathbf{B}\mathbf{x} + \boldsymbol{\mu}$. A typical example of an elliptical family is the multivariate normal family, obtained with $f(y) = (2\pi)^{-\frac{p}{2}} \exp(-\frac{1}{2}y^2)$.

When $X_1, X_2, \ldots, X_n$ are sampled from an elliptical distribution, and if the estimators $\mathbf{t}_n$ and $\mathbf{C}_n$ are affine equivariant, the investigation of the asymptotic properties of $\mathbf{t}(X_1, \ldots, X_n)$ and $\mathbf{C}(X_1, \ldots, X_n)$ often reduces to studying the estimators under the assumption that the elliptical distribution is spherical.

One of these properties is the asymptotic efficiency. In our situation we will then be dealing with an estimator that converges at a rate $\sqrt{n}$ towards a normal distribution. When the sample distribution is elliptical with parameters $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$, the limiting variances of the location estimators under consideration will be some multiple $\gamma\mathbf{\Sigma}$, and the limiting variance of the covariance estimators under consideration will be of the type

$$\sigma_1(\mathbf{I} + \mathbf{K}_{p,p})(\mathbf{\Sigma} \otimes \mathbf{\Sigma}) + \sigma_2 \mathrm{vec}(\mathbf{\Sigma})\,\mathrm{vec}(\mathbf{\Sigma})^T$$

where $\mathbf{K}_{p,p}$ is some fixed $p^2 \times p^2$-matrix and $\mathrm{vec}(\mathbf{\Sigma})$ is the p^2-vector consisting of the p columns of the matrix $\mathbf{\Sigma}$. By the asymptotic efficiency of the location estimator $\mathbf{t}_n$, we will generally mean asymptotic efficiency relative to the sample mean, i.e. the value γ for the sample mean divided by the value γ for $\mathbf{t}_n$. Sometimes the efficiency is measured relative to the maximum likelihood estimator for $\boldsymbol{\mu}$. Of course, at the normal distribution both ratios coincide. For covariance estimators the limiting variances only differ in σ_1 and σ_2. Tyler (1983) compared values of σ_1 for different covariance M-estimators with simulated values of a Monte Carlo study of robust covariance estimators in Devlin, Gnanadesikan and Kettenring (1981), and concluded that σ_1 suffices as an index for the asymptotic variance of the correlation estimator based upon the covariance M-estimator. In order to compare the asymptotic efficiency of our covariance estimators with that of others, we define the asymptotic efficiency of a covariance estimator $\mathbf{C}_n$ as the value σ_1 for the sample covariance divided by the value σ_1 of $\mathbf{C}_n$.

3. Breakdown Point and Influence Function

3.1 Breakdown Point

One way to investigate the robustness of an estimator is to study the global behaviour of the estimator under large pertubations of a given situation. Hodges (1967) proposed a simple finite sample measure of the degree to which an estimate of location is able to tolerate outliers. He studied univariate location estimators that are linear combinations of the order statistics $x_{1:n} \leq x_{2:n} \leq \cdots \leq x_{n:n}$, and defined the *left-* and *right tolerance* of an estimate t_n that is based on a collection $x_1, x_2 \ldots, x_n$, as follows. The left- and right tolerance of t_n are defined as the smallest integers $\alpha = \alpha(t_n)$ and $\beta = \beta(t_n)$, such that $x_{\alpha+1:n} \leq t_n \leq x_{n-\beta:n}$ and such that, whatever be the fixed values of $x_{\alpha+2:n}, \ldots, x_{n:n}$,

$$x_{\alpha+1:n} \to -\infty \text{ implies } t_n \to -\infty$$

and whatever be the fixed values of $x_{1:n}, \ldots, x_{n-\beta-1:n}$,

$$x_{n-\beta:n} \to \infty \text{ implies } t_n \to \infty.$$

In this context one can say that the estimate t_n can tolerate $\alpha(t_n)$ extreme values on the left and can tolerate $\beta(t_n)$ extreme values on the right. When $\alpha(t_n) = \beta(t_n) = \gamma(t_n)$ the estimate t_n was said to have tolerance $\gamma(t_n)$. Hodges mentions the mean and the median among other examples. The poor robustness of the mean is illustrated by having tolerance zero, whereas the median can tolerate $\lfloor \frac{n-1}{2} \rfloor$ extreme values, which is the maximum number that is possible.

Hampel (1968, 1971) proposed a more general but also more complicated concept, and was the first to use the name 'breakdown point' referring to the amount of extreme values for which an estimator completely collapses, or breaks down. He basically considered estimators as functionals $t_n = t(P_n)$ and studied the sensitivity of the functional $t(\cdot)$ under pertubations of some distribution P. In this setup,

Hampel showed that the insensitivity of the estimator t_n to pertubations in a sample generated by P is more or less equivalent with the uniform continuity of the functional $t(\cdot)$ (see Hampel 1971 for details). As a measure that tells us up to what distance from the assumed distribution P (or typically, up to what fraction of gross errors) the estimator still gives *some* indication about the distribution P, Hampel proposed the breakdown point ε^*, defined as

$$\varepsilon^*(\{t_n\}, P) = \sup \{\varepsilon \leq 1 : \exists \text{ a compact set } K = K(\varepsilon) \text{ such that } \pi(P,Q) < \varepsilon \Longrightarrow Q\{t_n \in K\} \to 1 \text{ as } n \to \infty\}$$

where $\pi(P,Q)$ denotes the Prohorov distance between the distributions P and Q. Among different examples also Hampel (1971) mentions the mean and the median. The sensitivity of the mean corresponds with breakdown point zero, the robustness of the median with breakdown point $\frac{1}{2}$.

Whereas Hampel's definition of the breakdown point may be too complicated to use, Hodges' tolerance concept is simple but not widely applicable. Donoho and Huber (1983) proposed a couple of finite-sample versions of the breakdown point. It is one of their proposals that we will use in this thesis as a measure for the global sensitivity of an estimator.

3.1.1 Multivariate Location

We first discuss the breakdown point of estimators of multivariate location.

DEFINITION 3.1: The *finite sample (replacement) breakdown point* of a location estimator $\mathbf{t}_n$ at a collection $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ is defined as the smallest fraction m/n of outliers that can take the estimator over all bounds :

$$\varepsilon^*(\mathbf{t}_n, \mathbf{X}) = \min_{1 \leq m \leq n} \left\{ \frac{m}{n} : \sup_{Y_m} \|\mathbf{t}_n(\mathbf{X}) - \mathbf{t}_n(\mathbf{Y}_m)\| = \infty \right\} \tag{3.1}$$

where the supremum in (3.1) is taken over all possible corrupted collections $\mathbf{Y}_m = (\mathbf{y}_1, \ldots, \mathbf{y}_m, \mathbf{x}_{i_m+1}, \ldots, \mathbf{x}_{i_n})$ that can be obtained from $\mathbf{X}$ by replacing any m points $\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_m}$ of $\mathbf{X}$ by arbitrary values $\mathbf{y}_1, \ldots, \mathbf{y}_m$.

This concept is simple and does not involve probability distributions. For a detailed discussion about the merits of this concept we refer to Donoho and Huber (1983). The breakdown point of the sample mean is $1/n$, the smallest possible value, illustrating the means sensitivity to outliers; the univariate sample median can easily be shown to have breakdown point $\lfloor \frac{n+1}{2} \rfloor / n$.

The computation of these breakdown points is easy since both estimators are nondecreasing in the observations. In the univariate case it is obvious that for such estimators the bias $\|\mathbf{t}_n(\mathbf{X}) - \mathbf{t}_n(\mathbf{Y}_m)\|$ is maximized by replacing the smallest sample values by $+\infty$, and it is straightforward to figure out the number m for

which this first becomes infinite. The computation of the breakdown point becomes more difficult when the estimators are no longer monotone, such as redescending M-estimators (see Section 4.1), or when more complicated multivariate estimators are considered.

The best possible value of the breakdown point among all estimators is 1. Indeed, consider an estimator which ignores $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ and attains a constant value. Of course, such an estimator does not make much sense. For instance, as an estimator of location it is not translation equivariant. Among all translation equivariant estimators of multivariate location the best possible breakdown point at any collection $\mathbf{X}$ is $\lfloor \frac{n+1}{2} \rfloor / n$ (see for instance Lopuhaä and Rousseeuw 1989). Intuitively this is clear, since if we replace half of the collection by translated copies of the other half, a translation equivariant estimator is not able to decide which half of the corrupted collection are the replacements and which half are the original points. Since, orthogonal (affine) equivariance implies translation equivariance, the upper bound $\lfloor \frac{n+1}{2} \rfloor / n$ also holds for the breakdown point of orthogonal (affine) equivariant estimators of multivariate location.

In Lopuhaä and Rousseeuw (1989) an example is given of a translation equivariant multivariate location estimator (which is not orthogonal equivariant) with breakdown point $\lfloor \frac{n+1}{2} \rfloor / n$, and an example of an orthogonal equivariant multivariate location estimator (which is not affine equivariant) with breakdown point $\lfloor \frac{n+1}{2} \rfloor / n$. Hence, for translation and orthogonal equivariant estimators of multivariate location the upper bound $\lfloor \frac{n+1}{2} \rfloor / n$ is sharp. Note that this bound is independent of the dimension p. So far, I do not know of any affine equivariant multivariate location estimator of which the breakdown point attains this value (except in the case $p = 1$, in which case the univariate sample median is affine equivariant and has breakdown point $\lfloor \frac{n+1}{2} \rfloor / n$). In Chapter 5 we will give examples of affine equivariant multivariate location estimators which have breakdown point $\lfloor \frac{n-p+1}{2} \rfloor / n$ at any collection $\mathbf{X}$ that is in general position. It seems that this is the best we can do for the moment.

At first glance, the breakdown point appears to depend on the collection $\mathbf{X}$. However, as we have seen above the breakdown point of the sample mean and of the univariate median are independent of $\mathbf{X}$. This behaviour turns out to be a rule rather than an exception. Most estimators have a breakdown point that does not depend on $\mathbf{X}$. Nevertheless, there exist location estimators with a breakdown point that does depend on the actual structure of the collection. For instance, consider a univariate location M-estimator, defined as the value t_n that minimizes

$$\sum_{i=1}^{n} \rho(x_i - t) \tag{3.2}$$

where $\rho : \mathbf{R} \to \mathbf{R}$. Huber (1984) showed that if ρ is symmetric, bounded and nondecreasing towards both sides with $\rho(y) \to 0$ as $|y| \to \infty$, the breakdown point of the resulting M-estimator depends on $\mathbf{X}$. This can best be understood if we consider a function $\rho(y)$ that is constant for $|y| \geq c$. If the width of such a function ρ is small compared to the distances between the sample points, for instance if all

the observations are at least $2c$ apart, one needs to replace only one observation by $+\infty$ in order to make one solution break down (in this case there are multiple solutions for minimizing (3.2)). On the other hand, if the width of the function ρ is large compared to the distances between the sample points, for instance if all sample points are the same, one needs to replace at least $\lfloor \frac{n+1}{2} \rfloor$ points in order to make the estimator break down.

Note that Huber (1984) also showed that location M-estimators defined by minimizing (3.2) using a function $\rho(y)$ that tends to ∞ at a sufficiently moderate rate when $|y| \to \infty$, have a breakdown point $\lfloor \frac{n+1}{2} \rfloor / n$ (in fact Huber considered the ε-addition breakdown point (see Section 3.1.3), but his proof can easily be adjusted for the ε-replacement breakdown point). This property will be important for the multivariate location estimator considered in Section 7.1.

Finally, we mention two properties of location estimators in relation to the breakdown point. First a relation with the so called *exact fit property*, introduced by Rousseeuw (1984). A location estimator $\mathbf{t}_n$ is said to satisfy the exact fit property for k points, if

$$\mathbf{t}(\underbrace{\boldsymbol{\mu}, \ldots, \boldsymbol{\mu}}_{n-k \text{ times}}, \mathbf{y}_1, \ldots, \mathbf{y}_k) = \boldsymbol{\mu}$$

for all $\mathbf{y}_1, \ldots, \mathbf{y}_k$. Define the *exact fit point* of a location estimator $\mathbf{t}_n$ as

$$\delta^*(\mathbf{t}_n, \boldsymbol{\mu}) = \min_{1 \le k \le n} \left\{ \frac{k}{n} : \exists \mathbf{Y}_k \text{ such that } \mathbf{t}_n(\mathbf{Y}_k) \neq \boldsymbol{\mu} \right\}$$

where $\mathbf{Y}_k = (\boldsymbol{\mu}, \ldots, \boldsymbol{\mu}, \mathbf{y}_1, \ldots, \mathbf{y}_k)$ is any collection of n points with $n-k$ points equal to $\boldsymbol{\mu}$ and k arbitrary points $\mathbf{y}_1, \ldots, \mathbf{y}_k$. Hence, $\delta^*(\mathbf{t}_n, \boldsymbol{\mu})$ is $1/n$ times the smallest k for which $\mathbf{t}_n$ does no longer satisfy the exact fit property. For any translation equivariant location estimator $\mathbf{t}_n$, that is also *scale equivariant*, i.e. $\mathbf{t}(\lambda \mathbf{x}_1, \ldots, \lambda \mathbf{x}_n) = \lambda \mathbf{t}(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ for any $\lambda > 0$, it holds that

$$\delta^*(\mathbf{t}_n, \boldsymbol{\mu}) \ge \varepsilon^*\big(\mathbf{t}_n, (\underbrace{\boldsymbol{\mu}, \ldots, \boldsymbol{\mu}}_{n \text{ times}})\big).$$

This can easily be shown along the lines of Remark 1 in Rousseeuw and Leroy (1987, p.123). This result becomes particularly useful if $\mathbf{t}_n$ is well behaved so that $\varepsilon^*(\mathbf{t}_n, \mathbf{X})$ is the same at all $\mathbf{X}$. When $\mathbf{t}_n$ is translation and scale equivariant with a breakdown point m^*/n that is independent of $\mathbf{X}$, it follows that if a collection of n points contains at least $n - m^* + 1$ copies of one point $\boldsymbol{\mu}$, the location estimate based on this collection will be $\boldsymbol{\mu}$.

Secondly, the relation with a measure of large deviations. Consider a location-family of univariate distributions $\{P_\mu : \mu \in \mathbf{R}\}$, that are symmetric around μ, with a density $f(x - \mu)$. As a measure of the tailperformance of a univariate location estimator $t_n = t(X_1, \ldots, X_n)$, Jurečková (1981) considered

$$B(a, t_n) = \frac{-\log P_\mu(|t_n - \mu| > a)}{-\log P_\mu(|X_1 - \mu| > a)}.$$

For an efficient estimator, the inaccuracy $P_\mu(|t_n - \mu| > a)$ is expected to tend to 0 as fast as possible as $a \to \infty$, i.e. the distribution of $t(X_1, \ldots, X_n)$ will have the smallest possible tails. Jurečková showed that the tails of a univariate translation equivariant estimator of location decrease at most n times faster than the tails of the sample distribution P_μ and at least as fast as the tails of P_μ, i.e.

$$1 \leq \liminf_{a\to\infty} B(a, t_n) \leq \limsup_{a\to\infty} B(a, t_n) \leq n.$$

For exponentially tailed distributions, the sample mean $\bar{X}_n$ performes optimally with $B(a, \bar{X}_n)$ tending to n, while for algebraically tailed distributions the lack of robustness of $\bar{X}_n$ is expressed by $B(a, \bar{X}_n)$ tending to 1. Let $X_1, X_2, \ldots, X_n$ be a sample generated by a member of the family $\{P_\mu : \mu \in \mathbf{R}\}$, and let $t_n = t(X_1, \ldots, X_n)$ be a translation equivariant univariate location estimator which is monotone in each argument X_i for $i = 1, 2, \ldots, n$. He, Jurečková, Koenker and Portnoy (1988) first discovered the following striking relation between the finite sample replacement breakdown point of such estimators t_n and the measure $B(a, t_n)$. Suppose that at any collection $\mathbf{X}$ the breakdown point $\varepsilon^*(t_n, \mathbf{X}) = m^*/n$ is independent of $\mathbf{X}$. Then

$$m^* \leq \liminf_{a\to\infty} B(a, t_n) \leq \limsup_{a\to\infty} B(a, t_n) \leq n - m^* + 1. \tag{3.3}$$

Hence, it turns out that the finite sample replacement breakdown point is not just an appealing and simple robustness concept, but it also has a stochastic interpretation. Relation (3.3) indicates that location estimators with a high breakdown point necessarily must sacrifice tailperformance. However, (3.3) also implies that estimators with maximal breakdown point $\lfloor \frac{n+1}{2} \rfloor / n$ satisfy a minimax property in the sense that they maximize least favorable tailperformance. The relation (3.3) is extended to multivariate location estimators in Lopuhaä and Rousseeuw (1989).

3.1.2 Covariance

A location estimator breaks down if contamination can drive the estimator to the boundary of the (location) parameter space. Because scale estimators are inherently nonnegative it makes sense to say that scale estimators break down if contamination can drive it either to ∞ or to 0. Huber (1981) refers to these possibilities as 'explosion' or 'implosion' of the scale estimator respectively. This brings us to the following definition of the finite sample breakdown point for covariance estimators.

DEFINITION 3.2: The finite sample (replacement) breakdown point of a covariance estimator $\mathbf{C}_n$ at a collection $\mathbf{X}$ is defined as the smallest fraction m/n of outliers that can either take the largest eigenvalue $\lambda_1(\mathbf{C}_n)$ over all bounds, or take the smallest eigenvalue $\lambda_p(\mathbf{C}_n)$ arbitrarily close to zero :

$$\varepsilon^*(\mathbf{C}_n, \mathbf{X}) = \min_{1\leq m\leq n} \left\{ \frac{m}{n} : \sup_{\mathbf{Y}_m} D(\mathbf{C}_n(\mathbf{X}), \mathbf{C}_n(\mathbf{Y}_m)) = \infty \right\}$$

where the supremum is taken over the same corrupted collections $\mathbf{Y}_m$ as in (3.2), and where $D(\mathbf{A}, \mathbf{B}) = \max\{|\lambda_1(\mathbf{A}) - \lambda_1(\mathbf{B})|, |\lambda_p(\mathbf{A})^{-1} - \lambda_p(\mathbf{B})^{-1}|\}$.

The univariate sample variance can easily be seen to have the same breakdown point $1/n$ as the sample mean, which illustrates its sensitivity to outliers. The *Median Absolute Deviation* (MAD) is an outlier resistant univariate scale estimator. It is defined as

$$\sigma_{\text{MAD}} = \underset{1\le i\le n}{\text{med}} \left| x_i - \underset{1\le j\le n}{\text{med}} x_j \right|$$

and has breakdown point $\lfloor \frac{n}{2} \rfloor / n$ at collections $x_1, x_2, \ldots, x_n$ that are in general position.

Among all affine equivariant covariance estimators the maximal breakdown point at collections $\mathbf{X}$ in general position is $\lfloor \frac{n-p+1}{2} \rfloor / n$ (see Davies 1987). Note that this bound depends on the dimension p. In Chapter 5 we will give some examples of affine equivariant covariance estimators with a breakdown point that attains this value at collections $\mathbf{X}$ that are in general position. Hence, this upper bound is sharp for any p and $n \geq p+1$. If a collection is not in general position, the upper bound will be smaller. Let $k_n = k_n(\mathbf{X})$ be the maximum number of sample points that are contained in some hyperplane of dimension less than $p-1$. Obviously, $k_n \geq p$ with equality if the collection is in general position. By a straightforward adjustment of Davies' proof one may show that for general collections $\mathbf{X}$ the breakdown point of any affine equivariant covariance estimator is at most $\lfloor \frac{n-k_n+1}{2} \rfloor / n$.

3.1.3 Alternative Definitions

The finite sample breakdown point as defined in Definition 3.1 is the so called ε-replacement version of Donoho and Huber (1983) for location estimators. They also considered two other versions.

The first one is *ε-addition* breakdown (in fact it was called ε-contamination breakdown, but that name seems a bit confusing). It is defined as the fraction $m/(n+m)$, where m is the minimum number of outliers $\mathbf{y}_1, \ldots, \mathbf{y}_m$ that one has to *add* to a collection $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ in order to make the maximum bias infinitely large. This type of breakdown has been used by several authors (Donoho 1982, Donoho and Huber 1983, Huber 1984, Tamura and Boos 1986, Tyler 1986), mostly because it is sometimes easier to compute than the ε-replacement breakdown point. For instance in the case of location M-estimators with a bounded loss function (see Section 3.1.1). However, the ε-replacement version seems more realistic and is generally applicable. Moreover, from an intuitive point of view, outliers are not some extreme observations that are added to the sample, but they 'hide' themselves by replacing some of the data points that should have been observed.

A second alternative to Definition 3.1 is the *ε-modification* breakdown point. Let π be an arbitrary distance function defined on the space of all empirical distributions. Let P_n be the empirical distribution corresponding to a given sample $\mathbf{X}$, and let $\mathbf{X}'$ be any other sample with empirical distribution $Q_{n'}$, such that $\pi(P_n, Q_{n'}) \leq \varepsilon$. The breakdown point is the smallest value of ε for which the maximum bias $\|\mathbf{t}(P_n) - \mathbf{t}(Q_{n'})\|$ becomes infinitely large. The ε-modification version has

never been very popular, although recently Davies (1989) used an affine equivariant asymptotic version of this concept.

Another asymptotic breakdown concept that is sometimes considered is the *gross-error breakdown point*, which corresponds with the so called gross-error model (Section 4.1). Consider mixtures of a distribution P and the Dirac measure $\delta_{\mathbf{x}}$

$$P_{\varepsilon,\mathbf{x}} = (1-\varepsilon)P + \varepsilon\delta_{\mathbf{x}}$$

with $0 < \varepsilon < 1$ and $\mathbf{x} \in \mathbf{R}^p$. Let $\mathbf{t}_n$ be an estimator that can be written by means of a functional $\mathbf{t}(\cdot)$, that is $\mathbf{t}_n = \mathbf{t}(P_n)$. The gross-error breakdown point of the estimator $\mathbf{t}$ at P, is the smallest ε for which

$$\sup_{\mathbf{x}} \left\| \mathbf{t}(P_{\varepsilon,\mathbf{x}}) - \mathbf{t}(P) \right\|$$

becomes infinite. This concept of breakdown was used for instance by Maronna (1976) and Huber (1977) to describe the global robustness of multivariate M-estimators.

3.2 Influence Function

The breakdown point measures the global behaviour of an estimator under *large* pertubations of a particular given situation. It tells us that the estimator will stay within finite bounds if a certain fraction of a given collection is replaced by outliers. Although it may give a good first impression about the robustness of an estimator, the breakdown point is a crude way to measure the sensitivity of an estimator. It does not tell us how much an estimator can be altered under small pertubations. It is necessary to have at least a positive breakdown point for a robust estimator. However, one should not judge an estimator's robustness merely on the basis of its breakdown point. To further investigate the robustness of an estimator, the breakdown point may be complemented by measures that describe the changes of an estimator under *small* pertubations of a given situation. In this section we discuss such a proposal.

Consider a location estimator $\mathbf{t}_n$, which can be written by means of a functional $\mathbf{t}(\cdot)$ defined on the space of all probability distributions on $\mathbf{R}^p$, i.e. $\mathbf{t}_n = \mathbf{t}(P_n)$.

DEFINITION 3.3: The *influence function* of an estimator $\mathbf{t}_n = \mathbf{t}(P_n)$ at a distribution P is defined as

$$\text{IF}(\mathbf{x}; \mathbf{t}, P) = \lim_{\varepsilon \downarrow 0} \frac{\mathbf{t}\big((1-\varepsilon)P + \varepsilon\delta_{\mathbf{x}}\big) - \mathbf{t}(P)}{\varepsilon} \tag{3.4}$$

if this limit exists for all $\mathbf{x} \in \mathbf{R}^p$.

The influence function, originally called the influence curve, is the Gateaux derivative of the functional $\mathbf{t}(\cdot)$ at a distribution P in the direction of the Dirac measure $\delta_{\mathbf{x}}$. The idea of differentiating statistical functionals (although for other purposes) goes back to von Mises (1937, 1947) and Filippova (1962). Hampel (1968, 1974) introduced the derivative (3.4) to investigate the sensitivity of estimators, and it became the cornerstone of the so called *infinitesimal approach* to robustness. An extensive account of this approach, as well as many references to related work, are given in Hampel, Ronchetti, Rousseeuw and Stahel (1986). We will have a very brief glance at this approach in Section 4.1 when it is discussed in relation with the minimax variance approach.

Whereas the continuity of the functional $\mathbf{t}(\cdot)$ at P is more or less equivalent with the global stability of the corresponding estimator $\mathbf{t}_n$, differentiability of the functional corresponds with the actual change of the estimator under small pertubations of P. The influence function describes approximately the changes of the estimator under a small pertubation of the distribution P, i.e. reallocation of a small fraction ε of the probability mass of P to a point $\mathbf{x}$. For instance, if for large n we replace P by P_{n-1} and put $\varepsilon = 1/n$, $\mathrm{IF}(\mathbf{x}; \mathbf{t}, P_n)$ can be interpreted as measuring n times the change of $\mathbf{t}_{n-1}$ caused by adding an observation $\mathbf{x}$ to the sample.

The influence function of the sample mean at any distribution P with zero mean, is given by

$$\mathrm{IF}(\mathbf{x}; \mathbf{t}, P) = \mathbf{x}$$

which illustrates the arbitrarily large influence of small pertubations at a point $\mathbf{x}$ that is far from the majority of the observations. The influence function of the univariate sample median at a distribution P that has a density f and for which $F(0) = \frac{1}{2}$, is given by

$$\mathrm{IF}(x; t, P) = \frac{\mathrm{sign}(x)}{2f(0)}$$

which says that every point x at the right (left) of the median of P has the same effect. Several other examples of influence functions of location estimators are given in Hampel (1974), Huber (1981) and Hampel et al. (1986).

A concept that is closely related to the influence function is the *gross-error sensitivity*, defined as

$$\gamma^*(\mathbf{t}, P) = \sup_{\mathbf{x}} \|\mathrm{IF}(\mathbf{x}; \mathbf{t}, P)\|. \tag{3.5}$$

It measures the worst approximate influence which a small amount of contamination can have on the value of an estimator. The gross-error sensitivity of the sample mean is infinite, whereas the gross-error sensitivity of the univariate sample median is $1/(2f(0))$. Note that among all univariate location M-estimators the gross-error sensitivity of the median is the smallest possible (see for instance Hampel et al. 1986, p.133).

Just as the breakdown point is only a measure of the global sensitivity that needs to be complemented by measures of the local sensitivity, the influence function and its corresponding gross-error sensitivity, should not be considered alone.

For instance multivariate location S-estimators have the same bounded influence function as corresponding location M-estimators. However, the breakdown point of the latter is at most $1/(p+1)$ (see Section 4.1.1), whereas the S-estimator can be constructed such that its breakdown point is $\lfloor \frac{n-p+1}{2} \rfloor / n$ (see Chapter 6). One can think of even more extreme examples (see for instance Donoho and Huber 1983, p.174) of a location estimator with a bounded influence function and breakdown point zero, while other estimators with the same influence function have a strictly positive breakdown point. By considering both $\varepsilon^*(\mathbf{t}_n, \mathbf{X})$ and $\mathrm{IF}(\mathbf{x}; \mathbf{t}, P)$ together one obtains a sensible approach, whereas optimization of one or the other alone is unwise.

The influence function of estimators of scale is defined similar to (3.4). Let $\mathbf{C}_n$ be a covariance estimator that can be written by means of a functional $\mathbf{C}(\cdot)$, i.e. $\mathbf{C}_n = \mathbf{C}(P_n)$.

DEFINITION 3.4: The influence function of a covariance estimator $\mathbf{C}_n$ is defined as

$$\mathrm{IF}(\mathbf{x}; \mathbf{C}, P) = \lim_{\varepsilon \downarrow 0} \frac{\mathbf{C}\big((1-\varepsilon)P + \varepsilon \delta_{\mathbf{x}}\big) - \mathbf{C}(P)}{\varepsilon}$$

if this limit exists for all $\mathbf{x} \in \mathbf{R}^p$.

The gross-error sensitivity of a covariance estimator is defined similar to (3.5). The influence function of the sample covariance at any distribution P with zero mean is

$$\mathrm{IF}(\mathbf{x}; \mathbf{C}, P) = \mathbf{x}\mathbf{x}^T.$$

Hence, for $\|\mathbf{x}\| \to \infty$, the influence function grows not only linearly, as with the sample mean, but even quadratically; obviously, the gross-error sensitivity of the sample covariance is infinite. The influence function of the MAD (see Section 3.1.2) at the univariate standard normal distribution Φ is

$$\mathrm{IF}(x; \sigma_{\mathrm{MAD}}, \Phi) = \frac{\mathrm{sign}\big(|x| - \Phi^{-1}(\frac{3}{4})\big)}{\beta}$$

where $\beta = 4\Phi^{-1}(\frac{3}{4})\phi\big(\Phi^{-1}(\frac{3}{4})\big)$. Note that the gross-error sensitivity of the MAD is the smallest possible among all univariate M-estimators of scale (Rousseeuw 1981, see also Hampel et al. 1986, p.142). Several other examples of influence functions of scale estimators are given in Hampel et al. (1986) and Huber (1981).

4. Other Approaches to Robustness

4.1 Minimax Variance

4.1.1 Introduction

The foundations of modern robustness theory were laid by P. J. Huber in his 1964 paper. He introduced the class of M-estimators for univariate location. These estimators, and several extensions of them, have become the most well known robust alternatives to the maximum likelihood estimators, in particular to the least squares estimator.

They arise as a generalization of maximum likelihood estimators as follows. The well known maximum likelihood estimator for the parameter $\boldsymbol{\theta}$ of a density $f_{\boldsymbol{\theta}}(\mathbf{x})$ is defined as the value $\boldsymbol{\theta}_n$ which maximizes $\prod_{i=1}^{n} f_{\boldsymbol{\theta}}(\mathbf{x}_i)$, or equivalently which minimizes

$$-\sum_{i=1}^{n} \log f_{\boldsymbol{\theta}}(\mathbf{x}_i).$$

Huber (1964) proposed to generalize this to minimizing

$$\sum_{i=1}^{n} \rho(\mathbf{x}_i, \boldsymbol{\theta}) \tag{4.1}$$

where ρ is some real valued function. When the function ρ has a partial derivative $\psi(\mathbf{x}, \boldsymbol{\theta}) = (\partial/\partial\boldsymbol{\theta})\rho(\mathbf{x}, \boldsymbol{\theta})$, the estimator $\boldsymbol{\theta}_n$ will satisfy

$$\sum_{i=1}^{n} \psi(\mathbf{x}_i, \boldsymbol{\theta}_n) = \mathbf{0}. \tag{4.2}$$

An estimator that is defined either by minimizing (4.1), or as a solution of (4.2), is called an *M-estimator* for the parameter θ ('M' stands for generalized Maximum likelihood).

Huber (1964) investigated robust estimation of a univariate location parameter, and determined the location M-estimators that are optimal in a minimax variance sense. For this purpose he considered the *gross-error model*, a kind of 'neighbourhood' of a fixed symmetric distribution. For simplicity, we confine ourself to the normal gross-error model

$$\mathcal{P}_\varepsilon = \{(1-\varepsilon)\Phi + \varepsilon H : H \text{ is any } \textit{symmetric} \text{ distribution}\} \tag{4.3}$$

where $0 < \varepsilon < 1$ is fixed. As a special case of (4.2), Huber considered the class of univariate location M-estimators defined as a solution of

$$\sum_{i=1}^{n} \psi(x_i - t) = 0 \tag{4.4}$$

where the function ψ was taken from some class Ψ, and showed that for all such functions the corresponding location M-estimator t_n is asymptotically normal. When we assume that $t_n \to 0$ and that $\mathbf{E}_P \psi(X_1) = 0$ for $P \in \mathcal{P}_\varepsilon$, this result can be made plausible as follows, using Taylor's formula :

$$\begin{aligned} 0 &= \sum_{i=1}^{n} \psi(X_i - t_n) \\ &= \sum_{i=1}^{n} \psi(X_i) - \sum_{i=1}^{n} \psi'(X_i)\, t_n + \text{ remainder term.} \end{aligned}$$

Hence, proving asymptotic normality boils down to showing that the remainder term is $o_P(1/\sqrt{n})$. In that case

$$\sqrt{n}\, t_n = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(X_i)}{\frac{1}{n} \sum_{i=1}^{n} \psi'(X_i)} + o_P(1) \to \mathcal{N}\big(0, V(\psi, P)\big)$$

where $V(\psi, P) = \mathbf{E}_P \psi^2(X_1) / \big(\mathbf{E}_P \psi'(X_1)\big)^2$, using the law of large numbers and the central limit theorem. Later, Huber (1967) provided sufficient conditions for consistency and asymptotic normality for general M-estimators defined either by minimizing (4.1) or as a solution of (4.2).

To find the most robust estimator, Huber proposed to minimize the maximal possible asymptotic variance $V(\psi, P)$ that one can have at a distribution P of $\mathcal{P}_\varepsilon$, i.e.

$$\sup_{P \in \mathcal{P}_\varepsilon} V(\psi, P)$$

among all M-estimators that are defined as a solution of (4.4) with the function ψ taken from the class Ψ. In case of the normal gross-error model (4.3) this minimax

variance problem yields the well known Huber estimator. It is defined as the solution of (4.4) using Huber's ψ-function

$$\psi_H(y;k) = \begin{cases} -k & \text{, for } y \leq -k \\ y & \text{, for } |y| \leq k \\ k & \text{, for } y \geq k \end{cases} \tag{4.5}$$

The value of k corresponds to the amount ε of gross-error contamination by means of the equation

$$2\Phi(k) - 1 + \frac{2\phi(k)}{k} = \frac{1}{1-\varepsilon}.$$

Hampel (1968) also considered the class of location M-estimators defined as solutions of (4.4), and showed that the influence function of these location M-estimators are proportional to the function ψ that defines the M-estimator, i.e.

$$\mathrm{IF}(x;t,P) = \frac{\psi(x)}{\beta}$$

where β is a positive constant that depends on ψ. Hampel proposed to find the M-estimator that minimizes the asymptotic variance $V(\psi,P)$ subject to a bound on the gross-error sensitivity : $\gamma^*(t,P) \leq k$. Huber's estimator also turned out to be optimal in this sense (see for instance Hampel et al. 1986, p.117). Moreover, one may show that this estimator has breakdown point $\lfloor \frac{n+1}{2} \rfloor / n$ (an adjusted version of Theorem 4.1 in Huber 1984).

As both minimizing $\sup_{\mathcal{P}_\varepsilon} V(\psi,P)$ and minimizing $V(\psi,P)$ under the constraint $\gamma^*(\psi,P) \leq k$ yield the same estimator, it is not surprising that both methodologies are related in some way. For a clear explanation of the relation between both approaches we refer to Hampel et al. (1986, Section 2.7).

Note that Huber's estimator corresponds with a function ψ in (4.4) that is monotone, so that the estimator is uniquely defined. When the function ψ in (4.4) is no longer monotone, (4.4) may have multiple solutions. To distinguish functions ψ that are zero for $|y|$ greater than some cutoff point c, from functions ψ that tend to zero for $|y| \to \infty$ but never become zero, we call the first type *strongly redescending* and the second type *weakly redescending*. Note that the influence function that corresponds with redescending M-estimators is also redescending, which illustrates the small, or even zero influence of outliers. Section 2.6 in Hampel et al. (1986) gives an overview of the infinitesimal approach for strongly redescending M-estimators.

4.1.2 Multivariate M-estimators

Huber's minimax variance approach was extended by Collins (1982) to M-estimators for multivariate location. These estimators are defined as solutions of

$$\sum_{i=1}^{n} \frac{\psi(\|\mathbf{x}_i - \mathbf{t}\|)}{\|\mathbf{x}_i - \mathbf{t}\|}(\mathbf{x}_i - \mathbf{t}) = \mathbf{0}.$$

Note that this reduces to (4.4) when $p = 1$ and ψ is skew-symmetric. Collins studied these estimators under the assumption that the sample distribution is spherically symmetric around some parameter vector μ, an assumption that is often not realistic. When the scale parameter is not assumed to be fixed, then it seems more natural to use affine equivariant estimators of location. Location M-estimators as defined above, do not satisfy this property.

Maronna (1976) introduced affine equivariant multivariate M-estimators for location and scatter. He considered simultaneous estimation of location and scatter, and defined the M-estimators of location and scatter as the vector $\mathbf{t}_n$ and the matrix $\mathbf{C}_n$ that are a solution of the simultaneous equations

$$\begin{aligned} &\sum_{i=1}^{n} u_1\big(d(\mathbf{x}_i, \mathbf{t}, \mathbf{C})\big)(\mathbf{x}_i - \mathbf{t}) = \mathbf{0} \\ &\sum_{i=1}^{n} \Big\{ u_2\big(d^2(\mathbf{x}_i, \mathbf{t}, \mathbf{C})\big)(\mathbf{x}_i - \mathbf{t})(\mathbf{x}_i - \mathbf{t})^T - \mathbf{C} \Big\} = \mathbf{0} \end{aligned} \tag{4.6}$$

where $u_1 : \mathbf{R} \to \mathbf{R}$ and $u_2 : \mathbf{R} \to \mathbf{R}$. As a special case one obtains the maximum likelihood estimators for the parameters μ and $\mathbf{\Sigma}$ of an elliptical distribution : $u_1(y) = u_2(y^2) = f'(y)/\big(yf(y)\big)$.

Later, Huber (1981) generalized (4.6) and defined multivariate M-estimators of location and scatter as the vector $\mathbf{t}_n$ and matrix $\mathbf{C}_n$ that are a solution of

$$\begin{aligned} &\sum_{i=1}^{n} v_1\big(d(\mathbf{x}_i, \mathbf{t}, \mathbf{C})\big)(\mathbf{x}_i - \mathbf{t}) = \mathbf{0} \\ &\sum_{i=1}^{n} \Big\{ v_2\big(d(\mathbf{x}_i, \mathbf{t}, \mathbf{C})\big)(\mathbf{x}_i - \mathbf{t})(\mathbf{x}_i - \mathbf{t})^T - v_3\big(d(\mathbf{x}_i, \mathbf{t}, \mathbf{C})\big)\mathbf{C} \Big\} = \mathbf{0} \end{aligned} \tag{4.7}$$

where $v_1 : \mathbf{R} \to \mathbf{R}$, $v_2 : \mathbf{R} \to \mathbf{R}$ and $v_3 : \mathbf{R} \to \mathbf{R}$. Huber (1981) gives the general expression for the influence function of these M-estimators. His results on the existence and uniqueness of solutions of (4.7) do not carry much further than those of Maronna (1976). For instance, if one wants the solution of (4.7) to be unique, Huber (1981) requires $v_3(y) = 1$, which yields equations (4.6) in return. Kent and Tyler (1989) also studied equations (4.6) in the case that $u_1(y) = u_2(y^2)$ and provided sufficient conditions for existence that are essentially the best possible.

The concept of affine equivariance as defined in Section 2.2 becomes a bit vague when $\mathbf{t}_n$ and $\mathbf{C}_n$ are not uniquely defined. However, these estimators will always be affine equivariant in the following sense. Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ be a collection and let V be set of solutions of (4.7). Let W be the set of solutions of equations (4.7) with $\mathbf{A}\mathbf{x}_1 + \mathbf{v}, \mathbf{A}\mathbf{x}_2 + \mathbf{v}, \ldots, \mathbf{A}\mathbf{x}_n + \mathbf{v}$ instead of $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$. Then $W = \{(\mathbf{A}\mathbf{t} + \mathbf{v}, \mathbf{A}\mathbf{C}\mathbf{A}^T) : (\mathbf{t}, \mathbf{C}) \in V\}$.

Multivariate M-estimators are probably the most well known affine equivariant robust alternatives to the multivariate sample mean and sample covariance. This is not surprising, since for a long time they seemed to outperform other affine

equivariant robust alternatives. By chosing suitable functions v_1, v_2 and v_3 in (4.7) one may obtain affine equivariant M-estimators that converge at rate $\sqrt{n}$, which are asymptotically normal with a reasonable efficiency, and which have a bounded influence function.

However, both Maronna (1976) and Huber (1977, 1981) also mention the poor breakdown behaviour of multivariate M-estimators. They both consider the gross-error breakdown point (see Section 3.1.3). It turns out that equations (4.7) will always have at least one solution with a gross-error breakdown point that is

$$\varepsilon^* \leq \frac{1}{p}.$$

Under certain monotonicity conditions on the functions in (4.7), the solution of these equations will be unique (see for instance Huber 1981, Kent and Tyler 1989) and must therefore have a breakdown point less than $1/p$. This means that multivariate M-estimators, despite the other nice properties they possess, will be extremely sensitive to outliers in higher dimensions. Tyler (1986) also studied the finite sample (addition) breakdown point of these M-estimators and showed that it is at most $1/(p+1)$. The breakdown is due to the covariance M-estimator, which is not able to resist so called *co-planar* contamination, i.e. outliers concentrated in a lower dimensional hyperplane.

4.2 Minimax Bias

The main approach to global robustness in recent years has been centered around the construction of estimators with a high breakdown point. The breakdown point approach is highly attractive for a number of reasons, not the least of which is the transparency of the concept. On the other hand it exhibits such a strong and crude 'distribution freeness', that this makes it quite unsuitable for optimizing global robustness. Nonethless one might wish to have a global optimality theory of robustness which emphasizes bias control also for fractions of contamination that are smaller than the breakdown point. In this context, Huber (1964) showed that the sample median minimizes the maximum asymptotic bias among all translation equivariant estimators of location, the maximum being over ε-contaminated neighbourhoods like (4.3). This approach to global robustness, i.e. the construction of minimax bias robust estimators has been essentially neglected until quite recently.

Consider an estimator $\mathbf{t}_n$ that can be written as $\mathbf{t}(P_n)$. The *bias curve* of $\mathbf{t}_n$ at P is a function of the amount ε of contamination. It is defined as

$$B(\varepsilon; \mathbf{t}, P) = \sup_{Q \in \mathcal{Q}} \|\mathbf{t}((1-\varepsilon)P + \varepsilon Q) - \mathbf{t}(P)\| \tag{4.8}$$

where the supremum is taken over some class $\mathcal{Q}$ consisting of distributions Q on $\mathbf{R}^p$. Similar to the minimax variance approach, consider some $0 < \varepsilon < 1$ to be fixed,

and let $\mathcal{T}$ be some class of estimators. The minimax bias location estimator in $\mathcal{T}$, would be the estimator $\mathbf{t}_n = \mathbf{t}(P_n)$ that minimizes $B(\varepsilon; \mathbf{t}, P_n)$ among all $\mathbf{t} \in \mathcal{T}$.

Recently, different authors have begun to study the robustness problem using this minimax bias approach. Martin and Zamar (1988) and Martin, Yohai and Zamar (1989) found minimax bias estimators for different classes of regression estimators. One drawback of these minimax bias estimators is their relatively poor asymptotic behaviour, e.g. a slow rate of convergence or a low efficiency at the assumed model distribution.

Maronna and Yohai (1989) investigated the bias curve for two classes of covariance estimators, which included covariance M-estimators and the minimum volume ellipsoid (MVE) covariance estimator (Rousseeuw 1983). As a special case they considered Tyler's (1987) distribution free M-estimator. As an M-estimator, it has a gross-error breakdown point of at most $1/p$, whereas the MVE covariance estimator has a high breakdown point (see Section 5.1). Nevertheless, it turns out that Tyler's estimator has a smaller maximum bias than the MVE covariance estimator, that is up to $\varepsilon = 1/p$ of course. The maximum bias of the MVE estimator will still be finite for $1/p \leq \varepsilon < 1/2$, whereas the maximum bias of the M-estimator then becomes infinite. This comparison indicates that further study of the bias curve is needed, and that the breakdown point does not always suffice to give complete information about the global sensitivity of an estimator.

5. Combining a high Breakdown Point with Affine Equivariance

Multivariate M-estimators are defined as solutions of (4.7). These equations will always have at least one solution that has a finite sample (addition) breakdown point of at most $1/(p+1)$. Donoho (1982) has listed several other affine equivariant 'robust' multivariate location estimators and has shown that the finite sample (addition) breakdown points of these procedures are also bounded by the 'mysterious' number $1/(p+1)$. This might tempt us to think that with an affine equivariant estimator one can do no better than a breakdown point of $1/(p+1)$. At least do these examples indicate that combining a high breakdown point with affine equivariance is not trivial, and this problem in higher dimensions seems to be essentially different from the univariate case. In the next sections we will discuss some of the first proposals of affine equivariant multivariate estimators with a breakdown point that attains the maximal possible value for covariance estimators.

5.1 Stahel-Donoho Estimator

Independent of each other, Stahel (1981) and Donoho (1982) constructed the first affine equivariant estimator of multivariate location and covariance with a high breakdown point. Their idea was to measure the 'outlyingness' of a point $\mathbf{x}_i$ relative to the center of a collection, and then to compute a weighted sample mean and sample covariance, where the points with a relatively large degree of outlyingness are downweighted.

They proposed to find the projection in which $\mathbf{x}_i$ appears to be most outlying, and to measure the degree of outlyingness of $\mathbf{x}_i$ by

$$r_i = \sup_{\|\mathbf{u}\|=1} \frac{|\mathbf{u}^T\mathbf{x}_i - \text{med}(\mathbf{u}^T\mathbf{X})|}{\text{MAD}(\mathbf{u}^T\mathbf{X})} \tag{5.1}$$

where $\text{med}(\mathbf{u}^T\mathbf{X})$ and $\text{MAD}(\mathbf{u}^T\mathbf{X})$ are the median and median absolute deviation of the univariate sample $\mathbf{u}^T\mathbf{x}_1, \ldots, \mathbf{u}^T\mathbf{x}_n$. Next, assign weights to each $\mathbf{x}_i$ according to its degree of outlyingness r_i :

$$w_i = w(r_i)$$

where $w : [0, \infty) \to [0, \infty)$ is assumed to be strictly positive, decreasing with $w(r) \to 0$ as $r \to \infty$, and such that $w(r)r$ is bounded. The estimators of multivariate location and covariance are then defined as

$$\begin{aligned} \mathbf{t}_w(\mathbf{X}) &= \frac{\sum_{i=1}^n w_i \mathbf{x}_i}{\sum_{i=1}^n w_i}. \\ \mathbf{C}_w(\mathbf{X}) &= \frac{\sum_{i=1}^n w_i^2 (\mathbf{x}_i - \mathbf{t}_w(\mathbf{X}))(\mathbf{x}_i - \mathbf{t}_w(\mathbf{X}))^T}{\sum_{i=1}^n w_i^2}. \end{aligned} \tag{5.2}$$

Donoho (1982) showed that these estimators are affine equivariant and have a finite sample breakdown point

$$\varepsilon^*(\mathbf{t}_w, \mathbf{X}) = \varepsilon^*(\mathbf{C}_w, \mathbf{X}) = \frac{\lfloor \frac{n+1}{2} \rfloor - p}{n} \tag{5.3}$$

at every collection $\mathbf{X}$ in general position. (In fact, Donoho computed the ε-addition breakdown point, but his proof can be adjusted easily for the ε-replacement breakdown point). Note that (5.3) is smaller than the maximal possible breakdown point for affine covariance estimators. The asymptotic properties of the Stahel-Donoho estimator, such as the rate of convergence, consistency or the limiting distribution, have not yet been investigated.

5.2 Minimum Volume Ellipsoid Estimator

Another proposal that combines affine equivariance with a high breakdown point is the minimum volume ellipsoid (MVE) estimator. It was originally introduced by Rousseeuw (1983), who defined

$$\begin{aligned} \mathbf{t}_h(\mathbf{X}) = &\text{ the center of the smallest ellipsoid that covers} \\ &\text{at least } h \text{ points of } \mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n). \end{aligned} \tag{5.4}$$

In principle, h may be any integer between $p + 1$ and n, but it is typically taken around $\frac{n}{2}$. The corresponding MVE covariance estimator is defined as the covariance structure of the same ellipsoid.

The MVE estimator is affine equivariant, and Rousseeuw (1983) showed that with $h = \lfloor \frac{n}{2} \rfloor + 1$, it has breakdown point

$$\varepsilon^*(\mathbf{t}_h, \mathbf{X}) = \frac{\lfloor \frac{n}{2} \rfloor - p + 1}{n} \tag{5.5}$$

at any collection $\mathbf{X}$ in general position. The breakdown point in (5.5) is smaller than the maximal possible breakdown point for covariance estimators. However, the MVE estimator can be adjusted by taking $h = \lfloor \frac{n+p+1}{2} \rfloor$, so that it has breakdown point $\lfloor \frac{n-p+1}{2} \rfloor / n$ at any collection in general position (see Lopuhaä and Rousseeuw 1989).

In the univariate case the estimators correspond with the midpoint and length of the shortest interval that covers at least $\lfloor \frac{n}{2} \rfloor + 1$ points. Rousseeuw (1983) sketched by means of heuristic arguments from Andrews et al. (1972) that in this case the MVE location estimator is $\sqrt[3]{n}$ consistent and converges weakly towards a limiting distribution that is nonnormal. Later this was made rigorous, for instance by Shorack and Wellner (1986) and by Kim and Pollard (1989). Somewhat surprisingly, Grübel (1988a) proved that the univariate MVE estimator of scale converges weakly to a normal distribution at rate $\sqrt{n}$.

For the cases $p \geq 2$, Davies (1987) showed that at an elliptical distribution $P_{\mu,\Sigma}$, the MVE estimators $\mathbf{t}_n$ and $\mathbf{C}_n$ are consistent for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ respectively. Recently, Davies (1989) has also obtained the limiting distribution. Both the location as well as the covariance MVE estimator converge weakly to a limiting distribution that is nonnormal at rate $\sqrt[3]{n}$. This may seem to be in conflict with Grübel's result for the univariate MVE estimator of scale. However, although the MVE covariance estimator converges at rate $\sqrt[3]{n}$, the trace of the covariance estimator turns out to be asymptotically normal at rate $\sqrt{n}$.

5.3 Minimum Covariance Determinant Estimator

To improve the poor rate of convergence of the MVE estimator, Rousseeuw (1983) also considered the minimum covariance determinant (MCD) estimator.

The MCD location estimator is defined as

$$\mathbf{t}_h(\mathbf{X}) = \text{the mean of the } h \text{ points of } \mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \text{ for which the determinant of the sample covariance is minimal.}$$

The corresponding MCD covariance estimator is proportional to the sample covariance of those h points. Typically, h was taken $h = \lfloor \frac{n}{2} \rfloor + 1$, which leads to the same breakdown point $(\lfloor \frac{n}{2} \rfloor - p + 1)/n$ as that of the MVE location estimator at any collection $\mathbf{X}$ in general position. Similar to the MVE estimator, one obtains a breakdown point $\lfloor \frac{n-p+1}{2} \rfloor / n$ at any collection in general position, by taking $h = \lfloor \frac{n+p+1}{2} \rfloor$.

In the univariate case the MCD estimator reduces to the sample mean and sample variance of the h points with the smallest sample variance. Rousseeuw (1983) showed that in this case the location estimator converges weakly to a normal distribution at rate $\sqrt{n}$. The asymptotic efficiency however, for instance at the standard normal, is disappointingly low. Recently, Butler and Juhn (1988) investigated the multivariate version of the MCD estimator and showed that it is asymptotically normal at rate $\sqrt{n}$.

6. Multivariate S-estimators of Location and Scatter

6.1 Introduction

Multivariate M-estimators for location and scatter are affine equivariant robust estimators that possess most of the basic desirable features, such as a $\sqrt{n}$ rate of convergence, a limiting normal distribution, good efficiency and a bounded influence function. Unfortunately, the breakdown point of these estimators is at most $1/(p+1)$, which illustrates their sensitivity for outliers in high dimensional samples. Stahel (1981), Donoho (1982) and Rousseeuw (1983) introduced affine equivariant multivariate estimators for location and scatter that have a high breakdown point for every p (when n is sufficiently large compared to p), but which on the other hand exhibit relatively poor asymptotic properties. It then becomes of interest whether there exist affine equivariant estimators, that are able to combine a high breakdown point and a bounded influence function together with a $\sqrt{n}$ rate of convergence towards a normal distribution and good efficiency. In this thesis three proposals are investigated, two of which will meet this objective. These estimators arise in two different ways.

First, one could try to combine multivariate M-estimators with a high breakdown estimator in a suitable way, such that the resulting procedure inherits the asymptotic properties from the M-estimator and the high breakdown point from the other estimator. In Chapter 7 we will discuss a few methods that are based on M-estimators.

Another way to improve the asymptotic behaviour of a high breakdown estimator such as the ones in Chapter 5, is to smoothen the estimator directly, in such a way that it has the same breakdown behaviour as the nonsmoothed version and such that it has better asymptotic properties. In this chapter we will discuss multivariate S-estimators, which form a class of estimators that are a first step in this direction.

Rousseeuw and Yohai (1984) originally introduced S-estimators in a regression context. How these estimators arise, may be best understood in the special case of estimating univariate location and scale.

Consider the least squares estimator for a univariate location parameter. It is defined by

$$\min_{t \in \mathbf{R}} \sum_{i=1}^{n} (x_i - t)^2. \tag{6.1}$$

On the other hand, consider the univariate MVE estimator, defined by

$$\begin{aligned} &\min_{t \in \mathbf{R}, \sigma > 0} \sigma^2 \\ &\text{subject to } \#\{i : t - \sigma \leq x_i \leq t + \sigma\} = \lfloor \tfrac{n}{2} \rfloor + 1. \end{aligned} \tag{6.2}$$

The key observation is that in both (6.1) and (6.2) one minimizes an M-estimator of scale $\sigma_n^2(t)$ as a function of t. Indeed, to obtain (6.1), first define the 'least squares' M-estimator of scale $\sigma_{\text{LS}}(t)$ as the solution of the equation

$$\frac{1}{n} \sum_{i=1}^{n} \rho_{\text{LS}}\left(\frac{x_i - t}{\sigma}\right) = 1 \tag{6.3}$$

with $\rho_{\text{LS}}(y) = y^2$, and then minimize $\sigma_{\text{LS}}^2(t) = \frac{1}{n} \sum_{i=1}^{n} (x_i - t)^2$ as a function of t. To obtain (6.2), first rewrite the constraint in (6.2) as

$$\frac{1}{n} \sum_{i=1}^{n} \rho_{\text{MVE}}\left(\frac{x_i - t}{\sigma}\right) = 1 - \frac{\lfloor \frac{n}{2} \rfloor + 1}{n} \tag{6.4}$$

with $\rho_{\text{MVE}}(y) = 1 - \{-1 \leq y \leq 1\}$, which is the indicator of the set $(-\infty, -1) \cup (1, \infty)$. Then define the 'minimum volume ellipsoid' M-estimator of scale $\sigma_{\text{MVE}}(t)$ as the solution of (6.4), where t is considered fixed. Minimizing $\sigma_{\text{MVE}}^2(t)$ over $t \in \mathbf{R}$ corresponds with (6.2). Hence, both minimization problems are a special case of

$$\begin{aligned} &\min_{t \in \mathbf{R}, \sigma > 0} \sigma^2 \\ &\text{subject to } \frac{1}{n} \sum_{i=1}^{n} \rho\left(\frac{x_i - t}{\sigma}\right) = b \end{aligned} \tag{6.5}$$

where b is some positive constant. That is, compute an M-estimator $\sigma_n(t)$ of *scale* (hence, the name S-estimators) by solving constraint (6.5), and then minimize $\sigma_n(t)$ over t.

The idea is now, to define estimators of univariate location and scale with a smooth function ρ, which is so to speak 'in between' $\rho_{\text{MVE}}(y)$ and $\rho_{\text{LS}}(y)$, and which is such that the resulting estimators have the same breakdown point as the MVE estimator and have asymptotic properties that are similar to that of the least squares estimator. Rousseeuw and Yohai (1984), aiming for both asymptotic normality and a high breakdown point, proposed to restrict attention to the class of functions ρ that satisfy

(R1) $\rho(0) = 0$, ρ is symmetric and ρ is twice continuously differentiable.

(R2) There exists a finite constant $c > 0$ such that ρ is strictly increasing on $[0, c]$ and constant on $[c, \infty)$.

The letter ρ, as well as the letter ψ for its derivative, was chosen on purpose to emphasize the relation with M-estimators of univariate location and scale that are defined as a solution of (4.7) with $u_1(y) = \psi(y)/y$ and $u_2(y^2) = \rho(y)/(by^2)$. One typically should think of a function ρ that is quadratic in the middle such as ρ_{LS} and which smoothly changes into a constant function such as ρ_{MVE}. A well known example of a function that satisfies these conditions is the biweight ρ-function

$$\rho_{\text{B}}(y;c) = \begin{cases} \dfrac{y^2}{2} - \dfrac{y^4}{2c^2} + \dfrac{y^6}{6c^4} & , \text{ for } |y| \le c \\ \dfrac{c^2}{6} & , \text{ for } |y| \ge c \end{cases} . \tag{6.6}$$

Its derivative is Tukey's biweight function

$$\psi_{\text{B}}(y;c) = \begin{cases} y\left(1 - \left(\dfrac{y}{c}\right)^2\right)^2 & , \text{ for } |y| \le c \\ 0 & , \text{ for } |y| \ge c \end{cases} .$$

This ψ-function was originally proposed by Tukey in the context of (redescending) M-estimators for univariate location.

6.2 Multivariate S-estimators

Multivariate S-estimators are defined in a similar way as described in the previous section, as a smoothed version of the multivariate MVE estimator. Recall that the volume of an ellipsoid is an increasing function of the determinant of the corresponding scatter matrix.

DEFINITION 6.1: Multivariate S-estimators for location and scatter are defined as the vector $\mathbf{t}_n$ and the positive definite symmetric matrix $\mathbf{C}_n$ that minimize the determinant $|\mathbf{C}|$ of the matrix $\mathbf{C}$, subject to

$$\frac{1}{n}\sum_{i=1}^{n} \rho\big(d(\mathbf{x}_i, \mathbf{t}, \mathbf{C})\big) = b \tag{6.7}$$

where $\rho : \mathbf{R} \to [0, \infty)$ is nondecreasing and b is a constant such that $0 < b < \sup \rho$.

In other words, S-estimators are the center and scatter matrix of the smallest ellipsoid that satisfies constraint (6.7). The MVE estimator is obtained with $\rho_{\text{MVE}}(y)$, as defined in the previous section, and with $b = 1 - \lfloor \frac{n+p+1}{2} \rfloor / n$. The sample mean and sample covariance are obtained with $\rho_{\text{LS}}(y) = y^2$ and $b = p$ (see for instance Grübel 1988b for $p \geq 2$). When the S-estimators are uniquely defined, they are obviously affine equivariant; if they are not uniquely defined they will always be affine equivariant in a similar way as nonuniquely defined multivariate M-estimators (see Section 4.1.1).

Davies (1987) first investigated some properties of multivariate S-estimators. His definition was slightly different, but is essentially equivalent to Definition 6.1. Instead of a nondecreasing function $\rho : \mathbf{R} \to [0, \infty)$, he considered a nonincreasing function $\kappa : [0, \infty) \to [0, 1]$, and defined S-estimators by minimizing $|\mathbf{C}|$ subject to

$$\frac{1}{n} \sum_{i=1}^{n} \kappa\big((\mathbf{x}_i - \mathbf{t})^T \mathbf{C}^{-1} (\mathbf{x}_i - \mathbf{t})\big) \geq 1 - \varepsilon \tag{6.8}$$

where ε is a constant between 0 and 1. Note that if κ is continuous, the solutions of this minimization problem satisfy (6.8) with equality. Hence, this definition is equivalent with Definition 6.1 for functions ρ that are continuous and bounded, i.e. $a = \sup \rho < \infty$, by taking

$$\kappa(y^2) = 1 - \frac{\rho(y)}{a}$$

and $\varepsilon = b/a$.

To obtain both a high breakdown point and asymptotic normality, the function ρ in Definition 6.1 is assumed to satisfy conditions (R1) and (R2). We prefer the function ρ in Definition 6.1 instead of κ because of a few minor reasons. For one thing, Definition 6.1 is a straightforward generalization of regression S-estimators as defined in Rousseeuw and Yohai (1984). Moreover, the letter ρ is chosen on purpose to emphasize the relation with multivariate M-estimators, defined by minimizing

$$\sum_{i=1}^{n} \rho\big(d(\mathbf{x}_i, \mathbf{m}, \mathbf{\Sigma})\big) \tag{6.9}$$

where $\mathbf{\Sigma}$ is the underlying covariance. Finally, with the function ρ it is easier to see that the sample mean and sample covariance may arise as limiting cases of S-estimators defined by a function ρ that satisfies (R1) and (R2). For instance, if we use the biweight ρ-function of (6.7) and let $c \to \infty$, we obtain ρ_{LS}, whereas the corresponding function κ tends to 1.

Since S-estimators are a smoothed version of the MVE estimator it is no surprise that its breakdown behaviour is similar. Davies (1987) extended the breakdown result for the MVE estimator to S-estimators that are zero in a neighbourhood of the origin. This result is complemented in Lopuhaä and Rousseeuw (1989) to encompass well known ρ-functions such as the biweight function of (6.6). The good breakdown properties basically follow from constraint (6.7), which guarantees a

sufficient number of points, namely $n - \lfloor nr \rfloor$ (where $r = b/a = b/\sup\rho$), inside the 'S-ellipsoid' and safeguards the covariance S-estimator against implosion, and from the fact that one minimizes a loss function $|\mathbf{C}|$ that is an increasing function of the magnitude of $\mathbf{C}$, which safeguards the covariance S-estimator against explosion. That the location S-estimator must also stay within bounds is then an immediate consequence. It turns out that the breakdown point depends on the constant b in (6.7), i.e.

$$\varepsilon^*(\mathbf{t}_n, \mathbf{X}) = \varepsilon^*(\mathbf{C}_n, \mathbf{X}) = \frac{\lceil nr \rceil}{n} \tag{6.10}$$

at any collection $\mathbf{X}$ in general position. For $b = \frac{n-p}{2n}a$, the S-estimators will have a breakdown point that attains the maximal possible value for covariance estimators.

Davies (1987) also proved existence and consistency for a class of S-estimators that was large enough to contain the MVE estimator, and sketched a proof for asymptotic normality under the assumption of κ having a third derivative. These results are extended in Lopuhaä (1989) to S-functionals corresponding to the estimators defined in Definition 6.1, i.e. $\mathbf{t}(P)$ and $\mathbf{C}(P)$ defined by minimizing $|\mathbf{C}|$ subject to

$$\int \rho\big(d(\mathbf{x}, \mathbf{t}, \mathbf{C})\big)\, dP(\mathbf{x}) = b \tag{6.11}$$

where the function ρ must satisfy (R1) and (R2). S-estimators defined with such a function ρ converge weakly to a normal distribution at rate $\sqrt{n}$, have a bounded influence function and have a high breakdown point for suitable choices of ρ.

Another interesting feature of these estimators is that (in addition to having a high breakdown point), their asymptotic behaviour is similar to that of multivariate M-estimators. In Lopuhaä (1989) it is shown that S-estimators are a solution of the simultaneous equations

$$\begin{aligned} &\sum_{i=1}^{n} u\big(d(\mathbf{x}_i, \mathbf{t}, \mathbf{C})\big)(\mathbf{x}_i - \mathbf{t}) = \mathbf{0} \\ &\sum_{i=1}^{n} \Big\{ pu\big(d(\mathbf{x}_i, \mathbf{t}, \mathbf{C})\big)(\mathbf{x}_i - \mathbf{t})(\mathbf{x}_i - \mathbf{t})^T - v\big(d(\mathbf{x}_i, \mathbf{t}, \mathbf{C})\big)\mathbf{C} \Big\} = \mathbf{0} \end{aligned} \tag{6.12}$$

where $u(y) = \psi(y)/y$, $v(y) = \psi(y)y - \rho(y) + b$, ψ is the derivative of ρ and b is the constant in (6.7). An important consequence of this is, that it shows that at least for some type of M-estimator score equations (4.7), there exist a solution which has a high breakdown point. In this context, S-estimators may be interpreted as a method to find a high breakdown solution of equations (6.12).

One should emphasize here, that S-estimators are only *a* solution of equations (6.12) among *other* solutions. Because equations (6.12) are a special case of M-estimator type of score equations (4.7), they will have at least one solution that has a breakdown point of at most $1/(p+1)$. The S-estimators are just one specific

solution of (6.12) that has a *high* breakdown point. To find it, one should therefore not try to solve (6.12) by means of some kind of fixed point or Newton-Raphson iteration process. By solving the minimization problem of Definition 6.1, a high breakdown solution of (6.12) is guaranteed. Nevertheless, S-estimators do satisfy (6.12) and hence, the asymptotic properties and the type of influence function of S-estimators are similar to those of multivariate M-estimators. In particular, if the underlying distribution is elliptical with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, the limiting distribution and the influence function of the location S-estimator are the same as those of the location M-estimator $\mathbf{m}_n$, defined by minimizing (6.9) using the same function ρ as in (6.7).

So far, it seems that S-estimators meet the objective formulated at the beginning of this chapter. However, it turns out that it is not possible to combine a high breakdown point with a good efficiency. To compare the asymptotic efficiency of S-estimators with that of the sample mean and sample covariance, one could consider an elliptical underlying distribution $P_{\mu,\Sigma}$. In this case a natural choice for b is

$$b = \int \rho\big(d(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})\big)\, dP_{\mu,\Sigma}(\mathbf{x}) = \int \rho(\|\mathbf{x}\|) f(\|\mathbf{x}\|)\, d\mathbf{x}. \tag{6.13}$$

This will still be a function of the tuning constant c of the function ρ, but for any value of c this choice for b guarantees consistency of $\mathbf{t}_n$ and $\mathbf{C}_n$ for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ respectively. It also means that the breakdown point in (6.10) will only depend on c. The asymptotic variances of the S-estimators will also be a function of c. In Lopuhaä (1989), S-estimators are investigated that are defined by the biweight function $\rho_{\text{B}}(y; c)$. For $c \to \infty$, the asymptotic efficiency relative to the sample mean and the sample variance tends to 1. Intuitively this is clear, since ρ_{LS} is the limiting case of $\rho_{\text{B}}(y; c)$ for $c \to \infty$. Unfortunately, a high breakdown point corresponds with small values of c. Hence, one still has to make a tradeoff between breakdown point and efficiency : a high breakdown point is counterbalanced by a low efficiency and vice versa. In Chapter 8 we will discuss a generalization of multivariate S-estimators that will enable us to avoid this tradeoff.

7. Using *M*-estimators to improve high Breakdown Estimators

An alternative method to improve the poor asymptotic properties of high breakdown estimators, is to use multivariate M-estimators in combination with a high breakdown estimator, in such a way that the resulting estimator retains the high breakdown point and improves the asymptotic properties. In this chapter we will discuss a few proposals.

7.1 Affinely scaled Location *M*-estimators

Consider univariate location M-estimators defined by minimizing (4.1) with $\rho(x,m) = \rho(x-m)$, where ρ is symmetric. A straightforward generalization to M-estimators of multivariate location would be to minimize

$$\sum_{i=1}^{n} \rho(\|\mathbf{x}_i - \mathbf{m}\|) \tag{7.1}$$

over $\mathbf{m}$ in $\mathbf{R}^p$. However, defined as such, the estimator would not be affine equivariant. To obtain affine equivariance together with a high breakdown point, a sort of scaled location M-estimator will be proposed in this section, where we use an affine equivariant covariance estimator with a high breakdown point to perform the scaling. This method is investigated in Lopuhaä (1988).

7.1.1 Weakly Redescending

Consider a random sample $X_1, X_2, \ldots, X_n$ from an elliptical distribution with a location parameter μ. Minimizing (7.1) would be sensible if the sample distribution

is spherically symmetric around μ. When the sample distribution has some known covariance structure $\Sigma = BB^T$, one could still make a sensible use of (7.1) to estimate μ. Indeed, if we scale the observations according to $X_i \longmapsto B^{-1}X_i$, the scaled observations are a sample from a distribution that is again spherically symmetric around μ. Hence, it would still make sense to estimate μ from the scaled sample by minimizing

$$\sum_{i=1}^{n} \rho(d(X_i, m, \Sigma))$$

over m in R^p. Often, Σ will be unknown, so that this procedure can not be performed in practice. However, instead of the true unknown Σ, we could use an affine equivariant estimator of it. This leads to the following proposal.

First use an affine equivariant covariance estimator $C_n = A_n A_n^T$ with a high breakdown point to scale the observations

$$x_1, \ldots, x_n \longmapsto A_n^{-1}x_1, \ldots, A^{-1}x_n.$$

Then compute an M-estimator of location by minimizing (7.1) based on the scaled observations $A_n^{-1}x_1, \ldots, A_n^{-1}x_n$. Finally, rescale the resulting M-estimator $m_n = m(A_n^{-1}x_1, \ldots, A_n^{-1}x_n)$, i.e.

$$t_n = t(x_1, \ldots, x_n) = A_n m_n.$$

Another way to formulate this, is to define the scaled location M-estimator directly, as the vector t_n that minimizes

$$\sum_{i=1}^{n} \rho(d(x_i, t, C_n)) \tag{7.2}$$

over t in R^p.

When minimization of (7.2) yields a unique solution t_n, this solution will be affine equivariant. If t_n is not uniquely defined, then the estimator is still affine equivariant in a similar sense as the nonuniquely defined multivariate M- and S-estimators.

When ρ is a symmetric function that *increases to* ∞ at a moderate rate towards both sides, then the scaled location M-estimator will inherit the breakdown point of the covariance estimator C_n. The main reason for this is that for such functions ρ, the unscaled location M-estimator defined by minimizing (7.1), has a breakdown point $\lfloor \frac{n+1}{2} \rfloor / n$ at any collection X (Huber 1984). Therefore, if we replace a sufficiently small fraction of a collection X, such that the covariance estimator C_n does not break down, the estimator m_n will stay within finite bounds, and hence the scaled location M-estimator t_n will stay within finite bounds.

When the covariance functional $C(\cdot)$, that corresponds with the estimator C_n, is continuous at a distribution P, the functional $t(\cdot)$ corresponding with t_n will also be continuous at P. Under further conditions on P, the influence function of

$\mathbf{t}_n$ exists. These conditions are weak enough to include elliptical distributions. In the special case of a spherically symmetric distribution P, the influence function of the scaled location M-estimator is the same as that of the unscaled location M-estimator defined by minimizing (7.1) using the same function ρ :

$$\text{IF}(\mathbf{x};\mathbf{t},P) = \frac{\psi(\|\mathbf{x}\|)}{\beta\|\mathbf{x}\|}\mathbf{x} \tag{7.3}$$

where β is a positive constant that depends on the function ρ and ψ is the derivative of ρ.

This behaviour is also typical for the asymptotic properties of $\mathbf{t}_n$. When the covariance estimator is consistent, the scaled location M-estimator will converge at rate $\sqrt{n}$. Under further conditions on the distribution P, which include all elliptical distributions with a finite second moment, the scaled location M-estimator is asymptotically normal. In the special case of a spherically symmetric distribution, the limiting variance of the scaled location M-estimator is the same as that of the corresponding unscaled location M-estimator. One should emphasize here, that these results hold regardless of the rate of convergence of the covariance estimator $\mathbf{C}_n$.

This enables us to combine a high breakdown point, inherited from the initial covariance estimator, with a bounded influence function, a $\sqrt{n}$ rate of convergence towards a normal distribution and good efficiency, given by the M-estimator. The conditions on the function ρ are weak enough to include the function

$$\rho_{\text{H}}(y;k) = \begin{cases} -\frac{1}{2}k^2 + ky & \text{, for } y \geq k \\ \frac{1}{2}y^2 & \text{, for } |y| \leq k \\ -\frac{1}{2}k^2 - ky & \text{, for } y \leq -k \end{cases} \tag{7.4}$$

corresponding with Huber's location M-estimator (Section 4.1). When we use $\rho_{\text{H}}(y;k)$ in (7.1), the asymptotic efficiency of the unscaled location M-estimator relative to the maximum likelihood estimator is known to be reasonable over a broad class of distributions (Maronna 1976), and its influence function in (7.3) is bounded. Hence, a typical example for the scaled location M-estimator would be to use the MVE covariance estimator in combination with the function $\rho_{\text{H}}(y;k)$ of (7.4). The resulting affinely scaled location M-estimator has breakdown point $\lfloor\frac{n-p+1}{2}\rfloor/n$, has a bounded influence function and converges weakly to a normal distribution at rate $\sqrt{n}$, where the efficiency relative to the maximum likelihood estimator is good over a broad class of distributions.

Since ρ must be increasing towards both sides, the influence function in (7.3) will be nonzero except for $\mathbf{x} = \mathbf{0}$. When $\|\mathbf{x}\|$ tends to ∞, $\text{IF}(\mathbf{x};\mathbf{t},P)$ will stay bounded; it may redescend to zero, although it will never be equal to zero. Similar to the function ψ that determines $\text{IF}(\mathbf{x};\mathbf{t},P)$, we call such influence functions ***weakly redescending***, to distinguish them from influence functions that actually are **0** for $\|\mathbf{x}\| \geq c$, which are called ***strongly redescending*** (see Section 4.1). Some people

prefer estimators with a strongly redescending influence function, because of the fact that beyond a certain boundary outliers will no longer have any effect on such an estimator.

In our situation a strongly redescending influence function would correspond with a function $\rho(y)$ in (7.2) that is constant for $|y| \geq c$. However, it is in general not possible to use a bounded function ρ in (7.2) and at the same time retain the high breakdown point of the initial covariance estimator. The reason for this is that the breakdown point at a collection $\mathbf{X}$, of location M-estimators defined by minimizing (7.1) using a bounded function ρ, will depend on the actual structure of the collection $\mathbf{X}$ (see Section 3.1). Nevertheless, when we use a specific covariance S-estimator, it is possible to construct an affinely scaled location M-estimator with an influence function that vanishes for $\|\mathbf{x}\| \geq c$. We discuss this proposal in the next subsection.

7.1.2 Strongly Redescending

Let $\rho_1 : \mathbf{R} \to [0, \infty)$ and $\rho_2 : \mathbf{R} \to [0, \infty)$ both satisfy (R1) and (R2) as defined in Chapter 6, and let $\rho_2(\cdot)$ be related to $\rho_1(\cdot)$ as follows :

$$\rho_1(y) \geq \rho_2(y) \tag{7.5}$$

and

$$\rho_1(c_1) = \sup \rho_1 = \sup \rho_2 = \rho_2(c_2).$$

Let $\mathbf{t}_{1,n} = \mathbf{t}_1(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ and $\mathbf{C}_{1,n} = \mathbf{C}_1(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ be the S-estimators defined with the function $\rho_1(\cdot)$ and constant b_1 in (6.7). Define the scaled location M-estimator as a vector $\mathbf{t}_{2,n}$ that minimizes

$$\sum_{i=1}^{n} \rho_2\big(d(\mathbf{x}_i, \mathbf{t}, \mathbf{C}_{1,n})\big) \tag{7.6}$$

over $\mathbf{t}$ in $\mathbf{R}^p$. In Lopuhaä (1988) a more general definition is given, which can be seen as the multivariate version of Yohai's (1987) regression MM-estimators. The definition given above is a special case, which suffices for our purposes.

Because the function ρ_2 is not convex, the scaled location M-estimator $\mathbf{t}_{2,n}$ will in general not be uniquely defined. However, it will always be affine equivariant in the same sense as nonuniquely defined M- and S-estimators.

For the breakdown behaviour of $\mathbf{t}_{2,n}$, recall that unscaled location M-estimators defined by minimizing (7.1) with a bounded function ρ have a breakdown point that depends on the structure of a collection $\mathbf{X}$ (Section 3.1.1). The closer the points of the collection are, compared to the width $2c$ of a function ρ that satisfies (R1) and (R2), the larger the breakdown point will be. In our situation, the fact that one minimizes (7.6) implies that

$$\sum_{i=1}^{n} \rho_2\big(d(\mathbf{x}_i, \mathbf{t}_{2,n}, \mathbf{C}_{1,n})\big) \leq \sum_{i=1}^{n} \rho_2\big(d(\mathbf{x}_i, \mathbf{t}_{1,n}, \mathbf{C}_{1,n})\big).$$

Together with condition (7.5) and (6.7) this means that

$$\frac{1}{n}\sum_{i=1}^{n}\rho_2\big(d(\mathbf{x}_i, \mathbf{t}_{2,n}, \mathbf{C}_{1,n})\big) \leq b_1$$

which, similar to multivariate S-estimators, will guarantee a sufficient number of points, namely $n - \lfloor nr_1 \rfloor$ (where $r_1 = b_1/a_1 = b_1/\sup\rho_1$), inside the ellipsoid $E(\mathbf{t}_{2,n}, \mathbf{C}_{1,n}, c_2)$. This means that at least $n - \lfloor nr_1 \rfloor$ points will be more or less close enough together compared to the width ρ_2. In fact, when the amount of contamination is sufficiently small so that the S-estimators do not break down, this property will also force at least one point $\mathbf{x}_i$ in the ellipsoid $E(\mathbf{t}_{2,n}, \mathbf{C}_{1,n}, c_2)$ that is not replaced. Hence, the estimator $\mathbf{t}_{2,n}$ must stay within finite bounds. Along these lines one may show rigorously that every $\mathbf{t}_{2,n}$ that minimizes (7.6), inherits the breakdown point of the initial S-estimator.

The results on the type of influence function and asymptotic normality are derived in a similar fashion as for the scaled location M-estimator that is discussed in the previous section. Again the type of influence function and the limiting distribution are the same as those of the unscaled location M-estimator defined by minimizing (7.1) with the bounded function ρ_2. In particular, the influence function at a spherically symmetric distribution will be of type (7.3), with $\psi = \psi_2$, and is therefore strongly redescending.

A typical example would be to use an S-estimator defined with the biweight function $\rho_1(y) = \rho_{\text{B}}(y; c_1)$ of (6.6), where c_1 is chosen small such that the S-estimator has a high breakdown point (see Chapter 6), together with another biweight function $\rho_2(y) = \rho_{\text{B}}(y; c_2)$ in (7.6), where $c_2 > c_1$. In case of a spherically symmetric underlying distribution, the limiting distribution is the same as that of the corresponding unscaled location M-estimator, and as that of the location S-estimator defined with the function ρ_2. We recall from Chapter 6 that for large values c_2 this gives good efficiency relative to the sample mean. Since the breakdown point does not depend on the function ρ_2, we would obtain an affinely scaled location M-estimator with a high breakdown point and a *strongly redescending* influence function, and which converges weakly to a normal distribution at rate $\sqrt{n}$, with a good efficiency relative to the sample mean.

7.2 One-step Reweighted Estimators

Instead of using robust estimators to estimate location and scatter directly, one could also use them as a diagnostic tool to *detect* outliers in the data. Once these points have been detected, one could downweight them or remove them completely, and then use a more efficient procedure on the remaining points, such as the classical sample mean and sample covariance.

To 'clean' the data, it has been proposed to identify outlying points $\mathbf{x}_i$ by means of the Mahalanobis distance with respect to the classical sample mean $\bar{\mathbf{x}}$ and the

sample covariance $\mathbf{C}_{sc}$

$$\sqrt{(\mathbf{x}_i - \bar{\mathbf{x}})\mathbf{C}_{sc}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})}. \tag{7.7}$$

Observations with relatively large distances would then be identified as outliers. However, the poor robustness of the sample mean and sample covariance will make this procedure useless if more than one outlier is present. Even worse, when multiple outliers are present, it is possible that the wrong points are flagged.

The Stahel-Donoho estimator can be seen as a robustification of such a procedure. Instead of using the Mahalanobis distance (7.7), outlying points are identified as points with a relatively large degree of outlyingness r_i. If one would then take an indicator function of an interval $[0, c]$ for the function $w(\cdot)$, then the Stahel-Donoho estimator would be the sample mean and sample covariance of all points with an outlyingness of at most c. However, note that for such functions $w(\cdot)$ the breakdown point in (5.3) is not valid.

In Lopuhaä and Rousseeuw (1989) a similar weighting procedure is investigated, that is based on the Mahalanobis distances with respect to an initial high breakdown estimator. Let $\mathbf{t}_{0,n} = \mathbf{t}_0(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ and $\mathbf{C}_{0,n} = \mathbf{C}_0(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ be estimators of location and scatter, where we would typically take estimators with a high breakdown point. Consider the Mahalanobis distances with respect to these estimators

$$d(\mathbf{x}_i, \mathbf{t}_{0,n}, \mathbf{C}_{0,n}) = \sqrt{(\mathbf{x}_i - \mathbf{t}_{0,n})\mathbf{C}_{0,n}^{-1}(\mathbf{x}_i - \mathbf{t}_{0,n})}.$$

The idea is, that high breakdown estimators will reflect the structure of the majority of the observations, such that if there are outliers present, they can identified by means of their relatively large $d(\mathbf{x}_i, \mathbf{t}_{0,n}, \mathbf{C}_{0,n})$. Let $w : [0, \infty) \to [0, \infty)$ be some weight function, and define a weighted sample mean and sample covariance

$$\begin{aligned} \mathbf{t}_{1,n}(\mathbf{X}) &= \frac{\sum_{i=1}^n w\big(d(\mathbf{x}_i, \mathbf{t}_{0,n}, \mathbf{C}_{0,n})\big)\mathbf{x}_i}{\sum_{i=1}^n w\big(d(\mathbf{x}_i, \mathbf{t}_{0,n}, \mathbf{C}_{0,n})\big)} \\ \mathbf{C}_{1,n}(\mathbf{X}) &= \frac{\sum_{i=1}^n w\big(d(\mathbf{x}_i, \mathbf{t}_{0,n}, \mathbf{C}_{0,n})\big)\big(\mathbf{x}_i - \mathbf{t}_{1,n}(\mathbf{X})\big)\big(\mathbf{x}_i - \mathbf{t}_{1,n}(\mathbf{X})\big)^T}{\sum_{i=1}^n w\big(d(\mathbf{x}_i, \mathbf{t}_{0,n}, \mathbf{C}_{0,n})\big)}. \end{aligned} \tag{7.8}$$

To keep the weighted estimators within finite bounds after contamination of some of the $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, we need some boundedness condition on the function w. To safeguard the weighted covariance estimator against implosion, we have to ensure that $w\big(d(\mathbf{x}_i, \mathbf{t}_{0,n}, \mathbf{C}_{0,n})\big) > 0$ for a sufficient number of points. Hence, we will assume that the function w satisfies

(W1) $w(y)$ and $w(y)y^2$ are bounded and $w(y)$ is nonincreasing.
(W2) There exists a c_0 such that $w(y) > 0$ for $y \in [0, c_0]$.

In addition, the constant c_0 must satisfy the following relation with the initial estimators $\mathbf{t}_{0,n}$ and $\mathbf{C}_{0,n}$. Consider the (robust) ellipsoid

$$E(\mathbf{t}_{0,n}, \mathbf{C}_{0,n}, c_0) \tag{7.9}$$

consisting of all points $\mathbf{x}$ with Mahalanobis distance $d(\mathbf{x}, \mathbf{t}_{0,n}, \mathbf{C}_{0,n})$ less than c_0. The constant c_0 must be such that (7.9) contains at least $\lfloor \frac{n+p+1}{2} \rfloor$ points. Given the estimators $\mathbf{t}_{0,n}$ and $\mathbf{C}_{0,n}$, one could take for c_0 any number that is greater than the $\lfloor \frac{n+p+1}{2} \rfloor$ largest Mahalanobis distance $d(\mathbf{x}_i, \mathbf{t}_{0,n}, \mathbf{C}_{0,n})$. However, typical choices for $\mathbf{t}_{0,n}$ and $\mathbf{C}_{0,n}$ would be the MVE estimator or an S-estimator with $r = (n-p)/(2n)$ (see Chapter 6). In that case, c_0 does not have to be defined by means of the sample values and could be any number that is greater than the tuning constant c either of ρ_{MVE} or of a function ρ that satisfies (R1) and (R2). Indeed, by definition this constant c will always be such that the minimum volume ellipsoid, or the 'S-ellipsoid' contains at least $\lfloor \frac{n+p+1}{2} \rfloor$ points.

A typical choice for $w(\cdot)$ is the indicator function of the interval $[0, c_1]$, where $c_1 \geq c_0$, in which case $\mathbf{t}_{1,n}$ and $\mathbf{C}_{1,n}$ are simply the sample mean and sample covariance of all points inside the ellipsoid

$$E(\mathbf{t}_{0,n}, \mathbf{C}_{0,n}, c_1)$$

which is of the same shape as the ellipsoid in (7.9). Here c_1 could be some quantile such that the 'true' ellipsoid $E(\boldsymbol{\mu}, \boldsymbol{\Sigma}, c_1)$ has high probability, and beyond which every point is identified as outlier.

Since the Mahalanobis distance with respect to affine equivariant estimators is invariant under affine transformations, it is immediately clear that affine equivariance of $\mathbf{t}_{0,n}$ and $\mathbf{C}_{0,n}$ carries over to $\mathbf{t}_{1,n}$ and $\mathbf{C}_{1,n}$. In Lopuhaä and Rousseeuw (1989) it is shown that under the conditions above, the weighted estimators also inherit the breakdown point of the initial estimators at any collection $\mathbf{X}$ in general position.

The asymptotic properties for these estimators are still under investigation. It seems that if the initial estimators are consistent, the weighted estimators are also consistent. However, weighting as defined in (7.8) does not seem to improve the rate of convergence. If one starts with estimators that have a rate of convergence which is slower than $\sqrt{n}$, such as the MVE estimator, the rate of convergence will stay the same. If, in addition, the initial estimators have a limiting distribution, the limiting distribution of the weighted estimators will be essentially the same. When one starts with a $\sqrt{n}$ consistent estimator, such as a smooth S-estimator, the weighted estimators will converge weakly to a normal distribution at rate $\sqrt{n}$. It is likely that the asymptotic efficiency relative to the sample mean and sample covariance will be improved considerably.

The weighted estimators as defined above are closely related to multivariate M-estimators that are a solution of equations (4.6). Solutions of this equation may be

found by means of a fixed point algorithm. Note that equations (4.6) can also be written as

$$\begin{aligned} \mathbf{t} &= \frac{\sum_{i=1}^{n} u_1\big(d(\mathbf{x}_i, \mathbf{t}, \mathbf{C})\big)\mathbf{x}_i}{\sum_{i=1}^{n} u_1\big(d(\mathbf{x}_i, \mathbf{t}, \mathbf{C})\big)} \\ \mathbf{C} &= \frac{\sum_{i=1}^{n} u_2\big(d^2(\mathbf{x}_i, \mathbf{t}, \mathbf{C})\big)(\mathbf{x}_i - \mathbf{t})(\mathbf{x}_i - \mathbf{t})^T}{\sum_{i=1}^{n} u_2\big(d^2(\mathbf{x}_i, \mathbf{t}, \mathbf{C})\big)}. \end{aligned} \tag{7.10}$$

The weighted estimators are almost the same as the estimators that one obtains by performing one iteration of this fixed point algorithm starting with initial high breakdown estimators $\mathbf{t}_{0,n}$ and $\mathbf{C}_{0,n}$, i.e. define

$$\begin{aligned} \mathbf{t}_{1,n} &= \frac{\sum_{i=1}^{n} u_1\big(d(\mathbf{x}_i, \mathbf{t}_{0,n}, \mathbf{C}_{0,n})\big)\mathbf{x}_i}{\sum_{i=1}^{n} u_1\big(d(\mathbf{x}_i, \mathbf{t}_{0,n}, \mathbf{C}_{0,n})\big)} \\ \mathbf{C}_{1,n} &= \frac{\sum_{i=1}^{n} u_2\big(d^2(\mathbf{x}_i, \mathbf{t}_{0,n}, \mathbf{C}_{0,n})\big)(\mathbf{x}_i - \mathbf{t}_{0,n})(\mathbf{x}_i - \mathbf{t}_{0,n})^T}{\sum_{i=1}^{n} u_2\big(d^2(\mathbf{x}_i, \mathbf{t}_{0,n}, \mathbf{C}_{0,n})\big)}. \end{aligned}$$

When we take $u_1(y) = u_2(y^2) = w(y)$, the estimators defined in (7.10) are almost the same as the weighted estimators defined in (7.8). Hence, we can interpret the weighted estimator as a sort of one-step M-estimator, based on a fixed point algorithm, with redescending $\psi_1(y) = w(y)y$ and $\psi_2(y^2) = w(y)y$.

David Tyler (personal communication) independently proposed (7.10) as a version of a one-step M-estimator. In the next section we will discuss the version that one usually has in mind if the term 'one-step M-estimator' is used. Tyler also provided a proof for the finite sample (addition) breakdown point of the one-step M-estimator as defined in (7.10) with *monotone* functions $\psi_1(y) = u_1(y)y$ and $\psi_2(y) = u_2(y)y$. However, this does not include functions w which are strongly redescending. The asymptotic behaviour of these one-step M-estimators is not yet investigated. However, the obvious conjecture is that the limiting behaviour will be the same as that of the weighted estimators defined in (7.8).

7.3 One-step Newton-Raphson M-estimators

Building on precursor ideas of Fisher, Neyman and others (Le Cam 1956), Bickel (1975) investigated an estimator that is defined as a one-step Newton-Raphson iteration of the M-estimator equation (4.2) (seen as a function of $\boldsymbol{\theta}$) starting from some initial estimator $\boldsymbol{\theta}_{0,n}$. Bickel studied these estimators in the linear model, and showed that the limiting behaviour of the one-step M-estimator is the same as that of the actual M-estimator, defined as a solution of (4.2).

For estimation of univariate location, the idea is the following. Let $t_{0,n}$ be some initial estimator of location and assume that $t_{0,n} \to 0$. Let ψ be a sufficiently

smooth function and be such that $E\psi(X_1) = 0$. Define the *one-step M-estimator* of univariate location by performing one Newton-Raphson iteration of (4.4) starting in $t_{0,n}$, i.e.

$$t_{1,n} = t_{0,n} + \frac{\sum_{i=1}^n \psi(X_i - t_{0,n})}{\sum_{i=1}^n \psi'(X_i - t_{0,n})}.$$

Then the estimator $t_{1,n}$ will have the same limiting behaviour as the actual fully iterated location M-estimator, which is a solution of (4.4) with the same function ψ. This can be made plausible as follows :

$$\begin{aligned} t_{1,n} &= t_{0,n} + \frac{\sum_{i=1}^n \psi(X_i - t_{0,n})}{\sum_{i=1}^n \psi'(X_i - t_{0,n})} \\ &= t_{0,n} + \frac{\sum_{i=1}^n \psi(X_i) - \sum_{i=1}^n \psi'(X_i)\, t_{0,n} + \text{ remainder}}{\sum_{i=1}^n \psi'(X_i - t_{0,n})} \\ &= t_{0,n} + \frac{\frac{1}{n}\sum_{i=1}^n \psi(X_i)}{\frac{1}{n}\sum_{i=1}^n \psi'(X_i)} - t_{0,n} + \text{ remainder} \qquad (7.11) \\ &= \frac{\frac{1}{n}\sum_{i=1}^n \psi(X_i)}{E\psi'(X_1)} + \text{ remainder} \qquad (7.12) \end{aligned}$$

The terms with $t_{0,n}$ in (7.11) cancel, and it remains to show that the remainder term in (7.12) is $o_P(1/\sqrt{n})$. Bickel (1975) proved this for a one-step M-estimator in the linear model, assuming that the initial estimator $t_{0,n}$ converges at rate $\sqrt{n}$. However, Rousseeuw and Leroy (1987, p.130) mention that according to personal communication with Bickel, this still holds when the rate of convergence of $t_{0,n}$ is better than $\sqrt[4]{n}$.

What makes this procedure particularly interesting for our purposes is, that if the one-step Newton-Raphson iteration retains the breakdown point of the initial estimator, it can be used as a method to improve the asymptotic properties of a high breakdown estimator. In the multivariate case we consider equations (4.7). Let us write these equations briefly as

$$\sum_{i=1}^n \boldsymbol{\Psi}(\mathbf{x}_i, \boldsymbol{\theta}) = \mathbf{0} \qquad (7.13)$$

where the parameter $\boldsymbol{\theta}$ is the pair $(\mathbf{t}, \mathbf{C})$, and where for every $\boldsymbol{\theta}$, $\boldsymbol{\Psi}(\cdot, \boldsymbol{\theta})$ is a mapping from $\mathbf{R}^p$ to $\mathbf{R}^p \times \mathrm{PDS}(p)$. Let $\boldsymbol{\theta}_{0,n} = (\mathbf{t}_{0,n}, \mathbf{C}_{0,n})$ be affine equivariant estimators of location and scatter. Of course, we typically think of estimators with a high breakdown point.

Similar to the univariate location case, the *one-step M-estimator for location and scatter* is the pair $\boldsymbol{\theta}_{1,n} = (\mathbf{t}_{1,n}, \mathbf{C}_{1,n})$, defined as a one-step Newton-Raphson iteration of (7.13), i.e.

$$\mathbf{0} = \sum_{i=1}^n \boldsymbol{\Psi}(\mathbf{x}_i, \boldsymbol{\theta}_{0,n}) + \mathbf{D}_{0,n}(\boldsymbol{\theta}_{1,n} - \boldsymbol{\theta}_{0,n}) \qquad (7.14)$$

where $D_{0,n}$ is the derivative

$$D_{0,n} = \frac{\partial}{\partial \theta}\left[\sum_{i=1}^{n} \Psi(\mathbf{x}_i, \theta)\right]_{\theta=\theta_{0,n}}. \tag{7.15}$$

If (7.15) is nonsingular, then

$$\theta_{1,n} = \theta_{0,n} - D_{0,n}^{-1}\left(\sum_{i=1}^{n} \Psi(\mathbf{x}_i, \theta_{0,n})\right). \tag{7.16}$$

The idea remains the same as in the univariate location case, except that the expression in (7.16) becomes much more complicated. It seems especially difficult to study the finite sample breakdown behaviour of these estimators; in particular to find sufficient conditions under which $D_{0,n}$ remains nonsingular if a part of the sample is contaminated. There are some variations on the definition given above, which make things a little easier.

For instance, one could consider a *location one-step M* and a *covariance one-step M* separately. For the one-step location M-estimator consider the first equation of (4.7) as a function of $\mathbf{t}$ and define $\mathbf{t}_{1,n}$ by

$$\mathbf{0} = \sum_{i=1}^{n} v_1\big(d(\mathbf{x}_i, \mathbf{t}_{0,n}, C_{0,n})\big)(\mathbf{x}_i - \mathbf{t}_{0,n}) + M_{0,n}(\mathbf{t}_{1,n} - \mathbf{t}_{0,n}) \tag{7.17}$$

where $M_{0,n}$ is the matrix

$$\begin{aligned} M_{0,n} &= \frac{\partial}{\partial \mathbf{t}}\left[\sum_{i=1}^{n} v_1\big(d(\mathbf{x}_i, \mathbf{t}, C_{0,n})\big)(\mathbf{x}_i - \mathbf{t})\right]_{\mathbf{t}=\mathbf{t}_{0,n}} \\ &= -\sum_{i=1}^{n} \frac{v_1'\big(d(\mathbf{x}_i, \mathbf{t}_{0,n}, C_{0,n})\big)}{d(\mathbf{x}_i, \mathbf{t}_{0,n}, C_{0,n})} C_{0,n}^{-1}(\mathbf{x}_i - \mathbf{t}_{0,n})(\mathbf{x}_i - \mathbf{t}_{0,n})^T \\ &\qquad - \sum_{i=1}^{n} v_1\big(d(\mathbf{x}_i, \mathbf{t}_{0,n}, C_{0,n})\big)\mathbf{I}. \end{aligned} \tag{7.18}$$

If $M_{0,n}$ is nonsingular, then

$$\mathbf{t}_{1,n} = \mathbf{t}_{0,n} + \sum_{i=1}^{n} v_1\big(d(\mathbf{x}_i, \mathbf{t}_{0,n}, C_{0,n})\big)\, M_{0,n}^{-1}(\mathbf{x}_i - \mathbf{t}_{0,n})$$

The finite sample breakdown behaviour of this one-step M-estimator of multivariate location basically depends on implosion or explosion of the matrix $M_{0,n}$ in (7.18) under contamination. This matrix is easier to study then the derivative $D_{0,n}$ of (7.15). Boundedness conditions on the functions v_1 and v_1' will keep the largest eigenvalue within bounds. To bound the smallest eigenvalue uniformly from below

is more difficult, especially if the function $\psi_1(y) = v_1(y)y$ is redescending. When the largest and the smallest eigenvalue can be bounded away from infinity and zero respectively, the one-step M-estimator as defined in (7.17) can easily be seen to inherit the breakdown point of the initial estimators $\mathbf{t}_{0,n}$ and $\mathbf{C}_{0,n}$ if $\psi(y) = v_1(y)y$ is bounded. For the covariance one-step M-estimator one could perform a Newton-Raphson iteration of the second equation of (4.7) seen as a function of $\mathbf{C}$. However, this will still yield complicated expressions.

Another alternative is to multiply (7.14) with $\frac{1}{n}$ and to replace the derivative $\frac{1}{n}\mathrm{D}_{0,n}$ by its limiting value

$$\mathrm{D}_0 = \mathrm{E}\frac{\partial}{\partial\boldsymbol{\theta}}\left[\boldsymbol{\Psi}(X_1,\boldsymbol{\theta})\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \tag{7.19}$$

where $\boldsymbol{\theta}_0$ is the limiting value of $\boldsymbol{\theta}_{0,n}$. Recently, Davies (1989) has investigated this one-step M-estimator for multivariate location and scatter. Especially at elliptical distributions the expression for (7.19) becomes easy. The one-step M procedure for location and scatter can be separated, and for location one has

$$\mathbf{t}_{1,n} = \mathbf{t}_{0,n} + \alpha^{-1}\sum_{i=1}^{n} v_1\big(d(\mathbf{x}_i, \mathbf{t}_{0,n}, \mathbf{C}_{0,n})\big)(\mathbf{x}_i - \mathbf{t}_{0,n}) \tag{7.20}$$

where α_1 is some positive constant that depends on v_1 and the underlying distribution P. A disadvantage is that this definition of the estimator, depends on the underlying distribution. On the other hand, it is easier to investigate the finite sample breakdown properties of (7.20).

8. Multivariate τ-estimators for Location and Scatter

8.1 Introduction

In the previous chapter we have discussed an affinely scaled M-estimator of multivariate location, which combines a high breakdown point and a bounded influence function with $\sqrt{n}$ asymptotic normality and good efficiency. It still remains of interest to construct an affine equivariant estimator of scatter that combines these properties. It is useless to extend the method of Section 7.1 to covariance estimators, i.e. estimate the location parameter affinely, center the observations and then compute a covariance M-estimator based on the centered observations. Simply because covariance M-estimators, even with a fixed location parameter, have a breakdown point that is at most $1/(p+1)$ (Tyler 1986).

Multivariate S-estimators retain the good breakdown properties of the MVE estimator, they have a bounded influence function, and they converge to a normal distribution at rate $\sqrt{n}$. Unfortunately, there still remains a trade-off between breakdown point and asymptotic efficiency. However, Yohai and Zamar (1988) investigated an extension of *regression* S-estimators, which retains the good breakdown properties and improves the asymptotic efficiency. In the special case of estimating univariate location and scale their proposal amounts to the following.

Recall the definition of univariate S-estimators of location and scale that are defined by means of a function ρ_1, i.e. first compute an M-estimate $\sigma_n(t)$ of scale by solving

$$\frac{1}{n}\sum_{i=1}^{n}\rho_1\left(\frac{x_i - t}{\sigma}\right) = b_1$$

where t is considered fixed, and then define the location S-estimator t_n as the value that minimizes $\sigma_n^2(t)$. Finally, take $\sigma_n(t_n)$ as the S-estimator of scale. To make the

M-estimator of scale more efficient, Yohai and Zamar (1988) consider an adaptive multiple of it, which they called a *τ-estimator of scale*. It is defined as

$$\tau_n^2(t) = \sigma_n^2(t)\frac{1}{nb_2}\sum_{i=1}^{n}\rho_2\left(\frac{x_i - t}{\sigma_n(t)}\right) \tag{8.1}$$

where b_2 is a normalizing constant and ρ_2 is a function that satisfies (R1) and (R2) and will generally differ from ρ_1. Instead of minimizing $\sigma_n^2(t)$ over t, Yohai and Zamar propose to minimize $\tau_n^2(t)$ over t. The minimizing t_n is taken as the τ-estimator of location and $\tau_n^2(t_n)$ is defined as the τ-estimator of scale.

That such a procedure may lead to estimators with good efficiency relative to the sample mean and sample variance can be seen as follows. Let $\rho_2(y)$ be a function that satisfies (R1)-(R2) and which tends to a function that is some multiple α of y^2, as the constant c_2 tends to infinity. An example is the biweight function $\rho_B(y; c_2)$ of (6.6). Then the limiting case of univariate τ-estimators would be to minimize

$$\tau_n^2(t) = \sigma_n^2(t)\frac{\alpha}{nb_2}\sum_{i=1}^{n}\left(\frac{x_i - t}{\sigma_n(t)}\right)^2 = \frac{\alpha}{nb_2}\sum_{i=1}^{n}(x_i - t)^2$$

over t, which yields the sample mean as location τ-estimator and the sample variance as the τ-estimator of scale, when b_2 is chosen suitably.

8.2 Multivariate τ-estimators

In Lopuhaä (1990), the multivariate version of regression τ-estimators are investigated. They are defined as follows.

DEFINITION 8.1: Multivariate τ-estimators of location and scatter are defined as the vector $\mathbf{t}_n$ and the matrix

$$\mathbf{V}_n = \mathbf{C}_n\frac{1}{nb_2}\sum_{i=1}^{n}\rho_2\big(d(\mathbf{x}_i, \mathbf{t}_n, \mathbf{C}_n)\big) \tag{8.2}$$

where $\mathbf{t}_n$ and $\mathbf{C}_n$ are the vector and the positive definite symmetric matrix that minimize

$$|\mathbf{C}|\left\{\sum_{i=1}^{n}\rho_2\big(d(\mathbf{x}_i, \mathbf{t}, \mathbf{C})\big)\right\}^p \tag{8.3}$$

subject to

$$\frac{1}{n}\sum_{i=1}^{n}\rho_1\big(d(\mathbf{x}_i, \mathbf{t}, \mathbf{C})\big) = b_1. \tag{8.4}$$

When this minimization problem has a unique solution, then the τ-estimators are affine equivariant; if the solution is not unique, the estimators are always affine equivariant in the same sense as nonuniquely defined S-estimators.

Note that constraint (8.4) is the same as constraint (6.7) of the minimization problem that defines multivariate S-estimators with the function ρ_1. In fact, multivariate S-estimators can be obtained as a special case of τ-estimators. Indeed, if $\rho_1 = \rho_2$ and $b_1 = b_2$, then $\mathbf{t}_n$ and $\mathbf{V}_n$ would just be the ordinairy S-estimators as defined in Chapter 6. Instead of minimizing the determinant of $\mathbf{C}$ over all pairs $\mathbf{t}$ and $\mathbf{C}$ that satisfy (8.4), we now minimize the determinant of an adaptive multiple of such $\mathbf{C}$, i.e. the determinant of the covariance τ-estimator $\mathbf{V}_n$ of (8.2). The sample mean and sample covariance can be obtained as a special case, namely with $\rho_1(y) = \rho_2(y) = y^2$ and $b_1 = b_2 = p$, as well as the MVE estimators with $\rho_1 = \rho_2$ an indicator function and $b_1 = b_2$ roughly $\frac{1}{2}$ (see also Chapter 6).

Similar to S-estimators, it is assumed that ρ_1 and ρ_2 both satisfy conditions (R1) and (R2) of Chapter 6, in order to get the good breakdown properties from the MVE estimator and a limiting behaviour that is similar to the sample mean and sample covariance. In addition, the following condition is imposed only on the function ρ_2.

(A) $2\rho_2(y) - \psi_2(y)y > 0$, for $y > 0$.

It will guarantee that the loss function in (8.3) is a strictly increasing function of the magnitude of $\mathbf{C}$. Similar to S-estimators, this property together with constraint (8.4) basically ensures the good breakdown properties of $\mathbf{t}_n$ and $\mathbf{C}_n$ and hence, of the τ-estimators. Constraint (8.4) guarantees a sufficient number of points, namely $n - \lfloor nr_1 \rfloor$ (where $r_1 = b_1/a_1 = b_1/\sup\rho_1$), inside the ellipsoid $E(\mathbf{t}_n, \mathbf{C}_n, c_1)$, which safeguards $\mathbf{C}_n$ against implosion. The fact that one minimizes a loss function in (8.3) that is an increasing function of the magnitude of $\mathbf{C}$, will safeguard $\mathbf{C}_n$ against explosion. That $\mathbf{t}_n$ must stay within bounds is then an immediate consequence. It turns out that the breakdown point only depends on the constant b_1 in (8.3), i.e.

$$\varepsilon^*(\mathbf{t}_n, \mathbf{X}) = \varepsilon^*(\mathbf{C}_n, \mathbf{X}) = \frac{\lceil nr_1 \rceil}{n} \tag{8.5}$$

at any collection $\mathbf{X}$ in general position. For $b_1 = \frac{n-p}{2n}a_1$, the τ-estimators will have a breakdown point $\lfloor \frac{n-p+1}{2} \rfloor / n$.

Also τ-estimators relate to multivariate M-estimators, but in a much more complicated way than S-estimators. To describe this, first consider a solution $\mathbf{t}_n$ and $\mathbf{C}_n$ of minimizing (8.3) subject to (8.4). It can be shown that $\mathbf{t}_n$ and $\mathbf{C}_n$ are solutions to M-estimator type of score equations (4.7), except that the functions v_1, v_2 and v_3 itself depend on the observations. One may show that $\mathbf{t}_n$ and $\mathbf{C}_n$ are a solution to the equations

$$\begin{aligned} &\sum_{i=1}^{n} u_n\big(d(\mathbf{x}_i, \mathbf{t}, \mathbf{C}); \mathbf{t}, \mathbf{C}\big)(\mathbf{x}_i - \mathbf{t}) = \mathbf{0} \\ &\sum_{i=1}^{n} \Big\{ p u_n\big(d(\mathbf{x}_i, \mathbf{t}, \mathbf{C}); \mathbf{t}, \mathbf{C}\big)(\mathbf{x}_i - \mathbf{t})(\mathbf{x}_i - \mathbf{t})^T - v_n\big(d(\mathbf{x}_i, \mathbf{t}, \mathbf{C}); \mathbf{t}, \mathbf{C}\big)\mathbf{C} \Big\} = \mathbf{0} \end{aligned} \tag{8.6}$$

where the functions $u_n(y;\mathbf{t},\mathbf{C}) = \psi_n(y;\mathbf{t},\mathbf{C})/y$ and $v_n(y;\mathbf{t},\mathbf{C})$ depend on the data:

$$\begin{aligned} \psi_n(y;\mathbf{t},\mathbf{C}) &= A_n(\mathbf{t},\mathbf{C})\psi_1(y) + B_n(\mathbf{t},\mathbf{C})\psi_2(y) \\ v_n(y;\mathbf{t},\mathbf{C}) &= \psi_n(y;\mathbf{t},\mathbf{C})y - 2b_2(\rho_1(y) - b_1) \end{aligned} \tag{8.7}$$

where the weights A_n and B_n are defined as

$$\begin{aligned} A_n(\mathbf{t},\mathbf{C}) &= \frac{1}{n}\sum_{i=1}^{n}\left\{2\rho_2\big(d(\mathbf{x}_i,\mathbf{t},\mathbf{C})\big) - \psi_2\big(d(\mathbf{x}_i,\mathbf{t},\mathbf{C})\big)\,d(\mathbf{x}_i,\mathbf{t},\mathbf{C})\right\} \\ B_n(\mathbf{t},\mathbf{C}) &= \frac{1}{n}\sum_{i=1}^{n}\psi_1\big(d(\mathbf{x}_i,\mathbf{t},\mathbf{C})\big)\,d(\mathbf{x}_i,\mathbf{t},\mathbf{C}) \end{aligned} \tag{8.8}$$

Hence, $\mathbf{t}_n$ and $\mathbf{C}_n$ are solutions to equations of type (6.12), except that the function $\psi(y) = u(y)y$ is now an adaptively weighted average $\psi_n(y;\mathbf{t},\mathbf{C})$ of the functions ψ_1 and ψ_2.

Although $\mathbf{t}_n$ and $\mathbf{C}_n$ are solutions of the complicated equations (8.6), their asymptotic behaviour is equivalent to that of regular multivariate M-estimators as defined with (4.7). One may first show that $\mathbf{t}_n$ and $\mathbf{C}_n$ are consistent, which then implies that the weights $A_n(\mathbf{t}_n,\mathbf{C}_n)$ and $B_n(\mathbf{t}_n,\mathbf{C}_n)$ in (8.8) tend to positive numbers A and B with probability one. This means that the function $\psi_n(y;\mathbf{t}_n,\mathbf{C}_n)$ converges pointwise to the function

$$\tilde{\psi}(y) = A\psi_1(y) + B\psi_2(y). \tag{8.9}$$

The limiting behaviour of $\mathbf{t}_n$ and $\mathbf{C}_n$ can then be shown to be equivalent to that of multivariate M-estimators defined as solutions of equations (8.6) with the function $\psi_n(y;\mathbf{t},\mathbf{C})$ replaced by the function $\tilde{\psi}(y)$. The asymptotic behaviour of the actual τ-estimators $\mathbf{t}_n$ and $\mathbf{V}_n$ can be obtained from that of $\mathbf{t}_n$ and $\mathbf{C}_n$.

In Lopuhaä (1990) a more general setup is used in terms of τ-functionals, in order to derive the influence function. Under weak assumptions on the sample distribution P, τ-estimators are shown to converge weakly to a normal distribution at rate $\sqrt{n}$, and to have an influence function that is bounded.

What remains, is whether we can combine a high breakdown point with good efficiency. To answer this, an elliptical underlying distribution is considered. In this case one should take b_1 and b_2 as in (6.13) to guarantee consistency of $\mathbf{t}_n$ and $\mathbf{V}_n$ for the parameters μ and Σ. Note that in this case b_1 and b_2 still depend on the constants c_1 and c_2 of the functions ρ_1 and ρ_2 respectively. In particular, the breakdown point in (8.5) will depend on c_1. However, as with S-estimators, c_1 can be chosen such that the τ-estimators have breakdown point $\lfloor\frac{n-p+1}{2}\rfloor/n$.

At an elliptical distribution, it also turns out that τ-estimators are asymptotically equivalent to multivariate S-estimators defined with a weighted ρ-function

$$\tilde{\rho}(y) = A\rho_1(y) + B\rho_2(y). \tag{8.10}$$

If the constant c_1 of the function ρ_1 remains fixed, for instance it could be chosen such that the τ-estimator has a high breakdown point, and if the constant c_2 of the function ρ_2 tends to infinity, the weight $A_n(\mathbf{t}, \mathbf{C})$ as well as its limiting value A tend to zero when ρ_2 is chosen suitably. For instance, if we take for $\rho_2(y)$ the biweight function $\rho_{\text{B}}(y; c_2)$, then $2\rho_2(y) - \psi_2(y)y$ tends to zero as $c_2 \to \infty$. In such a case, the function $\tilde{\rho}$ becomes more and more similar to the function y^2. Since with $\rho(y) = y^2$, the corresponding S-estimator coincides with the sample mean and the sample covariance, this indicates that for c_2 large one has good efficiency for the τ-estimators relative to the sample mean and the sample covariance. Because the breakdown point does not depend on the constant c_2, we are able to combine a high breakdown point and good efficiency with suitable choices for c_1 and c_2 respectively.

A typical example would be to take for $\rho_1(y)$ the biweight function $\rho_{\text{B}}(y; c_1)$, with c_1 chosen such that the breakdown point is $\lfloor \frac{n-p+1}{2} \rfloor / n$, and to take for $\rho_2(y)$ another biweight function $\rho_{\text{B}}(y; c_2)$, with c_2 chosen such that the estimators have good efficiency relative to the sample mean and sample covariance. The resulting estimators are affine equivariant, and combine a high breakdown point and a bounded influence function with a $\sqrt{n}$ rate of convergence towards a normal distribution and good efficiency for both the location as well as the covariance estimator.

9. References

ANDREWS, D.F., BICKEL, P.J., HAMPEL, F.R., HUBER, P.J., ROGERS, W.H. and TUKEY, J.W. (1972). *Robust estimates of Location: Survey and Advances.* Princeton University Press, Princeton, N.J..

BICKEL, P.J. (1975). One-step Huber estimates in the linear model. *J. Am. Statist. Assoc.* **70** 428–434.

BUTLER, R.W. and JHUN M. (1987). Asymptotic for trimmed multivariate data. Revised Preprint November 87, University of Michigan and University of Florida.

CAMBANIS, S., HUANG, S. and SIMONS, G. (1981). On the theory of elliptically contoured distributions. *J. Multivariate Analysis* **11** 368–385.

COLLINS, J.R. (1982). Robust M-estimators of location vectors. *J. Multivariate Analysis* **12** 480–492.

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

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

DEVLIN, S.J., GNANADESIKAN, R. and KETTENRING, J.R. (1981). Robust estimation of dispersion matrices and principal components. *J. Am. Statist. Assoc.* **76** 354-362.

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.

FILIPPOVA, A.A. (1961). Mises' theorem on the asymptotic behaviour of empirical distribution functions and its statistical applications. *Theory Prob. Appl.* **7** 24–57.

GRÜBEL, R. (1988a). The length of the shorth. *Ann. Stat.* **16** 619–628.

GRÜBEL, R. (1988b). A minimal characterization of the covariance matrix. *Metrika* **35** 49–52.

HAMPEL, F.R. (1968). Contributions to the Theory of Robust Estimation. Ph.D.thesis, University of California at Berkeley.

HAMPEL, F.R. (1971). A general qualitative definition of robustness. *Ann. Math. Statist.* **42** 1887–1896.

HAMPEL, F.R. (1974). The influence curve and its role in robust estimation. *J. Am. Statist. Assoc.* **69** 383–393.

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.

HE, X., JUREČKOVÁ, J., KOENKER, R. and PORTNOY, S. (1988). Tail behaviour of regression estimators and their breakdown points. Technical Report, University of Illinois at Urbana-Champaign.

HODGES, J.L. Jr. (1967). Efficiency in normal samples and tolerance of extreme values for some estimates of location. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (L.M.Le Cam and J.Neyman, eds.) 163–186. University of California Press, Berkeley.

HUBER, P.J. (1964). Robust estimation of a location parameter. *Ann. Statist.* **35** 73–101.

HUBER, P.J. (1967). The behaviour of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (L.M.Le Cam and J.Neyman, eds.) 221-223. University of California Press, Berkeley.

HUBER, P.J. (1977). Robust covariances. In *Statistical decision theory and related topics* (S.S. Gupta and D.S. Moore, eds.) 165–191. Academic Press, New York.

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

HUBER, P.J. (1984). Finite sample breakdown point of M- and P-estimators. *Ann. Statist.* **12** 119–126.

KENT, J.T. and TYLER, D.E. (1989). Redescending M-estimates of multivariate location and scatter. Technical Report, University of Leeds and Rutgers University.

JUREČKOVÁ, J. (1981). Tail-behaviour of location estimators. *Ann. Stat.* **9** 578–585.

KELKER, D. (1970). Distribution theory of spherical distributions and a location -scale parameter generalization. *Sankhya, Ser.A, 32 p.419–430.*

KIM, J. and POLLARD, D. (1989). Cube root asymptotics. Technical Report, Yale University. To appear Ann. Statist..

LE CAM, L. (1956). On the asymptotic theory of estimation and testing hypotheses. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability.* University of California Press, Berkeley. 129–156.

LOPUHAÄ, H.P. (1988). Highly efficient estimators of multivariate location with high breakdown point. Revised version of Technical Report 88-14, Delft University of Technology.

LOPUHAÄ, H.P. (1989). On the relation between S-estimators and M-estimators of multivariate location and covariance. *Ann. Statist.* **17** 1662–1683.

LOPUHAÄ, H.P. (1990). Multivariate τ-estimators for location and scatter. Technical Report 90-4, Delft University of Technology.

LOPUHAÄ, H.P. and ROUSSEEUW, P.J. (1989). Breakdown properties of affine equivariant estimators of multivariate location and covariance matrices. Revised version of Technical Report 87-14, Delft University of Technology. Tentatively accepted by Ann. Statist..

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

MARONNA, R.A. and YOHAI, V.J. (1989). The maximum bias of robust covariances. To appear in Communications in Statistics.

MARTIN, R.D. and ZAMAR, R.H. (1987). Mini-max bias robust M-estimates of scale. To appear J. Amer. Statist. Assoc..

MARTIN, R.D., YOHAI, V.J. and ZAMAR, R.H. (1989). Min-max bias robust regression. To appear Ann. Statist..

ROUSSEEUW, P.J. (1981). New infinitesimal methods in robust statistics. Ph.D.thesis, Vrije Universiteit, Brussels.

ROUSSEEUW, P.J. (1983). Multivariate estimation with high breakdown point. Paper presented at the Fourth Pannonian Symposium on Mathematical Statistics and Probability, Bad Tatzmannsdorf, Austria, September 4-9, 1983. In *Mathematical Statistics and Applications (1985)* (W.Grossmann, G.Pflug, I.Vincze and W.Wertz, eds.) 283–297. Reidel, Dordrecht, The Netherlands.

ROUSSEEUW, P.J. (1984). Least median of squares regression. *J. Am. Stat. Assoc.* **79** 871–880.

ROUSSEEUW, P.J. and LEROY, A.M. (1987). *Robust Regression and Outlier Detection.* Wiley, New York.

ROUSSEEUW, P.J. and YOHAI, V. (1984). Robust regression by means of S-estimators. In *Robust and Nonlinear Time Series Analysis* Lecture Notes in Statistics **26** 256–272. Springer Verlag, New York.

SHORACK, G.R. and WELLNER, J.A. (1986). *Empirical Processes with Applications to Statistics*. Wiley, New York.

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

TAMURA, R.N. and BOOS. D.D. (1986). Minimum Hellinger distance estimation for multivariate location and scatter. *J. Amer. Statist. Assoc.* **81** 223–229.

TYLER, D.E. (1983). Robustness and efficiency properties of scatter matrices. *Biometrika* **70** 411–420.

TYLER, D.E. (1986). Breakdown properties of the M-estimators of multivariate scatter. Technical Report, Department of Statistics, Rutgers University, New Jersey.

TYLER, D.E. (1987). A distribution-free M-estimator of multivariate scatter. *Ann. Statist.* **15** 234–251.

VON MISES, R. (1937). Sur les fonctions statistiques. In *Conf. de la Réunion Internationale des Math.* Gauthier-Villars, Paris. 1–8.

VON MISES, R. (1947). On the asymptotic distribution of differentiable statistical functions. *Ann. Math. Statist.* **18** 309–348.

YOHAI, V.J. (1987). High breakdown-point and high efficiency robust estimates for regression. *Ann. Statist.* **15** 642–656.

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

Copies of the four papers

[1] Tentatively accepted by the Annals of Statistics

[2] Appeared in the Annals of Statistics 1989 Vol. 17, No. 4, 1662-1683

BREAKDOWN POINTS OF AFFINE EQUIVARIANT ESTIMATORS OF MULTIVARIATE LOCATION AND COVARIANCE MATRICES

HENDRIK P. LOPUHAÄ AND PETER J. ROUSSEEUW

Technische Universiteit Delft and Vrije Universiteit Brussel

Finite-sample replacement breakdown points are derived for different types of estimators of multivariate location and covariance matrices. The role of various equivariance properties is illustrated. The breakdown point is related to a measure of performance based on large deviations probabilities. Finally, we show that one-step reweighting retains the breakdown point.

1. Introduction. Several notions of robustness have been considered for estimators of a multivariate location parameter $\mu \in \mathbf{R}^p$. One of these concepts is the breakdown point, a global measure of robustness suggested by Hodges (1967) and Hampel (1968). A simple and appealing finite-sample version of this concept was given by Donoho and Huber (1983). Roughly, this finite-sample replacement breakdown point measures the minimum fraction of outliers that will spoil the estimate completely. Estimators with zero breakdown point can therefore not be robust. Recently, He, Jurečková, Koenker and Portnoy (1988) established a relation between the replacement breakdown point and a measure of performance based on large deviations. Their result shows that the breakdown point is not just an attractive robustness concept, but that it also has a stochastic motivation.

A natural condition for multivariate estimators is equivariance under affine transformations. To combine affine equivariance with a high breakdown point is not trivial. Donoho (1982) discusses several affine equivariant multivariate methods, showing that their breakdown point goes down to zero as the dimension p increases. Stahel (1981) and Donoho (1982) independently introduced an affine estimator of multivariate location and covariance with a high breakdown point in any dimension. Another estimator with this combination of properties was the minimum volume ellipsoid estimator (Rousseeuw 1983).

But what is the best possible value of the breakdown point ? For covariance estimators the maximal breakdown point was derived by Davies (1987). In our paper we are mainly concerned with upper bounds for pure location estimators satisfying various equivariance properties. Section 2 discusses these types of equivariance and gives an upper bound on the breakdown point. In order to investigate to what extent this bound is sharp, Sections 2 and 3 study several examples including the L_1-estimator, Oja's (1983) generalized median, and smooth S-estimators in the sense of Rousseeuw and Yohai (1984).

Research of the first author was financially supported by NWO under grant 10-62-10.
1980 *Mathematics subject classifications* : 62F35, 62H12
Keywords : Breakdown point, Affine equivariance, Tailperformance, Weighted estimators.

Section 4 extends the result of He et al. (1988) to multivariate location estimators. Estimators with maximal breakdown point satisfy a minimax property in the sense that they maximize least favorable tail performance at algebraically tailed distributions.

To combine a high breakdown point with good asymptotic efficiency, it has often been suggested to start with a high breakdown estimator and then to take a one-step improvement which retains the breakdown point and obtains a better efficiency. Section 5 shows that the breakdown point is retained if one does a one-step reweighting by computing the usual weighted mean and covariance matrix, where the weights are based on the Mahalanobis distances with respect to the initial estimators.

2. Maximal breakdown point of equivariant estimators. Let $(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \mathbf{X}$ be a collection of n points in $\mathbf{R}^p$ and denote by $\mathbf{t}_n(\mathbf{X}) \in \mathbf{R}^p$ a location estimator based on $\mathbf{X}$. We say that $\mathbf{t}_n$ is *translation equivariant* if $\mathbf{t}_n(\mathbf{X} + \mathbf{v}) = \mathbf{t}_n(\mathbf{X}) + \mathbf{v}$ for all $\mathbf{v} \in \mathbf{R}^p$, where $\mathbf{X} + \mathbf{v} = (\mathbf{x}_1 + \mathbf{v}, \ldots, \mathbf{x}_n + \mathbf{v})$. When $\mathbf{t}_n$ is equivariant not only under shifts of $\mathbf{X}$ but also under affine transformations, then $\mathbf{t}_n$ is called *affine equivariant*, i.e.

$$\mathbf{t}_n(\mathbf{AX} + \mathbf{v}) = \mathbf{A}\mathbf{t}_n(\mathbf{X}) + \mathbf{v} \tag{2.1}$$

for all nonsingular $p \times p$-matrices $\mathbf{A}$ and $\mathbf{v} \in \mathbf{R}^p$, where $\mathbf{AX} + \mathbf{v} = (\mathbf{Ax}_1 + \mathbf{v}, \ldots, \mathbf{Ax}_n + \mathbf{v})$. Although this condition is quite natural, it turns out that some well-known estimators of multivariate location fail to satisfy it. The condition can be relaxed by requiring (2.1) only for orthogonal matrices, and it is then referred to as *orthogonal equivariance* or rigid motion equivariance. At the end of this section we shall consider a translation equivariant estimator which is not orthogonal equivariant, and also an orthogonal equivariant estimator which is not affine equivariant. A covariance estimator $\mathbf{C}_n(\mathbf{X}) \in \mathrm{PDS}(p)$, the class of all positive definite symmetric $p \times p$-matrices, is said to be affine equivariant if $\mathbf{C}_n(\mathbf{AX} + \mathbf{v}) = \mathbf{A}\mathbf{C}_n(\mathbf{X})\mathbf{A}^T$ for all $\mathbf{v} \in \mathbf{R}^p$ and nonsingular $\mathbf{A}$, where $\mathbf{A}^T$ denotes the transpose of $\mathbf{A}$.

We measure the robustness of $\mathbf{t}_n$ and $\mathbf{C}_n$ by means of the finite-sample replacement breakdown point (Donoho and Huber 1983). The breakdown point of a location estimator $\mathbf{t}_n$ at a collection $\mathbf{X}$ is defined as the smallest fraction m/n of outliers that can take the estimator over all bounds :

$$\varepsilon^*(\mathbf{t}_n, \mathbf{X}) = \min_{1 \le m \le n} \left\{ \frac{m}{n} : \sup_{Y_m} \|\mathbf{t}_n(\mathbf{X}) - \mathbf{t}_n(\mathbf{Y}_m)\| = \infty \right\} \tag{2.2}$$

where the supremum is taken over all possible corrupted collections $\mathbf{Y}_m$ that are obtained from $\mathbf{X}$ by replacing m points of $\mathbf{X}$ by arbitrary values. Although $\varepsilon^*(\mathbf{t}_n, \mathbf{X})$ appears to depend on $\mathbf{X}$, for most $\mathbf{t}_n$ this will not be the case. However, location estimators $\mathbf{t}_n$ with $\varepsilon^*(\mathbf{t}_n, \mathbf{X})$ depending on $\mathbf{X}$ do exist (see for instance Huber 1984). The breakdown point of a covariance estimator $\mathbf{C}_n$ at a collection $\mathbf{X}$ is defined as the smallest fraction m/n of outliers that can either take the largest eigenvalue

$\lambda_1(\mathbf{C}_n)$ over all bounds, or take the smallest eigenvalue $\lambda_p(\mathbf{C}_n)$ arbitrarily close to zero :

$$\varepsilon^*(\mathbf{C}_n, \mathbf{X}) = \min_{1 \le m \le n} \left\{ \frac{m}{n} : \sup_{\mathbf{Y}_m} D(\mathbf{C}_n(\mathbf{X}), \mathbf{C}_n(\mathbf{Y}_m)) = \infty \right\}$$

where the supremum is taken over the same corrupted collections $\mathbf{Y}_m$ as in (2.2), and where $D(\mathbf{A}, \mathbf{B}) = \max\{|\lambda_1(\mathbf{A}) - \lambda_1(\mathbf{B})|, |\lambda_p(\mathbf{A})^{-1} - \lambda_p(\mathbf{B})^{-1}|\}$, with $\lambda_1(\mathbf{A}) \ge \cdots \ge \lambda_p(\mathbf{A})$ being the ordered eigenvalues of the matrix $\mathbf{A}$.

Donoho and Huber (1983) also considered other finite-sample versions, such as addition breakdown. We personally prefer replacing observations to adding observations because replacement contamination is simple, realistic, and generally applicable. Indeed, from an intuitive point of view, outliers are not some faulty observations that are added at the end of the sample, but they treacherously hide themselves by replacing some of the data points that should have been observed. Moreover, as we will see in Section 4, the replacement breakdown point also has a stochastic interpretation.

First we show that the breakdown point of any affine equivariant estimator is itself invariant under affine transformations.

LEMMA 2.1. *Let* $\mathbf{X}$ *be a collection of* n *points in* $\mathbf{R}^p$, *and let* $\mathbf{t}_n(\mathbf{X}) \in \mathbf{R}^P$ *and* $\mathbf{C}_n(\mathbf{X}) \in PDS(p)$ *be location and covariance estimators based on* $\mathbf{X}$.

(i) When $\mathbf{t}_n$ *is translation equivariant, then for any* $\mathbf{v} \in \mathbf{R}^p$ *it holds that* $\varepsilon^*(\mathbf{t}_n, \mathbf{X} + \mathbf{v}) = \varepsilon^*(\mathbf{t}_n, \mathbf{X})$.

(ii) When $\mathbf{t}_n$ *is affine (orthogonal) equivariant, then for any* $\mathbf{v} \in \mathbf{R}^p$ *and for any nonsingular (orthogonal)* $p \times p$*-matrix* $\mathbf{A}$ *it holds that* $\varepsilon^*(\mathbf{t}_n, \mathbf{AX} + \mathbf{v}) = \varepsilon^*(\mathbf{t}_n, \mathbf{X})$.

(iii) When $\mathbf{C}_n$ *is affine equivariant, then for any* $\mathbf{v} \in \mathbf{R}^p$ *and for any nonsingular* $p \times p$*-matrix* $\mathbf{A}$ *it holds that* $\varepsilon^*(\mathbf{C}_n, \mathbf{AX} + \mathbf{v}) = \varepsilon^*(\mathbf{C}_n, \mathbf{X})$.

PROOF: Let $\mathbf{A}$ be a nonsingular $p \times p$-matrix and $\mathbf{v} \in \mathbf{R}^p$. Denote by $\mathbf{Y}_m$ a corrupted collection that differs from $\mathbf{X}$ in at most m points, so that $\mathbf{AY}_m + \mathbf{v}$ differs from $\mathbf{AX} + \mathbf{v}$ in at most m points. When $\mathbf{t}_n$ is affine equivariant we have that $\|\mathbf{t}_n(\mathbf{AX}+\mathbf{v}) - \mathbf{t}_n(\mathbf{AY}_m+\mathbf{v})\| = \|\mathbf{A}[\mathbf{t}_n(\mathbf{X}) - \mathbf{t}_n(\mathbf{Y}_m)]\|$. In that case, together with the fact that for symmetric $p \times p$-matrices $\mathbf{M}$ one has

$$\lambda_p(\mathbf{M}) = \inf_{\mathbf{y}} \frac{\mathbf{y}^T\mathbf{M}\mathbf{y}}{\mathbf{y}^T\mathbf{y}} \quad \text{and} \quad \lambda_1(\mathbf{M}) = \sup_{\mathbf{y}} \frac{\mathbf{y}^T\mathbf{M}\mathbf{y}}{\mathbf{y}^T\mathbf{y}} \tag{2.3}$$

we obtain

$$\lambda_p(\mathbf{A}^T\mathbf{A}) \le \frac{\|\mathbf{t}_n(\mathbf{AX}+\mathbf{v}) - \mathbf{t}_n(\mathbf{AY}_m+\mathbf{v})\|^2}{\|\mathbf{t}_n(\mathbf{X}) - \mathbf{t}_n(\mathbf{Y}_m)\|^2} \le \lambda_1(\mathbf{A}^T\mathbf{A}).$$

This means that $\sup_{\mathbf{Y}_m} \|\mathbf{t}_n(\mathbf{X}) - \mathbf{t}_n(\mathbf{Y}_m)\|$, taken over all possible $\mathbf{Y}_m$, is finite or infinite at the same time as $\sup_{\mathbf{Z}_m} \|\mathbf{t}_n(\mathbf{AX}+\mathbf{v}) - \mathbf{t}_n(\mathbf{Z}_m)\|$, taken over all corrupted collections $\mathbf{Z}_m$ that differ from $\mathbf{AX}+\mathbf{v}$ in at most m points. This proves (ii) for the case that $\mathbf{t}_n$ is affine equivariant. Clearly, if $\mathbf{A}$ is orthogonal the argument above

can be repeated for orthogonal equivariant $\mathbf{t}_n$, and if we take $\mathbf{A} = \mathbf{I}$ the argument can be repeated for translation equivariant $\mathbf{t}_n$. This leaves us with proving (iii).

In that case, (2.3) and affine equivariance of $\mathbf{C}_n$ imply that for any collection $\mathbf{S}$ of n points

$$\lambda_1(\mathbf{C}_n(\mathbf{S}))\lambda_p(\mathbf{A}\mathbf{A}^T) \leq \lambda_1(\mathbf{C}_n(\mathbf{A}\mathbf{S}+\mathbf{v})) \leq \lambda_1(\mathbf{C}_n(\mathbf{S}))\lambda_1(\mathbf{A}\mathbf{A}^T). \tag{2.4}$$

Apply (2.4) to $\mathbf{S} = \mathbf{X}$ and $\mathbf{S} = \mathbf{Y}_m$. If we write

$$\alpha = [\lambda_1(\mathbf{A}\mathbf{A}^T) - \lambda_p(\mathbf{A}\mathbf{A}^T)]\lambda_1(\mathbf{C}_n(\mathbf{X}))$$

and if we suppress the notation $\mathbf{C}_n$ for a moment, we find that

$$\lambda_1(\mathbf{A}\mathbf{X}+\mathbf{v}) - \lambda_1(\mathbf{A}\mathbf{Y}_m+\mathbf{v}) \leq \lambda_p(\mathbf{A}\mathbf{A}^T)\,[\lambda_1(\mathbf{X}) - \lambda_1(\mathbf{Y}_m)] + \alpha$$
$$\lambda_1(\mathbf{A}\mathbf{X}+\mathbf{v}) - \lambda_1(\mathbf{A}\mathbf{Y}_m+\mathbf{v}) \geq \lambda_1(\mathbf{A}\mathbf{A}^T)\,[\lambda_1(\mathbf{X}) - \lambda_1(\mathbf{Y}_m)] - \alpha.$$

Inequalities that relate $\lambda_p(\mathbf{A}\mathbf{X}+\mathbf{v}) - \lambda_p(\mathbf{A}\mathbf{Y}_m+\mathbf{v})$ to $\lambda_p(\mathbf{X})-\lambda_p(\mathbf{Y}_m)$ can be obtained in a similar way. As in the first part of the proof it then follows that $\sup_{\mathbf{Y}_m} D(\mathbf{C}_n(\mathbf{X}), \mathbf{C}_n(\mathbf{Y}_m))$ and $\sup_{\mathbf{Z}_m} D(\mathbf{C}_n(\mathbf{A}\mathbf{X}+\mathbf{v}), \mathbf{C}_n(\mathbf{Z}_m))$, taken over all corrupted collections $\mathbf{Z}_m$ that differ from $\mathbf{A}\mathbf{X}+\mathbf{v}$ in at most m points, are finite or infinite at the same time, which proves (iii). □

It is natural to ask for the maximal breakdown point of an estimator satisfying one of the equivariance properties mentioned above. The next theorem gives the upper bound for translation equivariant location estimators.

THEOREM 2.1. *Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ be a collection of n points in $\mathbf{R}^p$. When $\mathbf{t}_n$ is translation equivariant, then $\varepsilon^*(\mathbf{t}_n, \mathbf{X}) \leq \lfloor \frac{n+1}{2} \rfloor / n$, where $\lfloor y \rfloor$ denotes the largest integer less than or equal to y.*

PROOF: Because $\mathbf{t}_n$ is translation equivariant, according to Lemma 2.1 we may assume that $\mathbf{t}_n(\mathbf{X}) = \mathbf{0}$. Suppose that $\varepsilon^*(\mathbf{t}_n, \mathbf{X}) > \lfloor \frac{n+1}{2} \rfloor / n$. This means that there would exist a constant k such that

$$\|\mathbf{t}_n(\mathbf{Y})\| \leq k < \infty \tag{2.5}$$

for all corrupted collections $\mathbf{Y}$ obtained by replacing $\lfloor \frac{n+1}{2} \rfloor$ points of $\mathbf{X}$. Denote by $q = n - \lfloor \frac{n+1}{2} \rfloor$ the number of points of $\mathbf{X}$ that are not replaced. Since $2q \leq n$, for any $\mathbf{v} \in \mathbf{R}^p$ we can always construct a collection $\mathbf{Y}_\mathbf{v}$ containing $\mathbf{x}_1, \ldots, \mathbf{x}_q, \mathbf{x}_1 + \mathbf{v}, \ldots, \mathbf{x}_q + \mathbf{v}$, and also a corresponding collection $\mathbf{Z}_\mathbf{v} = \mathbf{Y}_\mathbf{v} - \mathbf{v}$ containing $\mathbf{x}_1 - \mathbf{v}, \ldots, \mathbf{x}_q - \mathbf{v}, \mathbf{x}_1, \ldots, \mathbf{x}_q$. Both collections contain at least q points of $\mathbf{X}$ so according to (2.5) we must have $\|\mathbf{t}_n(\mathbf{Y}_\mathbf{v})\| \leq k$ as well as $\|\mathbf{t}_n(\mathbf{Y}_\mathbf{v}) - \mathbf{v}\| = \|\mathbf{t}_n(\mathbf{Z}_\mathbf{v})\| \leq k$, using that $\mathbf{t}_n$ is translation equivariant. Clearly, for large $\mathbf{v} \in \mathbf{R}^p$ these two inequalities cannot both be true. □

As the class of affine (orthogonal) equivariant estimators is contained in the class of translation equivariant estimators, the upper bound $\lfloor \frac{n+1}{2} \rfloor / n$ obviously also holds for this smaller class. It then becomes of interest whether there exist estimators with these equivariance properties that attain this upper bound. The following two examples show that the upper bound $\lfloor \frac{n+1}{2} \rfloor / n$ is sharp for translation and orthogonal equivariant estimators.

COORDINATEWISE MEDIAN: A simple way to obtain a multivariate translation equivariant location estimator with a high breakdown point is to take a univariate translation equivariant location estimator with a high breakdown point and construct its multivariate analogue coordinatewise. Define $\mathbf{t}_n = (t_{n1} \cdots t_{np})^T$ coordinatewise by $t_{nj}(\mathbf{X}) = \text{median}\{x_{ij} : 1 \leq i \leq n\}$ for $j = 1, \ldots, p$, where $\mathbf{x}_i = (x_{i1} \cdots x_{ip})^T$ for $i = 1, \ldots, n$. Clearly, the breakdown point $\lfloor \frac{n+1}{2} \rfloor / n$ of the univariate median is retained. Note that $\mathbf{t}_n$ is translation equivariant but not orthogonal equivariant.

There are several other ways of generalizing the one-dimensional median to higher dimensions. One of the oldest generalized medians is the following example of an orthogonal equivariant estimator.

L_1-ESTIMATOR: Define the L_1-estimator as the vector $\mathbf{t}_n$ that minimizes

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

Because the Euclidean norm is invariant under orthogonal transformations it follows that the L_1-estimator is orthogonal equivariant. However, it is not affine equivariant. The breakdown point is independent of the dimension p and $\mathbf{X}$, and is equal to that of the univariate median.

THEOREM 2.2. *Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ be a collection of n points in $\mathbf{R}^p$. Then the L_1-estimator has breakdown point $\varepsilon^*(\mathbf{t}_n, \mathbf{X}) = \lfloor \frac{n+1}{2} \rfloor / n$.*

PROOF: Since $\mathbf{t}_n$ is translation equivariant, according to Lemma 2.1 we may assume $\mathbf{t}_n(\mathbf{X}) = \mathbf{0}$. Put $M = \max_{1 \leq i \leq n} \|\mathbf{x}_i\|$, and let $B(\mathbf{0}, 2M)$ be the sphere with center $\mathbf{0}$ and radius $2M$. Denote by $\mathbf{Y}_m = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ a corrupted collection obtained by replacing at most $m = \lfloor \frac{n-1}{2} \rfloor$ points of $\mathbf{X}$ and let $\mathbf{t}_n(\mathbf{Y}_m)$ minimize $\sum_{i=1}^{n} \|\mathbf{y}_i - \mathbf{t}\|$.

We show that $\sup_{\mathbf{Y}_m} \|\mathbf{t}_n(\mathbf{Y}_m)\|$, taken over all possible $\mathbf{Y}_m$, is finite. Denote by $d = \inf_{\mathbf{v} \in B(0,2M)} \|\mathbf{t}_n(\mathbf{Y}_m) - \mathbf{v}\|$ the distance between $\mathbf{t}_n(\mathbf{Y}_m)$ and $B(\mathbf{0}, 2M)$, so that $\|\mathbf{t}_n(\mathbf{Y}_m)\| \leq d + 2M$. Then for each of the $\lfloor \frac{n-1}{2} \rfloor$ replaced $\mathbf{y}_j$'s, it holds that

$$\|\mathbf{y}_j - \mathbf{t}_n(\mathbf{Y}_m)\| \geq \|\mathbf{y}_j\| - \|\mathbf{t}_n(\mathbf{Y}_m)\| \geq \|\mathbf{y}_j\| - (d + 2M). \tag{2.6}$$

Suppose that $\mathbf{t}_n(\mathbf{Y}_m)$ is outside $B(\mathbf{0}, 2M)$ and that the distance between $\mathbf{t}_n(\mathbf{Y}_m)$ and $B(\mathbf{0}, 2M)$ is large : $d > 2M \lfloor \frac{n-1}{2} \rfloor$. Since $\mathbf{X} \subset B(\mathbf{0}, M)$, for each of the $n - \lfloor \frac{n-1}{2} \rfloor$ original $\mathbf{x}_k$'s in $\mathbf{Y}_m$ we would then have that

$$\|\mathbf{x}_k - \mathbf{t}_n(\mathbf{Y}_m)\| \geq M + d \geq \|\mathbf{x}_k\| + d. \tag{2.7}$$

From (2.6) and (2.7) it would follow that

$$\sum_{i=1}^{n} \|\mathbf{y}_i - \mathbf{t}_n(\mathbf{Y}_m)\| \geq \sum_{i=1}^{n} \|\mathbf{y}_i\| + (n - \lfloor \tfrac{n-1}{2} \rfloor)d - \lfloor \tfrac{n-1}{2} \rfloor (d + 2M)$$
$$\geq \sum_{i=1}^{n} \|\mathbf{y}_i\| + d - 2M\lfloor \tfrac{n-1}{2} \rfloor > \sum_{i=1}^{n} \|\mathbf{y}_i\|.$$

This is a contradiction with the fact that $\mathbf{t}_n(\mathbf{Y}_m)$ minimizes $\sum_{i=1}^{n} \|\mathbf{y}_i - \mathbf{t}\|$. Therefore $d \leq 2M\lfloor \frac{n-1}{2} \rfloor$ and hence, $\sup_{\mathbf{Y}_m} \|\mathbf{t}_n(\mathbf{Y}_m)\| \leq d + 2M \leq 2M\lfloor \frac{n+1}{2} \rfloor$. We conclude that $\varepsilon^*(\mathbf{t}_n, \mathbf{X}) \geq \lfloor \frac{n+1}{2} \rfloor / n$. The other inequality is obtained directly from Theorem 2.1. □

3. Affine equivariance and breakdown point. Is the upper bound $\lfloor \frac{n+1}{2} \rfloor / n$ also sharp for affine equivariant estimators ? Davies (1987) showed that for covariance estimators this is no longer the case. When the collection $\mathbf{X}$ is in *general position*, i.e. no $p+1$ points are contained in some hyperplane of dimension smaller than p, and if $n \geq p+1$, the breakdown point of any affine equivariant covariance estimator $\mathbf{C}_n$ is at most $\lfloor \frac{n-p+1}{2} \rfloor / n$. Although the result is stated for pairs $(\mathbf{t}_n, \mathbf{C}_n)$, it is only shown that the covariance part might break down if one replaces $\lfloor \frac{n-p+1}{2} \rfloor$ points or more, regardless of what happens with the location part. This means that the upper bound $\lfloor \frac{n-p+1}{2} \rfloor / n$ does not have to apply to affine equivariant *location* estimators, especially not to those that are defined without a corresponding covariance part. Such an estimator is Oja's (1983) affine equivariant multivariate median.

OJA'S ESTIMATOR: Consider the volumes $\Delta(\mathbf{t}, \mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_p})$ of all simplices formed by $\mathbf{t} \in \mathbf{R}^p$ and all possible subcollections $\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_p}$ from $\mathbf{X}$. Oja's multivariate median is the vector $\mathbf{t}_n$ in $\mathbf{R}^p$ that minimizes

$$\sum_{\{\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_p}\} \subset \mathbf{X}} \Delta(\mathbf{t}, \mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_p}).$$

This location estimator is affine equivariant and is defined without any covariance part. In the simple case of four points in $\mathbf{R}^2$ (so that $n \geq p+1$ is satisfied) it is not difficult to see that when one point is replaced, Oja's solution will always stay within the convex hull of the remaining three original points. Hence, even if $\mathbf{X}$ is in general position, $\lfloor \frac{n-p+1}{2} \rfloor / n$ is not generally valid as an upper bound for the breakdown point of affine equivariant location estimators.

The example of Oja's estimator seems to suggest that affine equivariant location estimators may have a breakdown point greater than $\lfloor \frac{n-p+1}{2} \rfloor / n$. This may be due to the fact that location estimators only break down if we can make them infinitely large by replacing points of $\mathbf{X}$, whereas covariance estimators also break down if we can make them infinitely 'small'. Therefore we may have to replace more points in order to let a location estimator break down.

In any case, the upper bound $\lfloor \frac{n+1}{2} \rfloor / n$ of Theorem 2.1 still holds, and we want to know how close we can get to this bound. The first example of an affine equivariant multivariate estimator with a high breakdown point was the Stahel-Donoho estimator. Donoho (1982) showed that it is affine equivariant and computed the addition breakdown point. By a slight adjustment of his proof one can show that if the collection $\mathbf{X}$ is in general position, the *replacement* breakdown point is equal to $(\lfloor \frac{n+1}{2} \rfloor - p)/n$, which is smaller than the upper bound $\lfloor \frac{n-p+1}{2} \rfloor / n$ for affine equivariant covariance estimators. We give two examples of estimators with a breakdown point that is equal to this upper bound.

Rousseeuw (1983) introduced the minimum volume ellipsoid (MVE) estimator, and showed it to be affine equivariant with breakdown point $\big(\lfloor \frac{n}{2} \rfloor - p + 1\big)/n$. Also this breakdown point is smaller than the covariance upper bound $\lfloor \frac{n-p+1}{2} \rfloor / n$. We will adjust the MVE estimator such that it does attain this upper bound.

MINIMUM VOLUME ELLIPSOID ESTIMATOR: Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ have $n \geq p+1$ points. Find $\mathbf{t}_n \in \mathbf{R}^p$ and $\mathbf{C}_n \in \mathrm{PDS}(p)$ that minimize the determinant of $\mathbf{C}$ subject to

$$\#\Big\{ i : (\mathbf{x}_i - \mathbf{t})^T \mathbf{C}^{-1} (\mathbf{x}_i - \mathbf{t}) \leq c^2 \Big\} \geq \lfloor \tfrac{n+p+1}{2} \rfloor. \tag{3.1}$$

Hence, $\mathbf{t}_n$ and $\mathbf{C}_n$ determine the center and the covariance structure of the minimum volume ellipsoid covering at least $\lfloor \frac{n+p+1}{2} \rfloor$ points. When every subcollection of $\lfloor \frac{n+p+1}{2} \rfloor$ points of $\mathbf{X}$ contains at least $p+1$ points in general position, there exists at least one solution $(\mathbf{t}_n, \mathbf{C}_n)$ in $\mathbf{R}^p \times \mathrm{PDS}(p)$. Even if some $\lfloor \frac{n+p+1}{2} \rfloor$ points lie on a lower dimensional hyperplane H, then one can still define $\mathbf{t}_n \in \mathbf{R}^p$ as the center of the minimum volume ellipsoid inside H covering at least $\lfloor \frac{n+p+1}{2} \rfloor$ points.

The number c is a fixed chosen constant and has no influence on the value of $\mathbf{t}_n$, which is taken as the MVE estimator of location. However, the choice of c determines the magnitude of $\mathbf{C}_n$, which can be taken as the MVE estimator of covariance. The value of c can be chosen in agreement with an assumed underlying distribution in order to obtain a consistent covariance estimator. For instance, if one assumes $X_1, \ldots, X_n$ to be a sample from an elliptical distribution $P_{\mu,\Sigma}$ with density $|\mathbf{B}|^{-1} f\big(\|\mathbf{B}^{-1}(\mathbf{x} - \boldsymbol{\mu})\|\big)$, where $\mathbf{B}\mathbf{B}^T = \boldsymbol{\Sigma}$, then a natural choice for c would be the value for which $P_{\mu,\Sigma}\{(X-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (X - \boldsymbol{\mu}) \leq c^2\} = \int f(\|\mathbf{x}\|)\{\|\mathbf{x}\| \leq c\}\, d\mathbf{x}$ is equal to $\frac{1}{2}$. In case one assumes $X_1, \ldots, X_n$ to be a normal sample, c^2 will be $\chi^2_{0.50}(p)$. An algorithm to compute $\mathbf{t}_n$ and $\mathbf{C}_n$ is described in Rousseeuw and Leroy (1987, p.259).

Before we derive the breakdown point of the MVE estimators, we first prove the following property for ellipsoids

$$E(\mathbf{t}, \mathbf{C}) = \{\mathbf{x} : (\mathbf{x} - \mathbf{t})^T \mathbf{C}^{-1} (\mathbf{x} - \mathbf{t}) \leq 1\} \tag{3.2}$$

where $\mathbf{t} \in \mathbf{R}^p$ and $\mathbf{C} \in \mathrm{PDS}(p)$.

LEMMA 3.1. *Consider* $\mathbf{v}_1, \ldots, \mathbf{v}_{p+1} \in \mathbf{R}^p$ *that span a nonempty simplex. Let* $E(\mathbf{t}, \mathbf{C})$ *be an ellipsoid as in (3.2), which contains* $\mathbf{v}_1, \ldots, \mathbf{v}_{p+1}$. *Then for every* $V > 0$ *there exists a constant* $M > 0$, *which only depends on* $\mathbf{v}_1, \ldots, \mathbf{v}_{p+1}$, *such that if* $\|\mathbf{t}\| > M$ *it follows that the volume of* $E(\mathbf{t}, \mathbf{C})$ *is larger than* V.

PROOF: Denote by $0 < \lambda_p \leq \ldots \leq \lambda_1 < \infty$ the eigenvalues of $\mathbf{C}$. The volume of $E(\mathbf{t}, \mathbf{C})$ is equal to $\alpha_p \sqrt{\lambda_1 \ldots \lambda_p}$ where $\alpha_p = \pi^{p/2}/\Gamma(\frac{p}{2}+1)$, and the axes of $E(\mathbf{t}, \mathbf{C})$ have lengths $\sqrt{\lambda_j}$ for $j = 1, \ldots, p$.

Because $E(\mathbf{t}, \mathbf{C})$ contains the nonempty simplex spanned by $\mathbf{v}_1, \ldots, \mathbf{v}_{p+1}$, there exists a constant $\beta > 0$, which only depends on $\mathbf{v}_1, \ldots, \mathbf{v}_{p+1}$, such that all axes are longer than β, i.e. for all $j = 1, \ldots, p$

$$\lambda_j > \beta^2. \tag{3.3}$$

Without loss of generality we may assume that $\mathbf{0} \in E(\mathbf{t}, \mathbf{C})$. According to (2.3), for every $\mathbf{v} \in E(\mathbf{t}, \mathbf{C})$ we have that $\|\mathbf{v} - \mathbf{t}\|^2 \leq (\mathbf{v} - \mathbf{t})^T \mathbf{C}^{-1} (\mathbf{v} - \mathbf{t}) \lambda_1 \leq \lambda_1$. In particular this holds for $\mathbf{v} = \mathbf{0}$, so that $\|\mathbf{t}\|^2 \leq \lambda_1$. This means that if we take $M = V/(\alpha_p \beta^{p-1})$, then from (3.3) it follows that the volume of $E(\mathbf{t}, \mathbf{C})$ is equal to $\alpha_p \sqrt{\lambda_1 \ldots \lambda_p} \geq \alpha_p \beta^{p-1} > V$. □

THEOREM 3.1. *Let* $\mathbf{X}$ *be a collection of* $n \geq p+1$ *points in* $\mathbf{R}^p$ *in general position, and let* $\mathbf{t}_n$ *and* $\mathbf{C}_n$ *be the MVE estimators of location and covariance. If* $p = 1$, *then* $\varepsilon^*(\mathbf{t}_n, \mathbf{X}) = \lfloor \frac{n+1}{2} \rfloor / n$ *and* $\varepsilon^*(\mathbf{C}_n, \mathbf{X}) = \lfloor \frac{n}{2} \rfloor / n$. *When* $p \geq 2$, *then* $\varepsilon^*(\mathbf{t}_n, \mathbf{X}) = \varepsilon^*(\mathbf{C}_n, \mathbf{X}) = \lfloor \frac{n-p+1}{2} \rfloor / n$.

PROOF: We extend the proof of Proposition 3.1 in Rousseeuw (1983). Without loss of generality we may assume that $c = 1$ in (3.1). When $p = 1$, $\mathbf{t}_n$ is the midpoint of the shortest interval covering at least $\lfloor \frac{n}{2} \rfloor + 1$ points, and $\mathbf{C}_n$ is proportional to the length of this interval. Even if this interval would have length zero, $\mathbf{t}_n$ is always defined. It is not difficult to see that one needs to replace at least $\lfloor \frac{n+1}{2} \rfloor$ points to make $\|\mathbf{t}_n\|$ infinitely large. By placing $\lfloor \frac{n}{2} \rfloor$ points in one of the remaining $n - \lfloor \frac{n-1}{2} \rfloor$ points, $\mathbf{C}_n$ can be made zero.

For $p \geq 2$, we first show that $\varepsilon^*(\mathbf{t}_n, \mathbf{X})$ and $\varepsilon^*(\mathbf{C}_n, \mathbf{X})$ are at least $\lfloor \frac{n-p+1}{2} \rfloor / n$. Replace at most $m = \lfloor \frac{n-p+1}{2} \rfloor - 1$ points of $\mathbf{X}$. Because every subcollection of $\lfloor \frac{n+p+1}{2} \rfloor$ points of the corrupted collection $\mathbf{Y}_m$ contains at least $\lfloor \frac{n+p+1}{2} \rfloor - \left(\lfloor \frac{n-p+1}{2} \rfloor - 1 \right) = p + 1$ points $\mathbf{x}_i$ of the original collection $\mathbf{X}$ in general position, there exists at least one solution $\left(\mathbf{t}_n(\mathbf{Y}_m), \mathbf{C}_n(\mathbf{Y}_m)\right)$ in $\mathbf{R}^p \times \mathrm{PDS}(p)$. Denote by $E_m = E\left(\mathbf{t}_n(\mathbf{Y}_m), \mathbf{C}_n(\mathbf{Y}_m)\right)$ the minimum volume ellipsoid of type (3.2) covering at least $\lfloor \frac{n+p+1}{2} \rfloor$ points of $\mathbf{Y}_m$.

Let V denote the volume of the smallest sphere with center $\mathbf{0}$ containing all points of $\mathbf{X}$. The corrupted collection $\mathbf{Y}_m$ still contains at least $n - \left(\lfloor \frac{n-p+1}{2} \rfloor - 1 \right) \geq \lfloor \frac{n+p+1}{2} \rfloor$ points of $\mathbf{X}$. The smallest sphere with center $\mathbf{0}$ containing these $\lfloor \frac{n+p+1}{2} \rfloor$ points of $\mathbf{X}$ must then have a volume less than V. At the same time this sphere is also an ellipsoid containing at least $\lfloor \frac{n+p+1}{2} \rfloor$ points of $\mathbf{Y}_m$. Therefore E_m, being the smallest ellipsoid of this kind, must also have a volume less than V. On the other

hand the ellipsoid E_m covers some subcollection of $\lfloor \frac{n+p+1}{2} \rfloor$ points of $\mathbf{Y}_m$. As we have seen above, such a subcollection must contain $p+1$ points $\mathbf{x}_i$ of the original collection $\mathbf{X}$ in general position. Since these $p+1$ $\mathbf{x}_i's$ span a nonempty simplex, it follows from Lemma 3.1 that there exists a constant $M > 0$, which only depends on $\mathbf{X}$, such that $\|\mathbf{t}_n(\mathbf{Y}_m)\| > M$ would force the volume of E_m to be greater than V. As we have just seen that this cannot be the case we conclude that

$$\|\mathbf{t}_n(\mathbf{Y}_m)\| \leq M. \tag{3.4}$$

These considerations about E_m also show that the covariance estimator does not break down either. Similar to (3.3), the fact that E_m contains $p+1$ original $\mathbf{x}_i$ in general position implies that there exists a constant β, which only depends on $\mathbf{X}$, such that

$$\lambda_j\big(\mathbf{C}_n(\mathbf{Y}_m)\big) > \beta^2 > 0 \tag{3.5}$$

for $j = 1, \ldots, p$. Since the volume of E_m, which is proportional to the product of the eigenvalues, is always less than V, there must also exist a constant $0 < \alpha < \infty$, which only depends on $\mathbf{X}$, such that $\lambda_1\big(\mathbf{C}_n(\mathbf{Y}_m)\big) \leq \alpha$. Together with (3.4) and (3.5) this proves that both $\varepsilon^*(\mathbf{t}_n, \mathbf{X})$ and $\varepsilon^*(\mathbf{C}_n, \mathbf{X})$ are at least $\lfloor \frac{n-p+1}{2} \rfloor / n$.

For the affine equivariant covariance estimator $\mathbf{C}_n$, the value $\lfloor \frac{n-p+1}{2} \rfloor / n$ is also an upper bound, therefore $\varepsilon^*(\mathbf{C}_n, \mathbf{X}) = \lfloor \frac{n-p+1}{2} \rfloor / n$. For $\varepsilon^*(\mathbf{t}_n, \mathbf{X})$ the other inequality is obtained as follows. Take any p points of $\mathbf{X}$ and consider the $(p-1)$-dimensional hyperplane H they determine. Replace $m = \lfloor \frac{n-p+1}{2} \rfloor$ other points of $\mathbf{X}$ by points on H. Then H contains $\lfloor \frac{n-p+1}{2} \rfloor + p = \lfloor \frac{n+p+1}{2} \rfloor$ points of the corrupted collection $\mathbf{Y}_m$. The minimum volume ellipsoid covering these $\lfloor \frac{n+p+1}{2} \rfloor$ points has a zero volume. Because $\mathbf{X}$ is in general position we can construct $\mathbf{Y}_m$ such that no other lower dimensional hyperplane contains $\lfloor \frac{n+p+1}{2} \rfloor$ points of $\mathbf{Y}_m$, therefore $\mathbf{t}_n(\mathbf{Y}_m)$ must lie on H. By sending the contaminated points on H to infinity, one of the axes of E_m becomes infinitely large, and so that the center $\mathbf{t}_n(\mathbf{Y}_m)$ of E_m becomes infinitely large. This proves $\varepsilon^*(\mathbf{t}_n, \mathbf{X}) \leq \lfloor \frac{n-p+1}{2} \rfloor / n$. □

The MVE location estimator suffers from the same poor rate of convergence as the least median of squares (LMS) regression estimator (Rousseeuw 1984). In order to obtain $\sqrt{n}$-consistency, Rousseeuw and Yohai (1984) considered smoothed versions of the LMS estimator. These S-estimators generalize easily to multivariate location and covariance, in which case they become smoothed versions of the MVE estimator.

S-ESTIMATORS: Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ have $n \geq p+1$ points. Find $\mathbf{t}_n \in \mathbf{R}^p$ and $\mathbf{C}_n \in \mathrm{PDS}(p)$ that minimize the determinant of $\mathbf{C}$ subject to

$$\frac{1}{n} \sum_{i=1}^{n} \rho \left[\{(\mathbf{x}_i - \mathbf{t})^T \mathbf{C}^{-1} (\mathbf{x}_i - \mathbf{t})\}^{1/2} \right] \leq b. \tag{3.6}$$

Note that one obtains the MVE estimators when $nb = n - \lfloor \frac{n+p+1}{2} \rfloor$ and $\rho(\cdot) = 1 - \mathbf{1}_{[-c,c]}(\cdot)$. Rousseeuw and Yohai (1984), aiming at both asymptotic normality and a high breakdown point, assumed the following conditions on ρ :

(R1) ρ is symmetric, twice continuously differentiable, and $\rho(0) = 0$.

(R2) There exists a constant $c > 0$ such that ρ is strictly increasing on $[0, c]$ and constant on $[c, \infty)$.

A typical example of such a ρ-function is the biweight function $\rho_{\text{B}}(y; c)$, which is defined as

$$\rho_{\text{B}}(y; c) = \begin{cases} \dfrac{y^2}{2} - \dfrac{y^4}{2c^2} + \dfrac{y^6}{6c^4}, & |y| \leq c \\ \dfrac{c^2}{6}, & |y| \geq c \end{cases}$$

Let $r = b/\sup\rho$ and denote by $\lceil y \rceil$ the smallest integer greater than or equal to y. When every subcollection of $\lceil n - nr \rceil$ points of $\mathbf{X}$ contains at least $p + 1$ points in general position, there exists at least one solution $(\mathbf{t}_n, \mathbf{C}_n)$ in $\mathbf{R}^p \times \text{PDS}(p)$. The constant $0 < b < \sup\rho$ can be choosen in agreement with an assumed underlying distribution. If one assumes $X_1, \ldots, X_n$ to be a sample from an elliptical distribution $P_{\mu,\Sigma}$ a natural choice for b is $\mathbf{E}\rho[\{(X_1 - \mu)^T \Sigma^{-1} (X_1 - \mu)\}^{1/2}] = \int \rho(\|\mathbf{x}\|) f(\|\mathbf{x}\|)\, d\mathbf{x}$. The choice of the (tuning) constant c then determines the value of b.

The properties of S-estimators have been investigated by Davies (1987) and Lopuhaä (1989). S-estimators defined by ρ-functions satisfying (R1)-(R2) have exactly the same asymptotic behaviour as multivariate M-estimators defined with the same ρ-function (Lopuhaä 1989). However, in contrast with M-estimators, S-estimators have a high breakdown point in any dimension p. In order to encompass S-estimators defined by smooth ρ such as $\rho_{\text{B}}(y; c)$, we complement the breakdown result of Davies (1987), who considers functions ρ that are equal to 0 in a neighbourhood of the origin.

THEOREM 3.2. *Let $\mathbf{X}$ be a collection of $n \geq p + 1$ points in $\mathbf{R}^p$ in general position. Write $r = b/\sup\rho$. If $r \leq (n - p)/(2n)$ then S-estimators defined by a function ρ that satisfies (R1)-(R2) have breakdown point $\varepsilon^*(\mathbf{t}_n, \mathbf{X}) = \varepsilon^*(\mathbf{C}_n, \mathbf{X}) = \lceil nr \rceil / n$.*

PROOF: The proof is similar to that of Theorem 3.1. As we can always rescale the function ρ we may assume that $c = 1$ and that $\sup\rho = 1$, so that b in (3.6) equals r. We first show that $\varepsilon^*(\mathbf{t}_n, \mathbf{X})$ and $\varepsilon^*(\mathbf{C}_n, \mathbf{X})$ are at least $\lceil nr \rceil / n$. Replace at most $m = \lceil nr \rceil - 1$ points of $\mathbf{X}$. Because $r \leq (n-p)/(2n)$, every subcollection of $\lceil n - nr \rceil$ points of the corrupted collection $\mathbf{Y}_m$ contains at least $\lceil n - nr \rceil - (\lceil nr \rceil - 1) \geq p + 1$ points $\mathbf{x}_i$ of the original collection $\mathbf{X}$ in general position. So there exists at least one solution $\big(\mathbf{t}_n(\mathbf{Y}_m), \mathbf{C}_n(\mathbf{Y}_m)\big)$ in $\mathbf{R}^p \times \text{PDS}(p)$. Denote by $E_m = E\big(\mathbf{t}_n(\mathbf{Y}_m), \mathbf{C}_n(\mathbf{Y}_m)\big)$ the smallest ellipsoid of type (3.2) that satisfies (3.6).

Since $nr - \lceil nr \rceil + 1$ is always strictly positive and ρ is continuous, we can find a smallest sphere with center $\mathbf{0}$ and radius, say R, such that $\sum_{i=1}^{n} \rho(\|\mathbf{x}_i\|/R) = nr - \lceil nr \rceil + 1$. Denote by V the volume of this sphere. The collection $\mathbf{Y}_m$ contains $n - m$ points of $\mathbf{X}$, say $\mathbf{x}_1, \ldots, \mathbf{x}_{n-m}$. The smallest sphere with center $\mathbf{0}$ and radius M such that for these points $\sum_{i=1}^{n-m} \rho(\|\mathbf{x}_i\|/M) = nr - \lceil nr \rceil + 1$ must then have a

volume less than V. At the same time this sphere is an ellipsoid for which

$$\sum_{\mathbf{y}_i \in \mathbf{Y}_m} \rho(\|\mathbf{y}_i\|/M) \leq \sum_{i=1}^{n-m} \rho(\|\mathbf{x}_i\|/M) + \lceil nr \rceil - 1 = nr.$$

Therefore E_m, being the smallest ellipsoid of this kind, must also have a volume less than V. On the other hand, it follows from constraint (3.6) that E_m must cover some subcollection of $\lceil n - nr \rceil$ points of $\mathbf{Y}_m$. As we have seen above, such a subcollection must contain $p+1$ points $\mathbf{x}_i$ of the original collection $\mathbf{X}$ in general position. At this point, we invoke Lemma 3.1 and use exactly the same argument as is in the first part of the proof of Theorem 3.1 to conclude that $\varepsilon^*(\mathbf{t}_n, \mathbf{X})$ and $\varepsilon^*(\mathbf{C}_n, \mathbf{X})$ are at least $\lceil nr \rceil / n$.

The other inequalities are obtained as follows. Replace $m = \lceil nr \rceil$ points of $\mathbf{X}$. Without loss of generality denote the corrupted collection by $\mathbf{Y}_m = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$, with $\mathbf{y}_j = \mathbf{x}_j$ for $j = \lceil nr \rceil + 1, \ldots, n$. Let $E(\mathbf{t}, \mathbf{C})$ be any ellipsoid of type (3.2) that satisfies

$$\sum_{i=1}^{n} \rho\left[\{(\mathbf{y}_i - \mathbf{t})^T \mathbf{C}^{-1} (\mathbf{y}_i - \mathbf{t})\}^{1/2}\right] \leq nr \tag{3.7}$$

and suppose that all replaced points $\mathbf{y}_1, \ldots, \mathbf{y}_{\lceil nr \rceil}$ are outside $E(\mathbf{t}, \mathbf{C})$. Then $\sum_{i=1}^{n} \rho[\{(\mathbf{y}_i - \mathbf{t})^T \mathbf{C}^{-1} (\mathbf{y}_i - \mathbf{t})\}^{1/2}]$ would be equal to

$$\sum_{j=\lceil nr \rceil + 1}^{n} \rho\left[\{(\mathbf{x}_j - \mathbf{t})^T \mathbf{C}^{-1} (\mathbf{x}_j - \mathbf{t})\}^{1/2}\right] + \lceil nr \rceil. \tag{3.8}$$

When $nr \in \mathbf{N}$, it follows from $r \leq (n-p)/(2n)$ that $n - \lceil nr \rceil \geq p + nr \geq p + 1$. In that case the summation in (3.8) runs over at least $p+1$ points in general position. Since ρ is strictly increasing it then follows that this sum must be strictly positive. When $nr \notin \mathbf{N}$, then $\lceil nr \rceil > nr$. Either way, we would find that (3.8) is strictly greater than nr. This is a contradiction with the fact that $E(\mathbf{t}, \mathbf{C})$ satisfies (3.7), so we conclude that at least one replacement, say $\mathbf{y}_1$, must be inside $E(\mathbf{t}, \mathbf{C})$.

Similarly, suppose that all $n - \lceil nr \rceil$ original points $\mathbf{x}_{\lceil nr \rceil + 1}, \ldots, \mathbf{x}_n$ are outside $E(\mathbf{t}, \mathbf{C})$. In that case we would find that

$$\sum_{i=1}^{n} \rho\left[\{(\mathbf{y}_i - \mathbf{t})^T \mathbf{C}^{-1} (\mathbf{y}_i - \mathbf{t})\}^{1/2}\right] \geq n - \lceil nr \rceil. \tag{3.9}$$

However, from $r \leq (n-p)/(2n)$ it follows that $\lceil n - nr \rceil - \lceil nr \rceil \geq p$. Because always $n - \lceil n - nr \rceil > nr - 1$, we would find that the right hand side of (3.9) is strictly greater than nr. As $E(\mathbf{t}, \mathbf{C})$ satisfies (3.7) this cannot be the case, so apart from $\mathbf{y}_1$ the ellipsoid $E(\mathbf{t}, \mathbf{C})$ must also contain at least one original $\mathbf{x}_i$.

By sending $\mathbf{y}_1$ to infinity we can make one of the axes of $E(\mathbf{t}, \mathbf{C})$ infinitely large. This means that for every $\mathbf{t}$ and $\mathbf{C}$ that satisfy (3.7), we can make both $\|\mathbf{t}\|$ and the largest eigenvalue $\lambda_1(\mathbf{C})$ infinitely large. Since $\mathbf{t}_n(\mathbf{Y}_m)$ and $\mathbf{C}_n(\mathbf{Y}_m)$ must satisfy (3.7) both estimates break down. □

REMARK 3.2: The breakdown point in Theorem 3.2 attains its largest value when $r = (n-p)/(2n)$. In that case the S-estimators have breakdown point $\lceil \frac{n-p}{2} \rceil / n = \lfloor \frac{n-p+1}{2} \rfloor / n$.

4. Breakdown and large deviations. The replacement breakdown point as defined in Section 2 is not only a simple and appealing robustness concept. Recently, He et al. (1988) showed that it also has a stochastic interpretation. We extend their result to multivariate location estimators.

In this section we consider $\mathbf{x}_1, \ldots, \mathbf{x}_n$ as a sample $X_1, \ldots, X_n$ from a spherically symmetric distribution P_μ with a density of the form $f(\|\mathbf{x}-\mu\|)$, $f(y) > 0$. We say that P_μ is *algebraically* tailed, if for some $m > 0$

$$-\log P_\mu(\|X_1 - \mu\| > a) \sim m \log a \text{ , as } a \to \infty. \tag{4.1}$$

Examples are the multivariate Cauchy distribution and the multivariate Student distribution. We say that P_μ is *exponentially* tailed, if for some $b > 0$ and $r > 0$

$$-\log P_\mu(\|X_1 - \mu\| > a) \sim b\, a^r \text{ , as } a \to \infty.$$

The multivariate normal is an example of such a distribution.

Jurečková (1981) considered

$$B(a, \mathbf{t}_n) = \frac{-\log P_\mu(\|\mathbf{t}_n - \mu\| > a)}{-\log P_\mu(\|X_1 - \mu\| > a)}$$

as a measure of tailperformance for $\mathbf{t}_n = \mathbf{t}_n(X_1, \ldots, X_n)$, and showed that in the case $p = 1$, under certain conditions on $\mathbf{t}_n$, it holds that

$$1 \le \liminf_{a\to\infty} B(a, \mathbf{t}_n) \le \limsup_{a\to\infty} B(a, \mathbf{t}_n) \le n.$$

For exponentially tailed distributions, the sample mean $\bar{X}_n$ performes optimally with $B(a, \bar{X}_n)$ tending to n, while for algebraically tailed distributions the lack of robustness of $\bar{X}_n$ is illustrated by $B(a, \bar{X}_n)$ tending to 1. In the multivariate setting one has something similar : when $X_1, \ldots, X_n$ have a standard normal distribution, then $\bar{X}_n$ has the same distribution as $n^{-\frac{1}{2}} X_1$, so that $B(a, \bar{X}_n)$ tends to n, and when $X_1, \ldots, X_n$ have a multivariate Cauchy distribution, then $\bar{X}_n$ and X_1 are equally distributed, so that $B(a, \bar{X}_n) = 1$.

Let $\mathbf{t}_n$ be an estimator of multivariate location. We say that $\mathbf{t}_n$ is *scale equivariant*, when $\mathbf{t}_n(\lambda \mathbf{x}_1, \ldots, \lambda \mathbf{x}_n) = \lambda \mathbf{t}_n(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ for all $\lambda > 0$. This type of equivariance is satisfied by most location estimators.

LEMMA 4.1. *Let $\mathbf{t}_n$ be translation and scale equivariant with a breakdown point $\varepsilon^*(\mathbf{t}_n, \mathbf{X}) = m^*/n$ that does not depend on $\mathbf{X}$ and suppose that for all $\mathbf{v} \in \mathbf{R}^p$*

$$\mathbf{t}_n(\mathbf{v}, \ldots, \mathbf{v}) = \mathbf{v}. \tag{4.2}$$

Then, for every collection $\mathbf{X}$ *with at least* $n - m^* + 1$ *points equal to some* $\mathbf{v} \in \mathbf{R}^p$, *it holds that* $\mathbf{t}_n(\mathbf{X}) = \mathbf{v}$.

PROOF: Since $\mathbf{t}_n$ is translation equivariant we may assume that $\mathbf{v} = \mathbf{0}$. Without loss of generality we may assume that $\mathbf{x}_{m^*} = \cdots = \mathbf{x}_n = \mathbf{0}$. Suppose that $\mathbf{t}_n(\mathbf{x}_1, \ldots, \mathbf{x}_{m^*-1}, \mathbf{0}, \ldots, \mathbf{0}) \neq \mathbf{0}$. Then for every $\lambda > 0$ we have that

$$\begin{aligned} &\|\mathbf{t}_n(\lambda\mathbf{x}_1, \ldots, \lambda\mathbf{x}_{m^*-1}, \mathbf{0}, \ldots, \mathbf{0}) - \mathbf{t}_n(\mathbf{0}, \ldots, \mathbf{0})\| \\ &\quad = \|\mathbf{t}_n(\lambda\mathbf{x}_1, \ldots, \lambda\mathbf{x}_{m^*-1}, \mathbf{0}, \ldots, \mathbf{0})\| \\ &\quad = \lambda\|\mathbf{t}_n(\mathbf{x}_1, \ldots, \mathbf{x}_{m^*-1}, \mathbf{0}, \ldots, \mathbf{0})\| \end{aligned}$$

using (4.2) and scale equivariance of $\mathbf{t}_n$. When λ tends to infinity, the right hand side tends to infinity, which means that $\mathbf{t}_n$ breaks down at the collection $(\mathbf{0}, \ldots, \mathbf{0})$ by replacing $m^* - 1$ points. This is in contradiction to the definition of m^*. □

He et al. (1988) related $B(a, \mathbf{t}_n)$ to the finite-sample replacement breakdown point $\varepsilon^*(\mathbf{t}_n, \mathbf{X}) = m^*/n$ of univariate location estimators $\mathbf{t}_n$ that are monotone in each observation. A function $g : \mathbf{R}^p \to \mathbf{R}^p$ is called monotone when $g(\mathbf{x}) \geq g(\mathbf{y})$ for $\mathbf{x} \geq \mathbf{y}$, where $\mathbf{x} \geq \mathbf{y}$ means $x_j \geq y_j$ for $j = 1, \ldots, p$.

THEOREM 4.1. *Let* $\mathbf{X} = (X_1, \ldots, X_n)$ *be a sample from a spherically symmetric distribution* P_μ *with density* $f(\|\mathbf{x} - \mu\|) > 0$. *Let* $\mathbf{t}_n(X_1, \ldots, X_n)$ *be a translation and scale equivariant estimator that satisfies (4.2) and which has a breakdown point* $\varepsilon^*(\mathbf{t}_n, \mathbf{X}) = m^*/n$ *that is independent of* $\mathbf{X}$. *Assume that* $\mathbf{t}_n$ *is monotone in each observation. If* P_μ *is algebraically tailed, then it holds that*

$$m^* \leq \liminf_{a \to \infty} B(a, \mathbf{t}_n) \leq \limsup_{a \to \infty} B(a, \mathbf{t}_n) \leq n - m^* + 1.$$

PROOF: Since $\mathbf{t}_n$ is translation equivariant we can restrict ourselves to the case $\mu = \mathbf{0}$, and write P for P_0. Monotonicity of $\mathbf{t}_n$ and Lemma 4.1 yield the following property

$$\begin{aligned} &x_{ij} \geq c, \text{ for } n - m^* + 1 \text{ points } \mathbf{x}_i \Longrightarrow t_{nj}(\mathbf{x}_1, \ldots, \mathbf{x}_n) \geq c \\ &x_{ij} \leq c, \text{ for } n - m^* + 1 \text{ points } \mathbf{x}_i \Longrightarrow t_{nj}(\mathbf{x}_1, \ldots, \mathbf{x}_n) \leq c \end{aligned} \tag{4.3}$$

for all $j = 1, \ldots, p$ and $c \in \mathbf{R}$. Indeed, assume without loss of generality that $x_{ij} \geq c$ for $i = 1, \ldots, n - m^* + 1$. Define the vector $\mathbf{v} = (v_1 \cdots v_p)^T$ as follows : $v_j = c$ and $v_k = \min\{x_{1k}, \ldots, x_{nk}\}$ for $k \neq j$. Then $\mathbf{x}_i \geq \mathbf{v}$ for $i = 1, \ldots, n - m^* + 1$. Monotonicity together with Lemma 4.1 implies $\mathbf{t}_n(\mathbf{x}_1, \ldots, \mathbf{x}_n) \geq \mathbf{v}$ and hence, $t_{nj}(\mathbf{x}_1, \ldots, \mathbf{x}_n) \geq c$. The other inequality is obtained similarly.

Because $X_1 = (X_{11} \cdots X_{1p})^T$ is spherically symmetric distributed, it holds that

$$P\big(\|X_1\| > a\big) \leq 2pP(X_{11} > a/\sqrt{p}).$$

Hence,

$$P(\|X_1\| > a)^{n-m^*+1} \leq (2p)^{n-m^*+1} P\Big(\bigcap_{i=1}^{n-m^*+1} \{X_{i1} > a/\sqrt{p}\}\Big). \tag{4.4}$$

According to property (4.3), the event on the right hand side of (4.4) implies that $t_{n1}(X_1, \ldots, X_n) > a/\sqrt{p}$ and hence, $\|\mathbf{t}_n\| > a/\sqrt{p}$. It follows that

$$P(\|X_1\| > a\sqrt{p})^{n-m^*+1} \leq (2p)^{n-m^*+1} P(\|\mathbf{t}_n\| > a). \tag{4.5}$$

Taking logarithms yields

$$B(a, \mathbf{t}_n) \leq (n - m^* + 1)\frac{-\log P(\|X_1\| > a\sqrt{p})}{-\log P(\|X_1\| > a)} + \frac{(n - m^* + 1)\log(2p)}{-\log P(\|X_1\| > a)} \tag{4.6}$$

so that with (4.1) we obtain $\limsup_{a\to\infty} B(a, \mathbf{t}_n) \leq n - m^* + 1$.

The lower bound on $\liminf_{a\to\infty} B(a, \mathbf{t}_n)$ is obtained similarly. We have that $\|\mathbf{t}_n\| > a$ implies that $|t_{nj}| > a/\sqrt{p}$ for some $1 \leq j \leq p$. According to property (4.3), $t_{nj} > a/\sqrt{p}$ implies that $X_{ij} > a/\sqrt{p}$ at least m^* times and hence, $\|X_i\| > a/\sqrt{p}$ at least m^* times. Similarly, $t_{nj} < -a/\sqrt{p}$ also implies that $\|X_i\| > a/\sqrt{p}$ at least m^* times. Therefore,

$$P(\|\mathbf{t}_n\| > a) \leq 2p \sum_{k=m^*}^{n} \binom{n}{k} r^k (1-r)^{n-k} \leq r^{m^*}\, 2p \sum_{k=m^*}^{n} \binom{n}{k}$$

where $r = P(\|X_1\| > a/\sqrt{p})$. Taking logarithms yields an inequality similar to (4.6), in which we let a tend to ∞ and obtain $\liminf_{a\to\infty} B(a, \mathbf{t}_n) \geq m^*$. □

The inequality $\limsup_{a\to\infty} B(a, \mathbf{t}_n) \leq n - m^* + 1$ indicates that estimators with a high breakdown point necessarily must sacrifice tailperformance. However, both inequalities in Theorem 4.1 imply that estimators with maximal breakdown point satisfy a minimax property in the sense that they maximize least favorable tail performance at algebraically tailed distributions.

5. Breakdown Point of one-step reweighted estimators. A high breakdown point is often counterbalanced by a low asymptotic efficiency. A possible way to avoid this is to use robust estimators as a diagnostic tool to select the 'good' observations from a (corrupted) collection (see for instance Rousseeuw and van Zomeren 1990). Once the 'good' observations have been identified, classical methods could be applied to obtain final estimators of location and covariance. If the breakdown point of the initial robust estimators is retained, we may be able to combine a high breakdown point with high efficiency. In this section we will show that for the usual weighted sample mean and sample covariance the breakdown point of the initial estimators is retained. Asymptotic properties of these estimators are still under investigation.

Assume throughout this section that the collection $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ is in general position. Let $\mathbf{t}_{0,n}(\mathbf{X}) \in \mathbf{R}^p$ and $\mathbf{C}_{0,n}(\mathbf{X}) \in \mathrm{PDS}(p)$ denote initial (robust) estimators of location and covariance based on $\mathbf{X}$. For $i = 1, \ldots, n$ compute Mahalanobis distances

$$d_0(\mathbf{x}_i) = \sqrt{(\mathbf{x}_i - \mathbf{t}_{0,n}(\mathbf{X}))^T \mathbf{C}_{0,n}(\mathbf{X})^{-1} (\mathbf{x}_i - \mathbf{t}_{0,n}(\mathbf{X}))}.$$

Identify observations with relatively small $d_0(\mathbf{x}_i)$ as 'good' observations, and identify observations with relatively large $d_0(\mathbf{x}_i)$ as outliers. Next compute the weighted sample mean and sample covariance, by assigning smaller weights to outlying observations. Let $w : [0, \infty) \to [0, \infty)$ be a function satisfying :

(W1) $w(y)$ and $w(y)y^2$ is bounded and $w(y)$ is nonincreasing.
(W2) There exists a constant $c_0 > 0$ such that $w(y) > 0$ for $y \in [0, c_0]$.

Define weighted estimators by

$$\mathbf{t}_{1,n}(\mathbf{X}) = \frac{\sum_{i=1}^n w(d_0(\mathbf{x}_i))\mathbf{x}_i}{\sum_{i=1}^n w(d_0(\mathbf{x}_i))}$$

$$\mathbf{C}_{1,n}(\mathbf{X}) = \frac{\sum_{i=1}^n w(d_0(\mathbf{x}_i))(\mathbf{x}_i - \mathbf{t}_{1,n}(\mathbf{X}))(\mathbf{x}_i - \mathbf{t}_{1,n}(\mathbf{X}))^T}{\sum_{i=1}^n w(d_0(\mathbf{x}_i))}.$$

Additionally, to prevent the weighted covariance estimator from implosion, we need some relation between the constant c_0 and the initial estimators $\mathbf{t}_{0,n}$ and $\mathbf{C}_{0,n}$. Consider the (robust) ellipsoid

$$E(\mathbf{t}_{0,n}, \mathbf{C}_{0,n}, c_0) = \{\mathbf{x} : (\mathbf{x} - \mathbf{t}_{0,n})^T \mathbf{C}_{0,n}^{-1}(\mathbf{x} - \mathbf{t}_{0,n}) \le c_0^2\}.$$

Then c_0, $\mathbf{t}_{0,n}$ and $\mathbf{C}_{0,n}$ must be such that

$$\#\left\{i : \mathbf{x}_i \in E(\mathbf{t}_{0,n}, \mathbf{C}_{0,n}, c_0)\right\} \ge \lfloor \tfrac{n+p+1}{2} \rfloor. \tag{5.1}$$

Given any $\mathbf{t}_{0,n}$ and $\mathbf{C}_{0,n}$, one could take for c_0 any number that is greater than or equal to the $\lfloor \frac{n+p+1}{2} \rfloor$ largest Mahalanobis distance $d_0(\mathbf{x}_i)$. However, typical choices for $\mathbf{t}_{0,n}$ and $\mathbf{C}_{0,n}$ would be the MVE estimator or an S-estimator with $b/\sup\rho = (n-p)/(2n)$. In that case, c_0 does not have to be defined by the sample values, but could be any number that is greater than the tuning constant c of the function ρ in (3.6). Indeed, by definition the constant c will then always be such that the minimum volume ellipsoid or the 'S-ellipsoid' contains at least $\lfloor \frac{n+p+1}{2} \rfloor$ points. A typical choice for $w(\cdot)$ would be the function $\mathbf{1}_{[0,c_0]}(\cdot)$, in which case $\mathbf{t}_{1,n}$ and $\mathbf{C}_{1,n}$ are simply the sample mean and sample covariance of the 'good' observations.

It is not difficult to see that the affine equivariance of $\mathbf{t}_{0,n}$ and $\mathbf{C}_{0,n}$ carries over to $\mathbf{t}_{1,n}$ and $\mathbf{C}_{1,n}$. We will show that also the breakdown point of the initial estimators is retained. We need the following property of eigenvalues.

LEMMA 5.1. *For symmetric $p \times p$-matrices $\mathbf{A}$ and $\mathbf{B}$ it holds that $\lambda_1(\mathbf{A}+\mathbf{B}) \le \lambda_1(\mathbf{A}) + \lambda_1(\mathbf{B})$, and $\lambda_p(\mathbf{A}+\mathbf{B}) \ge \lambda_p(\mathbf{A}) + \lambda_p(\mathbf{B})$.*

PROOF: Apply (2.3) and use standard properties of infima and suprema. □

THEOREM 5.1. *Let $\mathbf{X}$ be a collection in general position with $n \ge p+1$ points in $\mathbf{R}^p$. Let $w(\cdot)$ satisfy (W1)-(W2) and let $\mathbf{t}_{0,n}(\mathbf{X}) \in \mathbf{R}^p$ and $\mathbf{C}_{0,n}(\mathbf{X}) \in PDS(p)$ be affine equivariant estimators of location and covariance that relate to c_0 as described above. Then*

$$\varepsilon^*(\mathbf{t}_{1,n}, \mathbf{X}) \ge \min\{\varepsilon^*(\mathbf{t}_{0,n}, \mathbf{X}), \varepsilon^*(\mathbf{C}_{0,n}, \mathbf{X})\} \tag{5.2}$$

and

$$\varepsilon^*(\mathbf{C}_{1,n}, \mathbf{X}) \geq \min\{\varepsilon^*(\mathbf{t}_{0,n}, \mathbf{X}), \varepsilon^*(\mathbf{C}_{0,n}, \mathbf{X})\}. \tag{5.3}$$

PROOF: Replace at most $m = n \min\{\varepsilon^*(\mathbf{t}_{0,n}, \mathbf{X}), \varepsilon^*(\mathbf{C}_{0,n}, \mathbf{X})\} - 1$ points of $\mathbf{X}$ and denote by $\mathbf{Y}_m = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ the new corrupted collection. Write $\mathbf{t}^*_{0,n}$ and $\mathbf{C}^*_{0,n}$ for $\mathbf{t}_{0,n}(\mathbf{Y}_m)$ and $\mathbf{C}_{0,n}(\mathbf{Y}_m)$, $\mathbf{t}^*_{1,n}$ and $\mathbf{C}^*_{1,n}$ similarly, and write

$$d_0^*(\mathbf{y}_i) = \sqrt{(\mathbf{y}_i - \mathbf{t}^*_{0,n})^T (\mathbf{C}^*_{0,n})^{-1} (\mathbf{y}_i - \mathbf{t}^*_{0,n})}.$$

Because $m \leq n\varepsilon^*(\mathbf{t}_{0,n}, \mathbf{X}) - 1$ and $m \leq n\varepsilon^*(\mathbf{C}_{0,n}, \mathbf{X}) - 1$ it follows that there exist constants k_0, k_1 and k_2, which only depend on $\mathbf{X}$, such that

$$\|\mathbf{t}^*_{0,n}\| \leq k_0 < \infty \quad \text{and} \quad 0 < k_1 \leq \lambda_p(\mathbf{C}^*_{0,n}) \leq \lambda_1(\mathbf{C}^*_{0,n}) \leq k_2 < \infty. \tag{5.4}$$

As $\mathbf{C}_{0,n}$ is an affine equivariant covariance estimator it holds that $m \leq \lfloor \frac{n-p+1}{2} \rfloor - 1$. It then follows from (5.1) that the corrupted ellipsoid $E(\mathbf{t}^*_{0,n}, \mathbf{C}^*_{0,n}, c_0)$ still covers $\lfloor \frac{n+p+1}{2} \rfloor - m \geq p + 1$ original points of $\mathbf{X}$. Without loss of generality assume that these points are $\mathbf{x}_1, \ldots, \mathbf{x}_{p+1}$. Because $w(\cdot)$ is nonincreasing

$$\sum_{i=1}^{n} w(d_0^*(\mathbf{y}_i)) \geq \sum_{j=1}^{p+1} w(d_0^*(\mathbf{x}_j)) \geq (p+1) w(c_0) > 0 \tag{5.5}$$

which means that the denominators of $\mathbf{t}^*_{1,n}$ and $\mathbf{C}^*_{1,n}$ will always be uniformly bounded away from zero.

We first show that $\|\mathbf{t}^*_{1,n}\|$ remains bounded. According to (2.3),

$$\begin{aligned} \left\| \sum_{i=1}^{n} w(d_0^*(\mathbf{y}_i)) \mathbf{y}_i \right\| &\leq \sum_{i=1}^{n} w(d_0^*(\mathbf{y}_i)) \|\mathbf{y}_i - \mathbf{t}^*_{0,n}\| + \sum_{i=1}^{n} w(d_0^*(\mathbf{y}_i)) \|\mathbf{t}^*_{0,n}\| \\ &\leq \sum_{i=1}^{n} w(d_0^*(\mathbf{y}_i)) d_0^*(\mathbf{y}_i) \lambda_1(\mathbf{C}^*_{0,n}) + \sum_{i=1}^{n} w(d_0^*(\mathbf{y}_i)) \|\mathbf{t}^*_{0,n}\| \end{aligned} \tag{5.6}$$

Because of (W1) and (5.4) it follows from (5.6) that there exists a constant A_0, which only depends on $\mathbf{X}$, such that

$$\|\mathbf{t}^*_{1,n}\| \leq A_0 < \infty \tag{5.7}$$

which proves (5.2).

Next we show that $\lambda_p(\mathbf{C}^*_{1,n})$ is uniformly bounded away from zero. Consider the numerator of $\mathbf{C}^*_{1,n}$ and write this as the sum $\mathbf{A} + \mathbf{B}$ of the matrices

$$\mathbf{A} = \sum_{i=1}^{p+1} w(d_0^*(\mathbf{x}_i))(\mathbf{x}_i - \mathbf{t}^*_{1,n})(\mathbf{x}_i - \mathbf{t}^*_{1,n})^T$$

$$\mathbf{B} = \sum_{i=p+2}^{n} w(d_0^*(\mathbf{y}_i))(\mathbf{y}_i - \mathbf{t}^*_{1,n})(\mathbf{y}_i - \mathbf{t}^*_{1,n})^T$$

where $\mathbf{x}_1, \ldots, \mathbf{x}_{p+1}$ have positive weight, as they are inside $E(\mathbf{t}^*_{0,n}, \mathbf{C}^*_{0,n}, c_0)$. Since both $\mathbf{A}$ and $\mathbf{B}$ are symmetric nonnegative matrices it follows from Lemma 5.1 that $\lambda_p(\mathbf{A}+\mathbf{B}) \geq \lambda_p(\mathbf{A})$. Also (2.3) implies that for $\alpha_1, \ldots, \alpha_k \geq \gamma > 0$ and $\mathbf{A}_1, \ldots, \mathbf{A}_k$ symmetric nonnegative it holds that $\lambda_p(\alpha_1\mathbf{A}_1 + \cdots + \alpha_k\mathbf{A}_k) \geq \gamma\lambda_p(\mathbf{A}_1 + \cdots + \mathbf{A}_k)$. Because $w(d^*_0(\mathbf{x}_i)) \geq w(c_0)$ for every $i = 1, \ldots, p+1$, it follows that $\lambda_p(\mathbf{A}) \geq w(c_0)\lambda_p(\mathbf{M})$, where $\mathbf{M} = \sum_{i=1}^{p+1}(\mathbf{x}_i - \mathbf{t}^*_{1,n})(\mathbf{x}_i - \mathbf{t}^*_{1,n})^T$. Write $\mathbf{M} = \mathbf{M}_1 + \mathbf{M}_2$, where

$$\mathbf{M}_1 = \sum_{i=1}^{p+1}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$

$$\mathbf{M}_2 = (p+1)(\bar{\mathbf{x}} - \mathbf{t}^*_{1,n})(\bar{\mathbf{x}} - \mathbf{t}^*_{1,n})^T$$

with $\bar{\mathbf{x}} = (p+1)^{-1}\sum_{i=1}^{p+1}\mathbf{x}_i$. Both $\mathbf{M}_1$ and $\mathbf{M}_2$ are symmetric and nonnegative, so that Lemma 5.1 yields $\lambda_p(\mathbf{M}) \geq \lambda_p(\mathbf{M}_1)$. The matrix $\mathbf{M}_1$ is proportional to the sample covariance matrix of $\mathbf{x}_1, \ldots, \mathbf{x}_{p+1}$ and as $\mathbf{X}$ is in general position, $\mathbf{M}_1$ must have a smallest eigenvalue $\lambda_p(\mathbf{M}_1) > 0$. We find that the smallest eigenvalue of the numerator of $\mathbf{C}^*_{1,n}$ is greater than $w(c_0)\lambda_p(\mathbf{M}_1) > 0$. It follows that there exists a constant $A_1 > 0$, which only depends on $\mathbf{X}$, such that $\lambda_p(\mathbf{C}^*_{1,n}) \geq A_1$.

Finally, for any $\mathbf{v} \in \mathbb{R}^p$ it holds that $\lambda_1(\mathbf{v}\mathbf{v}^T) = \|\mathbf{v}\|^2$. Together with Lemma 5.1 and (5.5), this implies that

$$\lambda_1(\mathbf{C}^*_{1,n}) \leq \big((p+1)w(c_0)\big)^{-1}\sum_{i=1}^{n} w\big(d^*_0(\mathbf{y}_i)\big)\|\mathbf{y}_i - \mathbf{t}^*_{1,n}\|^2.$$

We can bound $\|\mathbf{y}_i - \mathbf{t}^*_{1,n}\|^2$ from above by

$$\|\mathbf{y}_i - \mathbf{t}^*_{0,n}\|^2 + \|\mathbf{t}^*_{0,n} - \mathbf{t}^*_{1,n}\|^2 + 2\|\mathbf{y}_i - \mathbf{t}^*_{0,n}\| \cdot \|\mathbf{t}^*_{0,n} - \mathbf{t}^*_{1,n}\|$$

Then use (2.3) as in (5.6) and recall that $w(y)$ and $w(y)y^2$ are bounded. It follows together with (5.4) and (5.7) that there exists a constant A_2, which only depends on $\mathbf{X}$, such that $\lambda_1(\mathbf{C}^*_{1,n}) \leq A_2 < \infty$. This completes the proof. □

There are other suggestions to combine high asymptotic efficiency with a high breakdown point. Results of Beran (1977) show that if you estimate the center of symmetry of a univariate distribution by Minimum Hellinger Distance you get high efficiency combined with breakdown point 1/2 asymptotically. However, MHD estimation may not be practical in high dimensions because it depends on density estimation. Other Minimum Distance estimators are discussed by Donoho and Liu (1988). Yohai (1987) combined high breakdown point with high efficiency for regression estimators. This approach is extended in Lopuhaä (1988) to affine multivariate location estimators. Unfortunately, a similar approach for covariance estimators, i.e. first estimate the location parameter affinely with high breakdown point and then compute an M-estimate of covariance based on the recentered observations, would fail because covariance M-estimators have a low breakdown point (Tyler 1986).

Acknowledgments. We thank Laurie Davies, Dave Tyler and Bert van Zomeren for stimulating discussions, and the referee for helpful suggestions and remarks.

REFERENCES

BERAN, R.J. (1977). Minimum Hellinger distance estimates for parametric models. *Ann. Stat.* **5** 445–463.

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

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.

DONOHO, D.L. and LIU, R.C. (1988). The "automatic" robustness of minimum distance functionals. *Ann. Stat.* **16** 552–586.

HAMPEL, F.R. (1968). Contributions to the Theory of Robust Estimation. Ph.D.thesis, University of California at Berkeley.

HE, X., JUREČKOVÁ, J., KOENKER, R. and PORTNOY, S. (1988). Tail behaviour of regression estimators and their breakdown points. Technical Report, University of Illinois at Urbana-Champaign.

HODGES, J.L. Jr. (1967). Efficiency in normal samples and tolerance of extreme values for some estimates of location. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (L.M.Le Cam and J.Neyman, eds.) 163–186. University of California Press, Berkeley.

HUBER, P.J. (1984). Finite sample breakdown of M- and P-estimators. *Ann. Statist.* **12** 119–126.

JUREČKOVÁ, J. (1981). Tail-behaviour of location estimators. *Ann. Stat.* **9** 578–585.

LOPUHAÄ, H.P. (1988). Highly efficient estimators of multivariate location with high breakdown point. Revised version of Technical Report 88-14, Delft University of Technology.

LOPUHAÄ, H.P. (1989). On the relation between S-estimators and M-estimators of multivariate location and covariance. *Ann. Statist.* **17** 1662–1683.

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

OJA, H. (1983). Descriptive statistics for multivariate distributions. *Stat. and Prob. Letters* **1** 327–333.

ROUSSEEUW, P.J. (1983). Multivariate estimation with high breakdown point. Paper presented at the Fourth Pannonian Symposium on Mathematical Statistics and Probability, Bad Tatzmannsdorf, Austria, September 4-9, 1983. In *Mathematical Statistics and Applications (1985)* (W.Grossmann, G.Pflug, I.Vincze and W.Wertz, eds.) 283–297. Reidel, Dordrecht, The Netherlands.

ROUSSEEUW, P.J. (1984). Least median of squares regression. *J. Am. Stat. Assoc.* **79** 871–880.

ROUSSEEUW, P.J. and YOHAI, V. (1984). Robust regression by means of S-estimators. In *Robust and Nonlinear Time Series Analysis.* Lecture Notes in Statistics **26** 256–272. Springer Verlag, New York.

ROUSSEEUW, P.J. and LEROY, A.M. (1987). *Robust Regression and Outlier Detection.* Wiley, New York.

ROUSSEEUW, P.J. and VAN ZOMEREN, B.C. (1990). Unmasking multivariate outliers and leverage points. To appear J. Am. Stat. Assoc..

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

TYLER, D.E. (1986). Breakdown properties of the M-estimators of multivariate scatter. Technical Report, Rutgers University, New Jersey.

YOHAI, V.J. (1987). High breakdown-point and high efficiency robust estimates for regression. *Ann. Stat.* **15** 642–656.

ON THE RELATION BETWEEN S-ESTIMATORS AND M-ESTIMATORS OF MULTIVARIATE LOCATION AND COVARIANCE

HENDRIK P. LOPUHAÄ

Delft University of Technology

We discuss the relation between S-estimators and M-estimators of multivariate location and covariance. As in the case of the estimation of a multiple regression parameter, S-estimators are shown to satisfy first-order conditions of M-estimators. We show that the influence function $\mathrm{IF}(\mathbf{x}; \mathbf{S}, P)$ of S-functionals exists and is the same as that of corresponding M-functionals. Also, we show that S-estimators have a limiting normal distribution which is similar to the limiting normal distribution of M-estimators. Finally, we compare asymptotic variances and breakdown point of both types of estimators.

1. Introduction and preliminaries. Recently Rousseeuw and Yohai (1984) introduced S-estimators in the framework of multiple regression. These estimators were shown to have the same asymptotic properties as corresponding regression M-estimators, and also to have good robustness properties, as their breakdown point (which can be interpreted as the percentage of outliers in the sample that an estimator can handle) was shown to be 50% .

Davies (1987) investigated some properties of S-estimators of multivariate location and covariance. Using a slightly different definition from the one suggested in Rousseeuw and Yohai (1984) he treated existence, consistency, asymptotic normality and breakdown point. However, the close correspondence with multivariate M-estimators, as was found in the case of estimating a regression parameter, remained hidden.

In this paper multivariate S-estimators are related to multivariate M-estimators. First the definition of multivariate M- and S-estimators is discussed and it is shown that S-estimators of multivariate location and covariance satisfy the first-order conditions of multivariate M-estimators.

This will have the consequence that the asymptotic normality results and the expression for the influence function of multivariate S-estimators are the same as those of corresponding multivariate M-estimators.

Finally, we will compare asymptotic variances in relation with the breakdown point for both types of estimators. It turns out that S-estimators can achieve the variances attained by M-estimators, but they have the additional advantage that in high dimensions (at the same level of asymptotic variance) the breakdown point is considerably higher than that of the M-estimators. All proofs have been saved for an Appendix at the end of the paper.

This research is financially supported by NWO under Grant 10-62-10.

1980 *Mathematics subject classifications* : 62F35, 62H12.

Key words: S-estimators, M-estimators, Influence Function, Asymptotic normality, Efficiency.

We will denote p-vectors by $\mathbf{t} = (t_1 \cdots t_p)^T$ and $p \times p$-matrices by $\mathbf{M} = (m_{ij})$. For any $p \times p$-matrix $\mathbf{M}$, we write $\mathbf{D}_\mathbf{M}$ for the diagonal matrix consisting of the diagonal of $\mathbf{M}$, and eigenvalues are denoted by $\lambda_p(\mathbf{M}) \leq \ldots \leq \lambda_1(\mathbf{M})$. The class of positive definite symmetric matrices is written as $\mathrm{PDS}(p)$ and by $\Theta = \mathbf{R}^p \times \mathrm{PDS}(p)$ we denote the set of pairs $\boldsymbol{\theta} = (\mathbf{t}, \mathbf{C})$, which can be seen as an open subset of $\mathbf{R}^{p+\frac{1}{2}p(p+1)}$.

By $\mathbf{x}_1, \mathbf{x}_2, \ldots$ we will mean vectors in $\mathbf{R}^p$ and we will write $X_1, X_2, \ldots$ instead if an underlying distribution is assumed.

The Euclidean norm is denoted by $\|\cdot\|$, and because of the frequent appearance of quadratic forms $(\mathbf{x} - \mathbf{t})^T\mathbf{C}^{-1}(\mathbf{x} - \mathbf{t})$ we will sometimes abbreviate them by $d^2(\mathbf{x}, \mathbf{t}, \mathbf{C})$. Denote by $E(\mathbf{t}, \mathbf{C}, c)$ an ellipsoid $\{\mathbf{x} : (\mathbf{x} - \mathbf{t})^T\mathbf{C}^{-1}(\mathbf{x} - \mathbf{t}) \leq c^2\}$. Partial derivatives $\partial g(\mathbf{x}, \boldsymbol{\theta})/\partial\boldsymbol{\theta}$ will sometimes be abbreviated by $g_{\boldsymbol{\theta}}(\mathbf{x}, \boldsymbol{\theta})$.

We will focus on the estimation of the parameter $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ which characterizes an elliptical distribution $P_{\mu,\Sigma}$ with a density of the form

$$|\mathbf{B}|^{-1} f\big(\|\mathbf{B}^{-1}(\mathbf{x} - \boldsymbol{\mu})\|\big) \tag{1.1}$$

where $\mathbf{B}\mathbf{B}^T = \boldsymbol{\Sigma}$, $f : [0, \infty) \to [0, \infty)$ is a fixed function and $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \Theta$. Expectations with respect to these distributions are denoted by $\mathbf{E}_{\mu,\Sigma}$. Note that it is often easier to write $\mathbf{E}_{0,I}h(\|X\|)$ instead of $\mathbf{E}_{\mu,\Sigma}h\big(\|\mathbf{B}^{-1}(X - \boldsymbol{\mu})\|\big)$ for any real-valued function h.

2. *M*-estimators and *S*-estimators.

2.1 M-estimators. M-estimators were originally constructed by Huber (1964) for the estimation of a one-dimensional location parameter. Later Huber (1967) considered a very general framework in which consistency and asymptotic normality were proved under relatively weak conditions.

Maronna (1976) was the first to define M-estimators for multivariate location and covariance. Huber (1981) extended Maronna's definition by defining M-estimators based upon $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbf{R}^p$ as solutions of the simultaneous equations

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n} v_1(d_i)(\mathbf{x}_i - \mathbf{t}) = \mathbf{0} \\ &\frac{1}{n}\sum_{i=1}^{n}\Big\{v_2(d_i)(\mathbf{x}_i - \mathbf{t})(\mathbf{x}_i - \mathbf{t})^T - v_3(d_i)\mathbf{C}\Big\} = \mathbf{0} \end{aligned} \tag{2.1}$$

where $d_i = d(\mathbf{x}_i, \mathbf{t}, \mathbf{C})$ and v_1, v_2 and v_3 are real-valued functions on $[0, \infty)$.

EXAMPLE 2.1 HUBER'S PROPOSAL 2: Take $v_3(y) = 1$ and $v_i(y) = \psi_i(y)/y$, for $i = 1, 2$, where $\psi_1(y) = \psi_{\mathrm{H}}(y; k)$ and $\psi_2(y) = \psi_{\mathrm{H}}(y^2; k^2)$. The function $\psi_{\mathrm{H}}(y; k) = \max\{-k, \min\{y, k\}\}$ is known as Huber's psi function.

Both existence and uniqueness of solutions of (2.1) was only shown for v_3 equal to 1 (Maronna 1976, Huber 1981). For this case Maronna (1976) shows consistency and asymptotic normality by means of Huber's 1967 results.

Maronna (1976) and Huber (1981) consider the breakdown point ε^* and the influence function IF to measure the robustness of these estimators. They both indicate that for solutions of (2.1) the breakdown point is at most $1/(p+1)$. A detailed treatment on the (finite-sample) breakdown behaviour of this type of M-estimators is given in Tyler (1986). One should note that these results assume monotonicity of v_2, and v_3 to be constant. So from the viewpoint of breakdown, M-estimators become more sensitive to outliers in higher dimensions. From the viewpoint of the influence function (which describes the effect of *one* outlier on the estimator), M-estimators are robust, as their influence function remains bounded when v_1, v_2 and v_3 in (2.1) are chosen suitably (see Huber 1981).

2.2 S-estimators. Rousseeuw and Yohai (1984) introduced S-estimators in a regression context and defined them as the solution to the problem of minimizing σ subject to

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{y_i-\boldsymbol{\theta}^T\mathbf{x}_i}{\sigma}\right)=b \tag{2.2}$$

among all $(\boldsymbol{\theta},\sigma)\in\mathbf{R}^p\times(0,\infty)$, where $0<b<\sup\rho$. The special case $\rho(y)=y^2$ in (2.2) obviously leads to the least squares estimators. In order to obtain more robust estimators and preserve asymptotic normality the function ρ was assumed to satisfy

(R1) ρ is symmetric, has a continuous derivative ψ, and $\rho(0)=0$.
(R2) There exists a finite constant $c>0$ such that ρ is strictly increasing on $[0,c]$ and constant on $[c,\infty)$. (Put $a=\sup\rho$.)

A direct generalization to S-estimators of multivariate location and covariance is obtained simply by adjustment of (2.2).

DEFINITION 2.1: Let $\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n\in\mathbf{R}^p$ and let $\rho:\mathbf{R}\to[0,\infty)$ satisfy (R1)-(R2). Then the S-estimator of multivariate location and covariance is defined as the solution $\boldsymbol{\theta}_n=(\mathbf{t}_n,\mathbf{C}_n)$ to the problem of minimizing $|\mathbf{C}|$ subject to

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left[\{(\mathbf{x}_i-\mathbf{t})^T\mathbf{C}^{-1}(\mathbf{x}_i-\mathbf{t})\}^{1/2}\right]=b \tag{2.3}$$

among all $(\mathbf{t},\mathbf{C})\in\Theta$. Denote this minimization problem by $(\mathcal{P}_n)$.

The constant $0<b<a$ can be chosen in agreement with an assumed underlying distribution. For instance, when $\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n$ are assumed to be a sample $X_1,X_2,\ldots,X_n$ with an underlying elliptical distribution (1.1), then the constant b is generally chosen to be $\mathsf{E}_{0,I}\rho(\|X\|)$. In that case the constant c can be chosen such that $0<b/a=r\leq\frac{n-p}{2n}$, which leads to a (finite-sample) breakdown point $\varepsilon_n^*=\lceil nr\rceil/n$ (see Lopuhaä and Rousseeuw, 1989). For $r=\frac{n-p}{2n}$ one obtains

the maximal breakdown point $\lfloor \frac{n-p+1}{2} \rfloor / n$, or asymptotically 0.50. However, the constant c at the same time determines the asymptotic variance and, as we will see in Section 6, it is not possible to achieve small asymptotic variance and 50% breakdown point simultaneously.

It might be worthwhile to mention that S-estimators of location and covariance can also be seen as robustifications of the least squares method. When $b = p$, then using $\rho(y) = y^2$ in (2.3) yields the sample mean and the sample covariance as unique solutions of $(\mathcal{P}_n)$ (see for instance Grübel 1988).

EXAMPLE 2.2 TUKEY'S BIWEIGHT: An example of a rho-function for (2.3) is

$$\rho_{\text{B}}(y;c) = \begin{cases} \dfrac{y^2}{2} - \dfrac{y^4}{2c^2} + \dfrac{y^6}{6c^4} & \text{, for } |y| \leq c \\ \dfrac{c^2}{6} & \text{, for } |y| \geq c \end{cases}.$$

Its derivative, which is a redescending (psi-) function, is known as Tukey's biweight function $\psi_{\text{B}}(y;c) = y\big(1 - (\frac{y}{c})^2\big)^2 \mathbf{1}_{[-c,c]}(y)$.

Davies (1987) defines S-estimators similarly only instead of ρ he uses a non-increasing function $\kappa : \mathbf{R}_+ \to [0,1]$ in (2.3). It is related to ρ as $\kappa(y) = 1 - \rho(y^2)/a$. If 'continuous differentiability of ρ' is weakened to 'ρ being left-continuous on $(0,\infty)$ and continuous at 0', if 'strictly increasing' on $[0,c]$ is weakened to 'nondecreasing' and if '$=$' is replaced by '$\leq$' in (2.3), then the two definitions are equivalent. Under these weaker conditions Davies proves existence and consistency of S-estimators, and he obtains asymptotic normality assuming that the function κ has a third continuous derivative.

We will extend existence and consistency of S-estimators to existence and continuity of S-functionals and obtain the influence function, and we will extend the asymptotic normality result by considering S-estimators as a special type of M-estimators and use Huber's (1967) results.

2.3 Relationship between M- and S-estimators. A drawback of using κ instead of ρ is that the conjectured correspondence with M-estimators remains hidden. In this subsection we will show that a solution to the minimization problem $(\mathcal{P}_n)$ also satisfies the first-order M-estimator conditions (2.1).

Let $\boldsymbol{\theta}_n = (\mathbf{t}_n, \mathbf{C}_n)$ be a solution of $(\mathcal{P}_n)$. Then, if by λ_n we denote the corresponding Lagrange multiplier, the pair $(\boldsymbol{\theta}_n, \lambda_n)$ is a zero of all partial derivatives $\partial \mathrm{L}_n/\partial \mathbf{t}$, $\partial \mathrm{L}_n/\partial \mathbf{C}$ and $\partial \mathrm{L}_n/\partial \lambda$, where L_n is the Lagrangean

$$\mathrm{L}_n(\boldsymbol{\theta}, \lambda) = \log\big(|\mathbf{C}|\big) - \lambda \left\{ \frac{1}{n} \sum_{i=1}^{n} \rho \left[\big\{ (\mathbf{x}_i - \mathbf{t})^T \mathbf{C}^{-1} (\mathbf{x}_i - \mathbf{t}) \big\}^{1/2} \right] - b \right\}.$$

This means that besides constraint (2.3), $(\boldsymbol{\theta}_n, \lambda_n)$ satisfies the equations

$$\begin{aligned} &\frac{\lambda}{n}\sum_{i=1}^{n} u(d_i)\mathbf{C}^{-1}(\mathbf{x}_i - \mathbf{t}) = \mathbf{0} \\ &2\mathbf{C}^{-1} - \mathbf{D}_{\mathbf{C}^{-1}} + \frac{\lambda}{2n}\sum_{i=1}^{n} u(d_i)\left(2\mathbf{V}_i - \mathbf{D}_{\mathbf{V}_i}\right) = \mathbf{0} \end{aligned} \tag{2.4}$$

where $u(y) = \psi(y)/y, d_i = d(\mathbf{x}_i; \mathbf{t}, \mathbf{C})$ and $\mathbf{V}_i = \mathbf{C}^{-1}(\mathbf{x}_i - \mathbf{t})(\mathbf{x}_i - \mathbf{t})^T\mathbf{C}^{-T}$ (for derivatives with respect to symmetric matrices, see Graybill 1983). But the second (matrix) equation can be written as

$$\mathbf{I} + \frac{\lambda}{2n}\sum_{i=1}^{n} u(d_i)\mathbf{A}^{-1}(\mathbf{x}_i - \mathbf{t})(\mathbf{x}_i - \mathbf{t})^T\mathbf{A}^{-T} = \mathbf{0} \tag{2.5}$$

where $\mathbf{A}\mathbf{A}^T = \mathbf{C}$. When we take the trace we get $p + \frac{\lambda}{2n}\sum_{i=1}^{n}\psi(d_i)d_i = 0$. Obviously we can solve λ_n from this equation, yielding

$$\lambda_n = -\frac{2p}{\left(\frac{1}{n}\sum_{i=1}^{n}\psi(d_{i,n})d_{i,n}\right)}$$

where $d_{i,n} = d(\mathbf{x}_i, \mathbf{t}_n, \mathbf{C}_n)$. If we substitute this into (2.5), together with (2.4) and (2.3) we find that any solution $\boldsymbol{\theta}_n$ of $(\mathcal{P}_n)$ satisfies the following equations

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n} u(d_i)(\mathbf{x}_i - \mathbf{t}) = \mathbf{0} \\ &\frac{1}{n}\sum_{i=1}^{n}\left\{pu(d_i)(\mathbf{x}_i - \mathbf{t})(\mathbf{x}_i - \mathbf{t})^T - v(d_i)\mathbf{C}\right\} = \mathbf{0} \end{aligned} \tag{2.6}$$

where $v(y) = \psi(y)y - \rho(y) + b$. The term $-\rho(y) + b$ is added to $\psi(y)y$ because merely substituting λ_n into (2.5) would give us a system of linear dependent equations for any pair $(\mathbf{t}, \mathbf{C}) \in \Theta$.

Hence, any solution of $(\mathcal{P}_n)$ turns out to be also a solution of equations (2.6) which obviously are of M-estimator type (2.1). To match the notation used in Huber (1967) write (2.6) as

$$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\Psi}(\mathbf{x}_i, \boldsymbol{\theta}) = \mathbf{0} \tag{2.7}$$

where $\boldsymbol{\theta} = (\mathbf{t}, \mathbf{C}) \in \Theta$ and $\boldsymbol{\Psi} = (\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2)$ is the function

$$\begin{aligned} \boldsymbol{\Psi}_1(\mathbf{x}, \boldsymbol{\theta}) &= u(d)(\mathbf{x} - \mathbf{t}) \\ \boldsymbol{\Psi}_2(\mathbf{x}, \boldsymbol{\theta}) &= pu(d)(\mathbf{x} - \mathbf{t})(\mathbf{x} - \mathbf{t})^T - v(d)\mathbf{C} \end{aligned} \tag{2.8}$$

with $d = d(\mathbf{x}, \mathbf{t}, \mathbf{C})$. We conclude that S-estimators satisfy first-order conditions (2.1) of M-estimators as defined in Huber (1981), or rather equations (2.7) of the type considered in Huber (1967).

However, recall that S-estimators are originally defined by the minimization problem $(\mathcal{P}_n)$, which is not equivalent to (2.7), and that in any dimension they can still be constructed with high breakdown point. The cause of these differences might lie in the functions $v_2(\cdot)$ and $v_3(\cdot)$ of (2.1) and the functions $u(\cdot)$ and $v(\cdot)$ of (2.6). For instance, Huber (1981) chooses $v_2 > 0$ and $v_3 \geq 0$ to be monotone, and the latter even equal to a constant for proving both existence and uniqueness of solutions of (2.1). The functions $u(y) = \psi(y)/y$ and $v(y) = \psi(y)y - \rho(y) + b$ will never satisfy either condition.

One could call any solution of (2.7) an M-estimator. However, M-estimators are generally associated with low breakdown point and with implicit equations (2.1), with v_2 decreasing and v_3 being constant. As this is not the case for S-estimators we tend to consider these estimators to be of a different type.

Although the S-estimator is probably not the only solution of (2.7), it is a solution with high breakdown point. To find it, one must therefore solve $(\mathcal{P}_n)$ and not just equation (2.7). Nevertheless S-estimators do satisfy (2.7) which has the consequence that their asymptotic behaviour and their influence function are the same as for M-estimators.

3. S-functionals and influence function. For the derivation of the influence function we have to extend Definition 2.1 to a functional formulation. Denote by $\mathcal{F}$ the class of all distributions on $\mathbf{R}^p$. The functional analogue of the S-estimator of multivariate location and covariance is defined as follows.

DEFINITION 3.1: Let $\rho : \mathbf{R} \to [0, \infty)$ be a function satisfying (R1)-(R2). Then the S-functional $\mathbf{S} : \mathcal{F} \to \Theta$ is defined as the solution $\mathbf{S}(P) = \big(\mathbf{t}(P), \mathbf{C}(P)\big)$ to the problem of minimizing $|\mathbf{C}|$ subject to

$$\int \rho\left[\left\{(\mathbf{y} - \mathbf{t})^T \mathbf{C}^{-1} (\mathbf{y} - \mathbf{t})\right\}^{1/2}\right] dP(\mathbf{y}) = b \tag{3.1}$$

among all $(\mathbf{t}, \mathbf{C}) \in \Theta$, where $0 < b < a$. Denote this problem by $(\mathcal{P}_P)$.

Existence of solutions of $(\mathcal{P}_P)$ is ensured if there is not too much mass concentrated at arbitrarily thin strips $H(\boldsymbol{\alpha}, l, \delta) = \{\mathbf{y} : l \leq \boldsymbol{\alpha}^T \mathbf{y} \leq l + \delta\}$, where $\|\boldsymbol{\alpha}\| = 1, \delta \geq 0$ and $l \in \mathbf{R}$. Let $0 < \varepsilon < 1$ and consider the following property for the distribution P on $\mathbf{R}^p$

(C_ε) The value $\delta_\varepsilon = \inf\{\delta : P\big(H(\boldsymbol{\alpha}, l, \delta)\big) \geq \varepsilon, \|\boldsymbol{\alpha}\| = 1, \delta \geq 0, l \in \mathbf{R}\}$ is strictly positive.

THEOREM 3.1. *Let P satisfy property (C_ε) for some $0 < \varepsilon \leq 1 - r$, where $r = b/a$. Then $(\mathcal{P}_P)$ has at least one solution.*

Let $P_k, k \geq 0$, be a sequence of probability distributions on $\mathbf{R}^p$ that converges weakly to P as $k \to \infty$. The following theorem shows when S-functionals are continuous.

THEOREM 3.2. *Let $\mathcal{C}$ be the class of all measurable convex subsets of $\mathbf{R}^p$ and suppose that every $E \in \mathcal{C}$ is a P-continuity set, i.e. $P(\partial E) = 0$. Suppose that P satisfies (C_ϵ) for some $0 < \epsilon < 1 - r$, and assume that the solution $\mathbf{S}(P)$ of $(\mathcal{P}_P)$ is unique. Then for k sufficiently large $(\mathcal{P}_{P_k})$ has at least one solution $\mathbf{S}(P_k)$, and for any sequence of solutions $\mathbf{S}(P_k)$, $k \geq 0$, it holds that $\lim_{k\to\infty} \mathbf{S}(P_k) = \mathbf{S}(P)$.*

REMARK 3.1: In the proof of Theorem 3.1 strict monotonicity of ρ on $[0, c]$ is not needed and continuity of ρ is not essential. This means that Theorem 3.1 can easily be shown to hold also for S-functionals that correspond to the larger class of S-estimators considered by Davies (1987) (see Section 2.2). With a stronger condition on P, which will ensure $\int \rho(\|\mathbf{y}\|/(1+\eta))dP(\mathbf{y})$ to be strictly decreasing at $\eta = 0$, also Theorem 3.2 can be shown to hold for these S-functionals.

REMARK 3.2: A part of the proof of Theorem 3.2 consists of showing that solutions $\mathbf{S}(P_k)$ eventually stay inside a fixed compact set. For the special case

$$P_{h,\mathbf{x}} = (1-h)P + h\delta_{\mathbf{x}}$$

(see Definition 3.2) one can show that if $0 < r < \frac{1}{2}$ and if P only satisfies (C_ϵ) for $\epsilon = (1-2r)(1-r)^{-1}$, then for any $0 < \alpha < 1$ there exists a compact set $K(\alpha)$ independent of $\mathbf{x}$ such that for all $h \in [0, \alpha r]$ the problem $(\mathcal{P}_{P_{h,\mathbf{x}}})$ has at least one solution and all solutions are contained in $K(\alpha)$.

Condition $0 < r < \frac{1}{2}$ is similar to the condition $0 < r < \frac{n-p}{2n}$ which ensures a (finite-sample) breakdown point $\varepsilon_n^* = \lceil nr \rceil / n$ (see Section 2.2). The latter means that the S-estimator stays in some fixed compact subset of Θ when the amount of contamination is less than ε_n^*. This is in agreement with the statement above that when the amount of contamination at $\mathbf{x} \in \mathbf{R}^p$ is less than r, the S-functional stays within a compact subset of Θ.

The robustness of the S-estimator can be measured by means of the influence function (see Hampel 1974). It is defined in terms of the S-functional in the following manner.

DEFINITION 3.2: Let $\mathbf{S}(\cdot)$ be a vector-valued mapping from a subset of $\mathcal{F}$ into Θ and let P lie in the domain of $\mathbf{S}(\cdot)$. If $\delta_{\mathbf{x}}$ denotes the atomic probability distribution concentrated in $\mathbf{x} \in \mathbf{R}^p$, then the influence function of $\mathbf{S}(\cdot)$ at P is defined pointwise by

$$\text{IF}(\mathbf{x}; \mathbf{S}, P) = \lim_{h \downarrow 0} \frac{\mathbf{S}((1-h)P + h\delta_{\mathbf{x}}) - \mathbf{S}(P)}{h} \tag{3.2}$$

if this limit exists for every $\mathbf{x} \in \mathbf{R}^p$.

If we replace P by the empirical distribution P_{n-1} and h by $\frac{1}{n}$, we realize that the IF measures a weighted alteration of the value of the estimator when one additional observation is added to a large sample of size $n-1$. The importance of the influence function lies in its heuristic interpretation : it describes the effect of an infinitesimal contamination at a point $\mathbf{x}$ on the estimator. A bounded influence function is therefore considered to be a good robustness property.

To derive the IF at a distribution P we need to be sure that $\mathbf{S}(\cdot)$ is uniquely defined at $(1-h)P + h\delta_{\mathbf{x}}$, for all $\mathbf{x} \in \mathbf{R}^p$ at least for small h, and secondly that the limit (3.2) exists for all $\mathbf{x} \in \mathbf{R}^p$. Theorem 3.2 ensures that for all $\mathbf{x} \in \mathbf{R}^p$ and h sufficiently small the problem $(\mathcal{P}_{P_{h,\mathbf{x}}})$, with $P_{h,\mathbf{x}} = (1-h)P + h\delta_{\mathbf{x}}$, has at least one solution and that all solutions are continuous.

We conclude that for h sufficiently small there exist solutions $\boldsymbol{\theta}_{h,\mathbf{x}} = (\mathbf{t}_{h,\mathbf{x}}, \mathbf{C}_{h,\mathbf{x}})$ of $(\mathcal{P}_{P_{h,\mathbf{x}}})$ and that they all converge to the same limit $\big(\mathbf{t}(P), \mathbf{C}(P)\big)$ as h tends to 0. Therefore there exists an open neighbourhood N of $\mathbf{S}(P)$ which contains all solutions $\boldsymbol{\theta}_h$ for h sufficiently small.

Remember that (2.7) is obtained from differentiation of the Lagrangean corresponding with problem $(\mathcal{P}_n)$. Similarly one could now differentiate the Lagrangean corresponding with the problem $(\mathcal{P}_P)$. If we restrict to the neighbourhood N, we may interchange the order of differentiation and integration, and similar to (2.7) we obtain the equation

$$\int \boldsymbol{\Psi}(\mathbf{y}, \boldsymbol{\theta}) dP(\mathbf{y}) = \mathbf{0} \tag{3.3}$$

where $\boldsymbol{\Psi}(\mathbf{y}, \boldsymbol{\theta})$ is defined in (2.8).

Solutions $(\mathbf{t}_{h,\mathbf{x}}, \mathbf{C}_{h,\mathbf{x}})$ of $(\mathcal{P}_{P_{h,\mathbf{x}}})$ must be a solution (not necessarily the only one) of (3.3), at least for h sufficiently small. Note that if we only consider a functional $\mathbf{M} : \mathcal{F} \to \Theta$, defined as the solution of (3.3), we may have some problems to ensure the uniqueness, and therefore for obtaining the influence function $\mathrm{IF}(\mathbf{x}; \mathbf{S}, P)$ we explicitly consider the solution $\mathbf{S}(P)$ of (3.3). The implicit function theorem, applied to this equation will ensure the uniqueness of $\mathbf{S}(\cdot)$ at $P_{h,\mathbf{x}}$ for h sufficiently small, and also the existence of $\mathrm{IF}(\mathbf{x}; \mathbf{S}, P)$.

THEOREM 3.3. *Let $\rho : \mathbf{R} \to [0, \infty)$ satisfy (R1)-(R2). Assume that ρ has a second derivative ψ' and suppose that*

(R3) $\psi'(y)$ and $u(y) = \psi(y)/y$ are bounded and continuous.

Suppose that the conditions of Theorem 3.2 hold. Let $\boldsymbol{\Psi}$ be defined as in (2.8) and let $\boldsymbol{\lambda}_P(\boldsymbol{\theta}) = \mathbf{E}_P \boldsymbol{\Psi}(X, \boldsymbol{\theta})$. Suppose that $\boldsymbol{\lambda}_P(\cdot)$ has a nonsingular derivative $\boldsymbol{\Lambda}$ at $\mathbf{S}(P) = \big(\mathbf{t}(P), \mathbf{C}(P)\big)$. Then the influence function IF$(\mathbf{x}; \mathbf{S}, P)$ exists and satisfies

$$IF(\mathbf{x}; \mathbf{S}, P) = -\boldsymbol{\Lambda}^{-1} \boldsymbol{\Psi}\big(\mathbf{x}, \mathbf{S}(P)\big). \tag{3.4}$$

Huber (1981) showed that equation (3.3) has a unique solution when certain monotonicity conditions on the functions $u(\cdot)$ and $v(\cdot)$ are satisfied. One of these

conditions is that the function $v(\cdot)$ is constant. In our case $v(\cdot)$ is certainly not a constant function and so equation (3.3) may have many solutions. However, it is possible that there is a unique 'S-solution' among all solutions of (3.3). For this solution we have derived the IF and naturally the expression is of the same type as for multivariate M-estimators. Note that the properties of the function ρ imply that the influence function $\mathrm{IF}(\mathbf{x}; \mathbf{S}, P)$ is bounded.

4. Asymptotic normality of S-estimators. Let $X_1, X_2, \ldots$ be a sequence of independent random vectors $X_i = (X_{i1} \cdots X_{ip})^T$ with a distribution P on $\mathbf{R}^p$. Suppose that for $n \geq p+1$ the sample $X_1, \ldots, X_n$ is in *general position*, i.e. no $p+1$ points lie in some $(p-1)$-dimensional subspace, almost surely.

When P in Definition 3.1 is equal to the empirical distribution P_n, we get the definition of the S-estimator. Note that P_n satisfies (C_ε) for $\varepsilon = (p+1)/n$ almost surely, so as a special case of Theorem 3.1 we have that for $n(1-r) \geq p+1$ the problem $(\mathcal{P}_n)$ has at least one solution almost surely. When P satisfies the conditions of Theorem 3.2 one has consistency : $\boldsymbol{\theta}_n \to (\mathbf{t}(P), \mathbf{C}(P))$ almost surely.

As we have seen in Subsection 2.3 solutions $\boldsymbol{\theta}_n$ of $(\mathcal{P}_n)$ satisfy first-order conditions (2.7) of M-estimators. An immediate consequence is that the asymptotic behaviour of S-estimators is similar to that of M-estimators.

THEOREM 4.1. *Let $\rho : \mathbf{R} \to [0, \infty)$ satisfy (R1)-(R3) and suppose that the conditions of Theorem 3.2 hold. Let $\boldsymbol{\Psi}$ be defined as in (2.8) and let $\boldsymbol{\lambda}_P(\cdot)$ be defined as in Theorem 3.3. Suppose that the solution $\mathbf{S}(P)$ of $(\mathcal{P}_P)$ is unique and that $\boldsymbol{\lambda}_P(\cdot)$ has a nonsingular derivative $\boldsymbol{\Lambda}$ at $\boldsymbol{\theta}_0 = \mathbf{S}(P)$. Let $\boldsymbol{\theta}_n$ be a solution of $(\mathcal{P}_n)$. Then $\sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0)$ has a limiting normal distribution with zero mean and covariance matrix $\boldsymbol{\Lambda}^{-1}\mathbf{M}\boldsymbol{\Lambda}^{-T}$, where $\mathbf{M}$ stands for the covariance matrix of $\boldsymbol{\Psi}(X_1, \boldsymbol{\theta}_0)$.*

REMARK 4.1: One might try to prove asymptotic normality of S-estimators directly from Definition 2.1 and avoid (2.7). A first derivative ψ of ρ in (R1), needed to arrive at (2.7) is then no longer required. At least continuity of ρ seems necessary. This is indicated by the results of Kim and Pollard (1989) on Rousseeuw's (1983) minimum volume ellipsoid estimator, which can be seen as an S-estimator with a discontinuous ρ-function.

5. Elliptical distributions. Consider the case that $P = P_{\mu,\Sigma}$ is elliptical and therefore take $b = \mathbf{E}_{0,I}\rho(\|X\|)$ in (2.3) and (3.1). For this choice of P does Huber (1981) obtain the expression for the IF of M-functionals and does Maronna (1976) give a detailed description for the asymptotic covariance matrix of location M-estimators. We compare these results with Theorems 3.3 and 4.1 applied to $P_{\mu,\Sigma}$.

It is not difficult to show that $P_{\mu,\Sigma}$ satisfies property (C_ε) for any $0 < \varepsilon < 1$, so that according to Theorem 3.1 at least one solution of $(\mathcal{P}_{P_{\mu,\Sigma}})$ exists. Davies (1987) showed that it is even unique and *Fisher consistent*

$$\mathbf{S}(P_{\mu,\Sigma}) = (\boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{5.1}$$

The following corollary gives a detailed description of the limiting normal distribution of $\sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0)$. In particular, the asymptotic covariance of the location

S-estimator $\mathbf{t}_n$ as defined in Definition 2.1 is exactly the same as the asymptotic covariance found for the location M-estimator considered by Maronna (1976) if one chooses $v_1(y) = \psi(y)/y$ in (2.1).

To describe the asymptotic covariance matrix of $\sqrt{n}(\mathbf{C}_n - \boldsymbol{\Sigma})$ in condensed form, we represent $p \times p$-matrices $\mathbf{M}$ by $\text{vec}(\mathbf{M}) = (m_{11} \cdots m_{p1} \cdots m_{1p} \cdots m_{pp})^T$. The operator $\text{vec}(\cdot)$ just stacks the columns of $\mathbf{M}$ on top of each other. Magnus and Neudecker (1979) investigated algebraic properties of this operator in relation with the commutation matrix $\mathbf{K}_{m,n}$. Here we will only use the special case $\mathbf{K}_{p,p}$, which is a $p^2 \times p^2$-block matrix with the (i,j)-block being equal to $\boldsymbol{\Delta}_{ji}$. The latter is a $p \times p$-matrix which is 1 at entry (j,i) and 0 everywhere else. Finally, $\mathbf{M} \otimes \mathbf{N}$ denotes the Kronecker product of the matrices $\mathbf{M}$ and $\mathbf{N}$ which is a $p^2 \times p^2$-block matrix with $p \times p$-blocks, the (i,j)-block equal to $m_{ij}\mathbf{N}$.

COROLLARY 5.1. *Let $\rho : \mathbf{R} \to [0,\infty)$ satisfy (R1)-(R3) and let P be a elliptical distribution with parameter $\boldsymbol{\theta}_0 = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Suppose that*

$$\begin{aligned} &\mathbf{E}_{0,I}\psi'(\|X\|) > 0. \\ &\mathbf{E}_{0,I}[\psi'(\|X\|)\|X\|^2 + (p+1)\psi(\|X\|)\|X\|] > 0. \end{aligned} \tag{5.2}$$

When $\boldsymbol{\theta}_n = (\mathbf{t}_n, \mathbf{C}_n)$ is a solution of $(\mathcal{P}_n)$, then $\sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0)$ has a limiting normal distribution with zero mean and $\mathbf{t}_n$ and $\mathbf{C}_n$ are asymptotically independent. The covariance matrix of the limiting distribution of $\sqrt{n}(\mathbf{t}_n - \boldsymbol{\mu})$ is given by $(\alpha/\beta^2)\boldsymbol{\Sigma}$, where

$$\begin{aligned} \alpha &= \frac{1}{p}\mathbf{E}_{0,I}\psi^2(\|X\|) \\ \beta &= \mathbf{E}_{0,I}\left[\left(1 - \frac{1}{p}\right)u(\|X\|) + \frac{1}{p}\psi'(\|X\|)\right]. \end{aligned} \tag{5.3}$$

The covariance matrix of the limiting distribution of $\sqrt{n}(\mathbf{C}_n - \boldsymbol{\Sigma})$ is given by

$$\sigma_1(\mathbf{I} + \mathbf{K}_{p,p})(\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma}) + \sigma_2 \text{vec}(\boldsymbol{\Sigma})\,\text{vec}(\boldsymbol{\Sigma})^T. \tag{5.4}$$

where

$$\sigma_1 = \frac{p(p+2)\mathbf{E}_{0,I}\psi^2(\|X\|)\|X\|^2}{\left\{\mathbf{E}_{0,I}[\psi'(\|X\|)\|X\|^2 + (p+1)\psi(\|X\|)\|X\|]\right\}^2}.$$

$$\sigma_2 = -\frac{2}{p}\sigma_1 + \frac{4\mathbf{E}_{0,I}(\rho(\|X\|) - b)^2}{\left\{\mathbf{E}_{0,I}\psi(\|X\|)\|X\|\right\}^2}. \tag{5.5}$$

For the influence function, it is sufficient to give the expression of $\text{IF}(\mathbf{x}; \mathbf{S}, P_{0,I})$ because affine equivariance of $\mathbf{S}(\cdot)$ yields the general expressions

$$\begin{aligned} \text{IF}(\mathbf{x}; \mathbf{t}, P_{\mu,\Sigma}) &= \mathbf{B}\,\text{IF}(\mathbf{B}^{-1}(\mathbf{x} - \boldsymbol{\mu}); \mathbf{t}, P_{0,I}) \\ \text{IF}(\mathbf{x}; \mathbf{C}, P_{\mu,\Sigma}) &= \mathbf{B}\,\text{IF}(\mathbf{B}^{-1}(\mathbf{x} - \boldsymbol{\mu}); \mathbf{C}, P_{0,I})\mathbf{B}^T. \end{aligned} \tag{5.6}$$

We will describe $\mathrm{IF}(\mathbf{x}; \mathbf{S}, P_{0,I})$ such that it can be compared with the expressions found for M-functionals in Huber (1981).

COROLLARY 5.2. *Let $\rho : \mathbf{R} \to [0, \infty)$ satisfy (R1)-(R3) and suppose that conditions (5.2) hold. Let P be spherically symmetric. Then the influence function $IF(\mathbf{x}; \mathbf{S}, P)$ of the S-functional defined in Definition 3.2 exists and it holds that*

$$IF(\mathbf{x}; \mathbf{t}, P) = \frac{1}{\beta}\psi(\|\mathbf{x}\|)\frac{\mathbf{x}}{\|\mathbf{x}\|} \tag{5.7}$$

where β is defined in (5.3), and $IF(\mathbf{x}; \mathbf{C}, P)$ satisfies

$$IF(\mathbf{x}; \mathbf{C}, P) - \frac{1}{p}\mathrm{trace}\left[IF(\mathbf{x}; \mathbf{C}, P_{0,I})\right]\mathbf{I} = \frac{1}{\gamma}p\psi(\|\mathbf{x}\|)\|\mathbf{x}\|\left(\frac{\mathbf{x}\mathbf{x}^T}{\|\mathbf{x}\|^2} - \frac{1}{p}\mathbf{I}\right) \tag{5.8}$$

and

$$\frac{1}{p}\mathrm{trace}\left[IF(\mathbf{x}; \mathbf{C}, P)\right] = \frac{2}{\omega}(\rho(\|\mathbf{x}\|) - b) \tag{5.9}$$

where γ is defined in (A.11) and $\omega = \mathrm{E}_{0,I}\psi(\|X\|)\|X\|$.

6. Asymptotic variance in relation to breakdown point. We compute asymptotic variances of the S-estimator defined by the biweight function $\rho_{\mathrm{B}}(\cdot; c)$ of Example 2.2. The variances are computed for different values of p (=1,2 and 10) and for each p the constant c is given five different values that correspond with the values for r =0.1, 0.2, 0.3, 0.4 and 0.5, by means of the relation

$$\frac{\mathrm{E}_{\Phi}\rho_{\mathrm{B}}(\|X\|; c)}{(c^2/6)} = r \tag{6.1}$$

where the expectation is with respect to the standard normal distribution. The values of r are the limiting values of the finite sample breakdown point $\varepsilon_n^* = \lceil nr \rceil / n$. Denote the corresponding S-estimator by $S(r, p)$.

We compare these results with the asymptotic variances of the M-estimator defined by Huber's Proposal 2 of Example 2.1. The different choices of k correspond with 'winsorizing proportions' $w = \mathrm{P}_{\Phi}\{\|X\| > k\}$ (=0.3, 0.2, 0.1 and 0). Denote the corresponding M-estimator by $H(w, p)$. Note that in all cases $\sup \psi_2 = k^2 > p$, which is needed for the existence of $H(w, p)$. By $H(0, p)$ we mean the limiting case of $H(w, p)$ as $k \to \infty$. Note that $H(0, p)$ is also the limiting case of $S(r, p)$ as $c \to \infty$, namely the sample mean and the sample covariance.

Maronna (1976) already computed asymptotic variances for the $H(w, p)$-location estimator at the multivariate student and the multivariate normal distribution, and Tyler (1983) computed an index for the asymptotic variance of the $H(w, p)$-covariance estimator also at these distributions as well as at a symmetric contaminated normal distribution with thicker tails.

Table 1

Asymptotic variances of S(r,p) and H(w,p)

attained at NOR *and* SCN

		NOR $p=1$	SCN $p=1$	NOR $p=2$	SCN $p=2$	NOR $p=10$	SCN $p=10$
$S(0.5)$	λ	3.485	4.007	1.725	1.952	1.072	1.191
	η	3.711	4.301	2.656	3.020	1.093	1.215
$S(0.4)$	λ	2.165	2.499	1.356	1.542	1.036	1.152
	η	2.949	3.554	1.736	1.991	1.045	1.163
$S(0.3)$	λ	1.512	1.757	1.157	1.327	1.016	1.133
	η	2.467	3.174	1.298	1.516	1.020	1.140
$S(0.2)$	λ	1.181	1.392	1.055	1.232	1.006	1.139
	η	2.176	3.173	1.096	1.334	1.007	1.174
$S(0.1)$	λ	1.035	1.271	1.011	1.252	1.001	1.250
	η	2.035	3.919	1.018	1.430	1.001	1.527
$H(0.3)$	λ	1.100	1.327	1.048	1.260	1.009	1.185
	η	3.974	4.231	1.256	1.302	1.047	1.066
$H(0.2)$	λ	1.060	1.302	1.029	1.257	1.005	1.190
	η	3.186	3.536	1.171	1.246	1.030	1.060
$H(0.1)$	λ	1.026	1.299	1.012	1.272	1.002	1.203
	η	2.561	3.119	1.087	1.230	1.014	1.070
$H(0)$	λ	1.000	1.800	1.000	1.800	1.000	1.800
	η	2.000	7.333	1.000	2.778	1.000	2.778

We consider the multivariate normal (NOR) distribution $N(\mu, \Sigma)$ and the symmetric contaminated normal (SCN) distribution $0.9\,N(\mu, \Sigma) + 0.1\,N(\mu, 9\Sigma)$. Table 1 lists the asymptotic variances. It partly overlaps similar tables in Maronna (1976) and Tyler (1983).

In all cases the location estimator has an asymptotic covariance which is a certain multiple λ of Σ. The expression for λ for $S(r,p)$ is obtained from (5.3), and the expression for λ for $H(w,p)$ is given in Maronna (1976). The values of λ are listed in Table 1.

In all cases the covariance estimator has an asymptotic covariance that is of type (5.4) (Tyler 1982). To measure the asymptotic variance of the covariance estimators we distinguish the cases $p = 1$ and $p \geq 2$. If $p = 1$ then (1.1) reduces to $(1/\sigma) f\big((x-\mu)/\sigma\big)$ and we give the value $\eta = 2\sigma_1 + \sigma_2$ which represents the asymptotic variance of $\sqrt{n}(s_n^2 - \sigma^2)$, where s_n^2 denotes the estimator for the scale parameter σ^2 of the underlying distribution.

For $p \geq 2$ we give the value $\eta = \sigma_1$. Tyler (1983) compared values of σ_1 for different covariance M-estimators with simulated values of a Monte Carlo study of robust covariance estimators in Devlin et al. (1981). It turned out that σ_1 suffices as an index for the asymptotic variance of the correlation estimator based upon the robust covariance estimator. The expression for σ_1 for $S(r,p)$ is given in (5.4), and the expression for σ_1 for $H(w,p)$ is given in Tyler (1982).

From Table 1 we see that the asymptotic variances of $S(r,p)$ for r not too large are of similar magnitude as the asymptotic variances of $H(w,p)$, except at the SCN distribution where the $H(w,p)$-covariance estimator has a better performance. In general the asymptotic variance of $S(r,p)$ decreases simultaneously with the breakdown point r. However, in contrast with M-estimators, for every dimension p it is possible to construct an S-estimator with a high breakdown point.

It is interesting to compare the breakdown points of $S(r,p)$ and $H(w,p)$ at the same level of asymptotic variance. Table 2 gives such a comparison.

Table 2

Comparison of breakdown points of $S(r,p)$ *and* $H(w,p)$

at the same level of asymptotic variance attained at NOR *and* SCN

		NOR		NOR		SCN	
$H(w,p)$	δ^*	λ	$r(\lambda)$	σ_1	$r(\sigma_1)$	λ	$r(\lambda)$
$H(0.1,2)$	0.217	1.012	0.104	1.087	0.192	1.272	0.256
$H(0.223,2)$	0.333	1.033	0.162	1.190	0.256	1.257	0.239
$H(0.3,2)$	0.169	1.048	0.189	1.256	0.285	1.260	0.243
$H(0.1,10)$	0.063	1.002	0.124	1.014	0.263	1.203	0.500
$H(0.2,10)$	0.074	1.005	0.186	1.030	0.349	1.190	0.497
$H(0.3,10)$	0.085	1.009	0.238	1.047	0.406	1.185	0.487
$H(0.358,10)$	0.091	1.011	0.258	1.057	0.432	1.183	0.483

According to Tyler (1986), when $k^2 > p$ then the limiting value of the breakdown point of $H(w,p)$ equals $\delta^* = \min\{1/k^2, 1-p/k^2\}$, which is maximal when $k^2 = p+1$. The values $w = 0.223$ and 0.358 in Table 2 correspond with the values $k^2 = p+1$.

Given the asymptotic variance λ of the $H(w,p)$-location estimator at the NOR distribution, the constant c of $\rho_B(\cdot;c)$ is determined such that the $S(r,p)$-location estimator achieves the same level of λ. With this value of c the breakdown point $r(\lambda)$ is computed by means of (6.1). Next this procedure is repeated given the value σ_1 of the $H(w,p)$-covariance estimator at the NOR distribution, and finally the procedure is repeated given the value λ of the $H(w,p)$-location estimator at the SCN distribution.

We conclude that the S-estimator is able to achieve the asymptotic variances attained by the M-estimator, but in addition it has a breakdown point that becomes considerably higher when the dimension p increases.

Appendix. The proofs of existence and consistency of S-estimators in Davies (1987) extend fairly easily to existence and continuity of S-functionals. The following lemma is fundamental.

LEMMA 3.1. *Let* $(\mathbf{t}, \mathbf{C}) \in \Theta$, $0 < m_0 < \infty$, $0 < c < \infty$ *and* $0 < \varepsilon < 1$.

(i) If P *satisfies* (C_ε) *and* $P\big(E(\mathbf{t}, \mathbf{C}, c)\big) \geq \varepsilon$, *then* $\lambda_p(\mathbf{C}) \geq k_1 > 0$, *where* k_1 *only depends on* ε, P *and* c.

(ii) Assume $\lambda_p(\mathbf{C}) \geq k_1 > 0$. *If* $\int \rho\big(\|\mathbf{y}\|/m_0\big) dP(\mathbf{y}) \leq b$, *then any solution* $(\mathbf{t}, \mathbf{C})$ *of* $(\mathcal{P}_P)$ *must have* $\lambda_1(\mathbf{C}) \leq k_2 < \infty$, *where* k_2 *only depends on* k_1 *and* m_0.

(iii) Let P *satisfy* (C_ε) *and suppose that* $P\big(E(\mathbf{t}, \mathbf{C}, c)\big) \geq \varepsilon$. *If* $\lambda_1(\mathbf{C}) \leq k_2 < \infty$, *then* $(\mathbf{t}, \mathbf{C})$ *is contained in a compact set* $K \subset \Theta$, *which only depends on* ε, F, c *and* k_2.

PROOF: Because $E(\mathbf{t}, \mathbf{C}, c)$ is contained in some strip $H\big(\alpha, l, 2c\sqrt{\lambda_p(\mathbf{C})}\big)$ it follows from (C_ε) that $\lambda_p(\mathbf{C}) \geq (\delta_\varepsilon / c)^2 / 4 > 0$, which proves (i).

The function ρ is continuous and nondecreasing on $[0, \infty)$, so any solution of $(\mathcal{P}_P)$ is also a solution to the same minimization problem with constraint (3.1) replaced by

$$\int \rho\left[\left\{(\mathbf{y} - \mathbf{t})^T \mathbf{C}^{-1} (\mathbf{y} - \mathbf{t})\right\}^{1/2}\right] dP(\mathbf{y}) \leq b. \tag{A.1}$$

As the pair $(\mathbf{0}, m_0^2 \mathbf{I})$ satisfies (A.1) we conclude that any possible solution of $(\mathcal{P}_P)$ must have $|\mathbf{C}| \leq m_0^{2p}$. Because $\lambda_p(\mathbf{C}) \geq k_1 > 0$ we find that $\lambda_1(\mathbf{C}) \leq m_0^{2p} / k_1^{p-1} < \infty$ which proves (ii).

Since every probability measure on $\mathbf{R}^P$ is tight, there exists a compact set B_ε such that $P(B_\varepsilon) \geq 1 - \varepsilon$. Then $\|\mathbf{t} - \mathbf{y}\| \leq c\sqrt{k_2}$ for some $\mathbf{y} \in B_\varepsilon$. Otherwise $E(\mathbf{t}, \mathbf{C}, c)$ would be contained in B_ε^c which would be in contradiction with $P\big(E(\mathbf{t}, \mathbf{C}, c)\big) \geq \varepsilon$. Hence, $\|\mathbf{t}\|$ is bounded and together with (i) the lemma follows. □

PROOF OF THEOREM 3.1: Let $(\mathbf{t}, \mathbf{C}) \in \Theta$ satisfy constraint (3.1). Then we find

$$P\big(E(\mathbf{t}, \mathbf{C}, c)\big) \geq 1 - \frac{1}{a} \int \rho\left[\left\{(\mathbf{y} - \mathbf{t})^T \mathbf{C}^{-1} (\mathbf{y} - \mathbf{t})\right\}^{1/2}\right] dP(\mathbf{y}) = 1 - r \geq \varepsilon. \tag{A.2}$$

Lemma 3.1(i) implies that $\lambda_p(\mathbf{C}) \geq k_1 > 0$. Because $\lim_{m \to \infty} \int \rho\big(\|\mathbf{y}\|/m\big) dP(\mathbf{y}) = 0$, there exists an $m_0 > 0$ such that $\int \rho\big(\|\mathbf{y}\|/m_0\big) dP(\mathbf{y}) \leq b$. Lemma 3.1(ii) yields that $\lambda_1(\mathbf{C}) \leq k_2 < \infty$. Finally Lemma 3.1(iii) implies that for solving $(\mathcal{P}_P)$ one may restrict to a compact subset $K \subset \Theta$. As $|\mathbf{C}|$ is a continuous function of $(\mathbf{t}, \mathbf{C})$ it must attain a minimum on K. □

LEMMA 3.2. *Let* P_k, $k \geq 0$, *be a sequence of distributions on* $\mathbf{R}^p$ *that converges weakly to* P *as* $k \to \infty$. *Let* $\boldsymbol{\theta}_k, k \geq 0$, *be a sequence in* Θ *such that* $\boldsymbol{\theta}_k \to \boldsymbol{\theta}_L$ *as* $k \to \infty$. *If* $g(\mathbf{y}, \boldsymbol{\theta}) = \rho\big[\{(\mathbf{y} - \mathbf{t})^T \mathbf{C}^{-1} (\mathbf{y} - \mathbf{t})\}^{1/2}\big]$, *then*

$$\lim_{k \to \infty} \int g(\mathbf{y}, \boldsymbol{\theta}_k)\, dP_k(\mathbf{y}) = \int g(\mathbf{y}, \boldsymbol{\theta}_L)\, dP(\mathbf{y}). \tag{A.3}$$

PROOF: Put $g_k(\mathbf{y}) = g(\mathbf{y}, \boldsymbol{\theta}_k)$ and $g_L(\mathbf{y}) = g(\mathbf{y}, \boldsymbol{\theta}_L)$. Then for every sequence $\{\mathbf{y}_k\}$ with $\mathbf{y}_k \to \mathbf{y}$ we have that

$$\lim_{k\to\infty} g_k(\mathbf{y}_k) = g(\mathbf{y}).$$

Next apply Theorem 5.5 of Billingsley (1968). Let $\Gamma : [0,\infty) \to [0,\infty)$ be the function $\Gamma(\mathbf{y}) = \mathbf{y}\mathbf{1}_{[0,a]}(\mathbf{y}) + a\mathbf{1}_{(a,\infty)}(\mathbf{y})$, which is a bounded uniformly continuous function. Then as a consequence of $P_k \Rightarrow P$ we have that

$$\lim_{k\to\infty} \int \Gamma\big(g_k(\mathbf{y})\big) dP_k(\mathbf{y}) = \int \Gamma\big(g_L(\mathbf{y})\big) dP(\mathbf{y})$$

which proves (A.3). □

PROOF OF THEOREM 3.2: According to Rao (1962, Theorem 4.2) we have that

$$\sup_{E\in\mathcal{C}} |P_k(E) - P(E)| \longrightarrow 0 \qquad \text{, as } k \to \infty. \tag{A.4}$$

Because strips $H(\boldsymbol{\alpha}, l, \delta) \in \mathcal{C}$, (A.4) implies that for k sufficiently large every strip with $P_k\big(H(\boldsymbol{\alpha}, l, \delta)\big) \geq 1 - r$ must also satisfy $P\big(H(\boldsymbol{\alpha}, l, \delta)\big) \geq \varepsilon$. This means that $\inf\{\delta : P_k\big(H(\boldsymbol{\alpha}, l, \delta)\big) \geq 1 - r\} \geq \inf\{\delta : P\big(H(\boldsymbol{\alpha}, l, \delta)\big) \geq \varepsilon\} > 0$, so that for k sufficiently large P_k satisfies (C_{1-r}) and according to Theorem 3.1 at least one solution exists.

Denote $\mathbf{S}(P_k) = \boldsymbol{\theta}_k = (\mathbf{t}_k, \mathbf{C}_k)$. Because convex sets are transformed affinely into convex sets and because $\mathbf{S}(\cdot)$ is affine equivariant, we may assume that $\mathbf{S}(P) = (\mathbf{0}, \mathbf{I})$. Similar to (A.2) we find $P_k\big(E(\mathbf{t}_k, \mathbf{C}_k, c)\big) \geq 1 - r$, such that from (A.4) it follows that for k sufficiently large $P\big(E(\mathbf{t}_k, \mathbf{C}_k, c)\big) \geq \varepsilon$. Lemma 3.1(i) implies that $\lambda_p(\mathbf{C}_k) \geq k_1 > 0$ for k sufficiently large.

According to Lemma 3.2 for any $\eta > -1$, it holds that $\int \rho\big(\|\mathbf{y}\|/(1+\eta)\big) dP_k(\mathbf{y}) \to \int \rho\big(\|\mathbf{y}\|/(1+\eta)\big) dP(\mathbf{y})$, as $k \to \infty$. As the limit is strictly decreasing at $\eta = 0$ we see that for any $\eta > 0$,

$$\int \rho\left(\frac{\|\mathbf{y}\|}{1+\eta}\right) dP_k(\mathbf{y}) \leq \int \rho(\|\mathbf{y}\|) dP(\mathbf{y}) = b$$

for k sufficiently large. Then similar to the proof of Lemma 3.1(ii) we find that $|\mathbf{C}_k| \leq (1+\eta)^{2p}$ eventually. Because $\eta > 0$ may be taken arbitrarily small, we conclude that

$$\limsup_{k\to\infty} |\mathbf{C}_k| \leq 1 \tag{A.5}$$

and we find that $\lambda_1(\mathbf{C}_k) \leq 4^p / k_1^{p-1}$ eventually. With Lemma 3.1(iii) we see that there exists a compact set K such that for k sufficiently large the sequence $\{\boldsymbol{\theta}_k\} \subset K$.

Consider a convergent subsequence $\{\boldsymbol{\theta}_{k_j}\}$ with $\boldsymbol{\theta}_{k_j} \to \boldsymbol{\theta}_L = (\mathbf{t}_L, \mathbf{C}_L)$. With Lemma 3.2 we find that

$$\int \rho\left[d(\mathbf{y}, \mathbf{t}_L, \mathbf{C}_L)\right] dP(\mathbf{y}) = \lim_{j\to\infty} \int \rho\left[d(\mathbf{y}, \mathbf{t}_{k_j}, \mathbf{C}_{k_j})\right] dP_{k_j}(\mathbf{y}) = b.$$

Hence, $\boldsymbol{\theta}_L$ turns out to satisfy constraint (3.1) of $(\mathcal{P}_P)$, of which $(\mathbf{0}, \mathbf{I})$ a is solution. This means that $|\mathbf{C}_L| \geq 1$. Next, (A.5) yields $|\mathbf{C}_L| = 1$. But then uniqueness of $(\mathbf{0}, \mathbf{I})$ implies $\boldsymbol{\theta}_L = (\mathbf{0}, \mathbf{I})$. As $\{\boldsymbol{\theta}_k\}$ eventually stays in a compact set we must have $\lim_{k\to\infty} \boldsymbol{\theta}_k = (\mathbf{0}, \mathbf{I})$. □

LEMMA 3.3. *Let $\rho : \mathbf{R} \to [0, \infty)$ satisfy (R1)-(R3) and consider the function $\Psi(\mathbf{x}, \boldsymbol{\theta})$ of (2.8). Then*

(i) Ψ is bounded and continuous on $\mathbf{R}^p \times \Theta$.

(ii) $\partial\Psi/\partial\boldsymbol{\theta}$ is continuous on $\mathbf{R}^p \times \Theta$ and is bounded by a constant which depends only upon $\|\mathbf{C}\|$ and $\|\mathbf{C}^{-1}\|$.

PROOF: Continuity of Ψ is obvious and boundness of the functions $u(y)y$, $u(y)y^2$ and $v(y)$ proves (i).

For (ii) compute the derivative $\partial\Psi/\partial\boldsymbol{\theta}$:

$$\frac{\partial\Psi_1}{\partial\mathbf{t}} = -\left(\frac{u'(d)}{d}\mathbf{C}^{-1}(\mathbf{x}-\mathbf{t})(\mathbf{x}-\mathbf{t})^T + u(d)\mathbf{I}\right)$$

$$\frac{\partial\Psi_{1,j}}{\partial\mathbf{C}} = -\frac{u'(d)}{2d}(x_j - t_j)\,(2\mathbf{V} - \mathbf{D}_\mathbf{V})$$

$$\begin{aligned}\frac{\partial\Psi_{2,ij}}{\partial\mathbf{t}} = &-p\frac{u'(d)}{d}(x_i - t_i)(x_j - t_j)\mathbf{C}^{-1}(\mathbf{x}-\mathbf{t}) \\ &+ pu(d)\frac{\partial(x_i - t_i)(x_j - t_j)}{\partial\mathbf{t}} \\ &+ \frac{v'(d)}{d}c_{ij}\mathbf{C}^{-1}(\mathbf{x}-\mathbf{t})\end{aligned}$$

$$\begin{aligned}\frac{\partial\Psi_2}{\partial c_{ij}} = &-p\frac{u'(d)}{2d}\,(2\mathbf{V} - \mathbf{D}_\mathbf{V})_{ij}\,(\mathbf{x}-\mathbf{t})(\mathbf{x}-\mathbf{t})^T \\ &+ \frac{v'(d)}{2d}\,(2\mathbf{V} - \mathbf{D}_\mathbf{V})_{ij}\,\mathbf{C} \\ &- v(d)\frac{\partial\mathbf{C}}{\partial c_{ij}}\end{aligned}$$

where $d = d(\mathbf{x}, \mathbf{t}, \mathbf{C})$ and $\mathbf{V} = \mathbf{C}^{-1}(\mathbf{x}-\mathbf{t})(\mathbf{x}-\mathbf{t})^T\mathbf{C}^{-T}$.

Because of $\|\mathbf{x}-\mathbf{t}\|^2/d^2 \leq \|\mathbf{C}\|$ and (R3) the second statement (ii) follows. □

To proof Theorem 3.3 we will apply the implicit function theorem (see for instance Theorem 10.2.1 in Dieudonné 1969) to the function

$$\text{(A.6)} \quad \mathbf{W}(h, \boldsymbol{\theta}; \mathbf{x}) = \int \Psi(\mathbf{y}, \boldsymbol{\theta})\, dP_{h,\mathbf{x}}(\mathbf{y}) = (1-h)\int \Psi(\mathbf{y}, \boldsymbol{\theta})\, dP(\mathbf{y}) + h\Psi(\mathbf{x}, \boldsymbol{\theta})$$

where $\Psi = (\Psi_1, \Psi_2)$ is defined in (2.8). Solutions $\theta_{h,\mathbf{x}}$ of $(\mathcal{P}_{P_{h,\mathbf{x}}})$ eventually are contained in an open neighbourhood N of $\mathbf{S}(P)$, and they satisfy equation (3.3) or equivalently the pairs $(h, \theta_{h,\mathbf{x}})$ are a zero of the function $\mathbf{W}(\cdot;\mathbf{x})$. We apply the implicit function theorem to the function $\mathbf{W}(\cdot;\mathbf{x})$ considered on the open neighbourhood $A = \mathbf{R} \times N$ of $(h_0, \theta_0) = \big(0, \mathbf{S}(P)\big)$. According to this theorem, when $\mathbf{W}(\cdot;\mathbf{x})$ is continuously differentiable on A and $\partial\mathbf{W}/\partial\theta$ is nonsingular at (h_0, θ_0), there exists a neighbourhood U of h_0 on which there exists a unique function $\theta(\cdot;\mathbf{x})$ such that $\big(h, \theta(h;\mathbf{x})\big) \in A$ and $\mathbf{W}\big(h, \theta(h;\mathbf{x});\mathbf{x}\big) = \mathbf{0}$ for any $h \in U$. Moreover, $\theta(\cdot;\mathbf{x})$ is continuously differentiable in U, with

$$\theta'(0;\mathbf{x}) = -\left[\frac{\partial\mathbf{W}}{\partial\theta}(0, \theta_0;\mathbf{x})\right]^{-1} \frac{\partial\mathbf{W}}{\partial h}(0, \theta_0;\mathbf{x}). \tag{A.7}$$

PROOF OF THEOREM 3.3: Let $\theta_0 = \big(\mathrm{t}(P), \mathbf{C}(P)\big)$ be the unique solution of $(\mathcal{P}_P)$ so that $\mathbf{W}(0, \theta_0;\mathbf{x}) = \mathbf{0}$. According to Lemma 3.3(i) the function $\Psi(\mathbf{y}, \theta)$ is bounded and continuous, so we conclude that also $\int \Psi(\mathbf{y}, \theta)\, dP(\mathbf{y})$ is bounded and continuous. Hence, $\partial\mathbf{W}/\partial h$ is continuous on A. Lemma 3.3(ii) implies that $\partial\Psi/\partial\theta$ is bounded and continuous on $\mathbf{R}^p \times N$, so that

$$\frac{\partial\mathbf{W}}{\partial\theta}(h, \theta;\mathbf{x}) = (1-h)\int \frac{\partial\Psi}{\partial\theta}(\mathbf{y}, \theta)\, dP(\mathbf{y}) + h\frac{\partial\Psi}{\partial\theta}(\mathbf{x}, \theta)$$

is also a continuous function on A. Finally, we have that

$$\frac{\partial\mathbf{W}}{\partial\theta}(0, \theta_0) = \int \frac{\partial}{\partial\theta}\Psi(\mathbf{y}, \theta_0)\, dP(\mathbf{y}) = \mathbf{\Lambda} \tag{A.8}$$

which is nonsingular. We may apply the implicit function theorem.

Let h be nonnegative and sufficiently small such that $(\mathcal{P}_{P_{h,\mathbf{x}}})$ has at least one solution. Suppose that $\theta_{h,\mathbf{x},1}$ and $\theta_{h,\mathbf{x},2}$ would be two solutions of $(\mathcal{P}_{P_{h,\mathbf{x}}})$. Then according to Theorem 3.2 both $(h, \theta_{h,\mathbf{x},1})$ and $(h, \theta_{h,\mathbf{x},2})$ are contained in the neighborhood $U \times N$ for h sufficiently small. Hence, by uniqueness of the function $\theta(\cdot;\mathbf{x})$ we conclude that $\theta_{h,\mathbf{x},1} = \theta_{h,\mathbf{x},2} = \theta(h;\mathbf{x})$. For nonnegative h sufficiently small the functional $\mathbf{S}(\cdot)$ is thus uniquely defined as

$$\mathbf{S}(P_{h,\mathbf{x}}) = \mathbf{S}\big((1-h)P + h\delta_{\mathbf{x}}\big) = \theta(h;\mathbf{x}).$$

Since $\theta(\cdot;\mathbf{x})$ is also continuously differentiable at 0, $\mathrm{IF}(\mathbf{x};\mathbf{S},P)$ exists and the expression can be obtained from (A.7). As $\int \Psi(\mathbf{y}, \theta_0)\, dP(\mathbf{y}) = \mathbf{W}(0, \theta_0) = 0$, we find that at $(0, \theta_0)$ the derivative $\partial\mathbf{W}/\partial h$ is equal to $\Psi(\mathbf{x}, \theta_0)$ and the theorem follows. □

PROOF OF THEOREM 4.1: Put $U(\mathbf{x};\theta,\delta) = \sup_{\|\tau-\theta\|\le\delta} \|\Psi(\mathbf{x}, \tau) - \Psi(\mathbf{x}, \theta)\|$. According to Huber's (1967) Theorem 3 and its corollary it is sufficient to prove the following conditions :

(1) $\boldsymbol{\lambda}_P(\boldsymbol{\theta}_0) = \mathbf{0}$.

(2) There exist strictly positive constants b, c and d_0 such that

$$\text{(i) } \mathbf{E}_P U(X_1; \boldsymbol{\theta}, \delta) \leq b\delta \qquad \text{, for } \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| + \delta \leq d_0.$$

$$\text{(ii) } \mathbf{E}_P U^2(X_1; \boldsymbol{\theta}, \delta) \leq c\delta \qquad \text{, for } \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| + \delta \leq d_0.$$

(3) $\mathbf{E}_P \|\boldsymbol{\Psi}(X_1, \boldsymbol{\theta}_0)\|^2$ is finite.

According to Theorem 3.1 a solution $\mathbf{S}(P)$ of $(\mathcal{P}_P)$ exists and it must therefore satisfy equation (3.3). In other words $\boldsymbol{\theta}_0 = \big(\mathbf{t}(P), \mathbf{C}(P)\big)$ is a zero of $\boldsymbol{\lambda}_P(\boldsymbol{\theta})$ which proves (1). Lemma 3.3(i) yields condition (3).

Let K be any compact subset of Θ which contains $\boldsymbol{\theta}_0$. We will show that for all $\boldsymbol{\theta} \in K^o$ and δ sufficiently small, there exists a constant $b > 0$ such that

$$U(\mathbf{x}; \boldsymbol{\theta}, \delta) \leq b\delta. \tag{A.9}$$

This obviously yields condition (2). Let $\boldsymbol{\theta} = (\mathbf{t}, \mathbf{C}) \in K$. So both $\|\mathbf{C}\|$ and $\|\mathbf{C}^{-1}\|$ are bounded away from 0 and ∞.

Let δ be sufficiently small such that the ball $B_\delta(\boldsymbol{\theta}) \subset K$. Then the mean value theorem together with Lemma 3.3(ii) yield that there exists some constant $b > 0$ such that for $\boldsymbol{\tau} \in B_\delta(\boldsymbol{\theta})$ we have $\|\boldsymbol{\Psi}(\mathbf{x}, \boldsymbol{\tau}) - \boldsymbol{\Psi}(\mathbf{x}, \boldsymbol{\theta})\| \leq b\|\boldsymbol{\theta} - \boldsymbol{\tau}\| \leq b\delta$. This proves (A.9) and the theorem follows. □

Before proving Corollaries 5.1 and 5.2 we state three minor lemma's. The first one states a property of elliptical distributions.

LEMMA 5.1. *Let X have an elliptical distribution P with parameter $(\mathbf{0}, \mathbf{I})$. Then $U = X/\|X\|$ is independent of $\|X\|$, has mean zero and covariance matrix $\frac{1}{p}\mathbf{I}$. Furthermore, $\mathbf{E}_{0,I} UU^T U = \mathbf{0}$ and*

$$\mathbf{E}_{0,I} \mathrm{vec}(UU^T)\mathrm{vec}(UU^T) = \sigma_1(\mathbf{I} + \mathbf{K}_{p,p}) + \sigma_2 \mathrm{vec}(\mathbf{I})\mathrm{vec}(\mathbf{I})^T$$

where $\sigma_1 = \sigma_2 = \big(p(p+2)\big)^{-1}$.

PROOF OF LEMMA 5.1: To show independence of U and $\|X\|$ it is sufficient to show that $\|X\|$ and $(U_1, \ldots, U_{p-1})^T$ are independent. This can be proven immediately by performing the coordinate transformation $Y_i = U_i$, for $i = 1, 2, \ldots, p-1$ and $Y_p = \|X\|$, and computing the simultaneous density of $(Y_1, Y_2, \ldots, Y_p)^T$. The other results can be obtained by using spherical coordinates in a suitable manner. □

LEMMA 5.2. *Let $\mathbf{Z}$ be a random $p \times p$-matrix which has zero mean and covariance matrix $\mathbf{E}\,\mathrm{vec}(\mathbf{Z})\mathrm{vec}(\mathbf{Z})^T = \sigma_1(\mathbf{I} + \mathbf{K}_{p,p}) + \sigma_2 \mathrm{vec}(\mathbf{I})\,\mathrm{vec}(\mathbf{I})^T$. Suppose that $\mathbf{B}\mathbf{B}^T = \boldsymbol{\Sigma}$, then $\mathbf{B}\mathbf{Z}\mathbf{B}^T$ has zero mean and covariance matrix*

$$\sigma_1(\mathbf{I} + \mathbf{K}_{p,p})(\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma}) + \sigma_2 \mathrm{vec}(\boldsymbol{\Sigma})\,\mathrm{vec}(\boldsymbol{\Sigma})^T.$$

PROOF: We use two identities from Magnus and Neudecker (1979). First

$$\text{vec}(\mathbf{ABC}) = (\mathbf{C}^T \otimes \mathbf{A})\text{vec}(\mathbf{B})$$

which implies $\text{vec}(\mathbf{BZB}^T) = (\mathbf{B} \otimes \mathbf{B})\text{vec}(\mathbf{Z})$ and $\text{vec}(\mathbf{\Sigma}) = (\mathbf{B} \otimes \mathbf{B})\text{vec}(\mathbf{I})$. The second identity $\mathbf{K}_{m,n}(\mathbf{A} \otimes \mathbf{B}) = (\mathbf{B} \otimes \mathbf{A})\mathbf{K}_{m,n}$ yields

$$(\mathbf{B} \otimes \mathbf{B})\mathbf{K}_{p,p}(\mathbf{B} \otimes \mathbf{B})^T = \mathbf{K}_{p,p}(\mathbf{B} \otimes \mathbf{B})(\mathbf{B} \otimes \mathbf{B})^T.$$

As it is not difficult to see that $(\mathbf{B} \otimes \mathbf{B})(\mathbf{B} \otimes \mathbf{B})^T = \mathbf{\Sigma} \otimes \mathbf{\Sigma}$ the lemma follows. □

LEMMA 5.3. *Let* $\mathbf{1}$ *denote the* $p \times p$*-matrix with all entries equal to 1. For* $a, b, c, d \in \mathbf{R}$ *it holds that*

(i) If $a \neq 0$ *and* $a + pb \neq 0$, *then* $(a\mathbf{I} + b\mathbf{1})^{-1} = (1/a)\mathbf{I} + \big(b/(a(a+pb))\big)\mathbf{1}$.

(ii) $(c\mathbf{I} + d\mathbf{1})(a\mathbf{I} + b\mathbf{1})(c\mathbf{I} + d\mathbf{1}) = c^2a\mathbf{I} + \big(cad + ad(c+pd) + b(c+pd)^2\big)\mathbf{1}$

PROOF: Straightforward. □

PROOF OF COROLLARY 5.1: Affine equivariance of $\mathbf{t}_n$ and $\mathbf{C}_n$ and Lemma 5.2 imply that we may restrict to $\boldsymbol{\theta}_0 = (\mathbf{0}, \mathbf{I})$. Obviously the conditions of Theorem 3.2 hold for elliptical distributions. Therefore in order to apply Theorem 4.1 we are left with showing that $\boldsymbol{\lambda}_{P_{0,I}}(\cdot)$ has a nonsingular derivative at $\boldsymbol{\theta}_0$. To show this and to derive (5.4) we first consider the symmetric $p \times p$-matrix $\mathbf{C}$ as $\frac{1}{2}p(p+1)$-vector $(c_{11}, \ldots, c_{pp}, c_{12}, \ldots, c_{p-1,p})^T$ consisting of the upper right triangle elements of the matrix $\mathbf{C}$.

According to (A.8), $\mathbf{\Lambda} = \mathbf{E}_{0,I}\mathbf{\Psi}_{\boldsymbol{\theta}}(X_1, \boldsymbol{\theta}_0)$. With Lemma 5.1 it follows from the expressions found for $\partial\mathbf{\Psi}_1/\partial\boldsymbol{\theta}$ and $\partial\mathbf{\Psi}_2/\partial\boldsymbol{\theta}$ in the proof of Lemma 3.3 that $\mathbf{\Lambda}$ is of block form

(A.10)
$$\mathbf{\Lambda} = \begin{array}{|c|c|}\hline \mathbf{\Delta_t} & \\ \hline & \mathbf{\Delta_C} \\ \hline \end{array}$$

where $\mathbf{\Delta_t} = \mathbf{E}_{0,I}(\mathbf{\Psi}_1)_{\mathbf{t}}(X_1, \boldsymbol{\theta}_0)$ and $\mathbf{\Delta_C} = \mathbf{E}_{0,I}(\mathbf{\Psi}_2)_{\mathbf{C}}(X_1, \boldsymbol{\theta}_0)$.

Using Lemma 5.1 again we see that $\mathbf{\Delta_t} = -\beta\mathbf{I}$ and is therefore nonsingular. The matrix $\mathbf{\Delta_C}$ is a $\frac{1}{2}p(p+1) \times \frac{1}{2}p(p+1)$-matrix which consists of two nonzero block matrices on the main diagonal. The upper left is a $p \times p$-matrix $\mathbf{\Delta}_{\mathbf{C},1} = -\gamma\mathbf{I} + \eta\mathbf{1}$ and the lower right matrix is a diagonal $\frac{1}{2}p(p-1) \times \frac{1}{2}p(p-1)$-matrix $\mathbf{\Delta}_{\mathbf{C},2} = -\gamma\mathbf{I}$, where

(A.11)
$$\gamma = \frac{\mathbf{E}_{0,I}\big[\psi'(\|X_1\|)\|X_1\|^2 + (p+1)\psi(\|X_1\|)\|X_1\|\big]}{p+2}$$
$$\eta = \frac{\mathbf{E}_{0,I}\big[2\psi'(\|X_1\|)\|X_1\|^2 + p\psi(\|X_1\|)\|X_1\|\big]}{2p(p+2)}.$$

As the matrix $\Delta_{\mathbf{C},1}$ has determinant $(-\gamma + p\eta)(-\gamma)^{p-1}$, it follows from (5.2) that $\Delta_{\mathbf{C}}$ is also nonsingular and hence Theorem 3.1 applies.

To obtain the expressions for the asymptotic covariance matrices first compute the covariance matrix $\mathbf{M}$ of $(\Psi_1, \Psi_2)^T$, with the symmetric matrix Ψ_2 considered as a $\frac{1}{2}p(p+1)$-vector. Lemma 5.1 implies that $\mathbf{M}$ is also a block matrix

$$\mathbf{M} = \begin{array}{|c|c|} \hline \mathbf{M}_1 & \\ \hline & \mathbf{M}_2 \\ \hline \end{array}$$

where $\mathbf{M}_1 = \mathbf{E}_{0,I} u^2(\|X_1\|) X_1 X_1^T = \alpha \mathbf{I}$ and $\mathbf{M}_2 = \mathbf{E}_{0,I} \Psi_2(X_1, \theta_0)\, \Psi_2(X_1, \theta_0)^T$. The matrix $\mathbf{\Lambda}^{-1}$ is of the same structure as $\mathbf{M}$, which immediately implies that $\mathbf{t}_n$ and $\mathbf{C}_n$ are asymptotically independent and that $\mathbf{t}_n$ has asymptotic covariance matrix

$$\Delta_{\mathbf{t}}^{-1} \mathbf{M}_1 \Delta_{\mathbf{t}}^{-T} = \frac{\alpha}{\beta^2} \mathbf{I}.$$

To describe the $\frac{1}{2}p(p+1) \times \frac{1}{2}p(p+1)$-matrix $\Delta_{\mathbf{C}}^{-1} \mathbf{M}_2 \Delta_{\mathbf{C}}^{-1}$, consider the covariance matrix $\mathbf{M}_2$ of $\Psi_2(X_1, \theta_0)$. Because

$$\Psi_{2,ij}(\mathbf{x}, \theta_0) = p\psi(\|\mathbf{x}\|)\|\mathbf{x}\| \frac{x_i x_j}{\|\mathbf{x}\|^2} - v(\|\mathbf{x}\|)\delta_{ij}$$

Lemma 5.1 implies that $\mathbf{M}_2$ is of the same structure as $\Delta_{\mathbf{C}}$. It also consist of two nonzero block matrices on the main diagonal. The upper-left is the $p \times p$-covariance matrix $\mathbf{M}_{2,1}$ of the diagonal elements $\Psi_{2,ii}(X_1, \theta_0)$: $\mathbf{M}_{2,1} = \delta_1 \mathbf{I} + \delta_2 \mathbf{1}$, and the lower-right matrix is a diagonal $\frac{1}{2}p(p-1) \times \frac{1}{2}p(p-1)$-covariance matrix $\mathbf{M}_{2,2}$ of the off-diagonal elements $\Psi_{2,ij}(X_1, \theta_0)$ $(1 \le i \le j \le p)$: $\mathbf{M}_{2,2} = \frac{1}{2}\delta_1 \mathbf{I}$, where $\delta_1 = 2p(p+2)^{-1} \mathbf{E}_{0,I} \psi^2(\|X_1\|)\|X_1\|^2$ and $\delta_2 = -\delta_1/p + \mathbf{E}_{0,I}(\rho(\|X_1\|) - b)^2$. Therefore, $\Delta_{\mathbf{C}}^{-1}$ and $\mathbf{M}_2$ are of the same structure and hence $\Delta_{\mathbf{C}}^{-1} \mathbf{M}_2 \Delta_{\mathbf{C}}^{-1}$ is. It follows immediately that the lower-right matrix is a $\frac{1}{2}p(p-1) \times \frac{1}{2}p(p-1)$-diagonal matrix with diagonal element

$$\text{(A.12)} \qquad \sigma_1 = \frac{\delta_1}{2\gamma^2} = \frac{p(p+2)\mathbf{E}_{0,I} \psi^2(\|X_1\|)\|X_1\|^2}{\left\{ \mathbf{E}_{0,I}[\psi'(\|X_1\|)\|X_1\|^2 + (p+1)\psi(\|X_1\|)\|X_1\|] \right\}^2}.$$

The upper-left matrix is the $p \times p$-matrix $\Delta_{\mathbf{C},1}^{-1} \mathbf{M}_{2,1} \Delta_{\mathbf{C},1}^{-T}$. Using Lemma 5.2, easy but tedious computations show that this equals $2\sigma_1 \mathbf{I} + \sigma_2 \mathbf{1}$, with σ_2 as in (5.5).

The expressions that we have found for $\Delta_{\mathbf{C},1}^{-1} \mathbf{M}_{2,1} \Delta_{\mathbf{C},1}^{-T}$ and $\Delta_{\mathbf{C},2}^{-1} \mathbf{M}_{2,2} \Delta_{\mathbf{C},2}^{-T}$ tell us that $\sqrt{n}(\mathbf{C}_n - \mathbf{I})$ converges in distribution to a symmetric random matrix $\mathbf{Z}$ of which the off-diagonal elements are uncorrelated with each other and uncorrelated with the diagonal elements, of which each off-diagonal element has variance σ_1, and of which the diagonal elements all have variance $2\sigma_1 + \sigma_2$ with the covariance between any two diagonal elements being σ_2. In other words

$$\mathbf{E}\text{vec}(\mathbf{Z})\text{vec}(\mathbf{Z})^T = \sigma_1(\mathbf{I} + \mathbf{K}_{p,p}) + \sigma_2 \text{vec}(\mathbf{I})\,\text{vec}(\mathbf{I})^T$$

which proves the corollary for the case $\theta_0 = (\mathbf{0}, \mathbf{I})$. Lemma 5.2 then implies the general form (5.4). □

PROOF OF COROLLARY 5.2: The conditions of Theorem 3.2 hold for elliptical distributions. According to (5.1), $\mathbf{S}(P) = (\mathbf{0}, \mathbf{I})$ and from the proof of Corollary 5.1 we have that $\boldsymbol{\lambda}_P(\cdot)$ has the nonsingular derivative $\boldsymbol{\Lambda}$ of (A.10) at $(\mathbf{0}, \mathbf{I})$. Therefore Theorem 3.3 applies, which means that $\mathrm{IF}(\mathbf{x}; \mathbf{S}, P)$ exists and its expression can be obtained from (3.4). As $\boldsymbol{\Lambda}$ consists of the two block matrices $\boldsymbol{\Delta}_\mathbf{t}$ and $\boldsymbol{\Delta}_\mathbf{C}$ on the main diagonal, $\mathrm{IF}(\mathbf{x}; \mathbf{t}, P)$ and $\mathrm{IF}(\mathbf{x}; \mathbf{C}, P)$ can be treated separately.

Equations (3.4) and (2.8) give

$$\mathrm{IF}(\mathbf{x}; \mathbf{t}, P) = -\boldsymbol{\Delta}_\mathbf{t}^{-1} \boldsymbol{\Psi}_1(\mathbf{x}, (\mathbf{0}, \mathbf{I})) = \frac{1}{\beta} u(\|\mathbf{x}\|)\mathbf{x}$$

which proves (5.7). Let us denote by $\mathbf{IF} = (\mathrm{IF}_{11}, \ldots, \mathrm{IF}_{pp}, \mathrm{IF}_{12}, \ldots, \mathrm{IF}_{p-1,p})^T$ the influence function $\mathrm{IF}(\mathbf{x}; \mathbf{C}, P)$ of the covariance estimator. Then (3.4) and (2.8), together with the expression found for $\boldsymbol{\Delta}_\mathbf{C}$ in the proof of Corollary 5.1, yield

$$-\gamma \mathrm{IF}_{ii} + \eta \mathrm{trace}(\mathbf{IF}) = -pu(\|\mathbf{x}\|)x_i^2 + v(\|\mathbf{x}\|) \tag{A.13}$$

$$-\gamma \mathrm{IF}_{ij} = -pu(\|\mathbf{x}\|)x_i x_j \tag{A.14}$$

where γ and η are as in (A.11). Summation of (A.13) over $i = 1, 2, \ldots, p$ gives

$$\mathrm{trace}(\mathbf{IF}) = \frac{-p\psi(\|\mathbf{x}\|)\|\mathbf{x}\| + pv(\|\mathbf{x}\|)}{-\gamma + p\eta}. \tag{A.15}$$

From (A.11) we have $-\gamma + p\eta = -\frac{1}{2}\mathbf{E}_{0,I}\psi(\|X\|)\|X\|$, and when we put in $v(y) = \psi(y)y - \rho(y) + b$, we find (5.9). Finally substitute (A.15) into (A.13). Together with (A.14) this proves (5.8). □

Acknowledgment. I thank Peter Rousseeuw for stimulating discussions and helpful remarks. I also thank the two referees whose comments have lead to a considerable improvement of the initial version of this paper.

REFERENCES

BILLINGSLEY, P. (1968). *Convergence of Probability Measures*. Wiley, New York.

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

DEVLIN, S.J., GNANADESIKAN, R. and KETTENRING, J.R. (1981). Robust estimation of dispersion matrices and principal components. *J. Am. Statist. Assoc.* **76** 354-362.

DIEUDONNÉ, J. (1969). *Foundations of Modern Analysis*. Academic Press, New York.

GRAYBILL, F.A. (1983). *Matrices with Applications in Statistics*. Wadsworth, Belmont, California.

GRÜBEL, R. (1988). A minimal characterization of the covariance matrix. *Metrika* **35** 49–52.

HAMPEL, F.R. (1974). The influence curve and its role in robust estimation. *J. Am. Statist. Assoc.* **69** 383–393.

HUBER, P.J. (1964). Robust estimation of a location parameter. *Ann. Statist.* **35** 73–101.

HUBER, P.J. (1967). The behaviour of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (L.M.Le Cam and J.Neyman, eds.) 221–233. University of California Press, Berkeley.

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

KIM, J. and POLLARD, D. (1989). Cube root asymptotics. Technical Report (1987), Yale University, To appear in Ann. Statist..

LOPUHAÄ, H.P. and ROUSSEEUW, P.J. (1989). Breakdown properties of affine equivariant estimators of multivariate location and covariance matrices. Revised version of Technical 87-14, Delft University of Technology. Tentatively accepted by Ann. Statist..

MAGNUS, J.R and NEUDECKER, H. (1979). The commutation matrix: Some properties and applications. *Ann. Statist.* **7** 381–394.

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

ROUSSEEUW, P.J. (1983). Multivariate estimation with high breakdown point. Paper presented at the Fourth Pannonian Symposium on Mathematical Statistics and Probability, Bad Tatzmannsdorf, Austria, September 4-9, 1983. In *Mathematical Statistics and Applications (1985)* (W.Grossmann, G.Pflug, I.Vincze and W.Wertz, eds.) 283–297. Reidel, Dordrecht, The Netherlands.

ROUSSEEUW, P.J. and YOHAI, V. (1984). Robust regression by means of S-estimators. In *Robust and Nonlinear Time Series Analysis.* Lecture Notes in Statistics **26** 256–272. Springer Verlag, New York.

TYLER, D.E. (1982). Radial estimates and the test for sphericity. *Biometrika* **69** 429–436.

TYLER, D.E. (1983). Robustness and efficiency properties of scatter matrices. *Biometrika* **70** 411–420.

TYLER, D.E. (1986). Breakdown properties of the M-estimators of multivariate scatter. Technical Report, Rutgers University, New Jersey.

HIGHLY EFFICIENT ESTIMATORS OF MULTIVARIATE LOCATION WITH HIGH BREAKDOWN POINT

HENDRIK P. LOPUHAÄ

Delft University of Technology

We propose a 2-stage procedure to compute a robust estimator of multivariate location. The procedure consist of a first stage where an affine equivariant covariance estimator is computed, and a second stage that is similar to computing an M-estimator of location. The resulting location estimator will inherit the breakdown point of the initial covariance estimator, and within the location-covariance model only the second stage will determine the type of influence function and the asymptotic behaviour. This enables us to combine a high breakdown point and a bounded influence function with high asymptotic efficiency. In defining the procedure, we will distinguish between weakly redescending and strongly redescending influence functions. We obtain the breakdown point and the influence function, and prove consistency and asymptotic normality.

1. Introduction. Consider the standard location-covariance model, i.e. one observes p-dimensional $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ and assumes these are realizations of independent random vectors $X_1, X_2, \ldots, X_n$, with an elliptical distribution $P_{\mu,\Sigma}$ with unknown parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, that has a density

$$f_{\mu,\Sigma}(\mathbf{x}) = |\mathbf{B}|^{-1} f\big(\|\mathbf{B}^{-1}(\mathbf{x}-\boldsymbol{\mu})\|\big) \tag{1.1}$$

where $\mathbf{B}\mathbf{B}^T = \boldsymbol{\Sigma}$. Here $\mathbf{x} = (x_1 \cdots x_p)^T \in \mathbf{R}^p$, $\boldsymbol{\mu} \in \mathbf{R}^p$, $\boldsymbol{\Sigma}$ is a positive definite symmetric $p \times p$-matrix, and $f : [0,\infty) \to [0,\infty)$ is a known function.

A well known estimator for the location parameter $\boldsymbol{\mu}$ is the least squares estimator, defined as the value $\mathbf{t}_n \in \mathbf{R}^p$ that minimizes $\sum_{i=1}^n \|\mathbf{x}_i - \mathbf{t}\|^2$, which results in the sample mean. In case $P_{\mu,\Sigma}$ is a normal distribution, this estimator corresponds to the maximum likelihood estimator for $\boldsymbol{\mu}$ and is therefore asymptotically efficient at $P_{\mu,\Sigma}$. However, it is not robust at all, as a single outlier can have a large effect on the estimator.

To measure the degree of robustness of an estimator, Hampel (1968) introduced two concepts : the breakdown point ε^* and the influence function IF. Donoho and Huber (1983) introduced a finite-sample version of the breakdown point based on contamination of arbitrary subsets of the observations. This finite-sample (*replacement*) breakdown point ε_n^* may be interpreted as the minimum fraction of contamination that spoils the estimator completely and it must be seen as a global measure of robustness. The influence function $\mathrm{IF}(\mathbf{x}; \mathbf{t}, P)$ describes the effect on the estimator of a small pertubation locally at point $\mathbf{x}$ (see Hampel 1974 and Hampel et al. 1986 for a discussion). For example, the poor robustness of the sample mean

This research is financially supported by NWO under Grant 10-62-10.

AMS 1980 *subject classifications* : 62F35, 62H12.

Key words: Multivariate Location, High breakdown point, Bounded influence, High efficiency

is illustrated by a breakdown point $\varepsilon_n^* = 1/n$ and an unbounded influence function $\mathrm{IF}(\mathbf{x}; \mathbf{t}, P) = \mathbf{x}$ at any distribution P with mean zero.

We are interested in combining a high breakdown point and a bounded influence function with high asymptotic efficiency for a multivariate location estimator, that also commutes with affine transformations of the observations. To our knowledge, no estimator of a multivariate location parameter combines these four properties. Location M-estimators are either not affine equivariant (Huber 1964) or, if they are defined simultaneously with a covariance M-estimator (Maronna 1976), the breakdown point is at most $1/(p+1)$ (see Tyler 1986 for a detailed treatment). Location and covariance S-estimators are affine equivariant and have a bounded influence function, however one still has to make a tradeoff between breakdown point and asymptotic efficiency; unfortunately, a high breakdown point is counterbalanced by a low asymptotic efficiency and vice versa (Lopuhaä 1989).

In this paper we propose a 2-stage procedure to obtain affine equivariant highly efficient multivariate location estimators with a high breakdown point and a bounded influence function. We distinguish between a procedure that yields an influence function which is strictly positive outside zero, and a procedure that yields a strongly redescending influence function, i.e. $\mathrm{IF}(\mathbf{x}; \mathbf{t}, P) = \mathbf{0}$ for $\|\mathbf{x}\| \geq c$. Both proposals consist of a first stage in which an affine equivariant high breakdown covariance estimator is computed, and a second stage that is similar to computing an M-estimator of location. In both cases the resulting location estimator inherits the breakdown point of the initial covariance estimator, and if we estimate within the location-covariance model, only the second stage determines the type of influence function, the rate of convergence as well as the asymptotic efficiency independent of the initial covariance estimator as long as it is consistent. A typical combination will be a covariance S-estimator, either the minimum volume ellipsoid estimator or a smoothed version of it, followed by a location M-estimator.

In Section 2 we discuss the pros and cons of M- and S-estimators, and next we define the two proposals and briefly summarize the results concerning their robustness and their asymptotic behaviour. All results are then proven in subsequent sections. In Sections 3 and 4 we obtain the breakdown point and the expression of the influence function. In Section 5 we show consistency and asymptotic normality.

2. Definition and corresponding M- and S-estimators.

2.1 Location M-estimators. Location M-estimators are a well known robustification of the least squares method. Similar to Huber (1964), one may define an M-estimator of multivariate location as the vector of $\mathbf{m}_n \in \mathbf{R}^p$ that minimizes

$$\sum_{i=1}^{n} \rho(\|\mathbf{x}_i - \mathbf{m}\|). \tag{2.1}$$

Typically, $\rho(y)$ is chosen to be a symmetric function that is quadratic in the middle and which increases slower than y^2 as $y \to \infty$. An example is the function

$$\rho_{\mathrm{H}}(y; k) = \begin{cases} \frac{1}{2}y^2 & , |y| \leq k \\ -\frac{1}{2}k^2 + k|y| & , |y| \geq k \end{cases} \tag{2.2}$$

with a bounded monotone derivative

$$\psi_H(y;k) = \begin{cases} -k & \text{, for } y \leq -k \\ y & \text{, for } |y| \leq k \\ k & \text{, for } y \geq k \end{cases}$$

known as Huber's ψ-function.

At distributions P that are spherically symmetric around the origin, the influence function of the location M-estimator is equal to

$$\text{IF}(\mathbf{x};\mathbf{m},P) = \frac{\psi(\|\mathbf{x}\|)}{\beta\|\mathbf{x}\|}\mathbf{x} \tag{2.3}$$

where ψ is the derivative of ρ and where β is a positive constant. (see for instance Hampel et al. 1986). In general, an unbounded function ρ in (2.1) which does not increase too fast, may lead to location M-estimators with a high breakdown point (Huber 1984), with a bounded influence function, and with good asymptotic efficiency relative to the maximum likelihood estimator at several spherically symmetric distributions (Maronna 1976). Unfortunately, these location M-estimators are not equivariant with respect to affine transformations of the $\mathbf{x}_i$. Maronna (1976) solves this by defining M-estimators simultaneously for location and covariance, but these estimators become more sensitive to outliers as the dimension p increases : $\varepsilon_n^* \leq 1/(p+1)$, due to breakdown of the covariance M-estimator (Tyler 1986).

Alternatively, one may obtain affine equivariance and retain the good breakdown properties, by estimating the unknown covariance structure with some affine equivariant covariance estimator $\mathbf{C}_n = \mathbf{A}_n\mathbf{A}_n^T$ that has a high breakdown point, and then define the final location estimator as $\mathbf{t}_n = \mathbf{A}_n\tilde{\mathbf{m}}_n$, where $\tilde{\mathbf{m}}_n$ is a location M-estimator based upon scaled observations $\tilde{\mathbf{x}}_i = \mathbf{A}_n^{-1}\mathbf{x}_i$ in (2.1). A suitable class of affine equivariant covariance estimators with a high breakdown point are S-estimators. We briefly discuss them in the next subsection.

2.2 S-estimators Another robustification of the least squares method, which does give affine equivariance as well as a high breakdown point, are S-estimators. Rousseeuw and Yohai (1984) originally introduced them in a regression context, but these estimators easily generalize to multivariate location and covariance.

S-estimators for multivariate location and covariance are defined as the vector $\mathbf{t}_n$ and the positive definite symmetric matrix $\mathbf{C}_n$ that minimize the determinant of $\mathbf{C}$ subject to

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left[\{(\mathbf{x}_i-\mathbf{t})^T\mathbf{C}^{-1}(\mathbf{x}_i-\mathbf{t})\}^{1/2}\right] \leq b. \tag{2.4}$$

The constant $0 < b < \sup\rho$ can be chosen in agreement with an assumed underlying distribution. For instance, when $\mathbf{x}_1,\ldots,\mathbf{x}_n$ are assumed to be a sample $X_1,\ldots,X_n$ from an elliptical distribution (1.1), a natural choice for b is

$$b = \mathbf{E}\rho\left[\{(X_1-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(X_1-\boldsymbol{\mu})\}^{1/2}\right] = \int \rho(\|\mathbf{x}\|)f(\|\mathbf{x}\|)\,d\mathbf{x} \tag{2.5}$$

which is independent of μ and Σ.

Davies (1987) has investigated some properties of these estimators. When the function $\rho : \mathbf{R} \to [0, \infty)$ satisfies

(R1) $\rho(0) = 0$, ρ is symmetric.
(R2) ρ is continuous at 0 and left-continuous on $[0, \infty)$.
(R3) ρ is nondecreasing on $[0, \infty)$ and there exists a constant $c > 0$ such that ρ equals $\rho(c)$ on $[c, \infty)$.

then at least one pair $\boldsymbol{\theta}_n = (\mathbf{t}_n, \mathbf{C}_n)$ minimizes the determinant of $\mathbf{C}$ subject to (2.4), and consistency of all such pairs can be obtained. Furthermore, the corresponding S-functional is uniquely defined at elliptical distributions : $(\mathbf{t}(P_{\mu,\Sigma}), \mathbf{C}(P_{\mu,\Sigma})) = (\mu, \Sigma)$. Although the S-estimator $\boldsymbol{\theta}_n$ may not be uniquely defined, it is affine equivariant in the following sense. For any pair $(\mathbf{t}_n, \mathbf{C}_n)$ that minimizes the determinant of $\mathbf{C}$ subject to (2.4) it holds that $(\mathbf{A}\,\mathbf{t}_n + \mathbf{v}, \mathbf{A}\,\mathbf{C}_n\mathbf{A}^T)$ minimizes the determinant of $\mathbf{C}$ subject to constraint (2.4) with observations $\tilde{\mathbf{x}}_i = \mathbf{A}\mathbf{x}_i + \mathbf{v}$.

When we use a smooth function ρ that, in addition to (R1)-(R3), also satisfies

(R4) ρ is strictly increasing on [0,c].
(R5) ρ is twice continuously differentiable.

then this will lead to S-estimators with a bounded influence function, which are asymptotically normal at rate $\sqrt{n}$ (Lopuhaä 1989). When the sample distribution is elliptical, so that b is chosen according to (2.5), the limiting variance and the breakdown point $\lceil nr \rceil / n$ of $\boldsymbol{\theta}_n$ (where $r = b/\rho(c)$) will both depend on c (see Lopuhaä and Rousseeuw 1989). Unfortunately, a high breakdown point is counterbalanced by a low asymptotic efficiency and vice versa. A typical function ρ in (2.4) that satisfies (R1)-(R5) would be quadratic in the middle, but as opposed to $\rho_{\text{H}}(y; k)$, it is constant outside an interval $[-c, c]$. A well known example is the function

$$\rho_{\text{B}}(y; c) = \begin{cases} \dfrac{y^2}{2} - \dfrac{y^4}{2c^2} + \dfrac{y^6}{6c^4} & , \text{ for } |y| \leq c \\ \dfrac{c^2}{6} & , \text{ for } |y| \geq c \end{cases} \tag{2.6}$$

which has a redescending derivative

$$\psi_{\text{B}}(y; c) = \begin{cases} y\left(1 - \left(\dfrac{y}{c}\right)^2\right)^2 & , \text{ for } |y| \leq c \\ 0 & , \text{ for } |y| \geq c \end{cases}$$

known as Tukey's biweight.

Perhaps one does not immediately recognize the least squares estimator as a special case of an S-estimator. However, when we take $\rho(y) = \frac{1}{2}y^2$ in (2.4), the S-minimization problem has the sample mean and the sample covariance as a unique solution (see for instance Grübel 1988).

2.3 Definition and remarks We will not only study estimators $\mathbf{t}_n$ and $\mathbf{C}_n$ as functions of $\mathbf{x}_1, \ldots, \mathbf{x}_n$, but we will also consider them as images $\mathbf{t}(P_n)$ and $\mathbf{C}(P_n)$ of the functionals $\mathbf{t} : \mathcal{P} \to \mathbf{R}^p$ and $\mathbf{C} : \mathcal{P} \to \mathrm{PDS}(p)$, where P_n is the empirical distribution that assigns mass $1/n$ to each $\mathbf{x}_i$ for $1 \leq i \leq n$, $\mathcal{P}$ is the space of all probability measures on $\mathbf{R}^p$, and $\mathrm{PDS}(p)$ denotes the class of all positive definite symmetric $p \times p$-matrices. Write $\Theta = \mathbf{R}^p \times \mathrm{PDS}(p)$. Affine equivariance of $\mathbf{t}(\cdot)$ and $\mathbf{C}(\cdot)$ means $\mathbf{t}(P_{\mathbf{A}X+\mathbf{v}}) = \mathbf{A}\,\mathbf{t}(P_X) + \mathbf{v}$ and $\mathbf{C}(P_{\mathbf{A}X+\mathbf{v}}) = \mathbf{A}\,\mathbf{C}(P_X)\,\mathbf{A}^T$, where P_X denotes the distribution of a random vector X.

Location M-estimators defined by minimizing (2.1) are not affine equivariant. We can obtain affine equivariance by estimating the unknown covariance structure affinely.

METHOD 1

Let $\rho : \mathbf{R} \to [0, \infty)$ be symmetric, increasing towards both sides, with $\rho(0) = 0$. Assume that

$$\lim_{|y| \to \infty} \rho(y) = \infty. \tag{2.7}$$

Furthermore, assume that the functions $\psi = \rho'$ and $u(y) = \psi(y)/y$ are continuous, and that there exists a $y_0 > 0$ such that ψ is nondecreasing on $(0, y_0)$ and nonincreasing on (y_0, ∞).

Let $\mathbf{C}_n \in \mathrm{PDS}(p)$ be an affine equivariant covariance estimator based upon $\mathbf{x}_1, \ldots, \mathbf{x}_n$. Define $\mathbf{t}_n$ as the vector that minimizes the function

$$R_n(\mathbf{t}) = \frac{1}{n} \sum_{i=1}^{n} \rho\left[\{(\mathbf{x}_i - \mathbf{t})^T \mathbf{C}_n^{-1} (\mathbf{x}_i - \mathbf{t})\}^{1/2}\right]. \tag{2.8}$$

REMARK 2.1: The corresponding location functional $\mathbf{t}(\cdot)$ is defined at P as the vector $\mathbf{t}(P)$ that minimizes the function

$$R_P(\mathbf{t}) = \int \rho\left[\{(\mathbf{x} - \mathbf{t})^T \mathbf{C}(P)^{-1} (\mathbf{x} - \mathbf{t})\}^{1/2}\right] dP(\mathbf{x}) \tag{2.9}$$

where $\mathbf{C} : \mathcal{P} \to \mathrm{PDS}(p)$ is an affine equivariant covariance functional.

EXAMPLE 2.1: For $\mathbf{C}_n$ one may use the covariance minimum volume ellipsoid (MVE) estimator. It is defined as the covariance matrix of the smallest ellipsoid that covers at least $\lfloor \frac{n+p+1}{2} \rfloor$ observations. The MVE estimator can be seen as an S-estimator defined with the function $1 - \mathbf{1}_{[-c,c]}(y)$ in (2.4), which satisfies (R1)-(R3), and it has breakdown point $\varepsilon_n^* = \lfloor \frac{n-p+1}{2} \rfloor / n$ (Rousseeuw 1983, see also Lopuhaä and Rousseeuw 1989). A choice for the function ρ in (2.8) may be the function ρ_{H} of (2.2). The location M-estimator defined with ρ_{H} turned out to be Huber's (1964) robust minimax solution. It has good asymptotic efficiency relative to the maximum likelihood estimator at several spherically symmetric distributions (Maronna 1976), and a bounded influence function.

It is not difficult to show that for any P, there always exists at least one vector $\mathbf{t}(P)$ that minimizes (2.9) In particular, there exists at least one vector $\mathbf{t}_n = \mathbf{t}(P_n)$ that minimizes $R_n(\mathbf{t})$, but it may not be unique as a function of $\mathbf{x}_1, \ldots, \mathbf{x}_n$. If $\mathbf{t}_n$ is unique, it is obviously affine equivariant; if it is not unique, it will always be affine equivariant in the following sense. When all $\mathbf{x}_i$'s are transformed into $\tilde{\mathbf{x}}_i = \mathbf{A}\,\mathbf{x}_i + \mathbf{v}$, then affine equivariance of $\mathbf{C}_n$ implies that $\tilde{R}_n(\mathbf{t}) = (1/n)\sum_{i=1}^n \rho[\{(\tilde{\mathbf{x}}_i - \mathbf{t})^T \tilde{\mathbf{C}}_n^{-1}(\tilde{\mathbf{x}}_i - \mathbf{t})\}^{1/2}]$, where $\tilde{\mathbf{C}}_n = \mathbf{A}\mathbf{C}_n\mathbf{A}^T$, relates to $R_n(\mathbf{t})$ as $\tilde{R}_n(\mathbf{t}) = R_n\big((\mathbf{A}^{-1}(\mathbf{t} - \mathbf{v})\big)$. Hence, for any $\mathbf{t}_n$ that minimizes $R_n(\mathbf{t})$, there exists a value $\tilde{\mathbf{t}}_n = \mathbf{A}\,\mathbf{t}_n + \mathbf{v}$ that minimizes $\tilde{R}_n(\mathbf{t})$. We will show that $\mathbf{t}_n$ will inherit the breakdown point of $\mathbf{C}_n$, and that, if one assumes that the underlying distribution P is elliptical, the influence function as well as the limiting distribution of $\mathbf{t}_n$ will be similar to those of the corresponding location M-estimator defined by minimizing (2.1) with the same ρ-function.

The restrictions imposed on the function ρ are somewhat weaker than in Huber (1984) to include the function (2.2). In particular, (2.7) guarantees that the location estimator inherits the breakdown point of any initial covariance estimator. For instance, when the function ρ in (2.1) is bounded, the breakdown point of location M-estimators at a collection $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)$ turns out to depend on the actual structure of the collection $\mathbf{X}$ (Huber 1984). Therefore, in general one can not expect $\mathbf{t}_n$ to have a breakdown point that is independent of the breakdown behaviour of $\mathbf{C}_n$. Nonetheless, a bounded ρ in (2.8) may be of interest, since in that case the influence function in (2.3) will be redescending in $\|\mathbf{x}\|$. In particular, if we use a function ρ that is constant outside an interval $[-c, c]$, the influence function will be *strongly* redescending, i.e. it is zero outside the ball $\{\|\mathbf{x}\| \leq c\}$.

Although in general, with bounded ρ in (2.8), one may have difficulties with obtaining a high breakdown point, there is a specific combination with bounded ρ that avoids this. In that situation we can even differentiate (2.8) with respect to $\mathbf{t}$, and define the final location estimator as any zero of the derivative. This proposal then becomes the multivariate version of Yohai's (1987) regression MM-estimators.

METHOD 2

Let $\rho_j : \mathbf{R} \to [0, \infty), j = 1, 2$, be two functions that satisfy conditions (R1)-(R5). Let $\boldsymbol{\theta}_{1,n} = (\mathbf{t}_{1,n}, \mathbf{C}_{1,n})$ be an S-estimator based upon $\mathbf{x}_1, \ldots, \mathbf{x}_n$, defined with the function $\rho_1(\cdot)$ and constant $0 < b_1 < \rho_1(c_1)$ in (2.4). Let $\rho_2(\cdot)$ be related to $\rho_1(\cdot)$ as follows :

$$\rho_1(y) \geq \rho_2(y) \tag{2.10}$$

and

$$\rho_1(c_1) = \sup \rho_1 = \sup \rho_2 = \rho_2(c_2). \tag{2.11}$$

Let ψ_2 denote the derivative of ρ_2 and let $u_2(y) = \psi_2(y)/y$. Define $\mathbf{t}_{2,n}$ as a solution of

$$\sum_{i=1}^n u_2\left[\{(\mathbf{x}_i - \mathbf{t})^T \mathbf{C}_{1,n}^{-1}(\mathbf{x}_i - \mathbf{t})\}^{1/2}\right](\mathbf{x}_i - \mathbf{t}) = \mathbf{0} \tag{2.12}$$

which verifies

$$R_{2,n}(\mathbf{t}_{2,n}) \leq R_{2,n}(\mathbf{t}_{1,n}) \tag{2.13}$$

where

$$R_{2,n}(\mathbf{t}) = \frac{1}{n}\sum_{i=1}^{n} \rho_2 \left[\{(\mathbf{x}_i - \mathbf{t})^T \mathbf{C}_{1,n}^{-1} (\mathbf{x}_i - \mathbf{t})\}^{1/2}\right]. \tag{2.14}$$

Of course, the vector that minimizes the function $R_{2,n}(\mathbf{t})$ of (2.14) is one solution of (2.12). However, an advantage of the second proposal may be that, besides the vector that minimizes $R_{2,n}(\mathbf{t})$, every solution of (2.12) that verifies (2.13) will be an affine equivariant highly efficient location estimator with a high breakdown point and a bounded influence function.

REMARK 2.2: The corresponding location functional $\mathbf{t}_2(\cdot)$ is defined at P as any solution of $\int u_2[\{(\mathbf{x} - \mathbf{t})^T \mathbf{C}_1(P)^{-1}(\mathbf{x} - \mathbf{t})\}^{1/2}](\mathbf{x} - \mathbf{t})\, dP(\mathbf{x}) = \mathbf{0}$, which verifies

$$R_{2,P}(\mathbf{t}_2(P)) \leq R_{2,P}(\mathbf{t}_1(P)) \tag{2.15}$$

where

$$R_{2,P}(\mathbf{t}) = \int \rho \left[\{(\mathbf{x} - \mathbf{t})^T \mathbf{C}_1(P)^{-1}(\mathbf{x} - \mathbf{t})\}^{1/2}\right] dP(\mathbf{x}) \tag{2.16}$$

and $\boldsymbol{\theta}_1(P) = \big(\mathbf{t}_1(P), \mathbf{C}_1(P)\big)$ is the S-functional defined with the function $\rho_1(\cdot)$.

It is not difficult to show that for any P, at which $\boldsymbol{\theta}_1(\cdot)$ is defined, there exists at least one value $\mathbf{t}_2(P)$ that minimizes (2.16). Hence, there exists at least one value $\mathbf{t}_{2,n}$ that satisfies (2.12) and (2.13). Any estimator $\mathbf{t}_{2,n}$ that satisfies (2.12) and (2.13) will always be affine equivariant in the same sense as $\mathbf{t}_n$ of Method 1.

Condition (2.10) is only to avoid that the breakdown point of $\mathbf{t}_{2,n}$ depends on the structure of the $\mathbf{x}_i$. We will see that with this condition, $\mathbf{t}_{2,n}$ will inherit the breakdown point $\lceil nr_1 \rceil / n$ of $\boldsymbol{\theta}_{1,n}$. All other results can be obtained under weaker conditions on $\rho_1(\cdot)$, which allow the MVE estimator as initial S-estimator. Condition (2.11) is not restrictive as we can always multiply (2.4) by a nonzero constant without changing the S-estimator. Finally, (2.13) is to guarantee consistency of $\mathbf{t}_{2,n}$.

We will show that when the underlying distribution P is elliptical, the influence function as well as the limiting distribution of $\mathbf{t}_{2,n}$ are similar to that of the location M-estimator defined by minimizing (2.1) using ρ_2. In particular, when P is spherically symmetric around the origin, the influence function will be zero outside the ball $\{\|\mathbf{x}\| \leq c_2\}$.

EXAMPLE 2.2: One may choose $\rho_1(\cdot)$ and $\rho_2(\cdot)$ as follows. For $\rho_2(y)$ take the biweight function $\rho_B(y;c_2)$. Then for c_2 large, $\mathbf{t}_{2,n}$ will have good asymptotic efficiency relative to the sample mean (see Lopuhaä 1989). To obtain a high breakdown point, take $\rho_1(y) = (c_2/c_1)^2\rho_B(y;c_1)$ and choose $c_1 < c_2$ such that the breakdown point $\lceil nr_1 \rceil/n$ of the S-estimator is maximal.

3. Breakdown point. Let $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)$ be a collection of n points in $\mathbf{R}^p$ and let $\mathbf{t}_n(\mathbf{X})$ be some location estimator based upon $\mathbf{X}$. When $\mathbf{X}$ is contaminated, some of the points of $\mathbf{X}$ might be replaced by other points and one obtains a different corrupted collection of n points. Denote by $\mathbf{Y}_m$ such a corrupted collection, where m points of $\mathbf{X}$ have been replaced by arbitrary values. The location estimator based upon $\mathbf{Y}_m$ will generally differ from $\mathbf{t}_n(\mathbf{X})$, and if at most m points are contaminated the difference is at most

$$\sup_{Y_m} \|\mathbf{t}_n(\mathbf{X}) - \mathbf{t}_n(\mathbf{Y}_m)\| \tag{3.1}$$

where the supremum is taken over all possible corrupted collections $\mathbf{Y}_m$. When (3.1) becomes infinite, it means that $\|\mathbf{t}_n(\mathbf{Y}_m)\|$ can be made arbitrarily large by replacing m points of $\mathbf{X}$. In that case $\mathbf{t}_n(\mathbf{Y}_m)$ will no longer give us any information about $\mathbf{X}$, and we say that $\mathbf{t}_n$ breaks down. The *breakdown point* of a location estimator $\mathbf{t}_n$ at a collection $\mathbf{X}$ is now defined as the smallest fraction m/n for which breakdown occurs :

$$\varepsilon^*(\mathbf{t}_n, \mathbf{X}) = \min_{1 \le m \le n} \left\{ \frac{m}{n} : \sup_{Y_m} \|\mathbf{t}_n(\mathbf{X}) - \mathbf{t}_n(\mathbf{Y}_m)\| = \infty \right\}$$

(see Donoho and Huber 1983 for a discussion). Similarly, the breakdown point of a covariance estimator $\mathbf{C}_n$ at a collection $\mathbf{X}$ is defined as the smallest fraction m/n of outliers that can either take the largest eigenvalue $\lambda_1(\mathbf{C}_n)$ over all bounds, or take the smallest eigenvalue $\lambda_p(\mathbf{C}_n)$ arbitrarily close to zero :

$$\varepsilon^*(\mathbf{C}_n, \mathbf{X}) = \min_{1 \le m \le n} \left\{ \frac{m}{n} : \sup_{Y_m} D\big((\mathbf{C}_n(\mathbf{X}), \mathbf{C}_n(\mathbf{Y}_m)\big) = \infty \right\}$$

where $D(\mathbf{A}, \mathbf{B}) = \max\{|\lambda_1(\mathbf{A}) - \lambda_1(\mathbf{B})|, |\lambda_p(\mathbf{A})^{-1} - \lambda_p(\mathbf{B})^{-1}|\}$. We restrict our attention to affine equivariant covariance estimators $\mathbf{C}_n$. When the collection $\mathbf{X}$ is in *general position*, i.e. no $p+1$ points are contained in some hyperplane of dimension less than $p-1$, then the breakdown point of affine equivariant $\mathbf{C}_n$ is at most $\lfloor \frac{n-p+1}{2} \rfloor / n$ (Davies 1987).

The breakdown behaviour of a location M-estimator depends on whether one uses a bounded or unbounded function ρ in (2.1). Similar to Huber (1984), it is easy to show that under the restrictions imposed on ρ in Method 1, the breakdown point of a location M-estimator is independent of $\mathbf{X}$ and attains the maximal value possible for translation equivariant location estimators : $\varepsilon_n^* = \lfloor \frac{n+1}{2} \rfloor / n$. This property suggests that if we estimate the covariance structure with $\mathbf{C}_n = \mathbf{A}_n\mathbf{A}_n^T$, the resulting location M-estimator based on scaled observations $\tilde{\mathbf{x}}_i = \mathbf{A}_n^{-1}\mathbf{x}_i$, will have a breakdown point that is at least equal to $\varepsilon^*(\mathbf{C}_n, \mathbf{X})$.

THEOREM 3.1. *Let* $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)$ *be a collection of* n *points in* $\mathbf{R}^p$. *Let* $\mathbf{C}_n(\mathbf{X}) \in PDS(p)$ *be an affine equivariant covariance estimator based upon* $\mathbf{X}$, *and let* $\mathbf{t}_n(\mathbf{X})$ *be defined as in Method 1. Then*

$$\varepsilon^*(\mathbf{t}_n, \mathbf{X}) \geq \varepsilon^*(\mathbf{C}_n, \mathbf{X}). \tag{3.2}$$

The proof is an adjustment of the proof of Theorem 4.1 in Huber(1984). To prove (3.2) we first need two lemmas. Define

$$M(\mathbf{t}) = \sup_{\mathbf{x}} \Big| \rho(\|\mathbf{x} + \mathbf{t}\|) - \rho(\|\mathbf{x}\|) \Big|. \tag{3.3}$$

LEMMA 3.1. *The difference* $\eta(\mathbf{t}) = M(\mathbf{t}) - \rho(\|\mathbf{t}\|)$ *is bounded :* $0 \leq \eta(\mathbf{t}) \leq y_0 \psi(y_0)$.

PROOF: By symmetry, we may rotate the vectors $\mathbf{x}$ and $\mathbf{x}+\mathbf{t}$ in (3.3) and consider them as multiples of the vector $\mathbf{t}$. Since ρ is increasing on $[0, \infty)$ we may then write $M(\mathbf{t}) = \sup_{\alpha \geq 0} \{\rho((1+\alpha)\|\mathbf{t}\|) - \rho(\alpha\|\mathbf{t}\|)\}$. Clearly $\eta(\mathbf{t}) \geq 0$. Take $\mathbf{t} \neq \mathbf{0}$ fixed and let $g(\alpha) = \rho((1+\alpha)\|\mathbf{t}\|) - \rho(\alpha\|\mathbf{t}\|)$. We have that

$$g(\alpha) = \int_{\alpha\|\mathbf{t}\|}^{(1+\alpha)\|\mathbf{t}\|} \psi(y)\, dy.$$

Since ψ is nonincreasing on (y_0, ∞), it follows that $g(y_0/\|\mathbf{t}\|) \geq g(\alpha)$, for $\alpha \geq y_0/\|\mathbf{t}\|$. As g is continuous it follows that it attains its maximum at some α^*, $0 \leq \alpha^* \leq y_0/\|\mathbf{t}\|$. By means of the mean value theorem we then find that (for some $0 < \gamma < 1$)

$$\eta(\mathbf{t}) = \rho((1+\alpha^*)\|\mathbf{t}\|) - \rho(\|\mathbf{t}\|) - \rho(\alpha^*\|\mathbf{t}\|) \leq \psi\big((1+\gamma\alpha^*)\|\mathbf{t}\|\big)\alpha^*\|\mathbf{t}\| \leq \psi(y_0)y_0.$$

□

LEMMA 3.2. *Let* $2m \leq n$ *and let* $\mathbf{C}_n(\mathbf{Y}_m) \in \mathrm{PDS}(p)$ *be a covariance estimator based upon a corrupted collection* $\mathbf{Y}_m$. *Suppose that there exist* k_1, k_2 *independent of* $\mathbf{Y}_m$ *such that* $0 < k_1^2 \leq \lambda_p\big(\mathbf{C}_n(\mathbf{Y}_m)\big) \leq \lambda_1\big(\mathbf{C}_n(\mathbf{Y}_m)\big) \leq k_2^2 < \infty$. *Let* $\Delta_{Y_m}(\mathbf{t})$ *be the difference*

$$\sum_{i=1}^{n} \Big\{ \rho\big[\{(\mathbf{y}_i - \mathbf{t})^T \mathbf{C}_n(\mathbf{Y}_m)^{-1}(\mathbf{y}_i - \mathbf{t})\}^{1/2}\big] - \rho\big[\{\mathbf{y}_i^T \mathbf{C}_n(\mathbf{Y}_m)^{-1}\mathbf{y}_i\}^{1/2}\big] \Big\}.$$

Then there exists a constant K *that is independent of* $\mathbf{Y}_m$, *such that*

$$\Delta_{Y_m}(\mathbf{t}) \geq (n-2m)\rho\left(\frac{\|\mathbf{t}\|}{k_2}\right) - K.$$

PROOF: The collections $\mathbf{X}$ and $\mathbf{Y}_m$ differ in at most m points. We may assume that $(\mathbf{y}_1, \ldots, \mathbf{y}_{n-m}) = (\mathbf{x}_1, \ldots, \mathbf{x}_{n-m})$. For $i = 1, 2, \ldots, n$ let $\mathbf{y}_i^* = \mathbf{C}_n(\mathbf{Y}_m)^{-1/2}\mathbf{y}_i$

and let $\mathbf{t}^* = \mathbf{C}_n(\mathbf{Y}_m)^{-1/2}\mathbf{t}$. Then $\Delta_{Y_m}(\mathbf{t}) = \sum_{i=1}^{n}\{\rho(\|\mathbf{y}_i^* - \mathbf{t}^*\|) - \rho(\|\mathbf{y}_i^*\|)\}$, and $k_2^{-1}\|\mathbf{y}_i\| \le \|\mathbf{y}_i^*\| \le k_1^{-1}\|\mathbf{y}_i\|$ and $k_2^{-1}\|\mathbf{t}\| \le \|\mathbf{t}^*\| \le k_1^{-1}\|\mathbf{t}\|$. For the summation over $i = 1, \ldots, n-m$, we have that

$$\begin{aligned}\Delta_{X_{n-m}}(\mathbf{t}) &= \sum_{i=1}^{n-m}\{\rho(\|\mathbf{y}_i^* - \mathbf{t}^*\|) - \rho(\|\mathbf{y}_i^*\|)\} \\ &= (n-m)\rho(\|\mathbf{t}^*\|) + \sum_{i=1}^{n-m}\{\rho(\|\mathbf{y}_i^* - \mathbf{t}^*\|) - \rho(\|\mathbf{t}^*\|)\} - \sum_{i=1}^{n-m}\rho(\|\mathbf{y}_i^*\|).\end{aligned}$$

According to the mean value theorem

$$\left|\rho(\|\mathbf{y}_i^* - \mathbf{t}^*\|) - \rho(\|\mathbf{t}^*\|)\right| \le \psi(\|\mathbf{t}^* - \gamma_i\mathbf{y}_i^*\|)\,\|\mathbf{y}_i^*\| \le \psi(y_0)k_1^{-1}\|\mathbf{x}_i\|$$

for $i = 1, \ldots, n-m$ and $0 < \gamma_i < 1$, which implies that

$$\Delta_{X_{n-m}}(\mathbf{t}) \ge (n-m)\rho(\|\mathbf{t}^*\|) - \frac{\psi(y_0)}{k_1}\sum_{i=1}^{n-m}\|\mathbf{x}_i\| - \sum_{i=1}^{n-m}\rho\left(\frac{\|\mathbf{x}_i\|}{k_1}\right). \tag{3.4}$$

For the remaining summation over $i = n-m+1, \ldots, n$, we find that

$$\begin{aligned}\left|\Delta_{Y_m}(\mathbf{t}) - \Delta_{X_{n-m}}(\mathbf{t})\right| &\le \sum_{i=n-m+1}^{n}\left|\rho(\|\mathbf{y}_i^* - \mathbf{t}^*\|) - \rho(\|\mathbf{y}_i^*\|)\right| \\ &\le mM(\mathbf{t}^*) \\ &= m\rho(\|\mathbf{t}^*\|) + m\eta(\mathbf{t}^*).\end{aligned}$$

Then together with (3.4) and Lemma 3.1 we conclude that

$$\Delta_{Y_m}(\mathbf{t}) \ge (n-2m)\rho(\|\mathbf{t}^*\|) - K \ge (n-2m)\rho\left(\frac{\|\mathbf{t}\|}{k_2}\right) - K$$

where $K = k_1^{-1}\psi(y_0)\sum_{i=1}^{n}\|\mathbf{x}_i\| + \sum_{i=1}^{n}\rho(\|\mathbf{x}_i\|/k_1) + m\,y_0\psi(y_0)$. □

PROOF OF THEOREM 3.1: If we replace at most $m \le n\varepsilon^*(\mathbf{C}_n, \mathbf{X}) - 1$ points then $\mathbf{C}_n$ does not break down and we must show that $\|\mathbf{t}_n(\mathbf{Y}_m)\|$ stays bounded. Since $\mathbf{C}_n$ is affine equivariant, $\varepsilon^*(\mathbf{C}_n, \mathbf{X})$ is at most $\lfloor\frac{n-p+1}{2}\rfloor/n$. Hence, $2m \le n-1$ and according to Lemma 3.2 and (2.7), $\Delta_{Y_m}(\mathbf{t})$ is bounded away from 0 for $\|\mathbf{t}\|$ sufficiently large uniformly in $\mathbf{Y}_m$. Because $\Delta_{Y_m}(\mathbf{0}) = 0$ and since $\mathbf{t}_n(\mathbf{Y}_m)$ minimizes $\Delta_{Y_m}(\mathbf{t})$, $\mathbf{t}_n(\mathbf{Y}_m)$ must be within a bounded neighbourhood of $\mathbf{0}$, uniformly in $\mathbf{Y}_m$. □

For the second proposal of Section 2.3 the breakdown behaviour is somewhat different, which has to do with the breakdown behaviour of location M-estimators defined by minimizing (2.1) using a function ρ that is constant outside an interval $[-c, c]$. The breakdown point then depends heavily on the actual structure of the collection $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$. If the width $2c$ of the function ρ is small compared to the distances between the $\mathbf{x}_i$, for instance if the $\mathbf{x}_i$ are at least $2c$ apart, then replacing only one point already forces breakdown of the location M-estimator. On the other hand, if the width $2c$ of ρ is large compared to the distances between the $\mathbf{x}_i$, for instance if all $\mathbf{x}_i$ are the same, then one needs to replace at least half of the observations to make the M-estimator break down.

Condition (2.10) is to guarantee that one is in the second situation, after scaling with $\mathbf{C}_{1,n}$. For $\mathbf{a} \in \mathbf{R}^p$, $\mathbf{A} \in \mathrm{PDS}(p)$ and $k > 0$, define the ellipsoid

$$E(\mathbf{a}, \mathbf{A}, k) = \{\mathbf{x} : (\mathbf{x} - \mathbf{a})^T \mathbf{A}^{-1} (\mathbf{x} - \mathbf{a}) \leq k^2\}.$$

Conditions (2.10) and (2.13), together with constraint (2.4), imply $R_{2,n}(\mathbf{t}_{2,n}) \leq b_1$. This forces at least $n - \lfloor n r_1 \rfloor$ observations (where $r_1 = b_1/\rho_1(c_1)$) inside the ellipsoid $E(\mathbf{t}_{2,n}, \mathbf{C}_{1,n}, c_2)$, which means that at least $n - \lfloor n r_1 \rfloor$ points are more or less close to each other compared to the width of ρ_2.

THEOREM 3.2. *Let $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ be a collection of n points in $\mathbf{R}^p$ in general position. Let $\mathbf{t}_{2,n}$ be the estimator based on $\mathbf{X}$ as defined in Method 2, and let $\boldsymbol{\theta}_{1,n} = (\mathbf{t}_{1,n}, \mathbf{C}_{1,n})$ be the initial S-estimator. Suppose that $0 < b_1/\rho_1(c_1) < \frac{n-p}{2n}$, then*

$$\varepsilon^*(\mathbf{t}_{2,n}, \mathbf{X}) \geq \varepsilon^*(\boldsymbol{\theta}_{1,n}, \mathbf{X}). \tag{3.6}$$

PROOF: We may assume $\sup \rho_1 = \sup \rho_2 = 1$. Replace at most $m \leq n\varepsilon^*(\boldsymbol{\theta}_{1,n}, \mathbf{X}) - 1$ points, so that $\mathbf{C}_{1,n}$ does not break down. Denote by $\mathbf{Y}_m = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ a corrupted collection, then we must show that $\|\mathbf{t}_{2,n}(\mathbf{Y}_m)\|$ remains bounded. Choose $0 < \delta < (n-m)^{-1}$ and $k > 0$ such that $\rho_2(y) \geq (1-\delta)$ for $|y| \geq k$. Then $(n-m)(1-\delta) > n - n\varepsilon^*(\boldsymbol{\theta}_{1,n}, \mathbf{X})$.

When $\mathbf{t} \in \mathbf{R}^p$ is such that all $\mathbf{x}_i$ of $\mathbf{X}$ are outside the ellipsoid $E(\mathbf{t}, \mathbf{C}_{1,n}(\mathbf{Y}_m), k)$, then

$$\sum_{i=1}^{n} \rho_2 \left[\left\{ (\mathbf{y}_i - \mathbf{t})^T \mathbf{C}_{1,n}(\mathbf{Y}_m)^{-1} (\mathbf{y}_i - \mathbf{v}) \right\}^{1/2} \right] \geq (n-m)(1-\delta) > n\big(1 - \varepsilon^*(\boldsymbol{\theta}_{1,n}, \mathbf{X})\big). \tag{3.7}$$

On the other hand, because of (2.13),(2.10) and (2.4) we also have that

$$\sum_{i=1}^{n} \rho_2 \left[\left\{ \big(\mathbf{y}_i - \mathbf{t}_{2,n}(\mathbf{Y}_m)\big)^T \mathbf{C}_{1,n}(\mathbf{Y}_m)^{-1} \big(\mathbf{y}_i - \mathbf{t}_{2,n}(\mathbf{Y}_m)\big) \right\}^{1/2} \right] \leq n b_1. \tag{3.8}$$

Since $0 < b_1/\rho_1(c_1) < \frac{n-p}{2n}$ and $\sup \rho_1 = 1$, the S-estimator $\boldsymbol{\theta}_{1,n}$ has breakdown point $\varepsilon^*(\boldsymbol{\theta}_{1,n}, \mathbf{X}) = \lceil n b_1 \rceil / n$, so that $1 - \varepsilon^*(\boldsymbol{\theta}_{1,n}, \mathbf{X}) \geq 1 - b_1 \geq b_1$. Hence, from

(3.7) and (3.8) it follows that at least one $\mathbf{x}_i$ of $\mathbf{X}$ must be inside the ellpsoid $E(\mathbf{t}_{2,n}(\mathbf{Y}_m), \mathbf{C}_{1,n}(\mathbf{Y}_m), k)$. But then for this $\mathbf{x}_i$ it holds that

$$\|\mathbf{x}_i - \mathbf{t}_{2,n}(\mathbf{Y}_m)\|^2 \leq k^2/\lambda_p(\mathbf{C}_{1,n}(\mathbf{Y}_m)).$$

As $\mathbf{C}_{1,n}$ does not break down, $\lambda_p(\mathbf{C}_{1,n}(\mathbf{Y}_m))$ is bounded away from zero uniformly in $\mathbf{Y}_m$, so $\|\mathbf{t}_{2,n}(\mathbf{Y}_m)\|$ remains bounded, uniformly in $\mathbf{Y}_m$. □

REMARK 3.1: Whether inequalities (3.2) and (3.6) are sharp, will probably depend on the structure of the collection $\mathbf{X}$, although in general it seems very unlikely that one could mess up the $\mathbf{x}_i$ in such a way that the initial covariance estimator breaks down and simultaneously the final location estimator does not breakdown (see for instance Rousseeuw and Leroy 1988 p.270-273).

If one does not impose conditions (2.10) and (R4)-(R5), one is allowed to use the MVE estimator as initial S-estimator in Method 2. The breakdown behaviour of $\mathbf{t}_{2,n}$ will then differ from that of Theorem 3.2. Although $\mathbf{C}_{1,n}$ may not break down it may still be so small, that the width of ρ_2 is small compared to the distances between the scaled $\mathbf{C}_{1,n}^{-1/2}\mathbf{y}_i$, in which case $\|\mathbf{t}_{2,n}(\mathbf{Y}_m)\|$ may become infinitely large. If we relax (2.8) only a little and require

$$c_1 < c_2 \tag{3.9}$$

instead of (2.8), the MVE estimator is allowed as initial S-estimator in Method 2. Although $\varepsilon^*(\mathbf{t}_{2,n}, \mathbf{X})$ may be strictly smaller than $\varepsilon^*(\boldsymbol{\theta}_{1,n}, \mathbf{X})$, it has the following lower bound.

THEOREM 3.3. *Let $\rho_1 : \mathbf{R} \to [0, \infty)$ only satisfy conditions (R1)-(R3) and let $\rho_2 : \mathbf{R} \to [0, \infty)$ satisfy all conditions (R1)-(R5). Let $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ be a collection of n points in $\mathbf{R}^p$ in general position. Let $\mathbf{t}_{2,n}$ be the estimator based on $\mathbf{X}$ as defined in Method 2, where we suppose that (3.9) holds instead of (2.8). Let $\boldsymbol{\theta}_{1,n} = (\mathbf{t}_{1,n}, \mathbf{C}_{1,n})$ be the initial S-estimator defined with the function ρ_1 and constant $0 < b_1 < \sup \rho_1$. Let $m(c_1, c_2)$ be the largest integer satisfying*

$$m(c_1, c_2) < (n - \lfloor nr_1 \rfloor)\left(1 - \frac{\rho_2(c_1)}{\rho_2(c_2)}\right) \tag{3.10}$$

where $r_1 = b_1/\rho_1(c_1)$. Then

$$\varepsilon^*(\mathbf{t}_{2,n}, \mathbf{X}) \geq \min\left\{\varepsilon^*(\boldsymbol{\theta}_{1,n}, \mathbf{X}), \frac{m(c_1, c_2) + 1}{n}\right\}.$$

PROOF: We may assume $\sup \rho_1 = \sup \rho_2 = 1$. Because of (3.9), the number $m(c_1, c_2) \geq 0$ is well defined. Replace at most $m \leq \min\{n\varepsilon^*(\boldsymbol{\theta}_{1,n}, \mathbf{X}) - 1, m(c_1, c_2)\}$ points of $\mathbf{X}$. Then $\mathbf{C}_{1,n}$ does not break down. Denote by $\mathbf{Y}_m = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ a corrupted collection, then we must show that $\|\mathbf{t}_{2,n}(\mathbf{Y}_m)\|$ remains bounded.

When $\mathbf{t} \in \mathbf{R}^p$ is such that all $\mathbf{x}_i \in \mathbf{X}$ are outside the ellipsoid $E(\mathbf{t}, \mathbf{C}_{1,n}(\mathbf{Y}_m), c_2)$, then

$$\sum_{i=1}^{n} \rho_2 \left[\{(\mathbf{y}_i - \mathbf{t})^T \mathbf{C}_{1,n}(\mathbf{Y}_m)^{-1} (\mathbf{y}_i - \mathbf{t})\}^{1/2} \right] \geq n - m.$$

By definition of $\boldsymbol{\theta}_{1,n}(\mathbf{Y}_m)$ at least $n - \lfloor nb_1 \rfloor$ points of $\mathbf{Y}_m$ must be inside the ellipsoid $E(\mathbf{t}_{1,n}(\mathbf{Y}_m), \mathbf{C}_{1,n}(\mathbf{Y}_m), c_1)$. Together with (2.11) this implies that

$$\sum_{i=1}^{n} \rho_2 \left[\{(\mathbf{y}_i - \mathbf{t}_{2,n}(\mathbf{Y}_m))^T \mathbf{C}_{1,n}(\mathbf{Y}_m)^{-1} (\mathbf{y}_i - \mathbf{t}_{2,n}(\mathbf{Y}_m))\}^{1/2} \right]$$
$$\leq (n - \lfloor nb_1 \rfloor) \rho_2(c_1) + \lfloor nb_1 \rfloor$$

which according (3.10) is strictly smaller than $n - m$. Then by a similar argument as in the proof of Theorem 3.2 we see that at least one $\mathbf{x}_i \in \mathbf{X}$ must be inside the ellipsoid $E(\mathbf{t}_{2,n}(\mathbf{Y}_m), \mathbf{C}_{1,n}(\mathbf{Y}_m), c_2)$, and we conclude that $\|\mathbf{t}_{2,n}(\mathbf{Y}_m)\|$ is bounded, uniformly in $\mathbf{Y}_m$. □

REMARK 3.2: The lower bound is sharp in the sense that it is possible to have such a configuration of $\mathbf{X}$ that replacing of $m(c_1, c_2) + 1$ points will lead to breakdown of $\mathbf{t}_{2,n}$ even if $\mathbf{C}_{1,n}$ does not.

4. Influence function. Robustness of an estimator $\mathbf{T}_n = \mathbf{T}(P_n)$ can also be measured by means of the influence function of the corresponding functional $\mathbf{T}(\cdot)$. Suppose that we pertubate a distribution P by putting a fraction ε of its probability mass at a point $\mathbf{x} \in \mathbf{R}^p$. The breakdown point may then be interpreted as to describe how large ε may be, before the estimator $\mathbf{T}_n$ becomes completely useless. This global concept may be complemented by the influence function as a local concept, which describes the stability of $\mathbf{T}(\cdot)$ under small pertubations at $\mathbf{x}$. If $\delta_{\mathbf{x}}$ denotes the probability measure that assigns mass 1 to $\mathbf{x} \in \mathbf{R}^p$, then the *influence function* of $\mathbf{T}(\cdot)$ at P is defined pointwise as

$$\text{IF}(\mathbf{x}; \mathbf{T}, P) = \lim_{\varepsilon \downarrow 0} \frac{\mathbf{T}((1-\varepsilon)P + \varepsilon\delta_{\mathbf{x}}) - \mathbf{T}(P)}{\varepsilon} \tag{4.1}$$

if this limit exists for every $\mathbf{x} \in \mathbf{R}^p$ (Hampel 1974). Typically, robust estimators will have an influence function that is bounded.

We will obtain the IF for the functionals $\mathbf{t}(\cdot)$ defined in Remark 2.1, and $\mathbf{t}_2(\cdot)$ defined in Remark 2.2. For the latter we only require (R1)-(R3) and (3.9) instead of (2.10) for the function ρ_1. We will need that the initial covariance functional $\mathbf{C}(\cdot)$ in (2.9) and the initial S-functional $\boldsymbol{\theta}_1(\cdot)$ in Remark 2.2 are continuous at P, i.e.

$$\mathbf{C}(P_k) \to \mathbf{C}(P) \tag{4.2}$$

$$\boldsymbol{\theta}_1(P_k) \to \boldsymbol{\theta}_1(P) \tag{4.3}$$

as $P_k \to P$ weakly. S-functionals $\boldsymbol{\theta}(P) = (\mathbf{t}(P), \mathbf{C}(P))$ defined with a function ρ that satisfies (R1)-(R3) satisfy properties (4.3) and (4.2) as long as $\boldsymbol{\theta}(P)$ is uniquely defined (Lopuhaä 1989). When we use such S-functionals for our initial $\mathbf{C}(\cdot)$ or $\boldsymbol{\theta}_1(\cdot)$, then at least at elliptical P, $\mathbf{t}(P)$ and $\mathbf{t}_2(P)$ will be uniquely defined, as will follow from the next lemma.

LEMMA 4.1. *Let P be an elliptical distribution with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, and suppose that f in (1.1) is decreasing. Let $\rho : \mathbf{R} \to [0,\infty)$ be a symmetric function with $\rho(0) = 0$ and $\rho(\infty) > 0$, which is nondecreasing on $[0,\infty)$. Let $\mathbf{C} : \mathcal{P} \to \mathrm{PDS}(p)$ be an affine equivariant Fisher-consistent functional :*

$$\mathbf{C}(P) = \boldsymbol{\Sigma} \tag{4.4}$$

and define $R_P(\mathbf{t})$ as in (2.9). If $R_P(\mathbf{t})$ is well defined for all $\mathbf{t} \in \mathbf{R}^p$ then the value $\mathbf{t}(P)$ that minimizes $R_P(\mathbf{t})$ is unique : $\mathbf{t}(P) = \boldsymbol{\mu}$.

PROOF: Because $\mathbf{C}(\cdot)$ is affine equivariant, we may assume that $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\mathbf{0}, \mathbf{I})$. For $s \geq 0$ define $\rho^{-1}(s) = \inf\{y \geq 0 : \rho(y) \geq s\}$ and for $r > 0$ let $B(\mathbf{t}, r) = \{\mathbf{x} : \|\mathbf{x} - \mathbf{t}\| \leq r\}$. Let $\mathbf{t} \neq \mathbf{0}$, then by Fubini's theorem we have that

$$\begin{aligned} R_P(\mathbf{t}) &= \iint \{0 \leq s \leq \rho(\|\mathbf{x} - \mathbf{t}\|)\} f(\|\mathbf{x}\|)\, ds\, d\mathbf{x} \\ &= \int \int_{B\left(\mathbf{t}, \rho^{-1}(s)\right)^c} f(\|\mathbf{x}\|)\, d\mathbf{x}\, ds. \end{aligned} \tag{4.5}$$

For $\mathbf{A} \subset \mathbf{R}^p$ and $\mathbf{t} \in \mathbf{R}^p$, write $\mathbf{A} + \mathbf{t} = \{\mathbf{x} + \mathbf{t} : \mathbf{x} \in \mathbf{A}\}$. Then, for every $r > 0$ it follows from Anderson's theorem (see for instance Tong 1980) that

$$\int_{B(0,r)+\mathbf{t}} f(\|\mathbf{x}\|)\, d\mathbf{x} \leq \int_{B(0,r)} f(\|\mathbf{x}\|)\, d\mathbf{x} \tag{4.6}$$

with equality if and only if $\left[(B(\mathbf{0}, r)+\mathbf{t}) \cap D_u\right] = \left[(B(\mathbf{0}, r) \cap D_u) + \mathbf{t}\right]$, for every level set $D_u = \{\mathbf{x} : f(\|\mathbf{x}\|) \geq u\}$, $u \geq 0$. Since f is decreasing, for every $r > 0$ we can find a $u > 0$ such that $D_u = B(\mathbf{0}, r)$. As $\mathbf{t} \neq \mathbf{0}$, it follows that $\left[(B(\mathbf{0}, r)+\mathbf{t}) \cap B(\mathbf{0}, r)\right] \neq B(\mathbf{0}, r) + \mathbf{t}$. We conclude that inequality (4.6) is strict for $r > 0$, hence from (4.5) we have that $R_P(\mathbf{t}) > R_P(\mathbf{0})$. □

Lemma 4.1 applies to the function $\rho(\cdot)$ in (2.9). Hence, if the initial covariance functional satisfies (4.4), $\mathbf{t}(\cdot)$ will be uniquely defined at elliptical P. Lemma 4.1 also applies to the function $\rho_2(\cdot)$ in (2.16). According to Section 2.2, at elliptical distributions the S-functional satisfies $\boldsymbol{\theta}_1(P_{\mu,\Sigma}) = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Because $R_{2,P}\left(\mathbf{t}_2(P)\right) \leq R_{2,P}\left(\mathbf{t}_1(P)\right)$, it follows that $\mathbf{t}_2(P) = \mathbf{t}_1(P)$ is uniquely defined at elliptical P. The functionals $\mathbf{t}(\cdot)$ and $\mathbf{t}_2(\cdot)$ may not be uniquely defined at $(1-\varepsilon)P + \varepsilon\delta_{\mathbf{x}}$. However, the next two lemmas show that every possible sequence of values for $\mathbf{t}\left((1-\varepsilon)P+\varepsilon\delta_{\mathbf{x}}\right)$ or $\mathbf{t}_2\left((1-\varepsilon)P + \varepsilon\delta_{\mathbf{x}}\right)$, will converge to $\mathbf{t}(P)$ or $\mathbf{t}_2(P)$ respectively, as $\varepsilon \downarrow 0$.

For the sake of brevity, we will sometimes write

$$Pg(\cdot) = \int g(\mathbf{y})\, dP(\mathbf{y})$$

or just simply Pg.

LEMMA 4.2. *Let $\rho : \mathbf{R} \to [0,\infty)$ satisfy the conditions imposed in Method 1 and suppose that $\mathbf{E}_P\|X\| < \infty$. Let $\mathbf{C} : \mathcal{P} \to \mathrm{PDS}(p)$ be an affine equivariant covariance functional that satisfies (4.2) for the sequence $\{(1-\varepsilon)P + \varepsilon\delta_{\mathbf{x}}\}$ as $\varepsilon \downarrow 0$. Consider the location functional $\mathbf{t}(\cdot)$ defined in Remark 2.1 and suppose that the value $\mathbf{t}(P)$ that minimizes the function $R_P(\mathbf{t})$ of (2.9) is unique. Let $\mathbf{x} \in \mathbf{R}^p$, then*

$$\lim_{\varepsilon \downarrow 0} \mathbf{t}\big((1-\varepsilon)P + \varepsilon\delta_{\mathbf{x}}\big) = \mathbf{t}(P).$$

PROOF: For $\boldsymbol{\theta} = (\mathbf{t}, \mathbf{C}) \in \Theta$, write $h(\mathbf{x}; \boldsymbol{\theta}) = \rho[\{(\mathbf{x}-\mathbf{t})^T\mathbf{C}^{-1}(\mathbf{x}-\mathbf{t})\}^{1/2}]$ and let $H(\boldsymbol{\theta}) = Ph(\cdot;\boldsymbol{\theta})$. Let $P_{\varepsilon,\mathbf{x}} = (1-\varepsilon)P + \varepsilon\delta_{\mathbf{x}}$ and let $H_{\varepsilon,\mathbf{x}}(\boldsymbol{\theta}) = P_{\varepsilon,\mathbf{x}}h(\cdot;\boldsymbol{\theta}) = (1-\varepsilon)H(\boldsymbol{\theta}) + \varepsilon h(\mathbf{x},\boldsymbol{\theta})$. Because $\mathbf{C}(\cdot)$ is affine equivariant, we can restrict ourselves to $\big(\mathbf{t}(P), \mathbf{C}(P)\big) = (\mathbf{0}, \mathbf{I})$. Then, it follows from (4.2) that $\mathbf{C}(P_{\varepsilon,\mathbf{x}}) \to \mathbf{I}$, so that all eigenvalues of $\mathbf{C}(P_{\varepsilon,\mathbf{x}})$ are between 1/4 and 4, say, for ε sufficiently small.

We first show that $\mathbf{t}(P_{\varepsilon,\mathbf{x}})$ eventually stays inside a fixed bounded set. According to the mean value theorem, for every $M \geq 0$ we have that

$$\mathbf{E}_P\rho(2\|X\| + 2M) \leq 2(\sup\psi)(\mathbf{E}_P\|X\| + M) < \infty. \tag{4.7}$$

Let $M > 0$ be such that

$$\rho(M/4)P(\|X\| \leq M/2) > 2\mathbf{E}_P\rho(2\|X\|). \tag{4.8}$$

When $\mathbf{t} \in \mathbf{R}^p$ is such that $\|\mathbf{t}\| > M$, then for ε sufficiently small we have that

$$\begin{aligned} R_{\varepsilon,\mathbf{x}}(\mathbf{t}) &= P_{\varepsilon,\mathbf{x}}h\big(\cdot, \mathbf{t}, \mathbf{C}(P_{\varepsilon,\mathbf{x}})\big) \\ &\geq \int \rho(\|\mathbf{y}-\mathbf{t}\|/2)\, dP_{\varepsilon,\mathbf{x}}(\mathbf{y}) \\ &\geq \rho(M/4)P_{\varepsilon,\mathbf{x}}(\|X\| \leq M/2) \longrightarrow \rho(M/4)P(\|X\| \leq M/2) \end{aligned} \tag{4.9}$$

as $\varepsilon \downarrow 0$, where we use that for symmetric matrices $\mathbf{A}$, it holds that

$$\lambda_p(\mathbf{A}) \leq \frac{\mathbf{v}^T\mathbf{A}\mathbf{v}}{\|\mathbf{v}\|^2} \leq \lambda_1(\mathbf{A}). \tag{4.10}$$

On the other hand, $R_{\varepsilon,\mathbf{x}}(\mathbf{0}) \leq \int \rho(2\|\mathbf{y}\|)\, dP_{\varepsilon,\mathbf{x}}(\mathbf{y})$, which tends to $\mathbf{E}_P\rho(2\|X\|)$ as $\varepsilon \downarrow 0$. Because $\mathbf{t}(P_{\varepsilon,\mathbf{x}})$ minimizes $R_{\varepsilon,\mathbf{x}}(\mathbf{t})$ we conclude from (4.8) and (4.9) that eventually $\|\mathbf{t}(P_{\varepsilon,\mathbf{x}})\| \leq M$.

Define the set $T = \{\boldsymbol{\theta} : \|\mathbf{t}\| \leq M,\ 1/4 \leq \lambda_j(\mathbf{C}) \leq 4,\ \text{for } j = 1, \ldots, p\}$, and consider the class of functions $\mathcal{F} = \{h(\cdot;\boldsymbol{\theta}) : \boldsymbol{\theta} \in T\}$. Then for every $h(\cdot,\boldsymbol{\theta})$ in $\mathcal{F}$, it holds that

$$h(\mathbf{x},\boldsymbol{\theta}) \leq \rho(2\|\mathbf{x}\| + 2M)$$

where we use (4.10) and the fact that $\rho(y)$ is increasing in $|y|$. According to (4.7), $\mathbf{E}_P\rho(2\|X\| + 2M)$ is finite, so that

$$\begin{aligned} \gamma_{\varepsilon,\mathbf{x}} &= \sup_{\boldsymbol{\theta}\in T} \big|H_{\varepsilon,\mathbf{x}}(\boldsymbol{\theta}) - H(\boldsymbol{\theta})\big| \\ &\leq \varepsilon\,\mathbf{E}_P\rho(2\|X\| + 2M) + \varepsilon\,\rho(2\|\mathbf{x}\| + 2M) \longrightarrow 0 \end{aligned} \tag{4.11}$$

as $\varepsilon \downarrow 0$. Because $R_P(\mathbf{t})$ has a unique minimum at $\mathbf{t}(P) = \mathbf{0}$, for all $\delta > 0$ there exist $\alpha > 0$ and $\beta > 0$ such that

$$\inf_{\|\mathbf{t}\|>\delta} H(\mathbf{t}, (1+\alpha)\mathbf{I}) > H(\mathbf{0}, (1-\alpha)\mathbf{I}) + \beta. \tag{4.12}$$

Let ε be sufficiently small such that $2\gamma_{\varepsilon,\mathbf{x}} \leq \beta$ and such that

$$1 - \alpha \leq \lambda_p\big(\mathbf{C}(P_{\varepsilon,\mathbf{x}})\big) \leq \lambda_1\big(\mathbf{C}(P_{\varepsilon,\mathbf{x}})\big) \leq 1 + \alpha.$$

Then with (4.11) and (4.12), for ε sufficiently small we have that

$$\inf_{\|\mathbf{t}\|>\delta} H_{\varepsilon,\mathbf{x}}\big(\mathbf{t}, \mathbf{C}(P_{\varepsilon,\mathbf{x}})\big) > H_{\varepsilon,\mathbf{x}}\big((\mathbf{0}, \mathbf{C}(P_{\varepsilon,\mathbf{x}})\big). \tag{4.13}$$

Since $\mathbf{t}(P_{\varepsilon,\mathbf{x}})$ minimizes $H_{\varepsilon,\mathbf{x}}\big(\mathbf{t}, \mathbf{C}(P_{\varepsilon,\mathbf{x}})\big)$, it follows that $\|\mathbf{t}(P_{\varepsilon,\mathbf{x}})\| \leq \delta$. We conclude that $\mathbf{t}(P_{\varepsilon,\mathbf{x}}) \to \mathbf{0}$. □

LEMMA 4.3. *Let $\rho_j : \mathbf{R} \to [0,\infty)$, $j = 1, 2$, satisfy the conditions of Theorem 3.3 and suppose that (3.9) holds instead of (2.10). Let $\boldsymbol{\theta}_1 : \mathcal{P} \to \Theta$ be the S-functional defined with the function $\rho_1(\cdot)$ and the constant $0 < b_1 < \sup \rho_1$, and suppose that $\boldsymbol{\theta}_1(\cdot)$ satisfies (4.3) for the sequence $\{(1-\varepsilon)P + \varepsilon\delta_{\mathbf{x}}\}$ as $\varepsilon \downarrow 0$. Consider the location functional $\mathbf{t}_2(\cdot)$ defined in Remark 2.2 and suppose that the function $R_{2,P}(\mathbf{t})$ of (2.16) has a unique minimum at $\mathbf{t}_1(P)$. Then $\mathbf{t}_2(P) = \mathbf{t}_1(P)$, and for $\mathbf{x} \in \mathbf{R}^p$ it holds that*

$$\lim_{\varepsilon \downarrow 0} \mathbf{t}_2\big((1-\varepsilon)P + \varepsilon\delta_{\mathbf{x}}\big) = \mathbf{t}_2(P).$$

PROOF: Because $R_{2,P}(\mathbf{t})$ has a unique minimum at $\mathbf{t}_1(P)$, and since $\mathbf{t}_2(P)$ must satisfy (2.15), it follows that $\mathbf{t}_2(P) = \mathbf{t}_1(P)$. Because the S-functional is affine equivariant, we may restrict to $\mathbf{t}_2(P) = \mathbf{t}_1(P) = \mathbf{0}$ and $\mathbf{C}_1(P) = \mathbf{I}$. As we can always rescale the functions ρ_1 and ρ_2, we may assume that $\sup \rho_1 = \sup \rho_2 = 1$.

First we show that $\mathbf{t}_2(P_{\varepsilon,\mathbf{x}})$ eventually stays inside a fixed bounded set. Consider a generic distribution Q and the corresponding ellipsoid $E_1 = E\big(\mathbf{t}_1(Q), \mathbf{C}_1(Q), c_1\big)$. By definition of $\boldsymbol{\theta}_1(\cdot)$, it holds that the Q-measure of this set is at least $1 - b_1$. For $\boldsymbol{\theta} = (\mathbf{t}, \mathbf{C})$, write $h_2(\mathbf{x}, \boldsymbol{\theta}) = \rho_2\big[\{(\mathbf{x}-\mathbf{t})^T\mathbf{C}^{-1}(\mathbf{x}-\mathbf{t})\}^{1/2}\big]$. We have that

$$\begin{aligned} R_{2,Q}\big(\mathbf{t}_2(Q)\big) &\leq R_{2,Q}\big(\mathbf{t}_1(Q)\big) \\ &= \int_{E_1} h_2\big(\mathbf{y}, \boldsymbol{\theta}_1(Q)\big)\, dQ(\mathbf{y}) + \int_{E_1^c} h_2\big(\mathbf{y}, \boldsymbol{\theta}_1(Q)\big)\, dQ(\mathbf{y}) \\ &\leq (1-b_1)\rho_2(c_1) + \big(Q(E_1) - 1 + b_1\big)\rho_2(c_1) + 1 - Q(E_1) \end{aligned}$$

and since $Q(E_1) \geq 1 - b_1$, it follows that $R_{2,Q}\big(\mathbf{t}_2(Q)\big) \leq (1-b_1)\rho_2(c_1) + b_1$. Because $\sup \rho_2 = 1$, we find that

$$Q\big(E(\mathbf{t}_2(Q), \mathbf{C}_1(Q), c_2)\big) \geq 1 - R_{2,Q}\big(\mathbf{t}_2(Q)\big) \geq \delta \tag{4.14}$$

where $\delta = (1-b_1)\big(1-\rho_2(c_1)\big)$.

The inequality in (4.14) holds in particular for $Q = P_{\varepsilon,\mathbf{x}} = (1-\varepsilon)P + \varepsilon\delta_{\mathbf{x}}$. We also have that

$$\sup_{\mathbf{t},\mathbf{C}} \left| P_{\varepsilon,\mathbf{x}}\big(E(\mathbf{t},\mathbf{C},c_2)\big) - P\big(E(\mathbf{t},\mathbf{C},c_2)\big) \right| \leq 2\varepsilon \longrightarrow 0 \tag{4.15}$$

as $\varepsilon \downarrow 0$. Together with inequality (4.14), for $Q = P_{\varepsilon,\mathbf{x}}$ it follows that eventually

$$P\big(E(\mathbf{t}_2(P_{\varepsilon,\mathbf{x}}), \mathbf{C}_1(P_{\varepsilon,\mathbf{x}}), c_2)\big) \geq \tfrac{1}{2}\delta. \tag{4.16}$$

Suppose that for ε sufficiently small, all eigenvalues of $\mathbf{C}_1(P_{\varepsilon,\mathbf{x}})$ are between $1/4$ and 4, say. Let $M > 0$ be such that $P(\|X\| > 2Mc_2) < \frac{1}{2}\delta$. When $\mathbf{t} \in \mathbf{R}^p$ is such that $\|\mathbf{t}\| > 2(1+M)c_2$, we would find that

$$P\big(E(\mathbf{t}, \mathbf{C}(P_{\varepsilon,\mathbf{x}}), c_2)\big) \leq P\big(\|X - \mathbf{t}\| \leq 2c_2\big) \leq P(\|X\| > 2Mc_2) < \tfrac{1}{2}\delta.$$

Hence, according to (4.16), it follows that $\|\mathbf{t}_2(P_{\varepsilon,\mathbf{x}})\| \leq 2(1+M)c_2$ for ε sufficiently small.

Consider a convergent subsequence $\{\mathbf{t}_2(P_{\varepsilon_k,\mathbf{x}})\}$ with

$$\lim_{k\to\infty} \mathbf{t}_2(P_{\varepsilon_k,\mathbf{x}}) = \mathbf{t}_L.$$

Continuity of ρ_2 implies that

$$\left| Ph_2\big(\cdot, \mathbf{t}_2(P_{\varepsilon_k,\mathbf{x}}), \mathbf{C}(P_{\varepsilon_k,\mathbf{x}})\big) - \mathbf{E}_P\rho_2(\|X - \mathbf{t}_L\|) \right| \to 0 \tag{4.17}$$

as $k \to \infty$. Since $|h_2(\mathbf{x};\boldsymbol{\theta})| \leq 1$, we have that

$$\sup_{\boldsymbol{\theta}} \left| P_{\varepsilon,\mathbf{x}} h_2(\cdot;\boldsymbol{\theta}) - Ph_2(\cdot;\boldsymbol{\theta}) \right| \leq 2\varepsilon \to 0 \tag{4.18}$$

as $\varepsilon \downarrow 0$. Together with (4.17) this means

$$P_{\varepsilon_k,\mathbf{x}} h_2\big(\cdot; \mathbf{t}_2(P_{\varepsilon_k,\mathbf{x}}), \mathbf{C}_1(P_{\varepsilon_k,\mathbf{x}})\big) \to \mathbf{E}_P\rho_2(\|X - \mathbf{t}_L\|).$$

Similarly, we also have that $P_{\varepsilon_k,\mathbf{x}} h_2\big(\cdot; \mathbf{t}_1(P_{\varepsilon_k,\mathbf{x}}), \mathbf{C}_1(P_{\varepsilon_k,\mathbf{x}})\big) \to \mathbf{E}_P\rho_2(\|X\|)$. Then by using (2.15), we find that

$$\begin{aligned} R_P(\mathbf{t}_L) &= \mathbf{E}_P\rho_2(\|X - \mathbf{t}_L\|) \\ &= \lim_{k\to\infty} P_{\varepsilon_k,\mathbf{x}} h_2\big(\cdot; \mathbf{t}_2(P_{\varepsilon_k,\mathbf{x}}), \mathbf{C}_1(P_{\varepsilon_k,\mathbf{x}})\big) \\ &\leq \lim_{k\to\infty} P_{\varepsilon_k,\mathbf{x}} h_2\big(\cdot; \mathbf{t}_1(P_{\varepsilon_k,\mathbf{x}}), \mathbf{C}_1(P_{\varepsilon_k,\mathbf{x}})\big) \\ &= \mathbf{E}_P\rho_2(\|X\|) \\ &= R_{2,P}(\mathbf{0}). \end{aligned}$$

Because $R_{2,P}(\mathbf{t})$ has a unique minimum at $\mathbf{t}_2(P) = \mathbf{0}$, it follows that $\mathbf{t}_L = \mathbf{0}$. Hence, every convergent subsequence $\{\mathbf{t}_2(P_{\varepsilon_k,\mathbf{x}})\}$ converges to $\mathbf{0}$. Since the whole sequence is eventually inside a compact set, we must have $\mathbf{t}_2(P_{\varepsilon_k,\mathbf{x}}) \to \mathbf{0}$. □

REMARK 4.1: Note that in the proofs of Lemmas 4.2 and 4.3, we have only used the expression $(1-\varepsilon)P + \varepsilon\delta_{\mathbf{x}}$ explicitly in (4.11), (4.15) and (4.18). This means that if, instead of (4.11), one has

$$\sup_{\theta\in T} \left| P_k h(\cdot,\theta) - Ph(\cdot,\theta) \right| \to 0 \tag{4.19}$$

for a sequence of distributions $\{P_k\}$ for which (4.2) holds, one may show along the lines of the proof of Lemma 4.2 that $\mathbf{t}(P_k) \to \mathbf{t}(P)$. Similarly, if instead of (4.15) and (4.18), one has

$$\begin{aligned} &\sup_{\mathbf{t},\mathbf{C}} \left| P_k\big(E(\mathbf{t},\mathbf{C},c_2)\big) - P\big(E(\mathbf{t},\mathbf{C},c_2)\big) \right| \to 0 \\ &\sup_{\theta} \left| P_k h_2(\cdot,\theta) - Ph_2(\cdot,\theta) \right| \to 0 \end{aligned} \tag{4.20}$$

for a sequence of distributions $\{P_k\}$ for which (4.3) holds, one may show along the lines of the proof of Lemma 4.3 that $\mathbf{t}_2(P_k) \to \mathbf{t}_2(P)$.

We may now obtain the expressions for $\mathrm{IF}(\mathbf{x};\mathbf{t},P)$ and $\mathrm{IF}(\mathbf{x};\mathbf{t}_2,P)$. Consider the function $\rho(\cdot)$ in (2.9). The derivative of $\rho[\{(\mathbf{y}-\mathbf{t})^T\mathbf{C}^{-1}(\mathbf{y}-\mathbf{t})\}^{1/2}]$ with respect to $\mathbf{t}$ is equal to

$$-u\left[\{(\mathbf{y}-\mathbf{t})^T\mathbf{C}^{-1}(\mathbf{y}-\mathbf{t})\}^{1/2}\right]\mathbf{C}^{-1}(\mathbf{y}-\mathbf{t}) \tag{4.21}$$

where $u(y) = \psi(y)/y$. Since ψ is bounded, also (4.21) is bounded as a function of $\mathbf{t}$. This means that for any P, the function $R_P(\mathbf{t})$ of (2.9) has derivative

$$-\mathbf{C}(P)^{-1}\int u\left[\{(\mathbf{y}-\mathbf{t})^T\mathbf{C}(P)^{-1}(\mathbf{y}-\mathbf{t})\}^{1/2}\right](\mathbf{y}-\mathbf{t})\,dP(\mathbf{y}).$$

We conclude that $\boldsymbol{\theta}(P) = \big(\mathbf{t}(P),\mathbf{C}(P)\big)$ will always be a zero of the function

$$G(\boldsymbol{\theta}) = Pg(\cdot;\boldsymbol{\theta}) \tag{4.22}$$

where for $\boldsymbol{\theta} = (\mathbf{t},\mathbf{C})$

$$g(\mathbf{x};\boldsymbol{\theta}) = u\left[\{(\mathbf{x}-\mathbf{t})^T\mathbf{C}^{-1}(\mathbf{x}-\mathbf{t})\}^{1/2}\right](\mathbf{x}-\mathbf{t}). \tag{4.23}$$

A vector $\mathbf{v}\in\mathbf{R}^p$ is called a *point of symmetry of* P, if

> $P(\mathbf{v}+A) = P(\mathbf{v}-A)$, for all P-measurable sets $A\subset\mathbf{R}^p$, where for $\lambda\in\mathbf{R}$ and $\mathbf{v}\in\mathbf{R}^p$, $\mathbf{v}+\lambda A$ denotes the set $\{\mathbf{v}+\lambda\mathbf{x} : \mathbf{x}\in A\}$.

When $\mathbf{t}(P)$ is uniquely defined, and when it is also a point of symmetry of P, then the expression for $\mathrm{IF}(\mathbf{x};\mathbf{t},P)$ will be independent of the influence function of the initial covariance functional $\mathbf{C}(\cdot)$, as long as (4.2) holds.

THEOREM 4.1. *Let $\rho : \mathbf{R} \to [0,\infty)$ satisfy the conditions of Method 1. Let $\mathbf{C} : \mathcal{P} \to \mathrm{PDS}(p)$ be an affine equivariant covariance functional that satisfies (4.2) for the sequence $\{(1-\varepsilon)P+\varepsilon\delta_{\mathbf{x}}\}$ as $\varepsilon \downarrow 0$. Let $\mathbf{t}(\cdot)$ be the location functional defined in Remark 2.1, and suppose that $\mathbf{t}(P)$ minimizes the function $R_P(\mathbf{t})$ of (2.9) uniquely. Suppose that $\mathbf{t}(P)$ is a point of symmetry of P, and assume that $\mathbf{E}_P\|X\| < \infty$. Define $G(\boldsymbol{\theta})$ by (4.22) and (4.23). Suppose that G has a partial derivative $\partial G/\partial \mathbf{t}$ that is continuous at $\boldsymbol{\theta}_0 = \big(\mathbf{t}(P), \mathbf{C}(P)\big)$, and suppose that $\boldsymbol{\Lambda} = (\partial G/\partial \mathbf{t})(\boldsymbol{\theta}_0)$ is nonsingular. Then for $\mathbf{x} \in \mathbf{R}^p$ it holds that*

$$\mathrm{IF}(\mathbf{x};\mathbf{t},P) = -\boldsymbol{\Lambda}^{-1}\, u\left[\left\{\big(\mathbf{x}-\mathbf{t}(P)\big)^T \mathbf{C}(P)^{-1}\big(\mathbf{x}-\mathbf{t}(P)\big)\right\}^{1/2}\right]\big(\mathbf{x}-\mathbf{t}(P)\big)$$

where $u(y) = \psi(y)/y$ and $\psi = \rho'$.

PROOF: Write $\boldsymbol{\theta}_0 = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Because $\partial G/\partial \mathbf{t}$ is continuous at $\boldsymbol{\theta}_0$ we have that

$$G(\mathbf{t},\mathbf{C}) = G(\boldsymbol{\mu},\mathbf{C}) + \frac{\partial G}{\partial \mathbf{t}}(\boldsymbol{\mu},\mathbf{C})\,(\mathbf{t}-\boldsymbol{\mu}) + (\mathbf{t}-\boldsymbol{\mu})\,r(\boldsymbol{\theta}) \tag{4.24}$$

where $r(\boldsymbol{\theta}) \to 0$ as $\boldsymbol{\theta} \to \boldsymbol{\theta}_0$. Furthermore, since $\boldsymbol{\mu}$ is a point of symmetry of P, it holds that

$$G(\boldsymbol{\mu},\mathbf{C}) = \mathbf{0} \tag{4.25}$$

for all nonsingular $\mathbf{C}$. Write $P_{\varepsilon,\mathbf{x}} = (1-\varepsilon)P + \varepsilon\delta_{\mathbf{x}}$, and with a slight abuse of notation write $\boldsymbol{\theta}_\varepsilon = (\mathbf{t}_\varepsilon, \mathbf{C}_\varepsilon) = \big(\mathbf{t}(P_{\varepsilon,\mathbf{x}}), \mathbf{C}(P_{\varepsilon,\mathbf{x}})\big)$. Since $\boldsymbol{\theta}_\varepsilon$ is a zero of the function $P_{\varepsilon,\mathbf{x}}g(\cdot,\boldsymbol{\theta})$, together with (4.24) and (4.25) it follows that

$$\begin{aligned} \mathbf{0} &= (1-\varepsilon)G(\mathbf{t}_\varepsilon,\mathbf{C}_\varepsilon) + \varepsilon g(\mathbf{x};\boldsymbol{\theta}_\varepsilon) \\ &= (1-\varepsilon)\left\{\frac{\partial G}{\partial \mathbf{t}}(\boldsymbol{\mu},\mathbf{C}_\varepsilon)\,(\mathbf{t}_\varepsilon-\boldsymbol{\mu}) + (\mathbf{t}_\varepsilon-\boldsymbol{\mu})r(\boldsymbol{\theta}_\varepsilon)\right\} + \varepsilon g(\mathbf{x};\boldsymbol{\theta}_\varepsilon). \end{aligned} \tag{4.26}$$

Because $\partial G/\partial \mathbf{t}$ and g are continuous at $\boldsymbol{\theta}_0$ we obtain

$$\mathbf{0} = (1-\varepsilon)\big(\boldsymbol{\Lambda} + o(1)\big)(\mathbf{t}_\varepsilon - \boldsymbol{\mu}) + O(\varepsilon) \tag{4.27}$$

as $\varepsilon \downarrow 0$. Because $\boldsymbol{\Lambda}$ is nonsingular, we conclude that $\mathbf{t}_\varepsilon - \boldsymbol{\mu} = O(\varepsilon)$, as $\varepsilon \downarrow 0$. When we insert this into (4.26) and use that g is continuous, (4.26) reduces to

$$\frac{\mathbf{t}(P_{\varepsilon,\mathbf{x}}) - \boldsymbol{\mu}}{\varepsilon} = -\boldsymbol{\Lambda}^{-1} g(\mathbf{x};\boldsymbol{\theta}_0) + o(1)$$

as $\varepsilon \downarrow 0$, which finishes the proof. □

Next consider the location functional $\mathbf{t}_2(\cdot)$ of Remark 2.2. By definition it holds that the pair $(\mathbf{t}_2(P), \mathbf{C}_1(P))$ is a zero of the function

$$G_2(\boldsymbol{\theta}) = Pg_2(\cdot;\boldsymbol{\theta}) \tag{4.28}$$

where for $\boldsymbol{\theta} = (\mathbf{t}, \mathbf{C})$

$$g_2(\mathbf{x};\boldsymbol{\theta}) = u_2\left[\{(\mathbf{x}-\mathbf{t})^T\mathbf{C}^{-1}(\mathbf{x}-\mathbf{t})\}^{1/2}\right](\mathbf{x}-\mathbf{t}). \tag{4.29}$$

When $\mathbf{t}_2(P)$ is uniquely defined and if it is a point of symmetry of P, then the expression of $\mathrm{IF}(\mathbf{x};\mathbf{t}_2,P)$ will be independent of the influence function of the initial S-functional $\boldsymbol{\theta}_1(\cdot)$ as long as (4.3) holds.

THEOREM 4.2. *Let $\rho_j : \mathbf{R} \to [0,\infty)$, $j = 1,2$, satisfy the conditions of Theorem 3.3 and suppose that (3.9) holds instead of (2.10). Let $\boldsymbol{\theta}_1 : \mathcal{P} \to \Theta$ be the S-functional defined with the function $\rho_1(\cdot)$ and the constant $0 < b_1 < \sup\rho_1$. Consider the location functional $\mathbf{t}_2(\cdot)$ defined in Remark 2.2, and suppose that the function $R_{2,P}(\mathbf{t})$ of (2.16) has a unique minimum at $\mathbf{t}_1(P)$, so that $\mathbf{t}_2(P) = \mathbf{t}_1(P)$. Suppose that $\mathbf{t}_2(P)$ is a point of symmetry of P. Define $G_2(\boldsymbol{\theta})$ by (4.28) and (4.29) and suppose that at $\boldsymbol{\theta}_0 = (\mathbf{t}_2(P), \mathbf{C}_1(P))$, the partial derivative $\boldsymbol{\Lambda}_2 = (\partial G_2/\partial\mathbf{t})(\boldsymbol{\theta}_0)$ is nonsingular. Then for $\mathbf{x} \in \mathbf{R}^p$ it holds that*

$$\mathrm{IF}(\mathbf{x};\mathbf{t}_2;P) = -\boldsymbol{\Lambda}_2^{-1}\, u_2\left[\left\{\left(\mathbf{x}-\mathbf{t}_2(P)\right)^T\mathbf{C}_1(P)^{-1}\left(\mathbf{x}-\mathbf{t}_2(P)\right)\right\}^{1/2}\right]\left(\mathbf{x}-\mathbf{t}_2(P)\right).$$

where $u_2(y) = \psi_2(y)/y$ and $\psi_2 = \rho_2'$.

PROOF: The proof is similar to that of Theorem 4.1. We only have to show that the conditions on ρ_2 imply that $\partial G_2/\partial\mathbf{t}$ is continuous at $\boldsymbol{\theta}_0$.

The function $g_2(\mathbf{x};\boldsymbol{\theta})$ of (4.23) has a partial derivative

$$\frac{\partial g_2}{\partial\mathbf{t}}(\mathbf{x};\mathbf{t},\mathbf{C}) = -\frac{u_2'(d)}{d}\mathbf{C}^{-1}(\mathbf{x}-\mathbf{t})(\mathbf{x}-\mathbf{t})^T - u_2(d)\,\mathbf{I} \tag{4.30}$$

where $d = \sqrt{(\mathbf{x}-\mathbf{t})^T\mathbf{C}^{-1}(\mathbf{x}-\mathbf{t})}$. Because ρ_2 is symmetric and differentiable, its derivative satisfies $\psi_2(0) = 0$. Since the second derivative ψ_2' is continuous, we see that $\lim_{y\downarrow 0} u_2(y) = \lim_{y\downarrow 0}\psi_2(y)/y$ exists. It follows that $u_2(\cdot)$ is continuous and because it is 0 outside $[-c,c]$, it must be bounded. The same holds for the function ψ_2'. We also have that $yu_2'(y) = \psi_2'(y) - u_2(y)$, and because $\|\mathbf{y}-\mathbf{t}\|/d \le \|\mathbf{C}\|$, it follows that (4.30) is bounded as a function of $\mathbf{t}$. We conclude that

$$\frac{\partial G_2}{\partial\mathbf{t}}(\boldsymbol{\theta}) = \int\frac{\partial g_2}{\partial\mathbf{t}}(\mathbf{x};\boldsymbol{\theta})\,dP(\mathbf{x}) \tag{4.31}$$

and that (4.31) is continuous. From now on the proof is the same as that of Theorem 4.1. □

When X has an elliptical distribution P with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, and when the function ψ of Theorem 4.1 has a continuous bounded derivative, it is not difficult to see that

$$\boldsymbol{\Lambda} = -\beta \mathbf{I} \tag{4.32}$$

where

$$\begin{aligned} \beta &= \mathrm{E}\left[\left(1-\frac{1}{p}\right)u(\|\mathbf{B}(X-\boldsymbol{\mu})\|) + \frac{1}{p}\psi'(\|\mathbf{B}(X-\boldsymbol{\mu})\|)\right] \\ &= \int\left[\left(1-\frac{1}{p}\right)u(\|\mathbf{x}\|) + \frac{1}{p}\psi'(\|\mathbf{x}\|)\right] f(\|\mathbf{x}\|)\, d\mathbf{x}. \end{aligned} \tag{4.33}$$

When ψ is not differentiable, such as for instance Huber's ψ-function ψ_{H}, the matrix $\boldsymbol{\Lambda}$ may still of type (4.32) under suitable conditions on the function f in (1.1). From conditions (R1)-(R5), it follows that $\boldsymbol{\Lambda}_2 = -\beta_2 \mathbf{I}$ with β_2 as in (4.33), with $\psi = \psi_2$.

Hence, we see from Theorems 4.1 and 4.2 that the influence functions of $\mathbf{t}(\cdot)$ and $\mathbf{t}_2(\cdot)$ are the same as that of the corresponding affine equivariant location M-estimator considered in Maronna (1976); in particular, when P is spherically symmetric, it is the same as that of the location M-estimator defined by minimizing (2.1) with the same function ρ or ρ_2 respectively. For $\mathbf{t}(\cdot)$ the influence function is *weakly* redescending, i.e. nondecreasing in $\|\mathbf{x}\|$ and nonzero for $\mathbf{x} \neq \mathbf{0}$. For $\mathbf{t}_2(\cdot)$ the influence function is *strongly* redescending, i.e. it is zero for $\|\mathbf{x}\| \geq c_2$.

5. Asymptotic normality. Let $X_1, X_2, \ldots$ be a sequence of independent identically distributed random vectors $X_i = (X_{i1} \cdots X_{ip})^T$ with a distribution P on $\mathbf{R}^p$. Denote by P_n the empirical distribution corresponding with the sample $X_1, \ldots, X_n$.

We first prove consistency for the location estimators proposed in Section 2.3. We will need that the initial covariance estimator $\mathbf{C}_n$ in (2.8) and the initial S-estimator $\boldsymbol{\theta}_{1,n}$ in Method 2 are consistent for the values of their corresponding functionals $\mathbf{C}(\cdot)$ and $\boldsymbol{\theta}(\cdot)$ at P, i.e.

$$\lim_{n\to\infty} \mathbf{C}_n = \mathbf{C}(P) \tag{5.1}$$

$$\lim_{n\to\infty} \boldsymbol{\theta}_{1,n} = \boldsymbol{\theta}_1(P) \tag{5.2}$$

with probability one. S-estimators $\boldsymbol{\theta}_n = (\mathbf{t}_n, \mathbf{C}_n)$, defined with a function ρ that satisfies (R1)-(R3), satisfy (5.1) and (5.2).

We will apply a uniform strong law for empirical processes $(P_n - P)\phi$, indexed by functions ϕ in a class $\mathcal{F}$, as is given in Pollard (1984). By the *envelope* F of $\mathcal{F}$ is meant a function F for which $|\phi| \leq F$ for every $\phi \in \mathcal{F}$. For further concepts involved, we refer to Pollard (1984).

THEOREM 5.1. *Let $\rho : \mathbf{R} \to [0, \infty)$ satisfy the conditions imposed in Method 1, and suppose that $\mathrm{E}\|X_1\| < \infty$. Let $\mathbf{C}_n$, $n = 1, 2, \ldots$, be a sequence of affine equivariant covariance estimators that satisfies (5.1). Define $R_n(\mathbf{t})$ and $R_P(\mathbf{t})$ as in (2.8) and (2.9) respectively. Suppose that the value $\mathbf{t}(P)$ that minimizes $R_P(\mathbf{t})$ is unique.*

Then for any sequence $\mathbf{t}_n$, $n = 1, 2, \ldots$, *where every* $\mathbf{t}_n$ *minimizes the function* $R_n(\mathbf{t})$, *it holds that*

$$\lim_{n\to\infty} \mathbf{t}_n = \mathbf{t}(P)$$

with probability one.

PROOF: According to Remark 4.1, it is sufficient to show that

$$\sup_{\theta\in T} \left| P_n h(\cdot,\theta) - Ph(\cdot,\theta) \right| \to 0 \tag{5.3}$$

with probability one, where T and $h(\mathbf{x},\theta)$ are defined in the proof of Lemma 4.2. Consider the class $\mathcal{F} = \{h(\cdot,\theta) : \theta \in T\}$. Then $\mathcal{F}$ is permissible in the sense of Pollard (1984, Appendix C). Moreover, according to the proof of Lemma 4.2, it has a integrable envelope $F(\mathbf{x}) = \rho(2\|\mathbf{x}\| + 2M)$, where M is defined in (4.8). Because $\rho(y)$ is monotone in $|y|$, it is not difficult to see (see for instance Lemma 22 in Nolan and Pollard 1987), that the class of graphs of functions in $\mathcal{F}$ has polynomial discrimination (Pollard 1984, p.17). According to Theorem II.24 and Lemma II.25 in Pollard (1984), (5.3) holds with probability one. □

THEOREM 5.2. *Let* $\rho_j : \mathbf{R} \to [0,\infty)$, $j = 1, 2$, *satisfy the conditions of Theorem 3.3 and suppose that (3.9) holds instead of (2.10) Let* $\theta_{1,n}$, $n = 1, 2, \ldots$, *be a sequence of S-estimators that are defined with the function* ρ_1 *and the constant* $0 < b_1 < \sup\rho_1$, *and suppose that* $\theta_{1,n}$ *satisfies (5.2). Let* $R_{2,P}(\mathbf{t})$ *be defined by (2.16) and suppose that it has a unique minimum at* $\mathbf{t}_1(P)$. *Then* $\mathbf{t}_2(P) = \mathbf{t}_1(P)$ *and for any sequence* $\mathbf{t}_{2,n}$, $n = 1, 2, \ldots$, *where every* $\mathbf{t}_{2,n}$ *satisfies (2.12) and (2.13), it holds that*

$$\lim_{n\to\infty} \mathbf{t}_{2,n} = \mathbf{t}_2(P)$$

with probability one.

PROOF: According to Remark 4.1, it sufficient to show that

$$\sup_{\mathbf{t},\mathbf{C}} \left| P_n\big(E(\mathbf{t},\mathbf{C},c_2)\big) - P\big(E(\mathbf{t},\mathbf{C},c_2)\big) \right| \to 0 \tag{5.4}$$

$$\sup_{\theta} \left| P_n h_2(\cdot,\theta) - Ph_2(\cdot,\theta) \right| \to 0 \tag{5.5}$$

with probability one, where $h_2(\mathbf{x},\theta)$ is defined in the proof of Lemma 4.3. Because the class of ellipsoids has polynomial discrimination, it follows from Theorem II.14 in Pollard (1984) that (5.4) holds with probability one. Since $\rho_2(y)$ is monotone in $|y|$ and bounded, (5.5) can be shown similar to (5.3). □

We may now show that $\mathbf{t}_n$ and $\mathbf{t}_{2,n}$ converge at rate $\sqrt{n}$ towards a normal distribution. We will use the following tightness property from Pollard (1984). It is a combination of the Approximation Lemma (p.27), Lemma II.36 (p.36) and the Equicontinuity Lemma (p.150).

LEMMA 5.1. *Let $\mathcal{F}$ be a permissible class of real valued functions with envelope F and suppose that $0 < PF^2 < \infty$. If the class of graphs of functions in $\mathcal{F}$ has polynomial discrimination, then for each $\eta > 0$ and $\varepsilon > 0$ there exists a $\delta > 0$ for which*

$$\limsup_{n\to\infty} \mathbf{P}\left\{ \sup_{\phi_1,\phi_2\in[\delta]} \left|\sqrt{n}(P_n - P)(\phi_1 - \phi_2)\right| > \eta \right\} < \varepsilon$$

where $[\delta] = \{(\phi_1, \phi_2) : \phi_1, \phi_2 \in \mathcal{F} \text{ and } P(\phi_1 - \phi_2)^2 \leq \delta^2\}$.

Since Lemma 5.1 is only stated for real valued fucntions, we will apply it to each component of the functions $g(\mathbf{x}, \boldsymbol{\theta})$ of (4.23) and $g_2(\mathbf{x}, \boldsymbol{\theta})$ of (4.29) separately. The following lemma implies that the classes of graphs of the functions which are then involved, have polynomial discrimination.

LEMMA 5.2. *For $\mathbf{x} = (x_1 \cdots x_p)^T$ and $\boldsymbol{\theta} \in \Theta$ let $k(\mathbf{x}, \boldsymbol{\theta})$ be a real valued function. Consider the classes functions $\mathcal{F} = \{k(\mathbf{x}, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ and $\mathcal{F}_j = \{k(\mathbf{x}, \boldsymbol{\theta})x_j : \boldsymbol{\theta} \in \Theta\}$ for $j = 1, \ldots, p$. Denote by $\mathcal{G}$ and $\mathcal{G}_j$ the corresponding classes of graphs of functions in $\mathcal{F}$ and $\mathcal{F}_j$ respectively. When $\mathcal{G}$ has polynomial discrimination, then also $\mathcal{G}_j$ has polynomial discrimination for $j = 1, \ldots, p$.*

PROOF: Consider a finite set $N = \{(\mathbf{x}_1, s_1), \ldots, (\mathbf{x}_m, s_m)\}$ in $\mathbf{R}^p \times \mathbf{R}$. Points $(\mathbf{x}, s)$ with $x_j = 0$ and $s \neq 0$ can never be in the graph of a function $k(\mathbf{x}, \boldsymbol{\theta})x_j$, and points $(\mathbf{x}, 0)$ will always be in the graph of any real-valued function. Therefore without loss of generality we may assume that for points in N it holds that $x_{ij} \neq 0$ and $s_i \neq 0$ for $i = 1, \ldots, m$. Note that a point $(\mathbf{x}, s) \in \mathbf{R}^p \times \mathbf{R}$ with $x_j \neq 0$ is picked out by the graph of the function $k(\mathbf{x}, \boldsymbol{\theta})x_j$ if and only if the point $(\mathbf{x}, s/x_j)$ is picked out by the graph of the function $k(\mathbf{x}, \boldsymbol{\theta})$. Since $\mathcal{G}$ has polynomial discrimination it picks out at most a polynomial number of subsets of the set $\{(\mathbf{x}_1, s_1/x_{1j}), \ldots, (\mathbf{x}_m, s_m/x_{mj})\}$. Then $\mathcal{G}_j$ picks out at most a polynomial number of subsets of the set N. □

To apply Lemma 5.1 we will need that the function $u(y) = \psi(y)/y$ in (4.23) is of bounded variation. This holds for instance for the function $u(y)$ that corresponds with the function ρ_{H} of (2.2).

THEOREM 5.3. *Let $\rho : \mathbf{R} \to [0, \infty)$ satisfy the conditions imposed in Method 1, and suppose that $\mathbf{E}\|X_1\|^2 < \infty$. Let $g(\mathbf{x}, \boldsymbol{\theta})$ and $G(\boldsymbol{\theta})$ be defined in (4.23) and (4.22) respectively, and suppose that the function $u(y) = \psi(y)/y$ is of bounded variation. Let $\mathbf{C}_n$, $n = 1, 2, \ldots$, be a sequence of affine equivariant covariance estimators that satisfies (5.1). Let $\boldsymbol{\theta}_0 = \big(\mathbf{t}(P), \mathbf{C}(P)\big)$ and suppose that $\mathbf{t}(P)$ and $G(\boldsymbol{\theta})$ satisfy the conditions of Theorem 4.1. Let $\mathbf{t}_n$ minimize the function $R_n(\mathbf{t})$ of (2.8) for $n = 1, 2, \ldots$. Then $\sqrt{n}\big(\mathbf{t}_n - \mathbf{t}(P)\big)$ has a limiting normal distribution with zero mean and covariance matrix $\boldsymbol{\Lambda}^{-1}\mathbf{M}\boldsymbol{\Lambda}^{-T}$, where $\mathbf{M}$ is the covariance matrix of $g(X_1, \boldsymbol{\theta}_0)$.*

PROOF: Let $\boldsymbol{\theta}_n = (\mathbf{t}_n, \mathbf{C}_n)$. We first show that

$$\left|\sqrt{n}(P_n - P)\Big(g(\cdot, \boldsymbol{\theta}_n) - g(\cdot, \boldsymbol{\theta}_0)\Big)\right| \to 0 \tag{5.6}$$

with probability one.

Write $k(\mathbf{x},\boldsymbol{\theta}) = u\big[\{(\mathbf{x}-\mathbf{t})^T\mathbf{C}^{-1}(\mathbf{x}-\mathbf{t})\}^{1/2}\big]$, and consider the j-th component

$$g_j(\mathbf{x},\boldsymbol{\theta}) = k(\mathbf{x},\boldsymbol{\theta})(x_j - t_j)$$

of function $g(\mathbf{x},\boldsymbol{\theta})$. Put $\mathbf{t}(P) = \boldsymbol{\mu} = (\mu_1 \cdots \mu_p)^T$ and $\mathbf{t}_n = (t_{n1} \cdots t_{np})^T$. Decompose as follows

$$\begin{aligned} g_j(\mathbf{x},\boldsymbol{\theta}_n) - g_j(\mathbf{x},\boldsymbol{\theta}_0) &= k(\mathbf{x},\boldsymbol{\theta}_n)x_j - k(\mathbf{x},\boldsymbol{\theta}_0)x_j \\ &\quad - t_{nj}\big(k(\mathbf{x},\boldsymbol{\theta}_n) - k(\mathbf{x},\boldsymbol{\theta}_0)\big) \\ &\quad + (\mu_j - t_{nj})k(\mathbf{x},\boldsymbol{\theta}_0). \end{aligned} \tag{5.7}$$

Consider the second term on the right hand side. Because the function $u(y)$ is of bounded variation, it follows from Lemma 22 of Nolan and Pollard (1987) that the class of graphs of the functions $\{k(\mathbf{x},\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ has polynomial discrimination and a bounded envelope. It is also permissible in the sense of Pollard (1984), so that Lemma 5.1 applies. Because $\boldsymbol{\theta}_n \to \boldsymbol{\theta}_0$, for each $\delta > 0$, the functions $k(\mathbf{x},\boldsymbol{\theta}_n)$ and $k(\mathbf{x},\boldsymbol{\theta}_0)$ are in the class $[\delta]$ of Lemma 5.1 for n sufficiently large. This means that if we integrate (5.7) with respect to $(P_n - P)$, the second term on the right hand side is $o_P(1/\sqrt{n})$. Similarly, using Lemma 5.2 and the fact that $\mathbf{E}\|X_1\|^2$ is finite and that $k(\mathbf{x},\boldsymbol{\theta})$ is bounded, the first term on the right hand side will $o_P(1/\sqrt{n})$. Finally, also the last term on the right hand side of (5.7) will be $o_P(1/\sqrt{n})$, because $\mathbf{t}_n \to \boldsymbol{\mu}$ and $(P_n - P)k(\cdot,\boldsymbol{\theta}_0)$ is $O_P(1/\sqrt{n})$ according to the central limit theorem. It follows that $(P_n - P)\big(g_j(\cdot,\boldsymbol{\theta}_n) - g_j(\cdot,\boldsymbol{\theta}_0)\big) = o_P(1/\sqrt{n})$. Since this holds for every $j = 1,\ldots,p$ we conclude that

$$(P_n - P)\Big(g(\cdot,\boldsymbol{\theta}_n) - g(\cdot,\boldsymbol{\theta}_0)\Big) = o_P(1/\sqrt{n}). \tag{5.8}$$

For any vector $\mathbf{t}_n$ that minimizes the function $R_n(\mathbf{t})$, it holds that the pair $\boldsymbol{\theta}_n = (\mathbf{t}_n, \mathbf{C}_n)$ is a zero of the function $P_n g(\cdot;\boldsymbol{\theta})$. Hence, together with (5.8), it follows that

$$\begin{aligned} \mathbf{0} &= P_n g(\cdot,\boldsymbol{\theta}_n) \\ &= Pg(\cdot,\boldsymbol{\theta}_n) + (P_n - P)g(\cdot,\boldsymbol{\theta}_0) + (P_n - P)\big(g(\cdot,\boldsymbol{\theta}_n) - g(\cdot,\boldsymbol{\theta}_0)\big) \\ &= Pg(\cdot,\boldsymbol{\theta}_n) + (P_n - P)g(\cdot,\boldsymbol{\theta}_0) + o_P(1/\sqrt{n}) \end{aligned}$$

Then use expansion (4.24) for $Pg(\cdot,\boldsymbol{\theta}_n)$, together with property (4.25). This gives

$$\mathbf{0} = \frac{\partial G}{\partial \mathbf{t}}(\boldsymbol{\mu}, \mathbf{C}_n)\,(\mathbf{t}_n - \boldsymbol{\mu}) + (\mathbf{t}_n - \boldsymbol{\mu})\,r(\boldsymbol{\theta}_n) + (P_n - P)g(\cdot,\boldsymbol{\theta}_0) + o_P(1/\sqrt{n}). \tag{5.9}$$

Because $\partial G/\partial \mathbf{t}$ is continuous at $\boldsymbol{\theta}_0$ and since $r(\boldsymbol{\theta}_n) = o_P(1)$, (5.9) reduces to

$$\mathbf{0} = \big(\boldsymbol{\Lambda} + o_P(1)\big)(\mathbf{t}_n - \boldsymbol{\mu}) + (P_n - P)g(\cdot,\boldsymbol{\theta}_0) + o_P(1/\sqrt{n}). \tag{5.10}$$

According to the central limit theorem $(P_n - P)g(\cdot, \boldsymbol{\theta}_0) = O_P(1/\sqrt{n})$, and as $\boldsymbol{\Lambda}$ is nonsingular, it follows that $\mathbf{t}_n - \boldsymbol{\mu} = O_P(1/\sqrt{n})$. When insert this in (5.10), we find that

$$\mathbf{0} = \boldsymbol{\Lambda}(\mathbf{t}_n - \boldsymbol{\mu}) + (P_n - P)g(\cdot, \boldsymbol{\theta}_0) + o_P(1/\sqrt{n}).$$

Because $\boldsymbol{\theta}_0$ is a zero of (4.22), it follows that

$$\sqrt{n}(\mathbf{t}_n - \boldsymbol{\mu}) = -\boldsymbol{\Lambda}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g(X_i, \boldsymbol{\theta}_0) + o_P(1). \tag{5.11}$$

Finally, since $\mathbf{E}\|X_1\|^2$ is finite and $u(y)$ is bounded, also $\mathbf{E}\|g(X_1, \boldsymbol{\theta}_0)\|^2$ is finite. Hence, the theorem follows after applying the central limit theorem to (5.11). □

When P is elliptical with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, then the matrix $\mathbf{M}$ is a multiple $\alpha\boldsymbol{\Sigma}$, where

$$\alpha = \frac{1}{p}\mathbf{E}\psi^2(\|\mathbf{B}^{-1}(X_1 - \boldsymbol{\mu})\|) = \frac{1}{p}\int \psi^2(\|\mathbf{x}\|) f(\|\mathbf{x}\|)\, d\mathbf{x}. \tag{5.12}$$

When the matrix $\boldsymbol{\Lambda}$ is of type (4.32), the limiting covariance of $\sqrt{n}(\mathbf{t}_n - \boldsymbol{\mu})$ reduces to $(\alpha/\beta^2)\boldsymbol{\Sigma}$, so that α/β^2 suffices as an index for the asymptotic efficiency. Moreover, the limiting distribution is the same as that of the corresponding affine equivariant location M-estimator considered in Maronna (1976); in particular, when P is spherically symmetric, it is the same as that of the location M-estimator defined by minimizing (2.1) with the same function ρ. When we use the function $\rho_{\mathrm{H}}(y; k)$ in (2.8), the asymptotic efficiency relative to the sample mean tends to 1, when $k \to \infty$. The efficiency relative to the maximum likelihood estimator can be read from Table 1 in Maronna (1976). It is reasonable at the multivariate normal, as well at several multivariate student distributions, for moderate values of k.

THEOREM 5.4. *Let $\rho_j : \mathbf{R} \to [0, \infty)$, $j = 1, 2$, satisfy the conditions of Theorem 3.3 and suppose that (3.9) holds instead of (2.10). Suppose that $\mathbf{E}\|X_1\|^2 < \infty$. Let $g_2(\mathbf{x}, \boldsymbol{\theta})$ and $G_2(\boldsymbol{\theta})$ be defined in (4.29) and (4.28) respectively. Let $\boldsymbol{\theta}_{1,n}$, for $n = 1, 2, \ldots$, be a sequence of S-estimators, defined with the function $\rho_1(\cdot)$ and the constant $0 < b_1 < \sup \rho_1$. Let $R_{2,P}(\mathbf{t})$ be defined in (2.16) and suppose that it has a unique minimum at $\mathbf{t}_1(P)$. Let $\boldsymbol{\theta}_0 = (\mathbf{t}_2(P), \mathbf{C}_1(P))$ and suppose that $\mathbf{t}_2(P)$ and $G_2(\boldsymbol{\theta})$ satisfy the conditions of Theorem 4.2. Let $\mathbf{t}_{2,n}$, $n = 1, 2, \ldots$, be any sequence of location estimators, where $\mathbf{t}_{2,n}$ satisfies (2.12) and (2.13). Then $\sqrt{n}(\mathbf{t}_{2,n} - \mathbf{t}_2(P))$ has a limiting normal distribution with zero mean and covariance matrix $\boldsymbol{\Lambda}_2^{-1}\mathbf{M}_2\boldsymbol{\Lambda}_2^{-T}$, where $\mathbf{M}_2$ is the covariance matrix of $g_2(X_1, \boldsymbol{\theta}_0)$.*

PROOF: The proof is the same as that of Theorem 5.3. We only have to show that $u_2(y) = \psi_2(y)/y$ is of bounded variation and that $\partial G_2/\partial \mathbf{t}$ is continuous at $\boldsymbol{\theta}_0$. The latter has already been shown in the proof of Theorem 4.2, and conditions (R1)-(R5) imply that $u_2(y)$ is of bounded variation. □

When P is elliptical with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, then the matrix $\mathbf{M}_2$ is equal to $\alpha_2\boldsymbol{\Sigma}$, where α_2 is as in (5.12) with $\psi = \psi_2$. When the matrix $\boldsymbol{\Lambda}_2$ is of type (4.32), the limiting covariance of $\sqrt{n}(\mathbf{t}_{2,n} - \boldsymbol{\mu})$ reduces to $(\alpha_2/\beta_2^2)\boldsymbol{\Sigma}$. The limiting distribution is the same as that of the corresponding affine equivariant location M-estimator considered in Maronna (1976), or of the location S-estimator defined with the function ρ_2. When we use the biweight function $\rho_B(y; c_2)$ in (2.8), the asymptotic efficiency relative to the sample mean tends to 1, when $c_2 \to \infty$. The efficiency at the multivariate normal and at a contaminated normal relative to the maximum likelihood estimator can be read from Table 1 in Lopuhaä (1989).

REMARK 5.1: If we do not assume that $\mathbf{t}(P)$ and $\mathbf{t}_2(P)$ are a point of symmetry of P, $\mathbf{C}_n$ and $\boldsymbol{\theta}_{1,n}$ may influence the limiting behaviour of $\mathbf{t}_n$ and $\mathbf{t}_{2,n}$ respectively. This can easily be seen from expansion (4.24). If for instance G also has a nonzero partial derivative $\partial G/\partial \mathbf{C}$ at $\boldsymbol{\theta}_0$, it follows that $\mathbf{t}_n$ converges at the same rate as $\mathbf{C}_n$ does.

Acknowledgments. I thank Peter Rousseeuw and Rudolf Grübel for stimulating discussions and helpful suggestions and remarks.

REFERENCES

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

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.

GRÜBEL, R. (1988). A minimal characterization of the covariance matrix. *Metrika* **35** 49–52.

HAMPEL, F.R. (1968). Contributions to the Theory of Robust Estimation. Ph.D.thesis, University of California, Department of Statistics.

HAMPEL, F.R. (1974). The influence curve and its role in robust estimation. *J. Am. Statist. Assoc.* **69** 383–393.

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.

HUBER, P.J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35** 73–101.

HUBER, P.J. (1967). The behaviour of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (L.M.Le Cam and J.Neyman, eds.) 221–233. University of California Press, Berkeley.

HUBER, P.J. (1984). Finite sample breakdown point of M- and P-estimators. *Ann. Statist.* **12** 119–126.

LOPUHAÄ, H.P. (1989). On the relation between S-estimators and M-estimators of multivariate location and covariance. *Ann. Statist.* **17** 1662–1683.

LOPUHAÄ, H.P. and ROUSSEEUW, P.J. (1989). Breakdown properties of affine equivariant estimators of multivariate location and covariance matrices. Revised version of Technical Report 87-14, Delft University of Technology. Tentatively accepted by Ann. Statist..

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

NOLAN, D. and POLLARD, D. (1987). U-Processes: rates of convergence. *Ann. Statist.* **15** 780–799.

POLLARD, D. (1984). *Convergence of Stochastic Processes.* Springer, New York.

ROUSSEEUW, P.J. (1983). Multivariate estimation with high breakdown point. Paper presented at the Fourth Pannonian Symposium on Mathematical Statistics and Probability, Bad Tatzmannsdorf, Austria, September 4-9, 1983. In *Mathematical Statistics and Applications (1985)* (W.Grossmann, G.Pflug, I.Vincze and W.Wertz, eds.) 283–297. Reidel, Dordrecht, The Netherlands.

ROUSSEEUW, P.J. and LEROY, A.M. (1987). *Robust Regression and Outlier Detection.* Wiley, New York.

ROUSSEEUW, P.J. and VAN ZOMEREN, B.C. (1990). Identification of multivariate outliers and leverage points by means of robust covariance matrices. To appear J. Am. Statist. Assoc..

ROUSSEEUW, P.J. and YOHAI, V.J. (1984). Robust regression by means of S-estimators. In *Robust and Nonlinear Time Series Analysis.* Lecture Notes in Statistics **26** 256–272. Springer Verlag, New York.

TONG, Y.L. (1980). *Probability Inequalities in Multivariate Distributions.* Academic Press, New York.

TYLER, D.E. (1986). Breakdown properties of the M-estimators of multivariate scatter. Technical Report, Rutgers University, New Jersey.

YOHAI, V.J. (1987). High breakdown-point and high efficiency robust estimates for regression. *Ann. Statist.* **15** 642–656.

MULTIVARIATE τ-ESTIMATORS FOR LOCATION AND SCATTER

HENDRIK P. LOPUHAÄ

Delft University of Technology

We discuss the robustness and asymptotic behaviour of τ-estimators for multivariate location and scatter. We show that τ-estimators correspond with multivariate M-estimators defined by a weighted average of redescending ψ-functions, where the weights are adaptive. We prove consistency and asymptotic normality under weak assumptions on the underlying distribution P, show that τ-estimators have a high breakdown point and obtain the influence function at general distributions P. In the special case of a location-scatter family τ-estimators are asymptotically equivalent to multivariate S-estimators defined by means of a weighted ρ-function. This enables us to combine a high breakdown point and bounded influence with good asymptotic efficiency for the location and covariance estimator.

1. Introduction. The minimum volume ellipsoid (MVE) estimators are defined as the center and scatter matrix of the smallest ellipsoid containing at least half of the observations (Rousseeuw 1983). These estimators are known to have good robustness properties, but their limiting behaviour is relatively poor, as they converge with rate $\sqrt[3]{n}$ towards a nonnormal limiting distribution (Kim and Pollard 1989, Davies 1989). To retain the robustness and to improve the asymptotic properties one can 'smoothen' the condition of covering half of the observations. This may result in multivariate S-estimators, defined as the center and scatter matrix of the smallest ellipsoid that satisfies a condition on the average of smoothly weighted Mahalanobis distances (Davies 1987, Lopuhaä 1989). In the univariate case this is equivalent to computing an M-estimator of scale as a function of the location parameter μ and to minimize this over μ. The S-estimators converge with rate $\sqrt{n}$ towards a normal distribution. However, there is a trade-off between robustness and asymptotic efficiency : a high breakdown point corresponds with a low efficiency and vice versa.

Yohai and Zamar (1988) investigated an extension of regression S-estimators, which retains the good robustness and improves the asymptotic efficiency. In the special case of estimating univariate location and scale their proposal amounts to the following. To make the M-estimator of scale more efficient they consider an adaptive multiple of it, which they call a τ-estimator of scale, and minimize this as a function of the location parameter. Regression τ-estimators were studied under the assumption of the usual parametric regression model with random carriers independent of the error terms. Although the τ-estimator of the error scale was claimed to be highly efficient, only the limiting distribution of the τ-estimator of

This research is financially supported by NWO under Grant 10-62-10.

1980 *Mathematics subject classifications* : 62F35, 62H12.

Keywords : τ-estimators, High breakdown point, Bounded influence, High efficiency.

the regression coefficient was given. In the special case of univariate location and scale, the claimed high efficiency of the scale estimator will follow as a corollary of our results.

In this paper we study the robustness and asymptotic behaviour of τ-estimators for multivariate location and scatter under weak conditions on the underlying distribution P. In Section 2 we give the definition of multivariate τ-functionals and give sufficient conditions for their existence. Continuity of these functionals, and hence consistency of the τ-estimators, is shown in Section 3. In Section 4 we show that multivariate τ-estimators relate to multivariate M-estimators as defined in Huber (1981). The location τ-estimator is shown to be equivalent to a location M-estimator, defined by an adaptively weighted average of redescending ψ-functions; for the covariance τ-estimator something similar holds. The corresponding M-estimator type of score equations therefore, become too complicated to obtain a limit theorem by means of Huber's (1967) results. Instead we will use empirical process theory (Pollard 1984) to obtain the simultaneous limiting distribution for τ-estimators of location and scatter.

The robustness of these estimators will be measured by means of the breakdown point and the influence function. The breakdown point provides a global measure of the sensitivity of an estimator to outlying observations. It may be complemented by measures of the local sensitivity such as the influence function and the corresponding gross-error sensitivity. In Section 5 we show that τ-estimators have the same high breakdown point as S-estimators, and we obtain the general expression for the influence function.

In Section 6 we consider a parametric location-scatter family as a special case. It turns out that in this case the limiting normal distribution and the influence function of τ-estimators are the same as those of multivariate S-estimators that are defined by means of a weighted ρ-function. This enables us to combine a high breakdown point and a bounded influence function with good asymptotic efficiency.

2. Definition and existence.

2.1. Definition. We will define estimators by means of a functional that acts on the space $\mathcal{P}(\mathbf{R}^p)$ of all probability distributions on $\mathbf{R}^p$, evaluated at the empirical distribution. Denote by $|\mathbf{M}|$ the determinant of a $p \times p$-matrix $\mathbf{M}$, and denote by $\lambda_p(\mathbf{M}) \leq \cdots \leq \lambda_1(\mathbf{M})$ the eigenvalues of $\mathbf{M}$. Let ρ_1 and ρ_2 be nonnegative functions on $\mathbf{R}$, and let b_1 and b_2 be positive constants. We define τ-functionals for location and scatter as follows.

For $P \in \mathcal{P}(\mathbf{R}^p)$ let $\mathbf{t}(P)$ and $\mathbf{C}(P)$ be the vector and the positive definite symmetric $p \times p$-matrix that minimize

$$|\mathbf{C}|\left\{ \int \rho_2\left[\{(\mathbf{x}-\mathbf{t})^T\mathbf{C}^{-1}(\mathbf{x}-\mathbf{t})\}^{1/2}\right] dP(\mathbf{x})\right\}^p \tag{2.1}$$

subject to

$$\int \rho_1\left[\{(\mathbf{x}-\mathbf{t})^T\mathbf{C}^{-1}(\mathbf{x}-\mathbf{t})\}^{1/2}\right] dP(\mathbf{x}) = b_1. \tag{2.2}$$

Denote this minimization problem by $(\mathcal{P}_P)$. Take $\mathbf{t}(P)$ to be the location τ-functional and define the covariance τ-functional as

$$\mathbf{V}(P) = b_2^{-1}\mathbf{C}(P)\int \rho_2\left[\{(\mathbf{x}-\mathbf{t}(P))^T\mathbf{C}(P)^{-1}(\mathbf{x}-\mathbf{t}(P))\}^{1/2}\right] dP(\mathbf{x}).$$

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be n observations in $\mathbf{R}^p$ and denote by P_n the corresponding empirical distribution. Multivariate τ-estimators are defined as the vector $\mathbf{t}_n = \mathbf{t}(P_n)$ and the matrix

$$\mathbf{V}_n = \mathbf{V}(P_n) = b_2^{-1}\mathbf{C}_n\frac{1}{n}\sum_{i=1}^{n}\rho_2\left[\{(\mathbf{x}_i-\mathbf{t}_n)^T\mathbf{C}_n^{-1}(\mathbf{x}_i-\mathbf{t}_n)\}^{1/2}\right]$$

where $\mathbf{t}_n$ and $\mathbf{C}_n$ minimize $|\mathbf{C}|\left\{\sum_{i=1}^{n}\rho_2\left[\{(\mathbf{x}_i-\mathbf{t})^T\mathbf{C}^{-1}(\mathbf{x}_i-\mathbf{t})\}^{1/2}\right]\right\}^p$ subject to

$$\frac{1}{n}\sum_{i=1}^{n}\rho_1\left[\{(\mathbf{x}_i-\mathbf{t})^T\mathbf{C}^{-1}(\mathbf{x}_i-\mathbf{t})\}^{1/2}\right] = b_1. \tag{2.3}$$

Note that constraint (2.3) is the same as the constraint of the minimization problem that defines multivariate S-estimators with the function ρ_1. In fact, multivariate S-estimators arise as a special case of τ-estimators. Indeed, if $\rho_1 = \rho_2$ and $b_1 = b_2$, then $\mathbf{t}_n$ and $\mathbf{V}_n$ would just be the ordinairy S-estimators. Instead of minimizing the determinant of $\mathbf{C}$ over all pairs $(\mathbf{t}, \mathbf{C})$ that satisfy (2.3), we now minimize the determinant of an adaptive multiple of such $\mathbf{C}$, i.e. the determinant of the covariance τ-estimator $\mathbf{V}_n$.

The least squares estimators can also be obtained as a special case, namely with $\rho_1(y) = \rho_2(y) = y^2$ and $b_1 = b_2 = p$ (see for instance Grübel 1988), as well as the MVE estimators with $\rho_1 = \rho_2$ an indicator function and $b_1 = b_2$ roughly $\frac{1}{2}$. To get the good robustness from the MVE estimators and the good limiting properties from the least squares estimators, we will take functions ρ_1 and ρ_2 that are so to speak 'in between' these two cases. Throughout the paper we will assume that ρ_1 and ρ_2 both satisfy the following conditions.

(R1) $\rho_k(0) = 0$, ρ_k is symmetric and ρ_k is twice continuously differentiable. Denote by ψ_k the derivative of ρ_k.

(R2) There exists a finite constant $c_k > 0$ such that ρ_k is strictly increasing on $[0, c_k]$ and constant on $[c_k, \infty)$. Write $a_k = \rho_k(c_k)$.

In addition we impose the following condition only on the function ρ_2.

(A) $2\rho_2(y) - \psi_2(y)y > 0$, for $y > 0$.

It will guarantee that the loss function in (2.1) is a strictly increasing function of the magnitude of $\mathbf{C}$ (see Remark 2.1). Together with the boundedness condition in (R2), i.e. $a_k = \sup \rho_k < \infty$, this provides the good breakdown properties of the τ-estimators. To guarantee the existence of solutions of $(\mathcal{P}_P)$, the constant b_1 in (2.2) must be chosen such that $0 < b_1 < a_1$. A typical function ρ that satisfies all conditions above is Tukey's biweight function $\rho_{\text{B}}(y; c)$.

The breakdown point of the τ-estimators turns out to be an increasing function of the constant b_1. However, when $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are assumed to be a sample from an elliptical distribution with density $|\mathbf{B}|^{-1} f(\|\mathbf{B}^{-1}(\mathbf{x} - \boldsymbol{\mu})\|)$, where $\mathbf{B}\mathbf{B}^T = \boldsymbol{\Sigma}$, then a natural choice for b_1 is $b_1 = \int \rho_1(\|\mathbf{x}\|) f(\|\mathbf{x}\|)\, d\mathbf{x}$. In this case the breakdown point will only be a function of c_1, where small values of c_1 correspond with a high breakdown point and vice versa. The smoothness conditions on ρ_1 and ρ_2 are needed to obtain asymptotic normality and a bounded influence function. The constant $b_2 > 0$ is only a normalizing constant to obtain consistency of $\mathbf{V}_n$ for the 'true' scatter parameter. In case of elliptically distributed observations one should choose $b_2 = \int \rho_2(\|\mathbf{x}\|) f(\|\mathbf{x}\|)\, d\mathbf{x}$ for $\mathbf{V}_n$ to be consistent for $\boldsymbol{\Sigma}$. In this case the limiting variances of the τ-estimators turn out to depend on both c_1 and c_2. However, for any c_1 fixed and c_2 large these variances will be close to those corresponding with the sample mean and the sample covariance. This enables us to combine a high breakdown point and bounded influence with a good efficiency for both $\mathbf{t}_n$ and $\mathbf{V}_n$, for instance at the normal distribution. Possible choices for ρ_1 and ρ_2 are the biweight functions $\rho_1(y) = \rho_{\text{B}}(y; c_1)$ and $\rho_2(y) = \rho_{\text{B}}(y; c_2)$.

REMARK 2.1: When the distribution P does not have all its mass concentrated in one point, then any pair $(\mathbf{t}(P), \mathbf{C}(P))$ that is a solution of minimization problem $(\mathcal{P}_P)$ will also be a solution to the problem of minimizing the same loss function subject to

$$\int \rho_1\left[\{(\mathbf{x} - \mathbf{t})^T \mathbf{C}^{-1}(\mathbf{x} - \mathbf{t})\}^{1/2}\right] dP(\mathbf{x}) \leq b_1. \tag{2.4}$$

This property will be useful as it will be more convenient to deal with constraint (2.4) then with constraint (2.2). That this property holds can be seen as follows. Consider the function $h : (0, \infty) \to \mathbf{R}$

$$h(s) = |s\mathbf{C}|\left\{ \int \rho_2\left[\{(\mathbf{x} - \mathbf{t})^T (s\mathbf{C})^{-1}(\mathbf{x} - \mathbf{t})\}^{1/2}\right] dP(\mathbf{x}) \right\}^p. \tag{2.5}$$

Note that $|s\mathbf{C}| = s^p |\mathbf{C}|$ and that the derivative of $s\rho_2(ys^{-\frac{1}{2}})$ with respect to s is $\rho_2(ys^{-\frac{1}{2}}) - \frac{1}{2}\psi_2(ys^{-\frac{1}{2}})ys^{-\frac{1}{2}}$. Since P can not have all its mass at $\mathbf{t}$ condition (A) implies that $h'(s) > 0$, so that h is strictly increasing in $s > 0$. By means of a standard argument it follows that any solution of $(\mathcal{P}_P)$ will be a solution of the same minimization problem, except with (2.4) instead of (2.2).

2.2. Existence. Denote by PDS(p) the class of all positive definite symmetric $p \times p$-matrices and let Θ be the parameter space $\mathbf{R}^p \times$ PDS(p) which can be seen as

an open subset of $\mathbf{R}^{p+\frac{1}{2}p(p+1)}$. Solutions of $(\mathcal{P}_P)$ in Θ exist when P does not have too much mass concentrated at some hyperplane of dimension $\leq p-1$, that is when P satisfies the following property for small enough ε.

(H_ε) For every hyperplane H with $\dim(H) \leq p-1$, it holds that $P(H) < \varepsilon$.

This condition is equivalent with

(C_ε) The value $\delta_\varepsilon = \inf\{\delta : P\big(H(\boldsymbol{\alpha}, \mathbf{v}, \delta)\big) \geq \varepsilon, \|\boldsymbol{\alpha}\| = 1, \delta \geq 0, \mathbf{v} \in \mathbf{R}^p\}$ is strictly positive

which was used in Lopuhaä (1989), where $H(\boldsymbol{\alpha}, \mathbf{v}, \delta) = \{\mathbf{x} : \boldsymbol{\alpha}^T\mathbf{v} \leq \boldsymbol{\alpha}^T\mathbf{x} \leq \boldsymbol{\alpha}^T\mathbf{v}+\delta\}$ is a strip of width δ and with a direction perpendicular to $\boldsymbol{\alpha}$, which has the point $\mathbf{v}$ on its boundary. Since we will refer to some parts of the proofs in Lopuhaä (1989) we briefly show the equivalence of (H_ε) and (C_ε).

Clearly, condition (C_ε) implies (H_ε). That (H_ε) implies $\delta_\varepsilon > 0$ can be seen as follows. Suppose that $\delta_\varepsilon = 0$. Then there is a sequence of strips $S_k = H(\boldsymbol{\alpha}_k, \mathbf{v}_k, \delta_k)$ with $\delta_k \downarrow 0$ for which $P(S_k) \geq \varepsilon$. Because P is tight there exists a nested collection of compact balls $\{B_\gamma : 0 < \gamma \leq 1\}$ with $P(B_\gamma) \geq 1-\gamma$. First consider the ball $B_{\varepsilon/2}$. Then every S_k must intersect $B_{\varepsilon/2}$ and hence, we assume without loss of generality that the sequence $\{\mathbf{v}_k\}$ is contained in every B_γ for $0 < \gamma \leq \varepsilon/2$. The sequence $\{\boldsymbol{\alpha}_k\}$ is contained in the compact set $A = \{\|\boldsymbol{\alpha}\| = 1\}$ and the sequence $\{\delta_k\}$ is contained in a compact subset D of $[0, \infty)$. Therefore, the sequence $\{(\boldsymbol{\alpha}_k, \mathbf{v}_k, \delta_k)\}$ is contained in the compact set $A \times B_{\varepsilon/2} \times D$ and has a density point $(\boldsymbol{\alpha}, \mathbf{v}, 0)$ in $A \times B_{\varepsilon/2} \times D$ and hence, in every $A \times B_\gamma \times D$ for $0 < \gamma \leq \varepsilon/2$. This means that the hyperplane $H(\boldsymbol{\alpha}, \mathbf{v}, 0)$ has a nonempty intersection with each ball B_γ for $0 < \gamma \leq \varepsilon/2$. Let $\varepsilon/2 < \eta < \varepsilon$ and consider the ball $B_{\varepsilon-\eta}$. Then $P(S_k \cap B_{\varepsilon-\eta}) \geq \eta$ and $H(\boldsymbol{\alpha}, \mathbf{v}, 0)$ has a nonempty intersection with $B_{\varepsilon-\eta}$. By a standard argument it follows that $P\big(H(\boldsymbol{\alpha}, \mathbf{v}, 0) \cap B_{\varepsilon-\eta}\big) \geq \eta$, so that $P\big(H(\boldsymbol{\alpha}, \mathbf{v}, 0)\big) \geq \eta$. As this holds for η arbitrarily close to ε it follows that $P\big(H(\boldsymbol{\alpha}, \mathbf{v}, 0)\big) \geq \varepsilon$ which is in contradiction with (H_ε).

THEOREM 2.1. *Suppose that P satisfies property (H_ε) for some $0 < \varepsilon \leq 1 - r_1$, where $r_1 = b_1/a_1$. Then $(\mathcal{P}_P)$ has at least one solution.*

Before we prove Theorem 2.1 we first show some preliminary lemmas. These lemmas will imply that all possible solutions of $(\mathcal{P}_P)$ are contained in a compact subset of Θ. We will denote ellipsoids $\{\mathbf{x} : (\mathbf{x}-\mathbf{t})^T\mathbf{C}^{-1}(\mathbf{x}-\mathbf{t}) \leq c^2\}$ by $E(\mathbf{t}, \mathbf{C}, c)$.

LEMMA 2.1. *Suppose that $(\mathbf{t}, \mathbf{C}) \in \Theta$ satisfies constraint (2.2). Then there exists a constant $q > 0$, which only depends on the functions ρ_1 and ρ_2, and the constant b_1 such that $\int \rho_2\big[\{(\mathbf{x}-\mathbf{t})^T\mathbf{C}^{-1}(\mathbf{x}-\mathbf{t})\}^{1/2}\big]\, dP(\mathbf{x}) \geq q$.*

PROOF: Consider the set $B = \big\{\mathbf{x} : \sqrt{(\mathbf{x}-\mathbf{t})^T\mathbf{C}^{-1}(\mathbf{x}-\mathbf{t})} > \rho_1^{-1}(b_1/2)\big\}$. Then it holds that

$$b_1 = \int \rho_1\big[\{(\mathbf{x}-\mathbf{t})^T\mathbf{C}^{-1}(\mathbf{x}-\mathbf{t})\}^{1/2}\big]\, dP(\mathbf{x}) \leq \frac{b_1}{2}(1 - P(B)) + a_1 P(B).$$

Since b_1 satisfies $0 < b_1 < a_1$, it follows that $P(B) \geq b_1/(2a_1 - b_1) > 0$. This means that $\int \rho_2[\{(\mathbf{x}-\mathbf{t})^T\mathbf{C}^{-1}(\mathbf{x}-\mathbf{t})\}^{1/2}]\, dP(\mathbf{x}) \geq \rho_2(\rho_1^{-1}(b_1/2))b_1/(2a_1 - b_1) > 0$. □

LEMMA 2.2. *Let $(\mathbf{t}, \mathbf{C}) \in \Theta$, $0 < m_0 < \infty$, $0 < c < \infty$ and $0 < \varepsilon < 1$.*

(i) If P satisfies (H_ε) and if $P(E(\mathbf{t}, \mathbf{C}, c_1)) \geq \varepsilon$, then there exists a constant $k_1 > 0$, which only depends on ε, P and c_1, such that $\lambda_p(\mathbf{C}) \geq k_1$.

(ii) Suppose that $\int \rho_1(\|\mathbf{x}\|/m_0)\, dP(\mathbf{x}) \leq b_1$ and suppose that $\lambda_p(\mathbf{C}) \geq k_1 > 0$. Then there exists a constant $k_2 < \infty$, which only depends on k_1, m_0, ρ_1, ρ_2 and b_1, such that if $\lambda_1(\mathbf{C}) > k_2$, the pair $(\mathbf{t}, \mathbf{C})$ can not be a solution of $(\mathcal{P}_P)$.

(iii) Let P satisfy (H_ε) and suppose that $P(E(\mathbf{t}, \mathbf{C}, c)) \geq \varepsilon$. Suppose that $0 < k_1 \leq \lambda_p(\mathbf{C}) \leq \lambda_1(\mathbf{C}) \leq k_2 < \infty$. Then there exists a compact set $K \subset \Theta$, which only depends on ε, P, c_1, k_1 and k_2, such that $(\mathbf{t}, \mathbf{C})$ is contained in K.

PROOF: Since condition (H_ε) is equivalent with condition (C_ε), the proof is similar to the proof of Lemma 3.1 in Lopuhaä (1989). The proof of (i) and (iii) remains the same. For (ii) note that according to Lemma 2.1

$$q \leq \int \rho_2[\{(\mathbf{x}-\mathbf{t})^T\mathbf{C}^{-1}(\mathbf{x}-\mathbf{t})\}^{1/2}]\, dP(\mathbf{x}) \leq a_2.$$

Therefore, since $(\mathbf{0}, m_0^2\mathbf{I})$ satisfies constraint (2.3), according to Remark 2.1 every possible solution of $(\mathcal{P}_P)$ must satisfy $|\mathbf{C}| \leq (m_0^2 a_2/q)^p$, which means that $\lambda_1(\mathbf{C}) \leq (m_0^2 a_2/q)^p/k_1^{p-1} < \infty$. □

PROOF OF THEOREM 2.1: Along the lines of the proof of Theorem 3.1 in Lopuhaä (1989) it follows with Lemma 2.2 that there exists a compact subset $K \subset \Theta$ to which we can restrict ourselves for solving $(\mathcal{P}_P)$. Since the loss function in $(\mathcal{P}_P)$ is a continuous function of $\mathbf{t}$ and $\mathbf{C}$ it must attain a minimum on K. □

The finite sample situation is of course a special case of Theorem 2.1. Let k_{max} be the maximum number of $\mathbf{x}_i$'s that are contained in some hyperplane of dimension $\leq p-1$. Obviously, $k_{max} \geq p$ and if $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are in *general position*, i.e. no $p+1$ points lie in some lower dimensional hyperplane, then $k_{max} = p$. An immediate consequence of Theorem 2.1 is that if $n(1-r_1) \geq k_{max} + 1$, problem $(\mathcal{P}_{P_n})$ has at least one solution $(\mathbf{t}_n, \mathbf{C}_n)$. To show that every solution $(\mathbf{t}_n, \mathbf{C}_n)$ of $(\mathcal{P}_{P_n})$ converges to a solution $(\mathbf{t}(P), \mathbf{C}(P))$ of $(\mathcal{P}_P)$ we will need that $(\mathbf{t}(P), \mathbf{C}(P))$ is uniquely defined. This will be the case for instance for any elliptical distribution $P_{\mu,\Sigma}$, which satisfies the following condition.

(F) f is nonincreasing and has at least one point of decrease on $[0, \min(c_1, c_2)]$.

Note that $P_{\mu,\Sigma}$ satisfies property (H_ε) for every $0 < \varepsilon \leq 1$, so that according to Theorem 2.1 at least one solution of $(\mathcal{P}_{P_{\mu,\Sigma}})$ exists.

THEOREM 2.2. *Let $P_{\mu,\Sigma}$ be an elliptical distribution that satisfies (F). Choose $b_1 = \int \rho_1(\|\mathbf{x}\|) f(\|\mathbf{x}\|)\, d\mathbf{x}$ in (2.2). Then $(\mathcal{P}_{P_{\mu,\Sigma}})$ has a unique solution (μ, Σ).*

PROOF: First note that by means of a suitable rescaling it is sufficient to consider the problem $(\mathcal{P}_{P_{\beta,\Lambda}})$: find a vector β in $\mathbf{R}^p$ and a diagonal matrix $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_p)$ with $\lambda_j > 0$ for $j = 1, \ldots, p$ that minimize

$$\left(\prod_{j=1}^{p} \lambda_j\right) \left\{ \int \rho_2[\{(\mathbf{x}-\beta)^T \Lambda^{-1}(\mathbf{x}-\beta)\}^{1/2}] f(\|\mathbf{x}\|)\, d\mathbf{x} \right\}^p$$

subject to $\int \rho_1[\{(\mathbf{x}-\beta)^T \Lambda^{-1}(\mathbf{x}-\beta)\}^{1/2}] f(\|\mathbf{x}\|)\, d\mathbf{x} = b_1$. To show that $(\mathcal{P}_{P_{\mu,\Sigma}})$ has a unique solution (μ, Σ) it is equivalent to show that $(\mathcal{P}_{P_{\beta,\Lambda}})$ has a unique solution $(\mathbf{0}, \mathbf{I})$. The proof of this is a subtle variation on the proof of Theorem 1 of Davies (1987), who shows that the ordinairy S-minimization problem $(\mathcal{P}^S_{\beta,\Lambda})$ of minimizing $\prod_{j=1}^p \lambda_j$ over all $\beta \in \mathbf{R}^p$ and positive definite diagonal matrices Λ satisfying

$$\int \rho[\{(\mathbf{x}-\beta)^T \Lambda^{-1}(\mathbf{x}-\beta)\}^{1/2}] f(\|\mathbf{x}\|)\, d\mathbf{x} = \int \rho(\|\mathbf{x}\|) f(\|\mathbf{x}\|)\, d\mathbf{x} \tag{2.6}$$

has a unique solution $(\mathbf{0}, \mathbf{I})$. This holds under conditions on the function ρ in (2.6) which are weaker than (R1)-(R2) and f nonincreasing with at least one common point of decrease with the function $-\rho$. Note therefore, that according to condition (F), Davies' Theorem 1 applies to the S-minimization problems with the function ρ_1 or ρ_2 in (2.6).

Consider the loss function

$$\varphi(\beta, \Lambda) = \left(\prod_{j=1}^{p} \lambda_j\right) \left\{ \int \rho_2[\{(\mathbf{x}-\beta)^T \Lambda^{-1}(\mathbf{x}-\beta)\}^{1/2}] f(\|\mathbf{x}\|)\, d\mathbf{x} \right\}^p.$$

First note that the constraint in (2.2) of the problem $(\mathcal{P}_{P_{\beta,\Lambda}})$ is exactly the same as the constraint (2.6) of the S-minimization problem $(\mathcal{P}^S_{\beta,\Lambda})$ with the function ρ_1. Since this minimization problem has a unique solution $(\mathbf{0}, \mathbf{I})$, we may restrict to minimizing φ over pairs (β, Λ) that satisfy (2.2) and for which $\prod_{j=1}^p \lambda_j \geq 1$. Define the sets $A = \{(\beta, \Lambda) : \prod_{j=1}^p \lambda_j \geq 1$ and (β, Λ) satisfy constraint (2.2)$\}$ and $B = \{(\beta, \Lambda) : \prod_{j=1}^p \lambda_j = 1\}$. We are left with showing that the problem

$$\min_{(\beta,\Lambda)\in A} \varphi(\beta, \lambda) \tag{2.7}$$

has a unique solution $(\mathbf{0}, \mathbf{I})$. We will consider the smaller set $(A \cap B) \subset A$, first show that minimizing φ over $A \cap B$ has a unique solution $(\mathbf{0}, \mathbf{I})$ and then show that on the set $A \backslash B$ the loss function φ only takes on values that are strictly greater than $\varphi(\mathbf{0}, \mathbf{I})$.

Since $\prod_{j=1}^p \lambda_j = 1$ for $(\beta, \Lambda) \in B$, it follows that minimizing φ over $A \cap B$ is equivalent to

$$\min_{(\beta,\Lambda)\in A\cap B} \int \rho_2[\{(\mathbf{x}-\beta)^T \Lambda^{-1}(\mathbf{x}-\beta)\}^{1/2}] f(\|\mathbf{x}\|)\, d\mathbf{x}. \tag{2.8}$$

First consider the problem of minimizing φ over the larger set $B \supset (A \cap B)$:

$$\min_{(\beta,\Lambda)\in B} \int \rho_2[\{(\mathbf{x}-\beta)^T\Lambda^{-1}(\mathbf{x}-\beta)\}^{1/2}] f(\|\mathbf{x}\|)\, d\mathbf{x}. \tag{2.9}$$

The key observation is that this minimization problem is exactly the transformed maximization problem considered by Davies (1987, p.1275). It is derived from the original S-minimization problem with the function ρ_2 using that this S-minimization problem has solution $(\beta^*, \Lambda^*) = (\mathbf{0}, \mathbf{I})$. According to the proof of Theorem 1 of Davies (1987) the transformed problem has a unique solution $(\mathbf{0}, \mathbf{I})$, hence problem (2.9) has a unique solution $(\mathbf{0}, \mathbf{I})$. However, $(\mathbf{0}, \mathbf{I})$ is also an element of the set A, so that minimization problem (2.8) must also have a unique solution $(\mathbf{0}, \mathbf{I})$.

Therefore, for showing that the minimization problem (2.7) has a unique solution $(\mathbf{0}, \mathbf{I})$, we are left with showing that $\varphi(\mathbf{0}, \mathbf{I}) < \inf\{\varphi(\beta, \Lambda) : (\beta, \Lambda) \in A \backslash B\}$. Suppose there would exist a pair $(\tilde{\beta}, \tilde{\Lambda}) \in A \backslash B$, with $\varphi(\tilde{\beta}, \tilde{\Lambda}) \leq \varphi(\mathbf{0}, \mathbf{I})$. Then for some $0 < s < 1$ the pair $(\tilde{\beta}, s\tilde{\Lambda}) \in B$. The function $\varphi(\tilde{\beta}, s\tilde{\Lambda})$ is equal to the function $h(s)$ in (2.5) with $(\mathbf{t}, \mathbf{C}) = (\tilde{\beta}, \tilde{\Lambda})$ and P spherically symmetric. This function was already shown to be strictly increasing for $s > 0$. Therefore we would find $\varphi(\tilde{\beta}, s\tilde{\Lambda}) < \varphi(\tilde{\beta}, \tilde{\Lambda}) \leq \varphi(\mathbf{0}, \mathbf{1})$. But this would be in contradiction with the fact that $(\mathbf{0}, \mathbf{I})$ minimizes φ over B. □

3. Continuity of τ-functionals. Denote by $\boldsymbol{\theta}(P) = (\mathbf{t}(P), \mathbf{C}(P))$ a solution of $(\mathcal{P}_P)$. For a distribution $P \in \mathcal{P}(\mathbf{R}^p)$ and function $g : \mathbf{R}^p \to \mathbf{R}$ we will write $Pg(\cdot) = \int g(\mathbf{x})\, dP(\mathbf{x})$, or just briefly Pg if there can be no confusion about what the variable of integration is. Finally, for $\boldsymbol{\theta} = (\mathbf{t}, \mathbf{C})$ write

$$d(\mathbf{x}, \boldsymbol{\theta}) = \sqrt{(\mathbf{x}-\mathbf{t})^T\mathbf{C}^{-1}(\mathbf{x}-\mathbf{t})}. \tag{3.1}$$

We first show continuity of the functional $\boldsymbol{\theta}(\cdot)$.

THEOREM 3.1. *Let P_k, $k \geq 0$, be a sequence of distributions that converges weakly to P. Let $\mathcal{C}$ be the class of all measurable convex subsets of $\mathbf{R}^p$ and suppose that every $C \in \mathcal{C}$ is a P-continuity set, i.e. $P(\partial C) = 0$. Suppose that P satisfies (H_ϵ) for some $0 < \epsilon < 1 - r_1$ where $r_1 = b_1/a_1$, and suppose that $\boldsymbol{\theta}(P) = (\mathbf{t}(P), \mathbf{C}(P))$ is uniquely defined. Then for k sufficiently large $(\mathcal{P}_{P_k})$ has at least one solution $\boldsymbol{\theta}(P_k)$, and for any sequence of solutions $\boldsymbol{\theta}(P_k)$, $k \geq 0$, it holds that $\lim_{k\to\infty} \boldsymbol{\theta}(P_k) = \boldsymbol{\theta}(P)$.*

PROOF: The proof is along the lines of the proof of Theorem 3.2 in Lopuhaä (1989), so that a brief sketch suffices. Without loss of generality we may assume that $\boldsymbol{\theta}(P) = (\mathbf{0}, \mathbf{I})$. By means of Theorem 4.2 in Rao (1962) it follows that for k sufficiently large P_k satisfies (H_{1-r_1}), so that according to Theorem 2.1 at least one solution $\boldsymbol{\theta}(P_k) = \boldsymbol{\theta}_k = (\mathbf{t}_k, \mathbf{C}_k)$ exists. By using that $\boldsymbol{\theta}_k$ satisfies constraint (2.2), one can show that $P_k\big(E(\mathbf{t}_k, \mathbf{C}_k, c_1)\big) \geq 1 - r_1 > \epsilon$ and conclude with Theorem 4.2 in Rao (1962) that for k sufficiently large $P\big(E(\mathbf{t}_k, \mathbf{C}_k, c_1)\big) \geq \epsilon$. According to Lemma 2.2(i) this means that there exists a constant $k_1 > 0$ such that $\lambda_p(\mathbf{C}_k) \geq k_1$ eventually. By using that ρ_1 is strictly increasing on $[0, c_1]$ and that $P_k \to P$ weakly, it follows

that for each $\eta > 0$ and k sufficiently large $P_k\rho_1(\|\cdot\|/(1+\eta)) \leq b_1$. This means that the pair $(0,(1+\eta)^2\mathbf{I})$ satisfies (2.4) for k sufficiently large. Using that this holds for η arbitrarily close to 0, it follows from Remark 2.1 that

$$\limsup_{k\to\infty} |\mathbf{C}_k| \left\{ P_k\rho_2(d(\cdot,\boldsymbol{\theta}_k)) \right\}^p \leq \left\{ P\rho_2(\|\cdot\|) \right\}^p . \tag{3.2}$$

Since $\lambda_p(\mathbf{C}_k) \geq k_1$, we find with Lemma 2.1 that $\lambda_1(\mathbf{C}_k)$ is uniformly bounded above, so that with Lemma 2.2(iii) it follows that there exists a compact subset K of Θ, such that for k sufficiently large $\boldsymbol{\theta}_k$ will be in K. This means that we are finished if we can show that every convergent subsequence $\{\boldsymbol{\theta}_{k_j}\}$ has limit $(\mathbf{0},\mathbf{I})$.

Let $\boldsymbol{\theta}_{k_j}$, $j = 1,2,\ldots$, be a subsequence for which $\lim_{j\to\infty} \boldsymbol{\theta}_{k_j} = \boldsymbol{\theta}_L$. According to Lemma 3.2 in Lopuhaä (1989) it holds that $b_1 = \lim_{j\to\infty} P_{k_j}\rho_1(d(\cdot,\boldsymbol{\theta}_{k_j})) = P\rho_1(d(\cdot,\boldsymbol{\theta}_L))$. This means that $\boldsymbol{\theta}_L$ satisfies constraint (2.2) of $(\mathcal{P}_P)$. Since this problem has solution $(\mathbf{0},\mathbf{I})$ we must have $|\mathbf{C}_L|\{P\rho_2(d(\cdot,\boldsymbol{\theta}_L))\}^p \geq \{P\rho_2(\|\cdot\|)\}^p$. Then with (3.2) it follows that

$$|\mathbf{C}_L| \left\{ P\rho_2(d(\cdot,\boldsymbol{\theta}_L)) \right\}^p = \left\{ P\rho_2(\|\cdot\|) \right\}^p .$$

However, $(\mathbf{0},\mathbf{I})$ is the unique solution of $(\mathcal{P}_P)$ so that we conclude that $\boldsymbol{\theta}_L = (\mathbf{0},\mathbf{I})$. This finishes the proof. □

Continuity of the location τ-functional $\mathbf{t}(\cdot)$ is contained in Theorem 3.1. For the covariance τ-functional $\mathbf{V}(\cdot)$, continuity follows immediately from Theorem 3.1.

COROLLARY 3.1. *Under the conditions of Theorem 3.1, $\lim_{k\to\infty} \mathbf{V}(P_k) = \mathbf{V}(P)$.*

PROOF: By definition we have $\mathbf{V}(P_k) = b_2^{-1}\mathbf{C}(P_k)\{P_k\rho_2(d(\cdot,\boldsymbol{\theta}(P_k)))\}$. According to Lemma 3.2 in Lopuhaä (1989) it holds that $P_k\rho_2(d(\cdot,\boldsymbol{\theta}(P_k))) \to P\rho_2(d(\cdot,\boldsymbol{\theta}(P))$, so that by definition of $\mathbf{V}(P)$ the corollary immediately follows from Theorem 3.1. □

Consistency of the τ-estimators $(\mathbf{t}_n,\mathbf{V}_n)$ is a consequence of the continuity of the functionals $\mathbf{t}(\cdot)$ and $\mathbf{V}(\cdot)$. Let $X_1,X_2,\ldots$ be a sequence of independent random vectors in $\mathbf{R}^p$ with a distribution P. From now on denote by P_n the empirical distribution corresponding with $X_1,\ldots,X_n$.

COROLLARY 3.2. *Suppose that the distribution P satisfies the conditions of Theorem 3.1. Then $\lim_{n\to\infty}(\mathbf{t}_n,\mathbf{V}_n) = (\mathbf{t}(P),\mathbf{V}(P))$ with probability one.*

The condition that every convex set is a P-continuity set is in fact not needed in Corollary 3.2. This was needed in the proof of Theorem 3.1 merely to guarantee that the class $\mathcal{E}$ of all ellipsoids in $\mathbf{R}^p$ satisfies $\sup_{\mathcal{E}} |P_k(E) - P(E)| \to 0$. Since $\mathcal{E}$ has polynomial discrimination (Pollard 1984, p.17), for the empirical distribution P_n this property is a consequence of Theorem II.14 in Pollard (1984).

When P is an elliptical distribution with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ the conditions of Theorem 3.1 are satisfied and hence, $(\mathbf{t}_n, \mathbf{C}_n) \to (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with probability one. Therefore, if we want $\mathbf{V}_n$ to be consistent for $\boldsymbol{\Sigma}$ we must choose $b_2 = \int \rho_2(\|\mathbf{x}\|) f(\|\mathbf{x}\|)\, d\mathbf{x}$. Suppose that in general $\mathbf{C}(P)$ is considered to be the true scatter parameter to be estimated. Then one should choose

$$b_2 = \int \rho_2\left[\{(\mathbf{x} - \mathbf{t}(P))^T \mathbf{C}(P)^{-1} (\mathbf{x} - \mathbf{t}(P))\}^{1/2}\right] dP(\mathbf{x}) \tag{3.3}$$

for $\mathbf{V}_n$ to be consistent for $\mathbf{C}(P)$.

4. Limiting distribution. We first investigate the asymptotic behaviour of $(\mathbf{t}_n, \mathbf{C}_n)$. The limiting distribution of the actual τ-estimators $(\mathbf{t}_n, \mathbf{V}_n)$ will then follow from that of $(\mathbf{t}_n, \mathbf{C}_n)$. We assume that P satisfies property (H_ε) for some $0 < \varepsilon < 1 - r_1$ and that the minimization problem $(\mathcal{P}_P)$ has a unique solution $\boldsymbol{\theta}_0 = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$. In order to let $(\mathbf{t}_n, \mathbf{V}_n)$ be consistent for $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ we take b_2 as in (3.3).

To study the limiting behaviour of solutions $(\mathbf{t}_n, \mathbf{C}_n)$ of $(\mathcal{P}_{P_n})$ we first show that $(\mathbf{t}_n, \mathbf{C}_n)$ are related to multivariate M-estimators as defined in Huber (1981). To do so, it will be more convenient to consider $(\mathbf{t}_n, \mathbf{C}_n)$ as solutions to the problem of finding a vector $\mathbf{t}$ and a positive definite symmetric matrix $\mathbf{C}$ that minimize

$$\log(|\mathbf{C}|) + p \log\left\{\frac{1}{n}\sum_{i=1}^{n} \rho_2\left[\{(X_i - \mathbf{t})^T \mathbf{C}^{-1} (X_i - \mathbf{t})\}^{1/2}\right]\right\}$$

subject to

$$\frac{1}{n}\sum_{i=1}^{n} \rho_1\left[\{(X_i - \mathbf{t})^T \mathbf{C}^{-1} (X_i - \mathbf{t})\}^{1/2}\right] = b_1. \tag{4.1}$$

This problem is equivalent to the one considered in Section 2.1 and will be referred to as minimization problem $(\mathcal{P}_{P_n})$ from now on. Because of the frequent appearance of quadratic forms we will write d_i for $\sqrt{(X_i - \mathbf{t})^T \mathbf{C}^{-1} (X_i - \mathbf{t})}$ for $i = 1, \ldots, n$ and d for $\sqrt{(\mathbf{x} - \mathbf{t})^T \mathbf{C}^{-1} (\mathbf{x} - \mathbf{t})}$.

4.1. Relation to M-estimators. The Lagrangean corresponding to minimization problem $(\mathcal{P}_{P_n})$ is

$$L_n(\mathbf{t}, \mathbf{C}, \lambda) = \log(|\mathbf{C}|) + p \log\left\{\frac{1}{n}\sum_{i=1}^{n} \rho_2(d_i)\right\} - \lambda\left\{\frac{1}{n}\sum_{i=1}^{n} \rho_1(d_i) - b_1\right\}$$

Every solution $(\mathbf{t}_n, \mathbf{C}_n)$ of $(\mathcal{P}_{P_n})$ must be a zero of all partial derivatives of L_n. It then follows that besides constraint (4.1) every solution $(\mathbf{t}_n, \mathbf{C}_n)$ of $(\mathcal{P}_{P_n})$ must also be a solution of the simultaneous equations

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left\{\left(\frac{1}{n}\sum_{j=1}^{n}\rho_2(d_j)\right)^{-1}\frac{p\psi_2(d_i)}{d_i} - \frac{\lambda\psi_1(d_i)}{d_i}\right\}(X_i - \mathbf{t}) = \mathbf{0} \\
&\frac{1}{2n}\sum_{i=1}^{n}\left\{\left(\frac{1}{n}\sum_{j=1}^{n}\rho_2(d_j)\right)^{-1}\frac{p\psi_2(d_i)}{d_i} - \frac{\lambda\psi_1(d_i)}{d_i}\right\}(X_i - \mathbf{t})(X_i - \mathbf{t})^T = \mathbf{C}.
\end{aligned} \tag{4.2}$$

We can eliminate λ from (4.2) by multiplying the second (matrix) equation with $\mathbf{C}^{-1}$, take traces and solve for λ. We find

$$\lambda_n = \frac{(-2p/n)\sum_{i=1}^n \rho_2(d_i) + (p/n)\sum_{i=1}^n \psi_2(d_i)d_i}{(1/n)\sum_{i=1}^n \rho_2(d_i)(1/n)\sum_{i=1}^n \psi_1(d_i)d_i}.$$

After we substitute λ_n in (4.2) we obtain simultaneous equations in $\mathbf{t}$ and $\mathbf{C}$.

To keep things tidy we first have to fix some more notation before we list these equations. Define the functions

$$a(\mathbf{x}, \boldsymbol{\theta}) = 2\rho_2(d) - \psi_2(d)d$$
$$b(\mathbf{x}, \boldsymbol{\theta}) = \psi_1(d)d$$

where d is the abbreviation defined before. Let $A_n(\boldsymbol{\theta}) = P_n a(\cdot, \boldsymbol{\theta})$, $B_n(\boldsymbol{\theta}) = P_n b(\cdot, \boldsymbol{\theta})$ and $A(\boldsymbol{\theta}) = Pa(\cdot, \boldsymbol{\theta})$, $B(\boldsymbol{\theta}) = Pb(\cdot, \boldsymbol{\theta})$. Note that because P satisfies (H_ϵ) we have that $A(\boldsymbol{\theta}) > 0$, and since $\boldsymbol{\theta}_0$ satisfies (2.2) the ellipsoid $E(\boldsymbol{\mu}, \boldsymbol{\Sigma}, c_1)$ must have positive probability, which means that $B(\boldsymbol{\theta}_0) > 0$.

The simultaneous equations that arise after substitution of λ_n in equations (4.2) are perhaps described most conveniently with the function

$$\psi_n(\cdot, \boldsymbol{\theta}) = A_n(\boldsymbol{\theta})\psi_1(\cdot) + B_n(\boldsymbol{\theta})\psi_2(\cdot) \tag{4.3}$$

which is an adaptively weighted average of the functions ψ_1 and ψ_2. We obtain the equations :

$$\sum_{i=1}^n \frac{\psi_n(d_i, \boldsymbol{\theta})}{d_i}(X_i - \mathbf{t}) = \mathbf{0}$$
$$\sum_{i=1}^n \left\{ p\frac{\psi_n(d_i, \boldsymbol{\theta})}{d_i}(X_i - \mathbf{t})(X_i - \mathbf{t})^T - \psi_n(d_i, \boldsymbol{\theta})d_i\mathbf{C} \right\} = \mathbf{0}.$$

This is of course a system of linear dependent equations. However, by adding a suitable multiple of the constraint (4.1) to the second (matrix) equation we can avoid the linear dependence. It follows that every solution $\boldsymbol{\theta}_n = (\mathbf{t}_n, \mathbf{C}_n)$ of $(\mathcal{P}_{P_n})$ will always be a solution of the simultaneous equations

$$\sum_{i=1}^n \frac{\psi_n(d_i, \boldsymbol{\theta})}{d_i}(X_i - \mathbf{t}) = \mathbf{0}$$
$$\sum_{i=1}^n \left\{ p\frac{\psi_n(d_i, \boldsymbol{\theta})}{d_i}(X_i - \mathbf{t})(X_i - \mathbf{t})^T - \left(\psi_n(d_i, \boldsymbol{\theta})d_i - 2b_2(\rho_1(d_i) - b_1)\right)\mathbf{C} \right\} = \mathbf{0}. \tag{4.4}$$

These equations look like the M-estimator type score equations as defined in Huber (1981), except that the function $\psi_n(\cdot, \boldsymbol{\theta})$ of (4.3) is an adaptively weighted average which itself depends on $X_1, \ldots, X_n$.

Although $\boldsymbol{\theta}_n$ is a solution of equations (4.4) defined with the function $\psi_n(\cdot, \boldsymbol{\theta})$, only the limiting expression of $\psi_n(\cdot, \boldsymbol{\theta}_n)$ is of importance for the asymptotic behaviour of $\boldsymbol{\theta}_n$. We will see in the next subsection that the function $\psi_n(\cdot, \boldsymbol{\theta}_n)$ converges with probability one pointwise to the function

$$\tilde{\psi}(\cdot) = A(\boldsymbol{\theta}_0)\psi_1(\cdot) + B(\boldsymbol{\theta}_0)\psi_2(\cdot). \tag{4.5}$$

The limiting distribution of $(\mathbf{t}_n, \mathbf{C}_n)$ will then be shown to be the same as that of multivariate M-estimators that are a solution of equations exactly like (4.4) except with the function $\tilde{\psi}(\cdot)$ instead of the function $\psi_n(\cdot, \boldsymbol{\theta})$. The limiting distribution of the actual τ-estimators $(\mathbf{t}_n, \mathbf{V}_n)$ can then be obtained from that of $(\mathbf{t}_n, \mathbf{C}_n)$.

4.2. Asymptotic normality. We will use the following tightness property from Pollard (1984) for empirical processes $(P_n - P)\phi$ indexed by functions ϕ in a class $\mathcal{F}$. It is a combination of the Approximation Lemma (p.27), Lemma II.36 (p.36) and the Equicontinuity Lemma (p.150). By the *envelope* F of $\mathcal{F}$ is meant a function F for which $|\phi| \leq F$ for every $\phi \in \mathcal{F}$.

LEMMA 4.1. *Let $\mathcal{F}$ be a permissible class of real valued functions with envelope F and suppose that $0 < PF^2 < \infty$. If the class of graphs of functions in $\mathcal{F}$ has polynomial discrimination, then for each $\eta > 0$ and $\varepsilon > 0$ there exists a $\delta > 0$ for which*

$$\limsup_{n\to\infty} \mathbf{P}\left\{ \sup_{\phi_1,\phi_2\in[\delta]} \left|\sqrt{n}(P_n - P)(\phi_1 - \phi_2)\right| > \eta \right\} < \varepsilon$$

where $[\delta] = \{(\phi_1, \phi_2) : \phi_1, \phi_2 \in \mathcal{F} \text{ and } P(\phi_1 - \phi_2)^2 \leq \delta^2\}$.

The classes of functions that we will encounter will always be indexed by the parameter set Θ, which can be seen as a subset of $\mathbf{R}^{p+\frac{1}{2}p(p+1)}$. This means that these classes will always be permissible in the sense of Pollard (1984, Appendix C). Also the corresponding classes of graphs will have polynomial discrimination (Pollard 1984, p.17) as will follow from the next lemma.

LEMMA 4.2. *Let $d(\mathbf{x}, \boldsymbol{\theta})$ be defined as in (3.1). Consider the class of functions $\mathcal{A} = \{2\rho_2\big(d(\cdot,\boldsymbol{\theta})\big) - \psi_2\big(d(\cdot,\boldsymbol{\theta})\big)d(\cdot,\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$, and for $k = 1, 2$ the classes $\mathcal{R}_k = \{\rho_k\big(d(\cdot,\boldsymbol{\theta})\big) : \boldsymbol{\theta} \in \Theta\}$, $\mathcal{U}_k = \{\psi_k\big(d(\cdot,\boldsymbol{\theta})\big)/d(\cdot,\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ and $\mathcal{W}_k = \{\psi_k\big(d(\cdot,\boldsymbol{\theta})\big)d(\cdot,\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$. Then the classes of functions in $\mathcal{A}$, $\mathcal{R}_1$, $\mathcal{R}_2$, $\mathcal{U}_1$, $\mathcal{U}_2$, $\mathcal{W}_1$ and $\mathcal{W}_2$ have polynomial discrimination.*

PROOF: Apply Lemma 22 of Nolan and Pollard (1987), which states that for any function $g : [0, \infty) \to \mathbf{R}$, which is of bounded variation, the class of graphs of the functions $g\big((\mathbf{x}-\mathbf{t})^T\mathbf{C}^{-1}(\mathbf{x}-\mathbf{t})\big)$ for $(\mathbf{t}, \mathbf{C}) \in \Theta$ has polynomial discrimination. Since ρ_1 and ρ_2 can be written as the sum of two monotone functions, it follows immediately that the classes $\mathcal{R}_1$ and $\mathcal{R}_2$ have polynomial discrimination. From (R1)-(R2) it follows that the functions $\psi_k(y)y$ and $\psi_k(y)/y$ for $k = 1, 2$ are continuous, that they vanish for $y \to -\infty$ and that they are almost everywhere differentiable with an absolutely integrable derivative. This implies that these functions are also of bounded variation. Hence, $\mathcal{A}$, $\mathcal{U}_1$, $\mathcal{U}_2$, $\mathcal{W}_1$ and $\mathcal{W}_2$ have polynomial discrimination. □

For the weights $A_n(\boldsymbol{\theta}_n)$ and $B_n(\boldsymbol{\theta}_n)$ of the function $\psi_n(\cdot, \boldsymbol{\theta}_n)$ the following result is an immediate consequence of Lemma 4.2.

LEMMA 4.3. *Let $\boldsymbol{\theta}_n = (\mathbf{t}_n, \mathbf{C}_n)$ be a solution of $(\mathcal{P}_{P_n})$. Then $A_n(\boldsymbol{\theta}_n) \to A(\boldsymbol{\theta}_0)$ and $B_n(\boldsymbol{\theta}_n) \to B(\boldsymbol{\theta}_0)$ with probability one.*

PROOF: Consider the class $\mathcal{A} = \{a(\cdot, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$. It has a bounded envelope and according to Lemma 4.2 the corresponding class of graphs has polynomial discrimination. We may therefore apply Theorem II.24 in combination with the Approximation Lemma in Pollard (1984), and it follows that $(P_n - P)a(\cdot, \boldsymbol{\theta}_n) \to 0$ with probability one. Because $A_n(\boldsymbol{\theta}_n) = A(\boldsymbol{\theta}_n) + (P_n - P)a(\cdot, \boldsymbol{\theta}_n)$ and because $a(\mathbf{x}, \boldsymbol{\theta})$ is continuous it follows that $A_n(\boldsymbol{\theta}_n) \to A(\boldsymbol{\theta}_0)$ with probability one. The other statement can be shown similarly. □

We can write equations (4.4) briefly as

$$\sum_{i=1}^{n} \boldsymbol{\Psi}_n(X_i, \boldsymbol{\theta}) = \mathbf{0}. \tag{4.6}$$

The function $\boldsymbol{\Psi}_n$ is then a weighted average $A_n(\boldsymbol{\theta})\boldsymbol{\Psi}_1(\mathbf{x}, \boldsymbol{\theta}) + B_n(\boldsymbol{\theta})\boldsymbol{\Psi}_2(\mathbf{x}, \boldsymbol{\theta}) - 2b_2\mathbf{R}(\mathbf{x}, \boldsymbol{\theta})$ of the functions $\boldsymbol{\Psi}_k = (\boldsymbol{\Psi}_{k,\mathrm{loc}}, \boldsymbol{\Psi}_{k,\mathrm{cov}})$, where

$$\begin{cases} \boldsymbol{\Psi}_{k,\mathrm{loc}}(\mathbf{x}, \boldsymbol{\theta}) = \dfrac{\psi_k(d)}{d}(\mathbf{x} - \mathbf{t}) \\ \boldsymbol{\Psi}_{k,\mathrm{cov}}(\mathbf{x}, \boldsymbol{\theta}) = p\dfrac{\psi_k(d)}{d}(\mathbf{x} - \mathbf{t})(\mathbf{x} - \mathbf{t})^T - \psi_k(d)d\,\mathbf{C} \end{cases} \tag{4.7}$$

for $k = 1, 2$, and of the function $\mathbf{R} = (\mathbf{R}_{\mathrm{loc}}, \mathbf{R}_{\mathrm{cov}})$, where

$$\begin{cases} \mathbf{R}_{\mathrm{loc}}(\mathbf{x}, \boldsymbol{\theta}) = \mathbf{0} \\ \mathbf{R}_{\mathrm{cov}}(\mathbf{x}, \boldsymbol{\theta}) = (\rho_1(d) - b_1)\mathbf{C}. \end{cases} \tag{4.8}$$

Let $\tilde{\boldsymbol{\Psi}} = (\tilde{\boldsymbol{\Psi}}_{\mathrm{loc}}, \tilde{\boldsymbol{\Psi}}_{\mathrm{cov}})$ be the function that is exactly like $\boldsymbol{\Psi}_n$, except that instead of the function $\psi_n(\cdot, \boldsymbol{\theta})$ it is defined with the function $\tilde{\psi}(\cdot)$ of (4.5) :

$$\begin{cases} \tilde{\boldsymbol{\Psi}}_{\mathrm{loc}}(\mathbf{x}, \boldsymbol{\theta}) = \dfrac{\tilde{\psi}(d)}{d}(\mathbf{x} - \mathbf{t}) \\ \tilde{\boldsymbol{\Psi}}_{\mathrm{cov}}(\mathbf{x}, \boldsymbol{\theta}) = p\dfrac{\tilde{\psi}(d)}{d}(\mathbf{x} - \mathbf{t})(\mathbf{x} - \mathbf{t})^T - \Big(\tilde{\psi}(d)d - 2b_2\big(\rho_1(d) - b_1\big)\Big)\mathbf{C}. \end{cases} \tag{4.9}$$

REMARK 4.1: Consider the problem $(\mathcal{P}_P)$ with solution $\boldsymbol{\theta}(P)$. Similar to differentiating L_n we can also differentiate the Lagrangean L_P corresponding with $(\mathcal{P}_P)$. Because the functions $\psi_k(y)$ and $\psi_k(y)y$ for $k = 1, 2$ are bounded we may change the order of integration and differentiation at least on a bounded neighbourhood of $\boldsymbol{\theta}(P)$ and conclude that $\boldsymbol{\theta}(P)$ is a zero of the equation

$$A(\boldsymbol{\theta})P\boldsymbol{\Psi}_1(\cdot, \boldsymbol{\theta}) + B(\boldsymbol{\theta})P\boldsymbol{\Psi}_2(\cdot, \boldsymbol{\theta}) - 2b_2P\mathbf{R}(\cdot, \boldsymbol{\theta}) = \mathbf{0}.$$

In particular this means that $P\tilde{\boldsymbol{\Psi}}(\cdot, \boldsymbol{\theta}_0) = \mathbf{0}$.

Since Lemma 4.1 is stated only for real valued functions we will apply it to each component of the functions Ψ_1, Ψ_2 and $\mathbf{R}$ separately. The following lemma implies that the classes of graphs of the functions which are then involved have polynomial discrimination.

LEMMA 4.4. *For $\mathbf{x} = (x_1 \cdots x_p)^T$ and $\theta \in \Theta$ let $g(\mathbf{x}, \theta)$ be a real valued function. Consider the classes functions $\mathcal{F} = \{g(\mathbf{x}, \theta) : \theta \in \Theta\}$, $\mathcal{F}_j = \{g(\mathbf{x}, \theta)x_j : \theta \in \Theta\}$ and $\mathcal{F}_{ij} = \{g(\mathbf{x}, \theta)x_i x_j : \theta \in \Theta\}$ for $i, j = 1, \ldots, p$. Denote by $\mathcal{G}$, $\mathcal{G}_j$ and $\mathcal{G}_{ij}$ the corresponding classes of graphs of functions in $\mathcal{F}$, $\mathcal{F}_j$ and $\mathcal{F}_{ij}$ respectively. If $\mathcal{G}$ has polynomial discrimination, then also $\mathcal{G}_j$ and $\mathcal{G}_{ij}$ have polynomial discrimination for $i, j = 1, \ldots, p$.*

PROOF: The lemma is an immediate consequence of Lemma 5.2 in Lopuhaä (1988). □

LEMMA 4.5. *Suppose that $\mathbf{E}\|X_1\|^4$ is finite. Let $\Psi_k(\mathbf{x}, \theta)$ for $k = 1, 2$ and $\mathbf{R}(\mathbf{x}, \theta)$ be defined in (4.7) and (4.8). Then*

$$P_n\Psi_k(\cdot, \theta_n) = P\Psi_k(\cdot, \theta_n) + (P_n - P)\Psi_k(\cdot, \theta_0) + o_P(1/\sqrt{n}).$$
$$P_n\mathbf{R}(\cdot, \theta_n) = P\mathbf{R}(\cdot, \theta_n) + (P_n - P)\mathbf{R}(\cdot, \theta_0) + o_P(1/\sqrt{n}).$$

PROOF: We only prove the lemma for Ψ_1 as the other cases can be shown similarly. We first deal with the matrix part $\Psi_{1,\text{cov}}$ of the function Ψ_1. Consider the (i, j)-th element of $\Psi_{1,\text{cov}}$ (where d as defined earlier) :

$$\Psi_{1,\text{cov},ij}(\mathbf{x}, \theta) = p\frac{\psi_1(d)}{d}(x_i - t_i)(x_j - t_j) - \psi_1(d)d\,c_{ij}. \tag{4.10}$$

Write $g(\mathbf{x}, \theta) = p\,\psi_1(d)/d$ and $h(\mathbf{x}, \theta) = \psi_1(d)d$. Consider the first term of (4.10) and decompose $g(\mathbf{x}, \theta_n)(x_i - t_{ni})(x_j - t_{nj}) - g(\mathbf{x}, \theta_0)(x_i - \mu_i)(x_j - \mu_j)$ as follows :

$$\begin{aligned} &g(\mathbf{x}, \theta_n)x_i x_j - g(\mathbf{x}, \theta_0)x_i x_j \\ &- t_{ni}\{g(\mathbf{x}, \theta_n)x_j - g(\mathbf{x}, \theta_0)x_j\} - t_{nj}\{g(\mathbf{x}, \theta_n)x_i - g(\mathbf{x}, \theta_0)x_i\} \\ &+ t_{ni}t_{nj}\{g(\mathbf{x}, \theta_n) - g(\mathbf{x}, \theta_0)\} \\ &- (t_{ni} - \mu_i)g(\mathbf{x}, \theta_0)x_j - (t_{nj} - \mu_j)g(\mathbf{x}, \theta_0)x_i + (t_{ni}t_{nj} - \mu_i\mu_j)g(\mathbf{x}, \theta_0). \end{aligned} \tag{4.11}$$

Consider the first term of (4.11). According to Lemma 4.2 the class of graphs of the functions $\{g(\mathbf{x}, \theta) : \theta \in \Theta\}$ has polynomial discrimination. Hence, according to Lemma 4.4 the class of graphs of the functions $\{g(\mathbf{x}, \theta)x_i x_j : \theta \in \Theta\}$ has polynomial discrimination. Since g is bounded and $\mathbf{E}\|X_1\|^4$ is finite this means that Lemma 4.1 applies to the first term of (4.11). As g is continuous, $g(\mathbf{x}, \theta_n)x_i x_j$ tends to $g(\mathbf{x}, \theta_0)x_i x_j$, which means that for any $\delta > 0$ both functions are in the class $[\delta]$ of Lemma 4.1 for n sufficiently large. It follows that if we integrate (4.11) with respect to $P_n - P$, the first term is $o_P(1/\sqrt{n})$. Similarly the next three terms will be $o_P(1/\sqrt{n})$. Because $\mathbf{t}_n \to \boldsymbol{\mu}$, the central limit theorem implies that also the last three terms of (4.11) will be $o_P(1/\sqrt{n})$. By a similar reasoning one can show for the second term of (4.10) that $(P_n - P)\big(h(\cdot, \theta_n)c_{nij} - h(\cdot, \theta_0)\sigma_{ij}\big) = o_P(1/\sqrt{n})$ and

it follows that $(P_n - P)\big(\Psi_{\mathrm{cov},1,ij}(\cdot,\theta_n) - \Psi_{\mathrm{cov},1,ij}(\cdot,\theta_0)\big) = o_P(1/\sqrt{n})$. This holds for each pair (i,j). Similarly one can show for each component of the location part $\Psi_{1,\mathrm{loc}}$ of the function Ψ_1 that $(P_n - P)\big(\Psi_{1,\mathrm{loc},i}(\cdot,\theta_n) - \Psi_{1,\mathrm{loc},i}(\cdot,\theta_0)\big) = o_P(1/\sqrt{n})$. Putting all parts together gives

$$(P_n - P)\Big(\Psi_1(\cdot,\theta_n) - \Psi_1(\cdot,\theta_0)\Big) = o_P(1/\sqrt{n})$$

which proves the lemma for Ψ_1. The other two cases can be shown similarly. □

The following theorem states that the limiting behaviour of each solution θ_n of the problem $(\mathcal{P}_{P_n})$ is the same as that of multivariate M-estimators, defined as a solution of the equation $\sum_{i=1}^n \tilde{\Psi}(X_i,\theta) = 0$, where $\tilde{\Psi}$ is given in (4.9).

THEOREM 4.1. *Suppose that* $\mathbf{E}\|X_1\|^4$ *is finite and that the function* $P\tilde{\Psi}(\cdot,\theta)$ *has a nonsingular derivative* $\tilde{\Lambda}$ *at* θ_0. *Let* $P\Psi_k(\cdot,\theta_0) = 0$ *for* $k = 1,2$ *and let* $\theta_n = (\mathrm{t}_n, \mathrm{C}_n)$ *be a solution of* $(\mathcal{P}_{P_n})$ *for* $n = 1,2,\ldots$. *Then it holds that*

$$\tilde{\Lambda}(\theta_n - \theta_0) = -\frac{1}{n}\sum_{i=1}^n \tilde{\Psi}(X_i,\theta_0) + o_P(1/\sqrt{n}). \tag{4.12}$$

PROOF: We first determine the rate of convergence of θ_n. From equation (4.6) and definition (4.9) it follows with Lemma 4.3 that

$$\begin{aligned} 0 &= A_n(\theta_n)\,P_n\Psi_1(\cdot,\theta_n) + B_n(\theta_n)\,P_n\Psi_2(\cdot,\theta_n) - 2b_2\,P_n\mathbf{R}(\cdot,\theta_n) \\ &= P_n\tilde{\Psi}(\cdot,\theta_n) + o_P(1)\,P_n\Psi_1(\cdot,\theta_n) + o_P(1)\,P_n\Psi_2(\cdot,\theta_n). \end{aligned} \tag{4.13}$$

By assumption $P\tilde{\Psi}(\cdot,\theta)$ is differentiable at θ_0 and according to Remark 4.1 we have that $P\tilde{\Psi}(\cdot,\theta_0) = 0$. Hence, $P\tilde{\Psi}(\cdot,\theta_n) = \tilde{\Lambda}(\theta_n - \theta_0) + o_P\big(\|\theta_n - \theta_0\|\big)$. With Lemma 4.5 we then obtain

$$P_n\tilde{\Psi}(\cdot,\theta_n) = \big(\tilde{\Lambda} + o_P(1)\big)(\theta_n - \theta_0) + (P_n - P)\tilde{\Psi}(\cdot,\theta_0) + o_P(1/\sqrt{n}). \tag{4.14}$$

For $k = 1,2$ the conditions on ρ_k imply that $P\Psi_k(\cdot,\theta)$ is differentiable at θ_0, so that similar to (4.14) we find that

$$o_P(1)\,P_n\Psi_1(\cdot,\theta_n) + o_P(1)\,P_n\Psi_2(\cdot,\theta_n) = o_P\big(\|\theta_n - \theta_0\|\big) + o_P(1/\sqrt{n}) \tag{4.15}$$

where we use that $(P_n - P)\Psi_k(\cdot,\theta_0) = O_P(1/\sqrt{n})$ according to the central limit theorem. By putting together (4.13), (4.14) and (4.15) we find that

$$0 = \big(\tilde{\Lambda} + o_P(1)\big)(\theta_n - \theta_0) + (P_n - P)\tilde{\Psi}(\cdot,\theta_0) + o_P(1/\sqrt{n}). \tag{4.16}$$

According to the central limit theorem $(P_n - P)\tilde{\Psi}(\cdot,\theta_0) = O_P(1/\sqrt{n})$, so that (4.16) boils down to

$$0 = \big(\tilde{\Lambda} + o_P(1)\big)(\theta_n - \theta_0) + O_P(1/\sqrt{n}).$$

Since $\tilde{\Lambda}$ is nonsingular it follows that $\theta_n - \theta_0 = O_P(1/\sqrt{n})$.

If we put this into (4.16), this equation reduces to

$$0 = \tilde{\Lambda}(\theta_n - \theta_0) + (P_n - P)\tilde{\Psi}(\cdot,\theta_0) + o_P(1/\sqrt{n})$$

which proves the theorem as $P\tilde{\Psi}(\cdot,\theta_0) = 0$. □

The simultaneous limiting distribution of the actual τ-estimators $(\mathbf{t}_n, \mathbf{V}_n)$ may now be obtained from Theorem 4.1 by means of expressing $\mathbf{V}_n$ in terms of $\boldsymbol{\theta}_n - \boldsymbol{\theta}_0$. The latter is done in the next lemma.

LEMMA 4.6. *Let $R_2(\mathbf{x}, \boldsymbol{\theta}) = \rho_2(d) - b_2$ where d is the abbreviation defined before. Consider the function $P R_2(\cdot, \boldsymbol{\theta})$ and let $\boldsymbol{\Delta}_2$ be the derivative of $P R_2(\cdot, \boldsymbol{\theta})$ at $\boldsymbol{\theta}_0$. Then*

$$\mathbf{V}_n = \mathbf{C}_n + b_2^{-1}\boldsymbol{\Sigma}\,(P_n - P)R_2(\cdot, \boldsymbol{\theta}_0) + b_2^{-1}\boldsymbol{\Sigma}\,\boldsymbol{\Delta}_2(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0) + o_P(1/\sqrt{n}).$$

PROOF: By definition we have

$$\mathbf{V}_n = \mathbf{C}_n + b_2^{-1}\boldsymbol{\Sigma}\,P_n R_2(\cdot, \boldsymbol{\theta}_n) + b_2^{-1}(\mathbf{C}_n - \boldsymbol{\Sigma})\,P_n R_2(\cdot, \boldsymbol{\theta}_n). \tag{4.17}$$

According to Lemma 4.2 the class of graphs of the functions $\{\rho_2(d(\cdot, \boldsymbol{\theta})) : \boldsymbol{\theta} \in \Theta\}$ has polynomial discrimination. Because ρ_2 is bounded we may apply Lemma 4.1 and conclude that $P_n R_2(\cdot, \boldsymbol{\theta}_n) = P R_2(\cdot, \boldsymbol{\theta}_n) + (P_n - P)R_2(\cdot, \boldsymbol{\theta}_0) + o_P(1/\sqrt{n})$. Because b_2 is defined as in (3.3), we have that $P R_2(\cdot, \boldsymbol{\theta}_0) = \mathbf{0}$. Furthermore, $P R_2(\cdot, \boldsymbol{\theta})$ is differentiable at $\boldsymbol{\theta}_0$ and therefore it holds that

$$P_n R_2(\cdot, \boldsymbol{\theta}_n) = \boldsymbol{\Delta}_2(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0) + o_P(\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_0\|) + (P_n - P)R_2(\cdot, \boldsymbol{\theta}_0) + o_P(1/\sqrt{n}). \tag{4.18}$$

Theorem 4.1 implies $\boldsymbol{\theta}_n - \boldsymbol{\theta}_0 = O_P(1/\sqrt{n})$, so that the lemma follows from (4.17) and (4.18). □

To apply Theorem 4.1 and to obtain a limiting theorem for $(\mathbf{t}_n, \mathbf{V}_n)$ it will be more convenient to consider the derivatives $\tilde{\boldsymbol{\Lambda}}$ and $\boldsymbol{\Delta}_2$ as linear mappings on Θ. These mappings are described in the next lemma.

LEMMA 4.7. *Let $d_0 = \sqrt{(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})}$ and let $\omega_k = \int \psi_k(d_0) d_0\, dP(\mathbf{x}) > 0$ for $k = 1, 2$. Let $P\boldsymbol{\Psi}_k(\cdot, \boldsymbol{\theta}_0) = \mathbf{0}$ for $k = 1, 2$. Let $\tilde{\boldsymbol{\Lambda}}$ and $\boldsymbol{\Delta}_2$ be the derivatives defined in Theorem 4.1 and Lemma 4.6. Write $\tilde{u}(y) = \tilde{\psi}(y)/y$ and $\tilde{g}(y) = \tilde{\psi}(y)y - 2b_2(\rho_1(y) - b_1)$. Then $\tilde{\boldsymbol{\Lambda}}$ is the linear mapping $(\tilde{\boldsymbol{\Lambda}}_{\text{loc}}, \tilde{\boldsymbol{\Lambda}}_{\text{cov}})$, where $\tilde{\boldsymbol{\Lambda}}_{\text{loc}}$ is the linear mapping that maps $(\mathbf{t}, \mathbf{C})$ to*

$$-\int \frac{\tilde{u}'(d_0)}{2d_0}\left\{(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}\mathbf{C}\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) + 2(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}\mathbf{t}\right\}(\mathbf{x} - \boldsymbol{\mu})\, dP(\mathbf{x})$$
$$-\int \tilde{u}(d_0)\mathbf{t}\, dP(\mathbf{x})$$

$\tilde{\boldsymbol{\Lambda}}_{\text{cov}}$ is the linear mapping that maps $(\mathbf{t}, \mathbf{C})$ to

$$-\int \frac{p\tilde{u}'(d_0)}{2d_0}\left\{(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}\mathbf{C}\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) + 2(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}\mathbf{t}\right\}(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T\, dP(\mathbf{x})$$
$$+\int \frac{\tilde{g}'(d_0)}{2d_0}\left\{(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}\mathbf{C}\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) + 2(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}\mathbf{t}\right\}\boldsymbol{\Sigma}\, dP(\mathbf{x}) - 2b_2\omega_1\mathbf{C}$$

and Δ_2 is the linear mapping that maps $(\mathbf{t}, \mathbf{C})$ to $-(2p)^{-1}\omega_2 trace(\mathbf{C}\boldsymbol{\Sigma}^{-1})$.

PROOF: First consider the derivative with respect to $\boldsymbol{\theta}$ of the function $\tilde{\boldsymbol{\Psi}}$ at $\boldsymbol{\theta}_0$. This is the linear mapping $\mathrm{D}\tilde{\boldsymbol{\Psi}}(\mathbf{x}, \boldsymbol{\theta}_0) = \big(\mathrm{D}\tilde{\boldsymbol{\Psi}}_{\mathrm{loc}}(\mathbf{x}, \boldsymbol{\theta}_0), \mathrm{D}\tilde{\boldsymbol{\Psi}}_{\mathrm{cov}}(\mathbf{x}, \boldsymbol{\theta}_0)\big)$.

The first component is the linear mapping $\mathrm{D}\tilde{\boldsymbol{\Psi}}_{\mathrm{loc}}(\mathbf{x}, \boldsymbol{\theta}_0)$ which maps $(\mathbf{t}, \mathbf{C})$ to

$$-\frac{\tilde{u}'(d_0)}{2d_0}\Big\{(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\mathbf{C}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) + 2(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\mathbf{t}\Big\}(\mathbf{x}-\boldsymbol{\mu}) - \tilde{u}(d_0)\mathbf{t}.$$

The second component of $\mathrm{D}\tilde{\boldsymbol{\Psi}}(\mathbf{x}, \boldsymbol{\theta}_0)$ is the linear map $\mathrm{D}\tilde{\boldsymbol{\Psi}}_{\mathrm{cov}}(\mathbf{x}, \boldsymbol{\theta}_0)$ which maps $(\mathbf{t}, \mathbf{C})$ to

$$\begin{aligned}
&-\frac{p\tilde{u}'(d_0)}{2d_0}\Big\{(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\mathbf{C}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) + 2(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\mathbf{t}\Big\}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T \\
&- p\tilde{u}(d_0)\Big\{\mathbf{t}(\mathbf{x}-\boldsymbol{\mu})^T + (\mathbf{x}-\boldsymbol{\mu})\mathbf{t}^T\Big\} \\
&+ \frac{\tilde{g}'(d_0)}{2d_0}\Big\{(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\mathbf{C}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) + 2(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\mathbf{t}\Big\}\boldsymbol{\Sigma} - \tilde{g}(d_0)\mathbf{C}
\end{aligned}$$

If we integrate both components with respect to P the conditions on ψ_1 and ψ_2 ensure that we may interchange differentiation and integration. The expression for $\tilde{\boldsymbol{\Lambda}}_{\mathrm{loc}}$ follows immediately. According to Remark 4.1 $P\tilde{\boldsymbol{\Psi}}_{\mathrm{loc}}(\cdot, \boldsymbol{\theta}_0) = \mathbf{0}$ so that the second term of $\mathrm{D}\tilde{\boldsymbol{\Psi}}_{\mathrm{cov}}(\mathbf{x}, \boldsymbol{\theta}_0)$ vanishes if we integrate with respect to P. By definition of $\tilde{\psi}$ it follows that $\int \tilde{g}(d_0)\, dP(\mathbf{x})$ reduces to $2b_2\omega_1$.

Finally, the derivative with respect to $\boldsymbol{\theta}$ of the function $R_2(\mathbf{x}, \boldsymbol{\theta})$ of Lemma 4.6 at $\boldsymbol{\theta}_0$, is the linear mapping that maps $(\mathbf{t}, \mathbf{C})$ to

$$-\frac{\psi_2(d_0)}{2d_0}\Big\{(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\mathbf{C}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) + 2(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\mathbf{t}\Big\}. \tag{4.19}$$

If we integrate this with respect to P we may again interchange differentiation and integration. Note that

$$(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\mathbf{C}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = \mathrm{trace}\big((\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\mathbf{C}\boldsymbol{\Sigma}^{-1}\big) \tag{4.20}$$

and that $P\boldsymbol{\Psi}_2(\cdot, \boldsymbol{\theta}_0) = \mathbf{0}$. This means that if we integrate (4.19) with respect to P, the term with $\mathbf{t}$ vanishes whereas the first term reduces to $-(2p)^{-1}\omega_2\mathrm{trace}(\mathbf{C}\boldsymbol{\Sigma}^{-1})$. □

The limiting distribution of the τ-estimators will be an immediate consequence of the next theorem.

THEOREM 4.2. *Let $\mathbf{R}_{12}(\mathbf{x}, \boldsymbol{\theta}) = \omega_1\big(\rho_2(d) - b_2\big) - \omega_2\big(\rho_1(d) - b_1\big)$, where d as defined in the beginning of this section and ω_1 and ω_2 as defined in Lemma 4.7. Under the conditions of Theorem 4.1 it holds that*

$$\tilde{\boldsymbol{\Lambda}}(\boldsymbol{\tau}_n - \boldsymbol{\theta}_0) = -\frac{1}{n}\sum_{i=1}^{n}\mathbf{T}(X_i, \boldsymbol{\theta}_0) + o_P(1/\sqrt{n})$$

where $\mathbf{T}(\mathbf{x}, \boldsymbol{\theta})$ *is the function*

$$\tilde{\boldsymbol{\Psi}}(\mathbf{x}, \boldsymbol{\theta}) - (b_2\omega_1)^{-1}\mathbf{R}_{12}(\mathbf{x}, \boldsymbol{\theta})\,(\mathbf{m}_0, \mathbf{M}_0)$$

and $(\mathbf{m}_0, \mathbf{M}_0)$ *is the pair* $\tilde{\boldsymbol{\Lambda}}(\mathbf{0}, \boldsymbol{\Sigma}) = \big(\tilde{\boldsymbol{\Lambda}}_{\text{loc}}(\mathbf{0}, \boldsymbol{\Sigma}), \tilde{\boldsymbol{\Lambda}}_{\text{cov}}(\mathbf{0}, \boldsymbol{\Sigma})\big)$.

PROOF: With Lemma 4.7 we find that

$$\boldsymbol{\Delta}_2(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0) = -\frac{\omega_2}{2p}\text{trace}\big((\mathbf{C}_n - \boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}\big). \tag{4.21}$$

We first determine the asymptotic expansion of this by means of Theorem 4.1. Consider the covariance part of (4.12) :

$$\tilde{\boldsymbol{\Lambda}}_{\text{cov}}(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0) = -\frac{1}{n}\sum_{i=1}^{n}\tilde{\boldsymbol{\Psi}}_{\text{cov}}(X_i, \boldsymbol{\theta}_0) + o_P(1/\sqrt{n}). \tag{4.22}$$

If we multiply $\tilde{\boldsymbol{\Psi}}_{\text{cov}}(X_i, \boldsymbol{\theta}_0)$ by $\boldsymbol{\Sigma}^{-1}$ and take traces it follows from definition (4.9) that this reduces to $2b_2p\big(\rho_1(d_{i0}) - b_1\big)$, where $d_{i0} = \sqrt{(X_i - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(X_i - \boldsymbol{\mu})}$. Consider the left hand side of (4.22) of which the expression is given in Lemma 4.7. Multiply by $\boldsymbol{\Sigma}^{-1}$ and take traces. Because $-\frac{p}{2}\tilde{u}'(y)y + \frac{p}{2}\tilde{g}'(y)/y = p\tilde{\psi}(y)/y - b_2p\psi_1(y)/y$ we obtain

$$\int\left(\frac{p\tilde{\psi}(d_0)}{d_0} - \frac{b_2p\psi_1(d_0)}{d_0}\right)\Big\{(\mathbf{x} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{C}_n - \boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) + 2(\mathbf{x} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{t}_n - \boldsymbol{\mu})\Big\}\,dP(\mathbf{x})$$
$$- 2b_2\omega_1\text{trace}\big((\mathbf{C}_n - \boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}\big).$$

This reduces to $-b_2\omega_1\text{trace}\big((\mathbf{C}_n - \boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}\big)$, where we use (4.20) and the fact that $P\tilde{\boldsymbol{\Psi}}(\cdot, \boldsymbol{\theta}_0) = \mathbf{0}$ and $P\boldsymbol{\Psi}_1(\cdot, \boldsymbol{\theta}_0) = \mathbf{0}$. Hence, if we multiply (4.22) with $\boldsymbol{\Sigma}^{-1}$ and take traces we find that

$$-b_2\omega_1\text{trace}\big((\mathbf{C}_n - \boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}\big) = -\frac{2b_2p}{n}\sum_{i=1}^{n}\Big(\rho_1(d_i) - b_1\Big) + o_P(1/\sqrt{n}). \tag{4.23}$$

According to Lemma 4.6 and (4.21) this means that

$$\mathbf{V}_n - \boldsymbol{\Sigma} = \mathbf{C}_n - \boldsymbol{\Sigma} + (b_2\omega_1)^{-1}\frac{1}{n}\sum_{i=1}^{n}\mathbf{R}_{12}(X_i, \boldsymbol{\theta}_0)\,\boldsymbol{\Sigma} + o_P(1/\sqrt{n})$$

and hence,

$$\boldsymbol{\tau}_n - \boldsymbol{\theta}_0 = \boldsymbol{\theta}_n - \boldsymbol{\theta}_0 + (b_2\omega_1)^{-1}\frac{1}{n}\sum_{i=1}^{n}\mathbf{R}_{12}(X_i, \boldsymbol{\theta}_0)\,(\mathbf{0}, \boldsymbol{\Sigma}) + o_P(1/\sqrt{n}).$$

When we apply the linear map $\tilde{\boldsymbol{\Lambda}}$ to both sides, the theorem follows from Theorem 4.1. □

COROLLARY 4.1. *Let* $\mathbf{T}(\mathbf{x}, \boldsymbol{\theta})$ *be defined in Theorem 4.2. Then under the conditions of Theorem 4.1* $\sqrt{n}(\boldsymbol{\tau}_n - \boldsymbol{\theta}_0)$ *has a limiting normal distribution with zero mean and covariance matrix* $\tilde{\mathbf{A}}^{-1}\mathbf{M}\tilde{\mathbf{A}}^{-T}$, *where* $\mathbf{M}$ *stands for the covariance matrix of* $\mathbf{T}(X_1, \boldsymbol{\theta}_0)$.

5. Robustness. We measure the robustness of the τ-estimators by means of the finite-sample breakdown point as defined in Donoho and Huber (1983) and the influence function as defined in Hampel (1974).

5.1. Breakdown point. The breakdown point of a location estimator $\mathbf{t}_n$ at a collection $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ is defined as the smallest fraction m/n of outliers that can take the estimator over all bounds :

$$\varepsilon^*(\mathbf{t}_n, \mathbf{X}) = \min_{1 \le m \le n} \left\{ \frac{m}{n} : \sup_{Y_m} \|\mathbf{t}_n(\mathbf{X}) - \mathbf{t}_n(\mathbf{Y}_m)\| = \infty \right\} \tag{5.1}$$

where the supremum is taken over all possible corrupted collections $\mathbf{Y}_m$ that can be obtained from $\mathbf{X}$ by replacing m points of $\mathbf{X}$ by arbitrary values. The breakdown point of a covariance estimator $\mathbf{C}_n$ at a collection $\mathbf{X}$ is defined as the smallest fraction m/n of outliers that can either take the largest eigenvalue $\lambda_1(\mathbf{C}_n)$ over all bounds, or take the smallest eigenvalue $\lambda_p(\mathbf{C}_n)$ arbitrarily close to zero :

$$\varepsilon^*(\mathbf{C}_n, \mathbf{X}) = \min_{1 \le m \le n} \left\{ \frac{m}{n} : \sup_{Y_m} D\big(\mathbf{C}_n(\mathbf{X}), \mathbf{C}_n(\mathbf{Y}_m)\big) = \infty \right\}$$

where the supremum is taken over the same corrupted collections $\mathbf{Y}_m$ as in (5.1), and where $D(\mathbf{A}, \mathbf{B}) = \max\{|\lambda_1(\mathbf{A}) - \lambda_1(\mathbf{B})|, |\lambda_p(\mathbf{A})^{-1} - \lambda_p(\mathbf{B})^{-1}|\}$.

We have already seen in Section 2.2 that for $\mathbf{x}_1, \ldots, \mathbf{x}_n$ in general position at least one solution exists if $n(1 - r_1) \ge p + 1$. In order to guarantee that at least one solution exists when we also replace $\lceil nr_1 \rceil - 1$ points $\mathbf{x}_i$ by arbitrary points we need that every subsample of $\lceil n - nr_1 \rceil$ points contains at least $p + 1$ points in general position, where $\lceil y \rceil$ denotes the smallest integer greater than or equal to y. That this implies existence of a solution to the finite sample problem can be seen as follows. For each $(\mathbf{t}, \mathbf{C})$ that satisfies (2.3) it holds that $P_n\big(E(\mathbf{t}, \mathbf{C}, c_1)\big) \ge \lceil n - nr_1 \rceil / n$, and clearly does P_n satisfy condition (H_ε) for $\varepsilon = \lceil n - nr_1 \rceil / n$. Apply Lemma 2.2(i) and then use an argument similar to that in the proof of Theorem 2.1.

THEOREM 5.1. *Let* $\mathbf{X}$ *be a collection of* $n \ge p + 1$ *points in* $\mathbf{R}^p$ *in general position. Let* $r_1 = b_1/a_1$. *If* $r_1 \le (n - p)/(2n)$ *then the* τ*-estimators* $(\mathbf{t}_n, \mathbf{V}_n)$ *have breakdown point* $\varepsilon^*(\mathbf{t}_n, \mathbf{X}) = \varepsilon^*(\mathbf{V}_n, \mathbf{X}) = \lceil nr_1 \rceil / n$.

PROOF: As we can always rescale the function ρ_1 we may assume that $a_1 = 1$, so that $b_1 = r_1$ in (2.3). According to Lemma 2.1 there exists a constant $q > 0$ which only depends on ρ_1, ρ_2 and b_1, such that

$$\frac{1}{n} \sum_{i=1}^{n} \rho_2\big[\{(\mathbf{x}_i - \mathbf{t}_n)^T \mathbf{C}_n^{-1} (\mathbf{x}_i - \mathbf{t}_n)\}^{1/2}\big] \ge q.$$

Therefore $\mathbf{V}_n$ breaks down simultaneously with $\mathbf{C}_n$ and it suffices to consider breakdown of $\mathbf{t}_n$ and $\mathbf{C}_n$. The rest of the proof is along the lines of the proof of Theorem 3.2 in Lopuhaä and Rousseeuw (1989). We give a brief sketch of the argument.

First show that $\varepsilon^*(\mathbf{t}_n, \mathbf{X})$ and $\varepsilon^*(\mathbf{C}_n, \mathbf{X})$ are at least $\lceil nr_1 \rceil / n$. Replace at most $m = \lceil nr_1 \rceil - 1$ points of $\mathbf{X}$. Since $\lceil n - nr_1 \rceil - \lceil nr_1 \rceil \geq p$ it follows that at least one solution $(\mathbf{t}_n(\mathbf{Y}_m), \mathbf{C}_n(\mathbf{Y}_m))$ in Θ exists, and that there exists a constant $k_1 > 0$, which only depends on the collection $\mathbf{X}$, such that the smallest eigenvalue $\lambda_p(\mathbf{C}_n(\mathbf{Y}_m)) \geq k_1$. The collection $\mathbf{Y}_m$ contains $n - m$ points of $\mathbf{X}$, say $\mathbf{x}_1, \ldots, \mathbf{x}_{n-m}$. Since $nr_1 - \lceil nr_1 \rceil + 1$ is always strictly positive and because ρ_1 is continuous, we can find a constant $M > 0$ such that $\sum_{i=1}^{n-m} \rho_1(\|\mathbf{x}_i\|/M) = nr_1 - \lceil nr_1 \rceil + 1$. In that case

$$\sum_{\mathbf{y}_i \in \mathbf{Y}_m} \rho_1(\|\mathbf{y}_i\|/M) \leq \sum_{i=1}^{n-m} \rho_1(\|\mathbf{x}_i\|/M) + \lceil nr_1 \rceil - 1 = nr_1$$

which means that the pair $(\mathbf{0}, M^2\mathbf{I})$ satisfies constraint (2.4). According to Remark 2.1 this means that for the finite sample problem corresponding with the collection $\mathbf{Y}_m$, the value of the loss function at $(\mathbf{t}_n(\mathbf{Y}_m), \mathbf{C}_n(\mathbf{Y}_m))$ is at most $(M^2 a_2)^p$. With Lemma 2.1 it follows that $|\mathbf{C}_n(\mathbf{Y}_m)| \leq (M^2 a_2/q)^p$ and hence, the largest eigenvalue $\lambda_1(\mathbf{C}_n(\mathbf{Y}_m)) \leq (M^2 a_2/q)^p / k_1^{p-1}$ which only depends on $\mathbf{X}$. We conclude that $\mathbf{C}_n(\mathbf{Y}_m)$ does not break down so that $\varepsilon^*(\mathbf{C}_n, \mathbf{X}) \geq \lceil nr_1 \rceil / n$.

To obtain the same inequality for $\varepsilon^*(\mathbf{t}_n, \mathbf{X})$ as well as the two opposite inequalities, only the constraint (2.3) matters whereas the type of loss function in $(\mathcal{P}_{P_n})$ is of no importance. These inequalities can therefore be obtained similar to those for S-estimators for multivariate location and scatter (Lopuhaä and Rousseeuw 1989, Theorem 3.2). □

Obviously the optimal value for the breakdown point is obtained by chosing $r_1 = (n-p)/(2n)$ in which case $\varepsilon^*(\mathbf{t}_n, \mathbf{X})$ and $\varepsilon^*(\mathbf{V}_n, \mathbf{X})$ attain the maximal possible value for affine equivariant covariance estimators : $\lfloor \frac{n-p+1}{2} \rfloor / n$. Note that the breakdown point of the τ-estimators depends only on the constant b_1, or only on the constant c_1 if b_1 is chosen as in Theorem 2.2. This means that the tuning constant c_2 of the function ρ_2 can be varried without changing the value of the breakdown point.

5.2. Influence function. The breakdown point of an estimator $\boldsymbol{\tau}_n$ is only a global measure of robustness. To assess sensitivity of the corresponding functional $\boldsymbol{\tau}(\cdot)$ under small pertubations, Hampel (1974) defined the influence function as

$$\text{IF}(\mathbf{x}; \boldsymbol{\tau}, P) = \lim_{h \downarrow 0} \frac{\boldsymbol{\tau}\big((1-h)P + h\delta_{\mathbf{x}}\big) - \boldsymbol{\tau}(P)}{h} \tag{5.2}$$

where $\delta_{\mathbf{x}}$ denotes the Dirac measure concentrated in $\mathbf{x} \in \mathbf{R}^p$. A related measure of robustness is the *gross-error sensitivity* defined as $\gamma^*(\boldsymbol{\tau}, P) = \sup_{\mathbf{x}} \text{IF}(\mathbf{x}; \boldsymbol{\tau}, P)$.

We assume that P satisfies property (H_ε) for some $0 < \varepsilon < 1 - r_1$, that the minimization problem $(\mathcal{P}_P)$ has a unique solution $\boldsymbol{\theta}_0 = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and we take b_2 as

in (3.3). For $\mathbf{x} \in \mathbf{R}^p$ and $0 \leq h \leq 1$ write $P_{h,\mathbf{x}} = (1-h)P + h\delta_{\mathbf{x}}$. For $h \downarrow 0$ the distribution $P_{h,\mathbf{x}}$ converges weakly to P. According to Theorem 4.1 this means that at least one solution to the problem $(\mathcal{P}_{P_{h,\mathbf{x}}})$ exists for h sufficiently small and that $\boldsymbol{\theta}(P_{h,\mathbf{x}}) \rightarrow \boldsymbol{\theta}_0$. Before we obtain the influence function for the τ-functionals, we first need the following equivalent of Theorem 4.1.

LEMMA 5.1. *Let $\mathbf{x} \in \mathbf{R}^p$ and let $\boldsymbol{\theta}(P_{h,\mathbf{x}}) = \big(\mathbf{t}(P_{h,\mathbf{x}}), \mathbf{C}(P_{h,\mathbf{x}})\big)$ be a solution of $(\mathcal{P}_{P_{h,\mathbf{x}}})$ for h sufficiently small. Under the conditions of Theorem 4.1 it holds that for $h \downarrow 0$*

$$\tilde{\boldsymbol{\Lambda}}\big(\boldsymbol{\theta}(P_{h,\mathbf{x}}) - \boldsymbol{\theta}_0\big) = -h\tilde{\boldsymbol{\Psi}}(\mathbf{x}, \boldsymbol{\theta}_0) + o(h). \tag{5.3}$$

PROOF: We can use arguments that are similar to the one used in the proofs in Section 4, except that one should read $P_{h,\mathbf{x}}$ instead of P_n. We will go through them briefly. Suppress the dependence on $\mathbf{x}$ and write $\boldsymbol{\theta}_h$ instead of $\boldsymbol{\theta}(P_{h,\mathbf{x}})$. We first show that $\boldsymbol{\theta}_h - \boldsymbol{\theta}_0 = O(h)$.

Define $A_{h,\mathbf{x}}(\boldsymbol{\theta}) = P_{h,\mathbf{x}}a(\cdot, \boldsymbol{\theta})$ and $B_{h,\mathbf{x}}(\boldsymbol{\theta}) = P_{h,\mathbf{x}}b(\cdot, \boldsymbol{\theta})$. According to Remark 4.1 it holds that the solution $\boldsymbol{\theta}_h$ of the problem $(\mathcal{P}_{P_{h,\mathbf{x}}})$ satisfies

$$\mathbf{0} = A_{h,\mathbf{x}}(\boldsymbol{\theta}_h)P_{h,\mathbf{x}}\boldsymbol{\Psi}_1(\cdot, \boldsymbol{\theta}_h) + B_{h,\mathbf{x}}(\boldsymbol{\theta}_h)P_{h,\mathbf{x}}\boldsymbol{\Psi}_2(\cdot, \boldsymbol{\theta}_h) - 2b_2 P_{h,\mathbf{x}}\mathbf{R}(\cdot, \boldsymbol{\theta}_h). \tag{5.4}$$

Since the functions $a(\mathbf{x}, \boldsymbol{\theta})$ and $b(\mathbf{x}, \boldsymbol{\theta})$ are bounded and continuous, it follows immediately that $A_{h,\mathbf{x}}(\boldsymbol{\theta}_h) \rightarrow A(\boldsymbol{\theta}_0)$ and $B_{h,\mathbf{x}}(\boldsymbol{\theta}_h) \rightarrow B(\boldsymbol{\theta}_0)$. Hence, similar to (4.13) it follows together with (5.4) that

$$\mathbf{0} = P_{h,\mathbf{x}}\tilde{\boldsymbol{\Psi}}(\cdot, \boldsymbol{\theta}_h) + o(1)\, P_{h,\mathbf{x}}\boldsymbol{\Psi}_1(\cdot, \boldsymbol{\theta}_h) + o(1)\, P_{h,\mathbf{x}}\boldsymbol{\Psi}_2(\cdot, \boldsymbol{\theta}_h). \tag{5.5}$$

Similar to (4.15) one finds that the last two terms on the right hand side of (5.5) are $o(\|\boldsymbol{\theta}_h - \boldsymbol{\theta}_0\|) + o(h)$. The function $\tilde{\boldsymbol{\Psi}}(\mathbf{x}, \boldsymbol{\theta})$ is continuous and bounded on a bounded neighbourhood of $\boldsymbol{\theta}_0$. This implies that $\tilde{\boldsymbol{\Psi}}(\mathbf{x}, \boldsymbol{\theta}_h)$ tends to $\tilde{\boldsymbol{\Psi}}(\mathbf{x}, \boldsymbol{\theta}_0)$ as $h \downarrow 0$, and that $P\tilde{\boldsymbol{\Psi}}(\cdot, \boldsymbol{\theta}_h)$ tends to $P\tilde{\boldsymbol{\Psi}}(\cdot, \boldsymbol{\theta}_0)$, which is zero according to Remark 4.1. Hence, similar to (4.14) one finds that $P_{h,\mathbf{x}}\tilde{\boldsymbol{\Psi}}(\cdot, \boldsymbol{\theta}_h) = \big(\tilde{\boldsymbol{\Lambda}} + o(1)\big)(\boldsymbol{\theta}_h - \boldsymbol{\theta}_0) + h\tilde{\boldsymbol{\Psi}}(\mathbf{x}, \boldsymbol{\theta}_0) + o(h)$. If we put this into the first term on the right hand side of (5.5) it follows that

$$\mathbf{0} = \big(\tilde{\boldsymbol{\Lambda}} + o(1)\big)(\boldsymbol{\theta}_h - \boldsymbol{\theta}_0) + h\tilde{\boldsymbol{\Psi}}(\mathbf{x}, \boldsymbol{\theta}_0) + o(h). \tag{5.6}$$

Since $\tilde{\boldsymbol{\Lambda}}$ is nonsingular this means that $\boldsymbol{\theta}_h - \boldsymbol{\theta}_0 = O(h)$. If we substitute this in (5.6), this equation reduces to $\mathbf{0} = \tilde{\boldsymbol{\Lambda}}(\boldsymbol{\theta}_h - \boldsymbol{\theta}_0) + h\tilde{\boldsymbol{\Psi}}(\mathbf{x}, \boldsymbol{\theta}_0) + o(h)$, which proves the lemma. □

THEOREM 5.2. *Under the conditions of Theorem 4.1 it holds that the τ-functional $\boldsymbol{\tau}(\cdot) = \big(\mathbf{t}(\cdot), \mathbf{V}(\cdot)\big)$ has influence function $IF(\mathbf{x}; \boldsymbol{\tau}, P) = -\tilde{\boldsymbol{\Lambda}}^{-1}\mathbf{T}(\mathbf{x}, \boldsymbol{\theta}_0)$, where the function $\mathbf{T}(\mathbf{x}, \boldsymbol{\theta})$ is defined in Theorem 4.2.*

PROOF: If one reads $P_{h,\mathbf{x}}$ instead of P_n, the proof is along the lines of the proof of Theorem 4.2. Suppress the dependence on $\mathbf{x}$ and write $\mathbf{t}_h$, $\mathbf{C}_h$ and $\mathbf{V}_h$ instead

of $\mathbf{t}(P_{h,\mathbf{x}})$, $\mathbf{C}(P_{h,\mathbf{x}})$ and $\mathbf{V}(P_{h,\mathbf{x}})$. Consider the function $R_2(\mathbf{x},\boldsymbol{\theta})$ and its derivative $\boldsymbol{\Delta}_2$ at $\boldsymbol{\theta}_0$ as defined in Lemma 4.6. By definition we have that

$$\mathbf{V}_h = \mathbf{C}_h + b_2^{-1}\boldsymbol{\Sigma}P_{h,\mathbf{x}}R_2(\cdot,\boldsymbol{\theta}_h) + b_2^{-1}(\mathbf{C}_h - \boldsymbol{\Sigma})P_{h,\mathbf{x}}R_2(\cdot,\boldsymbol{\theta}_h). \tag{5.7}$$

By means of a similar reasoning that leads to (4.18) we obtain

$$P_{h,\mathbf{x}}R_2(\cdot,\boldsymbol{\theta}_h) = \boldsymbol{\Delta}_2(\boldsymbol{\theta}_h - \boldsymbol{\theta}_0) + o(\|\boldsymbol{\theta}_h - \boldsymbol{\theta}_0\|) + hR_2(\mathbf{x},\boldsymbol{\theta}_0) + o(h). \tag{5.8}$$

Lemma 5.1 implies that $\boldsymbol{\theta}_h - \boldsymbol{\theta}_0 = O(h)$, so that with (5.7) and (5.8) it follows that

$$\mathbf{V}_h = \mathbf{C}_h + hb_2^{-1}\boldsymbol{\Sigma}R_2(\mathbf{x},\boldsymbol{\theta}_h) + b_2^{-1}\boldsymbol{\Sigma}\boldsymbol{\Delta}_2(\boldsymbol{\theta}_h - \boldsymbol{\theta}_0) + o(h). \tag{5.9}$$

With Lemma 4.7 and the covariance part of Lemma 5.1, similar to (4.21) and (4.23) we find that

$$\boldsymbol{\Delta}_2(\boldsymbol{\theta}_h - \boldsymbol{\theta}_0) = -\frac{\omega_2}{2p}\operatorname{trace}\big((\mathbf{C}_h - \boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}\big) = -h\frac{\omega_2}{\omega_1}\Big(\rho_1\big(d(\mathbf{x},\boldsymbol{\theta}_0)\big) - b_1\Big) + o(h)$$

where $d(\mathbf{x},\boldsymbol{\theta})$ as defined in (3.1). Similar to the proof of Theorem 4.2 it follows from (5.9) that $\boldsymbol{\tau}_h - \boldsymbol{\theta}_0 = \boldsymbol{\theta}_h - \boldsymbol{\theta}_0 + h\,(b_2\omega_1)^{-1}\mathbf{R}_{12}(\mathbf{x},\boldsymbol{\theta}_0)(\mathbf{0},\boldsymbol{\Sigma}) + o(h)$. If we then apply the linear map $\tilde{\boldsymbol{\Lambda}}$ to both sides the theorem follows from Lemma 5.1. □

It follows immediately from the expression of $\mathbf{T}(\mathbf{x},\boldsymbol{\theta}_0)$ given in Theorem 4.2 that $\mathrm{IF}(\mathbf{x};\boldsymbol{\tau},P)$ is bounded and hence, that the gross-error sensitivity $\gamma^*(\boldsymbol{\tau},P)$ is finite. A more explicit expression for $\mathrm{IF}(\mathbf{x};\boldsymbol{\tau},P)$ and $\gamma^*(\boldsymbol{\tau},P)$ can be obtained at elliptical distributions. This will be done in the next section.

6. Elliptical distributions. As a special case we consider elliptical distributions. In this case it turns out that the limiting distribution of the τ-estimators is exactly the same as that of multivariate S-estimators defined with a weighted ρ-function

$$\tilde{\rho}(\cdot) = A(\boldsymbol{\theta}_0)\rho_1(\cdot) + B(\boldsymbol{\theta}_0)\rho_2(\cdot). \tag{6.1}$$

To describe the limiting covariance matrix of $\sqrt{n}(\mathbf{V}_n - \boldsymbol{\Sigma})$ we use the commutation matrix $\mathbf{K}_{p,p}$ and the operator $\mathrm{vec}(\cdot)$. The matrix $\mathbf{K}_{p,p}$ is a $p^2 \times p^2$-block matrix with the (i,j)-th block being equal to $\boldsymbol{\Delta}_{ji}$, which is a $p \times p$-matrix with entry 1 at (j,i) and 0 everywhere else. The operator $\mathrm{vec}(\cdot)$ is a p^2-vector that stacks the columns of a $p \times p$-matrix on top of each other. Finally, by $\mathbf{A} \otimes \mathbf{B}$ we mean the Kronecker product which is a $p^2 \times p^2$-block matrix with the (i,j)-th block $a_{ij}\mathbf{B}$.

COROLLARY 6.1. *Let P be an elliptical distribution with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ that satisfies condition (F) and has a finite fourth moment. Let $\boldsymbol{\theta}_0 = (\boldsymbol{\mu},\boldsymbol{\Sigma})$ and let $d_0 = \sqrt{(X_1 - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(X_1 - \boldsymbol{\mu})}$. Let $\tilde{\rho}$ and $\tilde{\psi} = \tilde{\rho}'$ be defined in (6.1) and (4.5), and suppose that*

$$\begin{aligned} &\mathbf{E}\tilde{\psi}'(d_0) > 0 \\ &\tilde{\gamma} = (p+2)^{-1}\mathbf{E}\big[\tilde{\psi}'(d_0)d_0^2 + (p+1)\tilde{\psi}(d_0)d_0\big] > 0. \end{aligned} \tag{6.2}$$

Then $\sqrt{n}(\tau_n - \theta_0)$ has a limiting normal distribution with zero mean and $\mathbf{t}_n$ and $\mathbf{V}_n$ are asymptotically independent. The covariance of the limiting distribution of $\sqrt{n}(\mathbf{t}_n - \boldsymbol{\mu})$ is given by $(\tilde{\alpha}/\tilde{\beta}^2)\boldsymbol{\Sigma}$, where

$$\tilde{\alpha} = \frac{1}{p}\mathbf{E}\tilde{\psi}^2(d_0)$$
$$\tilde{\beta} = \frac{1}{p}\mathbf{E}\left[(p-1)\tilde{\psi}(d_0)/d_0 + \tilde{\psi}'(d_0)\right].$$

The covariance matrix of the limiting distribution of the matrix $\sqrt{n}(\mathbf{V}_n - \boldsymbol{\Sigma})$ is given by $\tilde{\sigma}_1(\mathbf{I} + \mathbf{K}_{p,p})(\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma}) + \tilde{\sigma}_2 \mathrm{vec}(\boldsymbol{\Sigma})\,\mathrm{vec}(\boldsymbol{\Sigma})^T$ where

$$\tilde{\sigma}_1 = \frac{p\mathbf{E}\tilde{\psi}^2(d_0)d_0^2}{(p+2)\tilde{\gamma}^2}$$
$$\tilde{\sigma}_2 = -\frac{2}{p}\tilde{\sigma}_1 + \frac{4\mathbf{E}(\tilde{\rho}(d_0) - \tilde{b})^2}{\tilde{\omega}^2}$$

and where $\tilde{\gamma}$ is defined in (6.2) and $\tilde{\omega} = \mathbf{E}\tilde{\psi}(d_0)d_0$.

PROOF: It is not difficult to see that we may restrict to the case $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\mathbf{0}, \mathbf{I})$. (see for instance Lemma 4.1 in Lopuhaä 1989). We first determine the expression of the function $\mathbf{T}(\mathbf{x}, \boldsymbol{\theta}_0)$ of Theorem 4.2. With Lemma 4.7 it follows by symmetry that $\mathbf{m}_0 = \mathbf{0}$ and that

$$\mathbf{M}_0 = -\frac{p}{2}\mathbf{E}\tilde{u}'(\|X_1\|)\|X_1\|\,X_1X_1^T + \frac{1}{2}\mathbf{E}\tilde{g}'(\|X_1\|)\|X_1\|\,\mathbf{I} - 2b_2\omega_1\mathbf{I}.$$

Then use that for spherically symmetric distributed X_1 it holds that $X_1/\|X_1\|$ is independent of $\|X_1\|$ and has covariance matrix $(1/p)\mathbf{I}$. Since by definition $-\frac{1}{2}\tilde{u}'(y)y^3 + \frac{1}{2}\tilde{g}'(y)y = \tilde{\psi}(y)y - b_2\psi_1(y)y$, it follows that $\mathbf{M}_0 = -b_2\omega_1\mathbf{I}$. All together we find that $\mathbf{T}(\mathbf{x}, \boldsymbol{\theta}_0) = \tilde{\mathbf{S}}(\mathbf{x}, \boldsymbol{\theta}_0)$ where $\tilde{\mathbf{S}} = (\tilde{\mathbf{S}}_{\mathrm{loc}}, \tilde{\mathbf{S}}_{\mathrm{cov}})$ is the function

$$\text{(6.3)} \qquad \begin{cases} \tilde{\mathbf{S}}_{\mathrm{loc}}(\mathbf{x}, \boldsymbol{\theta}) = \dfrac{\tilde{\psi}(d)}{d}(\mathbf{x} - \mathbf{t}) \\ \tilde{\mathbf{S}}_{\mathrm{cov}}(\mathbf{x}, \boldsymbol{\theta}) = p\dfrac{\tilde{\psi}(d)}{d}(\mathbf{x} - \mathbf{t})(\mathbf{x} - \mathbf{t})^T - (\tilde{\psi}(d)d - \tilde{\rho}(d) + \tilde{b})\,\mathbf{C} \end{cases}$$

with $\tilde{\rho}$ defined in (6.1), $\tilde{\psi}$ defined in (4.5) and $\tilde{b} = (2b_2 - \omega_2)b_1 + \omega_1 b_2 = \mathbf{E}\tilde{\rho}(\|X_1\|)$.

At this point we recognize the function $\tilde{\mathbf{S}}$ as being the Θ-valued function that defines the M-estimator type of score equations $\sum_{i=1}^n \tilde{\mathbf{S}}(X_i, \boldsymbol{\theta}) = \mathbf{0}$, of which the multivariate S-estimators $\boldsymbol{\theta}_n^S = (\mathbf{t}_n^S, \mathbf{C}_n^S)$ defined by the function $\tilde{\rho}$ are a solution (Lopuhaä 1989). If we denote by $\tilde{\mathbf{D}}$ the derivative of the function $P\tilde{\mathbf{S}}(\cdot, \boldsymbol{\theta})$ at $\boldsymbol{\theta}_0$, this has the following two consequences. First, note that $\tilde{\rho}$ satisfies conditions (R1)-(R2), so that conditions (6.2) imply that $\tilde{\mathbf{D}}$ is nonsingular (see the proof of Corollary

5.1 in Lopuhaä 1989). Secondly, because $\tilde{\mathbf{D}}$ is nonsingular and since $\boldsymbol{\theta}_0$ is the unique solution of the S-minimization problem, the following expansion holds for $\boldsymbol{\theta}_n^S$:

$$\tilde{\mathbf{D}}(\boldsymbol{\theta}_n^S - \boldsymbol{\theta}_0) = -\frac{1}{n}\sum_{i=1}^{n} \tilde{\mathbf{S}}(X_i, \boldsymbol{\theta}_0) + o_P(1/\sqrt{n}) \tag{6.4}$$

(proof of Theorem 4.1 in Lopuhaä 1989 together with Theorem 3 in Huber 1967). However, it is easily seen that $\tilde{\mathbf{D}} = \tilde{\mathbf{A}}$. This means that also $\tilde{\mathbf{A}}$ is nonsingular. Because $\boldsymbol{\theta}_0$ is the unique solution of $(\mathcal{P}_P)$, Theorem 4.2 applies and we may conclude that

$$\tilde{\mathbf{D}}(\boldsymbol{\tau}_n - \boldsymbol{\theta}_0) = -\frac{1}{n}\sum_{i=1}^{n} \tilde{\mathbf{S}}(X_i, \boldsymbol{\theta}_0) + o_P(1/\sqrt{n}). \tag{6.5}$$

It follows from (6.4) and (6.5) that the limiting distribution of $\sqrt{n}(\boldsymbol{\tau}_n - \boldsymbol{\theta}_0)$ is the same as that of multivariate S-estimators $\sqrt{n}(\boldsymbol{\theta}_n^S - \boldsymbol{\theta}_0)$. This is a limiting normal distribution which is described in Corollary 5.1 in Lopuhaä (1989). Hence, the expressions for the scalars $\tilde{\alpha}$, $\tilde{\beta}$, $\tilde{\sigma}_1$ and $\tilde{\sigma}_2$ follow immediately from this corollary. □

Note that because $\mathrm{E}g(\sqrt{(X_1-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(X_1-\boldsymbol{\mu})}) = \int g(\|\mathbf{x}\|)f(\|\mathbf{x}\|)\,d\mathbf{x}$, the scalars in Corollary 6.1 do not depend on $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. When $c_2 \to \infty$ then $\tilde{\alpha}/\tilde{\beta}^2$, $\tilde{\sigma}_1$ and $\tilde{\sigma}_2$ will tend to the corresponding values for the sample mean and the sample covariance. This means that for large values of c_2 one has good asymptotic efficiency relative to the sample mean and sample covariance. This is true for any fixed value of c_1. Hence, we can choose c_1 such that $\mathbf{t}_n$ and $\mathbf{V}_n$ have a high breakdown point (Theorem 5.1) and then vary c_2 to obtain good efficiency for instance at the normal distribution.

For the influence function we only give the expression at spherically symmetric distributions. The expressions at general elliptical distributions can be found by using affine equivariance.

COROLLARY 6.2. *Let P be spherically symmetric. Under the conditions of Corollary 6.1 it holds that the location τ-functional has influence function*

$$IF(\mathbf{x}; \mathbf{t}, P) = \frac{\tilde{\psi}(\|\mathbf{x}\|)}{\tilde{\beta}\|\mathbf{x}\|}\mathbf{x}$$

where $\tilde{\beta}$ is defined in Corollary 6.1. The covariance τ-functional has influence function

$$IF(\mathbf{x}; \mathbf{V}, P) = \frac{p\tilde{\psi}(\|\mathbf{x}\|)}{\tilde{\gamma}\|\mathbf{x}\|}\mathbf{x}\mathbf{x}^T - \frac{\tilde{\psi}(\|\mathbf{x}\|)\|\mathbf{x}\|}{\tilde{\gamma}}\mathbf{I} + \frac{2(\tilde{\rho}(\|\mathbf{x}\|) - \tilde{b})}{\tilde{\omega}}\mathbf{I}$$

where $\tilde{\gamma}$ and $\tilde{\omega}$ are defined in Corollary 6.1.

PROOF: From the proof of Corollary 6.1 we know that the function $\mathbf{T}(\mathbf{x}, \boldsymbol{\theta}_0) = \tilde{\mathbf{S}}(\mathbf{x}, \boldsymbol{\theta}_0)$, and that $\tilde{\mathbf{A}}$ is equal to the derivative $\tilde{\mathbf{D}}$ which is nonsingular. Hence,

with Theorem 5.2 we find that $\mathrm{IF}(\mathbf{x}; \boldsymbol{\tau}, P) = -\tilde{\mathbf{D}}^{-1}\tilde{\mathbf{S}}(\mathbf{x}, \boldsymbol{\theta}_0)$. From Theorem 3.3 in Lopuhaä (1989) we see that the influence function of the τ-functional is the same as that of the S-functional defined by means of the function $\tilde{\rho}$. Therefore, the exact expressions for $\mathrm{IF}(\mathbf{x}; \mathbf{t}, P)$ and $\mathrm{IF}(\mathbf{x}; \mathbf{V}, P)$ in the case of a spherically symmetric distribution can be taken from Corollary 5.2 in Lopuhaä (1989). □

Acknowledgment. I thank Rudolf Grübel for helfpful suggestions and remarks.

References

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

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

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.

GRÜBEL, R. (1988). A minimal characterization of the covariance matrix. *Metrika* **35** 49-52.

HAMPEL, F.R. (1974). The influence curve and its role in robust estimation. *J. Am. Statist. Assoc.* **69** 383–393.

HUBER, P.J. (1967). The behaviour of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (L.Le Cam and J.Neyman, eds.) 221–233. University of California Press, Berkeley.

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

KIM, J. and POLLARD, D. (1989). Cube root asymptotics. Technical Report, Yale University. To appear Ann. Statist..

LOPUHAÄ, H.P. (1988). Highly efficient estimators of multivariate location with high breakdown point. Revised version of Technical Report 88-14, Delft University of Technology.

LOPUHAÄ, H.P. (1989). On the relation between S-estimators and M-estimators of multivariate location and covariance. *Ann. Statist.* **17** 1662–1683.

LOPUHAÄ, H.P. and ROUSSEEUW, P.J. (1989). Breakdown properties of affine equivariant estimators of multivariate location and covariance matrices. Revised version of Technical Report 87-14, Delft University of Technology. Tentatively accepted by Ann. Statist..

POLLARD, D. (1984). *Convergence of Stochastic Processes.* Springer, New York.

RAO, R.R. (1962). Relations between weak and uniform convergence of measures with applications. *Ann. Math. Statist.* **33** 659–680.

ROUSSEEUW, P.J. (1983). Multivariate estimation with high breakdown point. Paper presented at the Fourth Pannonian Symposium on Mathematical Statistics and Probability, Bad Tatzmannsdorf, Austria, September 4-9, 1983. In *Mathematical Statistics and Applications (1985)* (W.Grossmann, G.Pflug, I.Vincze and W.Wertz, eds.) 283–297. Reidel, Dordrecht, The Netherlands.

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

SCHATTING VAN LOKATIE EN COVARIANTIE MET HOOG BREEKPUNT

Samenvatting

In dit proefschrift worden de robuustheid en het asymptotisch gedrag van multivariate schatters voor lokatie en covariantie bestudeerd. Robuustheid van de schatters wordt op twee verschillende manieren gemeten. De *globale* robuustheid wordt gemeten aan de hand van het breekpunt, hetgeen ruwweg de kleinste fraktie van uitschieters in de collectie is die de schatter voorbij iedere grens kan trekken; het beschrijft het globale gedrag van een schatter onder grote verstoringen. De *lokale* robuustheid wordt gemeten aan de hand van de invloedsfunktie, die de invloed van één enkele uitschieter op de schatter beschrijft. Voor het onderzoeken van het asymptotisch gedrag, wordt de collectie opgevat als een steekproef, die gegenereerd is door een verdeling P op $\mathbf{R}^p$, en wordt het gedrag van de schatters bestudeerd als de steekproefgrootte n naar oneindig gaat. De meeste belangstelling gaat uit naar de snelheid van convergentie, de limietverdeling en de asymptotische efficiëntie. Als speciaal geval wordt het gebruikelijke lokatie-covariantie model beschouwd, waarbij wordt aangenomen dat de verdeling P elliptische contouren heeft met een onbekende lokatie en spreidingsparameter.

De aandacht richt zich op schatters die commuteren met affiene transformaties van de punten. De eerste multivariate affien equivariante schatters voor lokatie en spreiding met een hoog breekpunt werden geïntroduceerd in het begin van de jaren tachtig. Helaas hebben deze schatters relatief slechte asymptotische eigenschappen. De snelheid van convergentie is in het algemeen langzamer dan de gebruikelijke $\sqrt{n}$ snelheid, de limietverdeling is niet altijd normaal, of de asymptotische efficiëntie is teleurstellend laag.

Het belangrijkste doel van dit proefschrift is om affien equivariante schatters te construeren voor multivariate lokatie en spreiding die globale en lokale robuustheid combineren met goede asymptotische eigenschappen. De bevindingen zijn samengevat in vier verschillende artikelen. Deze artikelen zijn gereproduceerd aan het einde van het proefschrift en worden vooraf gegaan door een inleiding waarin de artikelen in een kader worden geplaatst.

In het eerste artikel wordt het breekpunt van diverse schatters voor multivariate lokatie en spreiding bestudeerd en wordt de rol van verschillende equivariantie eigenschappen geïllustreerd. Bovendien wordt een relatie tussen het breekpunt van univariate lokatie schatters en een maat voor grote afwijkingen uitgebreid tot multivariate lokatie schatters. Het tweede artikel is gewijd aan multivariate S-schatters. Deze schatters zijn een gladde versie van Rousseeuw's minimum volume ellipsoïde schatter en kunnen gezien worden als een eerste stap in de richting van het combineren van robuustheid met goede limiet eigenschappen.

In het derde artikel wordt een affien geschaalde lokatie M-schatter onderzocht. Dit is min of meer een lokatie M-schatter gebaseerd op de steekproef die onstaat

na schaling met een affien equivariante covariantie schatter met een hoog breekpunt. In het vierde artikel wordt de klasse van S-schatters uitgebreid tot de klasse van multivariate τ-schatters. Zowel van deze schatters als van de affien geschaalde lokatie M-schatters wordt aangetoond dat ze affien equivariant zijn en een hoog breekpunt alsmede een begrensde invloedsfunktie hebben. Bovendien wordt bewezen dat deze schatters met snelheid $\sqrt{n}$ convergeren naar een normale verdeling en dat ze een goede efficiëntie hebben met betrekking tot het steekproef gemiddelde en de steekproef covariantie.

Curriculum Vitae

De auteur van dit proefschrift werd geboren op 13 november 1957 te Amsterdam. Na het examen Atheneum-B aan het Spinoza Lyceum te Amsterdam in 1977 ging hij wiskunde studeren aan de Universiteit van Amsterdam. Van september 1983 tot april 1986 was hij kandidaats-assistent bij Prof. Dr. J. Th. Runnenburg. Op 23 april 1986 deed hij (cum laude) doctoraal examen bij Prof. Dr. P. Groeneboom met als specialisatie Mathematische Statistiek en Kansrekening. Zijn doctoraalscriptie was getiteld 'A Martingale associated with the Grenander estimator'. In mei 1986 trad hij in dienst bij de Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) als onderzoeksmedewerker op het project 'Multivariate Statistische Analyse met hoog Breekpunt' met als projectleider Prof. Dr. P. J. Rousseeuw. Het onderzoek voerde hij uit bij de vakgroep Statistiek, Stochastiek en Operationele Analyse (SSOR) van de Faculteit der Technische Wiskunde en Informatica aan de Technische Universiteit Delft en heeft geresulteerd in dit proefschrift.