Robust Regression and Outlier Detection
Peter J. Rousseeuw, Annick M. Leroy
Copyright © 1987 John Wiley & Sons, Inc.
ISBN: 0-471-85233-3

CHAPTER 7

Related Statistical Techniques

The preceding chapters all dealt with high-breakdown methods in regression analysis (including the special case of one-dimensional location). However, the same ideas can also be applied to other statistical procedures. In the first section of this chapter we shall focus our attention on multivariate location and covariance, a topic which is itself a key to various statistical techniques. Section 2 is about robust time series analysis, and Section 3 briefly discusses the merits of robustification in other situations.

1. ROBUST ESTIMATION OF MULTIVARIATE LOCATION AND COVARIANCE MATRICES, INCLUDING THE DETECTION OF LEVERAGE POINTS

Outliers are much harder to identify in multivariate data clouds than in the univariate case. Therefore, the construction of robust techniques becomes more difficult. We shall focus on the estimation of the "center" of a point cloud, in which all variables are treated in the same way (unlike regression analysis, where one tries to "explain" one variable by means of the remaining ones). We are also interested in the dispersion of the data about this "center."

Suppose we have a data set

$$\begin{aligned} X &= \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \\ &= \{(x_{11}, x_{12}, \ldots, x_{1p}), \ldots, (x_{n1}, x_{n2}, \ldots, x_{np})\} \end{aligned} \tag{1.1}$$

of n points in p dimensions, and we want to estimate its "center." (We have decided to denote the cases by *rows*, to keep the notation consistent with previous chapters.) For this purpose, we apply a multivariate location estimator, that is, a statistic T which is *translation equivariant*,

$$T(\mathbf{x}_1+\mathbf{b},\ldots,\mathbf{x}_n+\mathbf{b})=T(\mathbf{x}_1,\ldots,\mathbf{x}_n)+\mathbf{b}\,, \tag{1.2}$$

for any p-dimensional vector $\mathbf{b}$. This property is also referred to as *location equivariance*. Naturally, T also has to be *permutation invariant*,

$$T(\mathbf{x}_{\pi(1)},\ldots,\mathbf{x}_{\pi(n)})=T(\mathbf{x}_1,\ldots,\mathbf{x}_n)\,, \tag{1.3}$$

for any permutation π on $\{1,2,\ldots,n\}$.

Of course, not every such T will be useful, and additional conditions may be required, depending on the situation. The most well-known estimator of multivariate location is the arithmetic mean

$$T(X)=\bar{\mathbf{x}}=\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\,, \tag{1.4}$$

which is the least squares estimator in this framework because it minimizes $\Sigma_{i=1}^{n}\|\mathbf{x}_i-T\|^2$, where $\|\cdots\|$ is the ordinary Euclidean norm. However, it is well known that $\bar{\mathbf{x}}$ is not robust, because even a single (very bad) outlier in the sample can move $\bar{\mathbf{x}}$ arbitrarily far away. To quantify such effects, we slightly adapt the finite-sample breakdown point of Section 2 in Chapter 1 to the framework of multivariate location. We consider all corrupted samples X' obtained by replacing any m of the original data points by arbitrary values, and we define the maximal bias by

$$\operatorname{bias}(m;\,T,X)=\sup_{X'}\|T(X')-T(X)\| \tag{1.5}$$

so the breakdown point

$$\varepsilon_n^*(T,X)=\min\{m/n;\operatorname{bias}(m;\,T,X)\text{ is infinite}\} \tag{1.6}$$

is again the smallest fraction of contamination that can cause T to take on values arbitrarily far away. Obviously, the multivariate arithmetic mean possesses a breakdown point of $1/n$. (Therefore, it is not well suited for the detection of outliers, as will be seen in Subsection d below.) We often consider the limiting breakdown point for $n\rightarrow\infty$, so we say that the multivariate mean has 0% breakdown.

It is clear that no translation equivariant T can have a breakdown point larger than 50%, because one could build a configuration of outliers which is just a translation image of the "good" data points, making it impossible for T to choose. In one dimension, this upper bound of 50% can easily be attained, for instance, by the sample median. Therefore, several multivariate generalizations of the median have been constructed, as well as some other proposals to achieve a certain amount of robustness.

Before listing some of these robust alternatives, we shall distinguish between two classes of location estimators: those that are affine equivariant and those that are not. Indeed, in many situations one wants the estimation to commute with linear transformations (i.e., a reparametrization of the space of the $\mathbf{x}_i$ should not change the estimate). We say that T is *affine equivariant* if and only if

$$T(\mathbf{x}_1\mathbf{A}+\mathbf{b},\ldots,\mathbf{x}_n\mathbf{A}+\mathbf{b})=T(\mathbf{x}_1,\ldots,\mathbf{x}_n)\mathbf{A}+\mathbf{b} \qquad (1.7)$$

for any vector $\mathbf{b}$ and any nonsingular matrix $\mathbf{A}$. (Because $\mathbf{x}_i$ and T are denoted by rows, the matrix $\mathbf{A}$ has to stand on the right.) For instance, the arithmetic mean does satisfy this property, but not all robust estimators do. We shall first consider some nonequivariant proposals.

a. Robust Estimators That Are Not Affine Equivariant

The simplest idea is to consider each variable separately. Indeed, for each variable j the numbers $x_{1j}, x_{2j}, \ldots, x_{nj}$ can be considered as a one-dimensional data set with n points. One may therefore apply a univariate robust estimator to each such "sample" and combine the results into a p-dimensional estimate. This procedure inherits the breakdown point of the original estimator.

For instance, the *coordinatewise median* is defined as

$$(\operatorname*{med}_i x_{i1}, \operatorname*{med}_i x_{i2}, \ldots, \operatorname*{med}_i x_{ip}) \qquad (1.8)$$

and possesses a 50% breakdown point. This estimator is easily computed, but fails to satisfy some "natural" properties. For instance, it does not have to lie in the convex hull of the sample when $p \geq 3$. As an example, consider the p unit vectors $(1,0,\ldots,0)$, $(0,1,\ldots,0),\ldots,$ $(0,0,\ldots,1)$, the convex hull of which is a simplex not containing the coordinatewise median $(0,0,\ldots,0)$. (However, it does lie in the convex hull when $p \leq 2$, as can be shown by a geometrical argument.)

Nath (1971) proposed another coordinatewise technique, in which each component is investigated separately, and a certain fraction of the largest and the smallest observations are not used. This method is very simple, but many outliers may go unnoticed because multivariate outliers do not necessarily stick out in any of their components (see Figure 4 of Chapter 1).

Note that the multivariate arithmetic mean, although affine equivariant, can also be computed coordinatewise. However, this is an exception. Indeed, Donoho (1982, Proposition 4.6) showed that the only (measurable) location estimator that is both affine equivariant and computable as a vector of one-dimensional location estimators is the arithmetic mean.

Sometimes one does not wish equivariance with respect to all affine transformations, but only for those that preserve Euclidean distances, that is, transformations of the type $\mathbf{x} \rightarrow \mathbf{x}\boldsymbol{\Gamma} + \mathbf{b}$ where $\boldsymbol{\Gamma}$ is an orthogonal matrix (this means that $\boldsymbol{\Gamma}' = \boldsymbol{\Gamma}^{-1}$). This includes translations, rotations, and reflections. For instance, the L_1 location estimator, given as the solution T of

$$\underset{T}{\text{Minimize}} \sum_{i=1}^{n} \|\mathbf{x}_i - T\| \tag{1.9}$$

is orthogonal equivariant, because the objective function only depends on Euclidean distances. In operations research, this estimate is called the Weber point, and it corresponds to the optimal location to build a factory when the customers sit at the $\mathbf{x}_i$ and the total transportation cost is to be minimized. Several routines are available to compute T. The L_1 estimator is also a generalization of the univariate median, and its breakdown point is still 50%. (Some people call it the "spatial median" or the "median center.")

Orthogonal equivariant estimators are often used to estimate the center of a spherically symmetric density

$$f_{\boldsymbol{\mu}}(\mathbf{x}) = g(\|\mathbf{x} - \boldsymbol{\mu}\|) . \tag{1.10}$$

This symmetry assumption may be reasonable when the data are points in some physical space (such as the customers in industrial location, or stars in three-dimensional space), but one should be careful when applying such procedures to variables of different types, in which case the choice of measurement units becomes important. For instance, when the variables are height and time, it makes a lot of difference whether these are expressed in feet and minutes or in centimeters and hours. (In such situations, it may be safer to apply either coordinatewise techniques or affine equivariant ones.)

When the spherically symmetric density is Gaussian, the maximum likelihood estimator becomes the arithmetic mean. This suggests a (modest) generalization of M-estimators, given by

$$\underset{T}{\text{Minimize}} \sum_{i=1}^{n} \rho(\|\mathbf{x}_i - T\|) . \tag{1.11}$$

Huber (1967) considered the asymptotic behavior of these estimators, and Collins (1982) obtained some results on minimax variance.

b. Affine Equivariant Estimators

Let us now consider estimators that do satisfy the affine equivariance condition. Such estimators are particularly useful in a model with so-called *elliptical symmetry*, with density

$$|\det(\mathbf{A})|^{-1} g(\|(\mathbf{x} - \boldsymbol{\mu})\mathbf{A}^{-1}\|) . \tag{1.12}$$

This model is obtained when starting from a spherically symmetric density, to which an affine transformation is applied. Unfortunately, the pair $(\boldsymbol{\mu}, \mathbf{A})$ is not a suitable parametrization because both $\mathbf{A}$ and $\boldsymbol{\Gamma}\mathbf{A}$ (with orthogonal $\boldsymbol{\Gamma}$) lead to the same distribution. However, the symmetric and positive definite matrix $\boldsymbol{\Sigma} = \mathbf{A}'\mathbf{A}$ is the same for all $\mathbf{A}$ leading to the same density, because $\|(\mathbf{x} - \boldsymbol{\mu})\mathbf{A}^{-1}\| = [(\mathbf{x} - \boldsymbol{\mu})(\mathbf{A}'\mathbf{A})^{-1}(\mathbf{x} - \boldsymbol{\mu})']^{1/2}$. Therefore, we parametrize the model as

$$f_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\mathbf{x}) = (\det(\boldsymbol{\Sigma}))^{-1/2} g([(\mathbf{x} - \boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})']^{1/2}) . \tag{1.13}$$

The typical example is the multivariate normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is the (variance–) covariance matrix. Therefore, $\boldsymbol{\Sigma}$ is called a (*pseudo-*) *covariance matrix*, or *scatter matrix*, even in the general situation. Both $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ must be estimated. For covariance estimators, affine equivariance means that

$$\mathbf{C}(\{\mathbf{x}_1\mathbf{A} + \mathbf{b}, \ldots, \mathbf{x}_n\mathbf{A} + \mathbf{b}\}) = \mathbf{A}'\mathbf{C}(\{\mathbf{x}_1, \ldots, \mathbf{x}_n\})\mathbf{A} , \tag{1.14}$$

where $\mathbf{A}$ is any nonsingular p-by-p matrix and $\mathbf{b}$ is any vector. At $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, maximum likelihood yields the equivariant estimators

$$T(X) = \bar{\mathbf{x}} \quad \text{and} \quad \mathbf{C}(X) = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})'(\mathbf{x}_i - \bar{\mathbf{x}}) \tag{1.15}$$

(to obtain an unbiased estimator of $\boldsymbol{\Sigma}$, the denominator of $\mathbf{C}(X)$ must be replaced by $n - 1$).

In order to generalize the maximum likelihood approach, Hampel (1973) suggested an affine equivariant iterative procedure to estimate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Maronna (1976) formally introduced affine equivariant M-estimators of location and scatter, defined as simultaneous solutions of systems of equations of the form

$$\begin{cases} \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} u_1([(\mathbf{x}_i - T)\mathbf{C}^{-1}(\mathbf{x}_i - T)^t]^{1/2})(\mathbf{x}_i - T) = \mathbf{0} \\ \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} u_2((\mathbf{x}_i - T)\mathbf{C}^{-1}(\mathbf{x}_i - T)^t)(\mathbf{x}_i - T)^t(\mathbf{x}_i - T) = \mathbf{C}\,, \end{cases} \tag{1.16}$$

where u_1 and u_2 must satisfy certain assumptions, and he considered the problems of existence, uniqueness, consistency, and asymptotic normality. Huber (1977, 1981) and Stahel (1981) computed influence functions and showed that the breakdown point of all affine equivariant M-estimators is at most $1/(p+1)$, which is disappointingly low. In a numerical study, Devlin et al. (1981, p. 361) found that M-estimators could tolerate even fewer outliers than indicated by this upper bound. Recently, Tyler (1986) found that the estimated scatter matrix becomes singular (which is a form of breaking down) when the contaminated data lie in some lower-dimensional subspace.

Affine equivariant M-estimators can be computed recursively, for instance by means of ROBETH (Marazzi 1980) or COVINTER (Dutter 1983b). A survey of affine equivariant M-estimators can be found in Chapter 5 of Hampel et al. (1986), and some recent results are given by Tyler (1985a,b).

Donoho (1982) lists some other well known affine equivariant techniques and shows that they all have a breakdown point of at most $1/(p+1)$. These proposals include:

1. *Convex Peeling* (Barnett 1976, Bebbington 1978; based on an idea of Tukey). This proceeds by discarding the points on the boundary of the sample's convex hull, and this is repeated until a sufficient number of points have been peeled away. On the remaining data, classical estimators can be applied. This procedure appeals to the intuition, and it is indeed equivariant because the convex hull is preserved by affine transformations. However, the breakdown point is quite low (even when the "good" data come from a multivariate normal distribution) because each step of peeling removes at least $p+1$ points from the sample, and often only one of these is really an outlier, so the stock of "good" points is being exhausted too fast.

2. *Ellipsoidal Peeling* (Titterington 1978, Helbling 1983). This is simi-

lar to convex peeling, but removes all observations on the boundary of the minimum volume ellipsoid containing the data. (Note that the convex hull is also the smallest convex set containing all the points.) Affine transformations $\mathbf{x} \rightarrow \mathbf{x}\mathbf{A} + \mathbf{b}$ map ellipsoids onto ellipsoids, and the volume of the image is just $|\det(\mathbf{A})|$ times the original volume, so ellipsoidal peeling is again affine equivariant. However, its breakdown point also tends to $1/(p+1)$ for the same reasons.

3. *Classical Outlier Rejection*. The (squared) Mahalanobis distance

$$\mathrm{MD}^2(\mathbf{x}_i, X) = (\mathbf{x}_i - T(X))\mathbf{C}(X)^{-1}(\mathbf{x}_i - T(X))' \qquad (1.17)$$

is computed for each observation, where $T(X)$ is the arithmetic mean and $\mathbf{C}(X)$ is the classical covariance estimate (1.15) with denominator $n-1$ instead of n. Points for which $\mathrm{MD}^2(\mathbf{x}_i, X)$ is large are deleted from the sample, and one processes the "cleaned" data in the usual way. (The result will be affine equivariant because the $\mathrm{MD}^2(\mathbf{x}_i, X)$ do not change when the data are subjected to an affine transformation.) This approach works well if only a single outlier is present (Barnett and Lewis 1978, David 1981), but suffers from the masking effect otherwise, because one far-away outlier can make all other outliers have small $\mathrm{MD}^2(\mathbf{x}_i, X)$. (In other words, the breakdown point is only $2/n$.) Therefore, some refinements of this technique have been proposed.

4. *Iterative Deletion*. This consists of finding the most discrepant observation according to (1.17), deleting it, recomputing $T(X)$ and $\mathbf{C}(X)$ for the remaining data and using it to find the most discrepant $\mathbf{x}_i$ among those $n-1$ points, and so on. Several rules are possible to decide how many observations are to be removed, but at any rate the breakdown point of the mean of the remainder can be no better than $1/(p+1)$.

5. *Iterative Trimming* (Gnanadesikan and Kettenring 1972, Devlin et al. 1975). This starts with $X^{(1)} = X$ and defines $X^{(k+1)}$ recursively as the set of observations with the $(1-\alpha)n$ smallest values in $\{\mathrm{MD}^2(\mathbf{x}_i, X^{(k)}), \mathbf{x}_i \in X\}$. (Note that $X^{(2)}, X^{(3)}, \ldots$ all have the same number of points.) This iterative process is halted when both $T(X^{(k)})$ and $\mathbf{C}(X^{(k)})$ stabilize. By means of a heuristic reasoning, Donoho (1982) concluded that the breakdown point of the final $T(X)$ is at most about $1/p$.

6. *Depth Trimming*. This is based on the concept of depth (Tukey 1974), which provides a kind of "rank statistic" for multivariate data sets. The depth of $\mathbf{x}_i$ is the smallest number of data points in any half-space containing it. Donoho (1982) defines a depth-trimmed mean as the average of all points with depth at least k. The higher the k, the better

the resulting breakdown point will be. The trimmed mean with $k = \max_i$ depth $(\mathbf{x}_i)$, which may be called the *deepest points estimator*, generalizes the univariate sample median. Unfortunately, not all data sets contain very deep points. In fact, Donoho proves that the maximal depth (in any data set) satisfies

$$[n/(p+1)] \leq \max_i \text{depth}\,(\mathbf{x}_i) \leq [n/2]\,.$$

Both bounds are sharp. The upper bound is achieved when there is a point about which the sample is centrosymmetric, and then the breakdown point of the deepest points estimator is about $1/3$. On the other hand, the lower bound is reached when the data are evenly distributed over small clusters on the $p+1$ vertices of a simplex, in which case the breakdown point of any depth-trimmed mean is at most $1/(p+2)$.

Recently, Oja (1983) has put forward another affine equivariant proposal, which he calls the *generalized median*. It is defined by minimization of

$$\sum \Delta(\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_p}, T)\,, \tag{1.18}$$

where the summation is over $1 \leq i_1 < \cdots < i_p \leq n$, and $\Delta(\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_p}, T)$ is the volume of the simplex with vertices $\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_p}$ and T. Figure 1 illustrates the meaning of (1.18) in two dimensions: Instead of minimizing the total length of the lines linking T to each data point [as in (1.9), the spatial median], we now minimize the total area of all triangles formed by T and any *pair* of data points.

This is already the fourth generalization of the univariate median that we encounter. Like Donoho's deepest points estimator, also Oja's generalized median is affine equivariant. Indeed, any affine transformation $\mathbf{x} \rightarrow \mathbf{x}\mathbf{A} + \mathbf{b}$ merely multiplies the volumes $\Delta(\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_p}, T)$ by the constant factor $|\det(\mathbf{A})|$, which is immaterial to the minimization. However, this equivariance costs a lot, because the algorithm mentioned by Oja and Niinimaa (1985) for computing $T(X)$ appears to involve the computation of the objective function (1.18), which contains C_n^p terms, in each of

$$C^p_{C^p_n}$$

candidate estimates. Oja and Niinimaa (1985) showed that the generalized median is asymptotically normal, with good statistical efficiency. It was recently found that the breakdown point equals $1/3$ in the bivariate case, and it is assumed that in the p-variate case it is $1/(p+1)$ (Oja 1986,

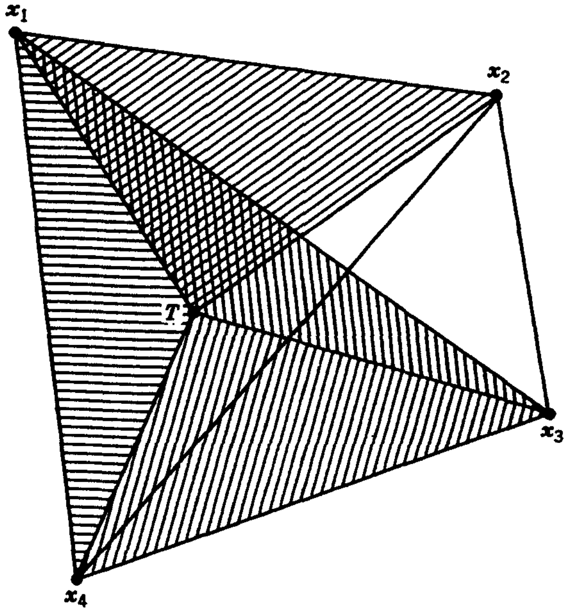


Figure 1. An illustration of Oja's generalized median, which minimizes the total area of all triangles formed with any pair of observations. (Not all such triangles are drawn.)

personal communication). Brown and Hettmansperger (1985) constructed affine invariant bivariate rank tests based on this estimator.

To conclude, all these affine equivariant estimators are facing the ubiquitous upper bound $1/(p+1)$ on their breakdown point. Fortunately, affine equivariant estimators with a higher breakdown point have recently been introduced.

c. Affine Equivariant Methods with High Breakdown Point

The first affine equivariant multivariate location estimator with a 50% breakdown point was obtained independently by Stahel (1981) and Donoho (1982). This estimator, called *outlyingness-weighted mean*, is defined as follows. For each observation $\mathbf{x}_i$, one looks for a one-dimensional projection leaving it most exposed. This is done by computing the following measure of the "outlyingness" of $\mathbf{x}_i$:

$$u_i = \sup_{\|\mathbf{v}\|=1} \frac{|\mathbf{x}_i\mathbf{v}^t - \operatorname*{med}_j (\mathbf{x}_j\mathbf{v}^t)|}{\operatorname*{med}_k |\mathbf{x}_k\mathbf{v}^t - \operatorname*{med}_j (\mathbf{x}_j\mathbf{v}^t)|}\,, \qquad (1.19)$$

where $\text{med}_j\,(\mathbf{x}_j\mathbf{v}^t)$ is the median of the projections of the data points $\mathbf{x}_j$ on the direction of the vector $\mathbf{v}$, and the denominator is the median absolute deviation of these projections. To compute u_i, one must (in principle) search over all possible directions. Then one estimates location by the weighted mean

$$T(X) = \frac{\sum_{i=1}^{n} w(u_i)\mathbf{x}_i}{\sum_{i=1}^{n} w(u_i)}\,, \tag{1.20}$$

where $w(u)$ is a strictly positive and decreasing function of $u \geq 0$, such that $uw(u)$ is bounded. [The latter bound is dictated by the one-dimensional case, where (1.20) reduces to a one-step W-estimator starting from the median (Hampel et al. 1986, p. 116), the influence function of which is proportional to $uw(u)$.] Analogously, these weights can also be used to compute a robust covariance matrix.

Donoho (1982) motivates (1.19) by noting that the classical (squared) Mahalanobis distance (1.17) can also be written as

$$\text{MD}^2(\mathbf{x}_i, X) = \left(\sup_{\|\mathbf{v}\|=1} \frac{\left|\mathbf{x}_i\mathbf{v}^t - \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{v}^t\right|}{\text{SD}\,(\mathbf{x}_1\mathbf{v}^t, \ldots, \mathbf{x}_n\mathbf{v}^t)} \right)^2 , \tag{1.21}$$

which is quite vulnerable to outliers, because they affect both the mean $(1/n)\,\Sigma_{i=1}^{n}\,\mathbf{x}_i\mathbf{v}^t$ and the standard deviation $\text{SD}\,(\mathbf{x}_1\mathbf{v}^t, \ldots, \mathbf{x}_n\mathbf{v}^t)$ of the projected data. Therefore it seems natural to replace these by robust estimators such as the median and the median absolute deviation, whereas Stahel (1981) proposed to insert also other univariate M-estimators. Points with large u_i are unlikely to belong to the "good" data, so their downweighting makes sense. It is shown that (1.20) is affine equivariant, because the u_i do not change when the $\mathbf{x}_i$ are transformed to $\mathbf{x}_i\mathbf{A} + \mathbf{b}$. Donoho (1982) also showed that the breakdown point of (1.20) is high (supposing that n is larger than $2p + 1$) and tends to 50% for $n \rightarrow \infty$. In his proof, he assumed that X is in *general position*, which, in the context of multivariate location, means that no more than p points of X lie in any $(p - 1)$-dimensional affine subspace. (For two-dimensional data, this says that there are no more than two points of X on any line, so any three points of X determine a triangle with nonzero area.) The asymptotic behavior of this estimator has recently been investigated by Donoho (1986, personal communication).

The Stahel–Donoho estimator downweights any point that is many robust standard deviations away from the sample in some projection. Therefore, it is related to the projection pursuit principle discussed in Section 5 of Chapter 3. The projections are on one-dimensional subspaces, and their "interestingness" is measured by the objective function in (1.19). For each point, the corresponding "least favorable" projection must be found, leading to u_i.

Rousseeuw (1983, 1984) introduced a second affine equivariant estimator with maximal breakdown point, by putting

$$\begin{aligned} T(X) = {} & \text{center of the minimal volume ellipsoid} \\ & \text{covering (at least) } h \text{ points of } X , \end{aligned} \tag{1.22}$$

where h can be taken equal to $[n/2]+1$. This is called the *minimum volume ellipsoid estimator* (MVE). The corresponding covariance estimator is given by the ellipsoid itself, multiplied by a suitable factor to obtain consistency. Affine equivariance of the MVE follows from the fact that the image of an ellipsoid through a nonsingular affine transformation $\mathbf{x} \rightarrow \mathbf{x}\mathbf{A} + \mathbf{b}$ is again an ellipsoid, with volume equal to $|\det(\mathbf{A})|$ times the original volume. Because $|\det(\mathbf{A})|$ is a constant, the relative sizes of ellipsoids do not change under affine transformations.

Theorem 1. At any p-dimensional sample X in general position, the breakdown point of the MVE estimator equals

$$\varepsilon_n^*(T, X) = ([n/2] - p + 1)/n ,$$

which converges to 50% as $n \rightarrow \infty$.

Proof. Without loss of generality, let $T(X) = \mathbf{0}$. We put $M :=$ volume of the smallest ellipsoid with center zero containing *all* points of X. Because X is in general position, each of its C_n^{p+1} subsets of $p+1$ points (indexed by some $J = \{i_1, \ldots, i_{p+1}\}$) determines a simplex with nonzero volume. Therefore, for each such J there exists a bound d_J such that any ellipsoid with center $\|\mathbf{c}\| > d_J$ and containing $\{\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_{p+1}}\}$ has a volume strictly larger than M. Put $d := \max_J d_J < \infty$.

Let us first show that $\varepsilon_n^*(T, X) \geq ([n/2] - p + 1)/n$. Take any sample X' obtained by replacing at most $[n/2] - p$ points of X. Suppose $\|T(X')\| > d$, and let E be the corresponding smallest ellipsoid containing (at least) $[n/2]+1$ points of X'. But then E contains at least $([n/2]+1) - ([n/2] - p) = p + 1$ points of X, so volume$(E) > M$. This is a contradiction, because the smallest ellipsoid with center zero around the

$n - ([n/2] - p) \geq [n/2] + 1$ "good" points of X' has a volume of at most M. Therefore, $\|T(X')\| \leq d$. [Even if $T(X')$ is not unique, then $\|T(X')\| \leq d$ still holds for all solutions.]

On the other hand, $\varepsilon_n^*(T, X) \leq ([n/2] - p + 1)/n$. Indeed, take any p points of X and consider the $(p-1)$-dimensional affine subspace H they determine. Now replace $[n/2] - p + 1$ other points of X by points on H. Then H contains $[n/2] + 1$ points of the new sample X', so the minimal volume ellipsoid covering these $[n/2] + 1$ points degenerates to zero volume. Because X is in general position, no ellipsoid covering another subset of $[n/2] + 1$ points of X' can have zero volume, so $T(X')$ lies on H. Finally, we note that $T(X')$ is not bounded because the $[n/2] - p + 1$ contaminated data points on H may have arbitrarily large norms. □

In most applications it is not feasible to actually consider all "halves" of the data and to compute the volume of the smallest ellipsoid around each of them. As in the case of least median of squares (LMS) regression, we therefore resort to an approximate algorithm not unlike the one described in Section 1 of Chapter 5. We start by drawing a subsample of $(p + 1)$ different observations, indexed by $J = \{i_1, \ldots, i_{p+1}\}$. For this subsample we determine the arithmetic mean and the corresponding covariance matrix, given by

$$\bar{\mathbf{x}}_J = \frac{1}{p+1} \sum_{i \in J} \mathbf{x}_i \quad \text{and} \quad \mathbf{C}_J = \frac{1}{p} \sum_{i \in J} (\mathbf{x}_i - \bar{\mathbf{x}}_J)'(\mathbf{x}_i - \bar{\mathbf{x}}_J)\,, \tag{1.23}$$

where $\mathbf{C}_J$ is nonsingular whenever $\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_{p+1}}$ are in general position. The corresponding ellipsoid should then be inflated or deflated to contain exactly h points, which corresponds to computing

$$m_J^2 = \operatorname*{med}_{i=1,\ldots,n} (\mathbf{x}_i - \bar{\mathbf{x}}_J)\mathbf{C}_J^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}_J)' \tag{1.24}$$

because m_J is the right magnification factor. The volume of the resulting ellipsoid, corresponding to $m_J^2\mathbf{C}_J$, is proportional to

$$(\det(m_J^2\mathbf{C}_J))^{1/2} = (\det(\mathbf{C}_J))^{1/2}(m_J)^p\,. \tag{1.25}$$

This has to be repeated for many J, after which the one with the lowest objective function (1.25) is retained. We then compute

$$T(X) = \bar{\mathbf{x}}_J \quad \text{and} \quad \mathbf{C}(X) = (\chi^2_{p,0.50})^{-1} m_J^2 \mathbf{C}_J\,, \tag{1.26}$$

where $\chi^2_{p,0.50}$ is the median of the chi-squared distribution with p degrees of freedom. (This correction factor is for consistency at multivariate normal data.) The number of random subsamples J that are needed can be determined as in Section 1 of Chapter 5. Indeed, the probability that at least one out of m subsamples consists exclusively of "good" points is approximately

$$1-(1-(1-\varepsilon)^{p+1})^{m} \tag{1.27}$$

when the original data contains a fraction ε of outliers. By imposing that this probability exceeds a given value, one obtains the number m of replications.

The complexity of this algorithm is similar to that of LMS regression, because for each of the m replications a p-by-p matrix $\mathbf{C}_J$ must be inverted (this can even be done somewhat faster because $\mathbf{C}_J$ is symmetric and positive definite). Then a median of the n "squared residuals" must be computed in (1.24). Note that the calculation of det $(\mathbf{C}_J)$ does not cost extra, because it is obtained as a by-product of the computation of $\mathbf{C}_J^{-1}$. Also note that this algorithm is itself affine equivariant and that it may easily be parallelized by treating many subsamples simultaneously.

The (approximate) MVE estimates can also be used as initial solutions on which to base a one-step improvement. As in the case of reweighted least squares regression (see Section 2 of Chapter 1) we can assign a weight w_i to each observation by means of the rule

$$w_i=\begin{cases}1 & \text{if} \quad (\mathbf{x}_i-T(X))\mathbf{C}(X)^{-1}(\mathbf{x}_i-T(X))' \le c\\ 0 & \text{otherwise}\,,\end{cases} \tag{1.28}$$

where the cut-off value c might be taken equal to $\chi^2_{p,0.975}$. Then one can apply the reweighted estimators

$$T_1(X)=\frac{\sum_{i=1}^{n} w_i\mathbf{x}_i}{\sum_{i=1}^{n} w_i} \tag{1.29}$$

and

$$\mathbf{C}_1(X)=\frac{\sum_{i=1}^{n} w_i(\mathbf{x}_i-T_1(X))'(\mathbf{x}_i-T_1(X))}{\sum_{i=1}^{n} w_i-1} \tag{1.30}$$

which simply means that the classical computations are carried out on the set of points for which $w_i = 1$.

The above algorithm (both MVE and reweighted) has been implemented in a program called PROCOVIEV (Rousseeuw and van Zomeren 1987). This abbreviation stands for Program for RObust COVariance and Identification of Extreme Values, indicating its ability to discover multivariate outliers.

Example

Let us consider the Hertzsprung–Russell data of Table 3 of Chapter 2. This time we treat both variables (log temperature and log light intensity) in the same way, so we have 47 points in the plane. Figure 2 shows the classical 97.5% tolerance ellipse, obtained from the usual mean and covariance matrix (dashed line). Note that this ellipse is very big, because it is attracted by the four outliers (indeed, the tolerance ellipse tries to engulf them). On the other hand, the tolerance ellipse based on the MVE estimates (solid line) is much smaller and essentially fits the main sequence. The four giant stars, lying far away from this ellipse, can then be identified as not belonging to the same population as the bulk of the

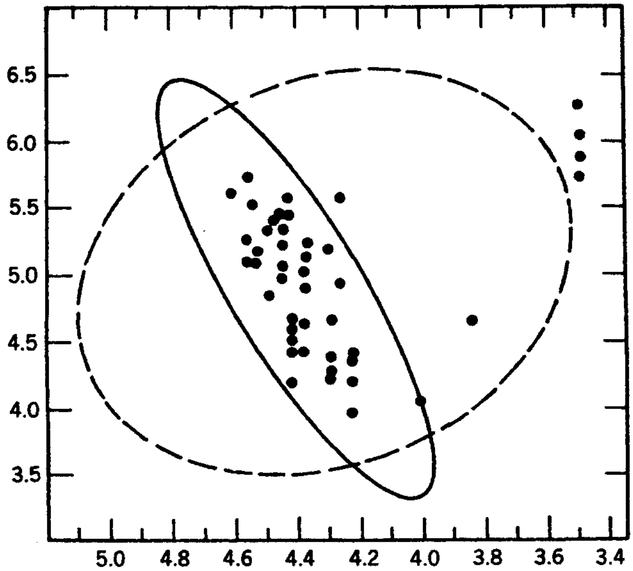


Figure 2. Hertzsprung–Russell data of Chapter 2, with 97.5% tolerance ellipse obtained from the classical estimator (dashed line) and based on the MVE estimator (solid line).

data. Note that choosing other measurement units for the variables would yield essentially the same result, because of affine equivariance. Also note that this analysis is basically different from that presented in Figure 4 of Chapter 2, because there the aim was to write the log light intensity as a linear function of the log temperature, and the LMS tried to find the narrowest *strip* covering half of the data, whereas we now made use of the smallest *ellipse* covering half of the data. However, we could construct another robust regression line of y on x based on this ellipse, by putting

$$\hat{y} = f(x) = \text{midpoint of the intersections of the vertical line through } x \text{ with the ellipse}$$

in the same way that the LS line is derived from the classical covariance matrix. This could easily be generalized to multiple regression.

For $p = 1$ (univariate location) the minimum volume ellipsoid reduces to the shortest half, so $T(X)$ becomes the one-dimensional LMS discussed in Chapter 4. In particular, Theorem 3 in Section 4 of Chapter 4 shows that this estimator converges as $n^{-1/3}$, which is abnormally slow. It is assumed that the multivariate MVE will not have a better rate, so it is useful to perform a one-step reweighting [based on formula (1.28) above] to improve its statistical efficiency. (Alternatively, a one-step M-estimator might be applied.)

Another approach would be to generalize the least trimmed squares (LTS) estimator (which converges like $n^{-1/2}$ according to Theorem 4 in Section 4 of Chapter 4) to multivariate location. This yields

$$T(X) = \text{Mean of the } h \text{ points of } X \text{ for which the determinant of the covariance matrix is minimal} \tag{1.31}$$

(Rousseeuw 1983, 1984). We call this the *minimum covariance determinant* estimator (MCD). It corresponds to finding the h points for which the classical tolerance ellipsoid (for a given level) has minimum volume, and then taking its center. This estimator is also affine equivariant because the determinant of the covariance matrix of the transformed data points equals

$$\det(\mathbf{A}^t\mathbf{C}\mathbf{A}) = (\det(\mathbf{A}))^2 \det(\mathbf{C})\,. \tag{1.32}$$

The MCD has the same breakdown point as the MVE, because of the same reasoning as in Theorem 1 above. Like the MVE, the MCD also yields a robust covariance estimate at the same time: One only has to use the (classical) covariance matrix of the selected h observations, multiplied by a constant to obtain consistency in the case of multivariate normality.

Both the MVE and the MCD are very drastic, because they are intended to safeguard against up to 50% of outliers. If one is certain that the fraction of outliers is at most α (where $0 < \alpha \leq \frac{1}{2}$), then one can work with the estimators MVE(α) and MCD(α) obtained by replacing h by $k(\alpha) = [n(1-\alpha)] + 1$ in (1.22) and (1.31). The breakdown point of these estimators is equal to α (for $n \rightarrow \infty$). For $\alpha \rightarrow 0$, the MVE yields the center of the smallest ellipsoid covering all the data, whereas the MCD tends to the arithmetic mean.

In the regression context, the LMS has been generalized to *S-estimators*, which are described in Section 4 of Chapter 3. These S-estimators can also be extended to multivariate location and covariance, by adapting formulas (4.28)–(4.30) of Chapter 3. This means that one must find a vector T and a symmetric positive definite matrix $\mathbf{C}$ such that

$$\begin{gathered}\det(\mathbf{C}) \text{ is minimized subject to} \\ \frac{1}{n}\sum_{i=1}^{n} \rho(\{(\mathbf{x}_i - T)\mathbf{C}^{-1}(\mathbf{x}_i - T)^t\}^{1/2}) = K\,, \end{gathered} \qquad (1:33)$$

where K is often put equal to the expected value $E[\rho(\{\mathbf{z}\mathbf{z}^t\}^{1/2})]$ in which $\mathbf{z}$ follows a standard multivariate normal distribution $N(\mathbf{0}, I)$, and where ρ satisfies the same conditions as before. These S-estimators are obviously affine equivariant, and for well-chosen ρ their breakdown point is again that of Theorem 1 above (this can be shown by means of a reasoning analogous to Theorem 8 in Section 4 of Chapter 3). Note that the MVE can actually be viewed as a special case of S-estimators, obtained by inserting the *discontinuous* ρ-function

$$\rho(u) = \begin{cases} 0 & \text{if } \ |u| < (\chi^2_{p,0.50})^{1/2} \\ 1 & \text{otherwise} \end{cases} \qquad (1.34)$$

and putting $K = \frac{1}{2}$. As in the case of regression analysis, choosing a *smooth* ρ-function greatly improves the asymptotic behavior. Davies (1987) indeed proves that S-estimators for multivariate location and covariance, constructed from a smooth ρ-function, are asymptotically normal, which means that

$$n^{1/2}(T_n(\mathbf{x}_1, \ldots, \mathbf{x}_n) - \boldsymbol{\mu})$$

and

$$n^{1/2}(\mathbf{C}_n(\mathbf{x}_1, \ldots, \mathbf{x}_n) - \mathbf{\Sigma})$$

converge in law to multivariate normal distributions with zero mean. (He writes this down with a function κ, which is related to ρ through $\kappa(u) = [\rho(\infty) - \rho(\sqrt{u})]/\rho(\infty)$.) In actual applications, Davies (1987, Section 5) does not use S-estimators with smooth ρ-function, for which the computations become too complex. Instead, he applies the original MVE estimator, for which he proposes an algorithm that differs from the one described above in that (a) more than $p + 1$ points may be drawn at each replication and (b) reweighting is continued as long as the volume of the 50% coverage ellipsoid decreases at each step. He then runs the MVE on a five-dimensional example in which the outliers are not detectable by means of classical methods, not even on the basis of the 10 two-dimensional plots of the projections onto the eigenvectors of the ordinary covariance matrix.

Remark. Both Davies (1987) and Lopuhaä and Rousseeuw (1987) independently found a way to increase the finite-sample breakdown point of the MVE to its best possible value. As in the case of the LMS regression estimator, this is done by letting h depend on the dimensionality. For the MVE (and the MCD), the optimal variant is to consider "halves" containing

$$h = [(n + p + 1)/2] \tag{1.35}$$

observations. The resulting breakdown value (which may be verified as in Theorem 1) is then

$$\varepsilon_n^* = \frac{[(n - p + 1)/2]}{n}\,. \tag{1.36}$$

Another way to construct affine equivariant estimators with high breakdown point (though usually not 50%) is to make use of the *minimum distance* principle. In this approach, one assumes that the data are sampled from an unknown member of a parametric family $\{F_\theta, \theta \in \Theta\}$ of model distributions. Let ν be some kind of distance between probability distributions. For any sample $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, one then constructs the empirical distribution F_n, and the minimum distance estimator $T(X)$ is defined as the value of θ that

$$\underset{\theta}{\text{Minimizes}}\ \nu(F_n, F_\theta)\,. \tag{1.37}$$

[Sometimes one does not use F_n but other estimators of the underlying density or cumulative distribution function (cdf)]. Minimum distance estimators go back to Wolfowitz (1957), who showed that they are "automatically" consistent. By means of the half-space metric, which is affine invariant, Donoho (1982) constructed an affine equivariant minimum distance estimator with 25% breakdown point. Donoho and Liu (1986) showed that any minimum distance estimator is "automatically" robust over contamination neighborhoods defined by the metric on which the estimator is based. Tamura and Boos (1986) used the minimum Hellinger distance approach to construct affine equivariant estimators of location and covariance, which are asymptotically normal and possess a 25% breakdown point in certain situations. Unfortunately, the minimum distance approach has not yet been developed to the stage where practical algorithms are proposed for multivariate applications.

d. The Detection of Multivariate Outliers, with Application to Leverage Points

It is very important to be able to identify outliers in multivariate point clouds. Such outliers do stick out in *certain* projections, but this does not make them easy to find because most projections do not reveal anything (like the projections on the coordinate axes in Figure 4 of Chapter 1).

The classical approach to outlier detection (see, e.g., Healy 1968) has been to compute the squared Mahalanobis distance for each observation as in (1.17), based on the arithmetic mean $T(X) = (1/n) \Sigma_{i=1}^{n} \mathbf{x}_i$ and the unbiased covariance estimator $\mathbf{C}(X) = (1/(n-1)) \Sigma_{i=1}^{n} (\mathbf{x}_i - T(X))'(\mathbf{x}_i - T(X))$. Points with large $\mathrm{MD}_i^2 = \mathrm{MD}^2(\mathbf{x}_i, X)$ (possibly compared to some χ_p^2 quantile) are then considered outliers.

However, as already explained above, this approach suffers from the fact that it is based on exactly those statistics that are most sensitive to outliers. This is particularly acute when there are several outliers forming a small cluster, because they will move the arithmetic mean toward them and (even worse) inflate the classical tolerance ellipsoid in their direction. As in Figure 2, the tolerance ellipsoid will do its best to encompass the outliers, after which their MD_i^2 won't be large at all.

Let us look at the modified wood gravity data in Table 8 of Chapter 6 in order to have a multivariate example. By considering only the explanatory variables, we obtain a data cloud of $n = 20$ points in $p = 5$ dimensions. We know that there are four outliers, but they do not show up in any of the coordinates separately. By means of the classical mean and covariance matrix, the squared Mahalanobis distance of each case may be computed. Actually, the values of MD_i^2 were already listed in Table 9 of Chapter 6. All MD_i^2 are smaller than the 95% quantile of the χ_5^2

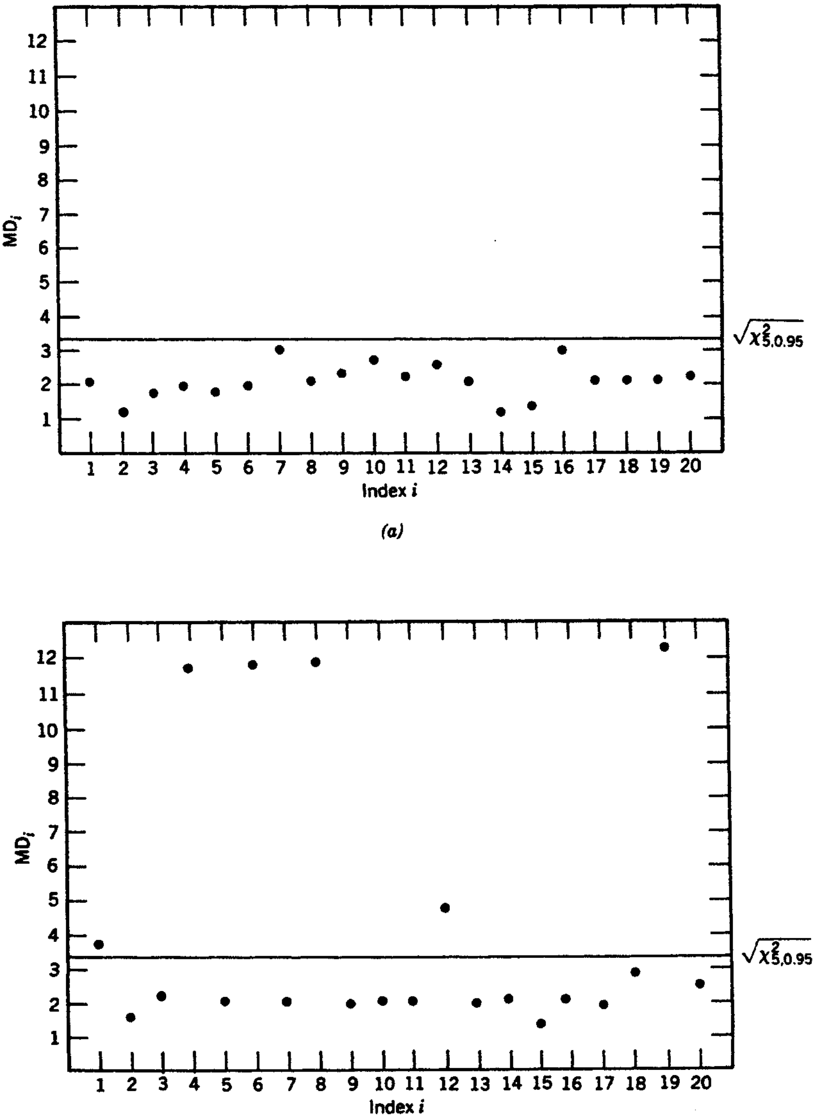
distribution, which equals about 11.07. The largest MD_i^2 belong to cases 7 and 16, but these are really good observations. An index plot of the $\mathrm{MD}_i = \sqrt{\mathrm{MD}_i^2}$ is shown in Figure 3*a*, which looks very regular.

In other applications it may happen that one (very bad) outlier is visible, but at the same time masks all the others. For an illustration of the masking effect, let us consider the Hawkins–Bradu–Kass data set described in Section 3 of Chapter 3. If we restrict our attention to the **x**-part of the data, we obtain 75 points in three dimensions. By construction, the first 14 points are outliers. However, the MD_i^2 (listed in Table 3 of Chapter 6) do not convey the whole picture. Indeed, the only MD_i exceeding $\sqrt{\chi^2_{3,0.95}} = 2.8$ are those of cases 12 and 14 (and case 13 comes close), whereas the other outliers yield quite inconspicuous values, as can be seen from Figure 4*a*. Only *after* the deletion of cases 12, 13, and 14 would the other outliers begin to obtain larger MD_i.

To overcome this weakness of the classical Mahalanobis distance, it is necessary to replace $T(X)$ and $\mathrm{C}(X)$ by robust estimators of location and scatter. A first step toward this goal was the use of affine equivariant M-estimators, as advocated by Campbell (1980), who gives some interesting examples. There is no doubt that this proposal constitutes a considerable improvement over the classical approach, but the low breakdown point of M-estimators limits its applicability. (Indeed, in the modified wood gravity data we have four outliers in a sample of size 20, which is already more than $1/(p+1) = 16.7\%$.)

To go any further, we need affine equivariant estimators with a high breakdown point. Therefore, we propose to compute the squared Mahalanobis distance relative to the MVE estimates of location and scatter, that is, to apply (1.17) where $T(X)$ and $C(X)$ are given by (1.22). In actual computations, one may use the results (1.26) of the resampling algorithm, as implemented in the program PROCOVIEV. For the explanatory part of the modified wood gravity data, this yields the robust MD_i displayed in Figure 3*b*, in which the four outliers are clearly visible. If we run the same program on the **x**-part of the Hawkins–Bradu–Kass data, we obtain Figure 4*b*. From this picture (constructed from a single program run!) it is at once evident that there are 14 far outliers. The fact that the robust MD_i consume more computer time than their classical counterparts is more than compensated by the reliability of the new method and the resulting gain of the statistician's time.

It is no coincidence that these examples were connected with regression data sets. Indeed, in regression analysis it is very important to discover leverage points, which are exactly those cases for which the $\mathbf{x}_i$-part is outlying. In retrospect, these examples also explain why the diagonal elements h_{ii} of the hat matrix are unable to cope with multiple



(a)

(b)

Figure 3. Index plot of Mahalanobis distances $\text{MD}_i = \sqrt{\text{MD}_i^2}$ of 20 points in five dimensions: (*a*) with the classical formula, (*b*) based on the MVE estimates.

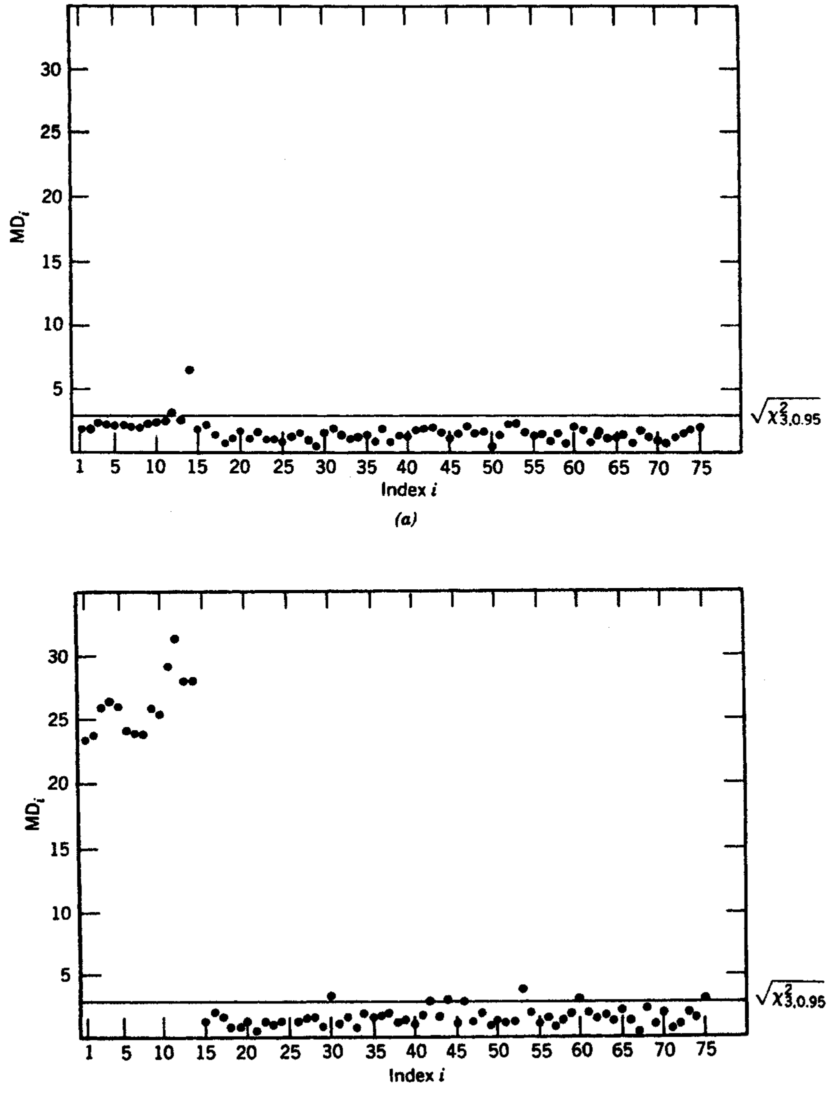


Figure 4. Index plot of Mahalanobis distances $MD_i = \sqrt{MD_i^2}$ of 75 points in three dimensions: (*a*) with the classical formula, (*b*) based on the MVE estimates.

leverage points. Indeed, these h_{ii} are related [through (2.15) of Chapter 6] to the Mahalanobis distance based on the classical estimates with zero breakdown point. An MVE analysis of the explanatory variables is very useful, because robust MD_i are more reliable *leverage diagnostics* than the h_{ii}. On the other hand, to know that $\mathbf{x}_i$ is a leverage point is not enough, because one also has to take the response y_i into account in order to find out if $(\mathbf{x}_i, y_i)$ is actually a regression outlier [i.e., whether $(\mathbf{x}_i, y_i)$ deviates from the linear pattern set by the majority of the data]. Hence, the question of whether $\mathbf{x}_i$ is a "good" or a "bad" leverage point can only be settled by means of a robust regression technique such as the LMS. Therefore, a combination of LMS and MVE gives more insight into regression data.

Also from a computational point of view, the LMS and the MVE fit together nicely. Indeed, their resampling algorithms are so similar that they could be run at the same time. The number of observations in each LMS subsample equals the number of "real" explanatory variables plus 1 (for the intercept term). This happens to be exactly the number of points that are needed in each MVE subsample. Moreover, the number of subsamples for the MVE can be taken equal to that for the LMS, because the probability of finding at least one "good" subsample is the same for both algorithms. Therefore, the whole resampling scheme is common to LMS and MVE, and the particular computations (finding a median squared residual for LMS, and the volume of a 50% ellipsoid for MVE) could be incorporated in the same program (and even run in a parallel way). Only a slight adaptation is necessary in the case of regression through the origin, because then one must restrict consideration to ellipsoids with center zero.

Gasko and Donoho (1982) provided another robustification of the MD_i to be used as a leverage diagnostic. Instead of first computing a robust estimator of location and scatter, they looked at the alternative formula (1.21) for the squared Mahalanobis distance and replaced it by the Stahel–Donoho measure of outlyingness (1.19).

e. Outlook

Multivariate location and covariance matrices are cornerstones of general multivariate statistics. Now that we have robust and affine equivariant estimators of location and scatter at our disposal, we can use them to robustify many classical techniques. There are basically two ways to do this:

(i) One may (in one way or other) insert robust estimators $T(X)$ and

C(X) instead of the classical mean and empirical covariance matrices.

(ii) Otherwise, one may first compute the robust estimators in order to identify outliers, which have to be corrected or deleted. Afterwards, the usual multivariate analyses may be carried out on the "cleaned" data. For instance, the correlation matrix of these points may be computed.

Maronna (1976), Campbell (1980), and Devlin et al. (1981) proposed to robustify *principal component analysis* by inserting robust covariance matrices obtained from affine equivariant M-estimators. Campbell (1982) likewise robustified *canonical variate analysis*. Our own proposal would be to apply the MVE, because of its better breakdown properties. Other fields in which the MVE could take the place of classical normal-theory covariance matrices are *cluster analysis* and *factor analysis*.

f. Some Remarks on the Role of Affine Equivariance

We have already seen that affine equivariance (1.7) is a very natural condition. Nevertheless, it is not so easy to combine with robustness. Indeed, the only affine equivariant multivariate location/covariance estimators with high breakdown point known so far (the Stahel–Donoho weighted mean and the MVE and its relatives) need substantially more computation time than the classical estimators.

The maximal breakdown point of affine equivariant estimators (1.7) is slightly lower than that of the larger class of translation equivariant estimators (1.2). Indeed, the breakdown point of any translation equivariant estimator of location is at most

$$\frac{[(n+1)/2]}{n}, \tag{1.38}$$

which can be proven as in Theorem 7 of Chapter 4 by replacing closed intervals by closed Euclidean balls. Note that this bound does not depend on p, the number of dimensions. Moreover, this bound is sharp because it is attained by the coordinatewise median, the L_1 estimator (1.9), and the estimator given by

$$\underset{T}{\text{Minimize}}\ \underset{i}{\text{med}}\ \|\mathbf{x}_i - T\|^2, \tag{1.39}$$

which corresponds to the center of the smallest sphere covering at least half of the points (Rousseeuw 1984). If we restrict our attention to the smaller class of all orthogonal equivariant estimators, the bound (1.38) is still sharp because both the L_1 estimator and (1.39) are orthogonal equivariant. It is only when we switch to affine equivariance that the best possible breakdown point of any pair $(T, \mathbf{C})$ goes down to

$$\frac{[(n-p+1)/2]}{n},$$

which is attained by the variants of the MVE and the MCD with $h = [(n+p+1)/2]$.

Some people have expressed the opinion that it should be very easy to obtain high-breakdown estimators that are affine equivariant, by first rescaling the observations. They proposed the following procedure: Calculate the ordinary covariance matrix $\mathbf{C}$ given by (1.15), and take a root $\mathbf{S}$ (i.e., $\mathbf{S}'\mathbf{S} = \mathbf{C}$). Transform the data as $\tilde{\mathbf{x}}_i = \mathbf{x}_i\mathbf{S}^{-1}$, apply to these $\tilde{\mathbf{x}}_i$ an easily computable estimator $\tilde{T}$ with $\varepsilon^* = 50\%$ which is not affine equivariant, and then transform back by putting $T = \tilde{T}\mathbf{S}$. First, we observe that this construction gives a unique result if and only if $\tilde{T}$ is equivariant for orthogonal transformations, because for any orthogonal matrix $\mathbf{\Gamma}$ the product $\mathbf{\Gamma}\mathbf{S}$ is also a root of $\mathbf{C}$. Also, $\tilde{T}$ has to be translation equivariant to ensure affine equivariance of T. Therefore, it seems that taking the L_1 estimator for $\tilde{T}$ will do the job. However, it turns out that its good breakdown behavior does not carry over.

Example

To show this, consider a two-dimensional example where a fraction $(1-\varepsilon)$ of the data is spherically bivariate normal (we assume that both coordinates follow a standard normal distribution $N(0, 1)$ and are independent of each other), and there is a fraction ε of outliers that are concentrated at the point with coordinates $(0, u)$. Here, $u > 0$ and $0 < \varepsilon < \frac{1}{2}$. (This configuration is sketched in Figure 5*a*.) The (usual) covariance matrix becomes

$$\mathbf{C} = \begin{bmatrix} 1-\varepsilon & 0 \\ 0 & (1-\varepsilon)(1+\varepsilon u^2) \end{bmatrix}.$$

Therefore, we can easily construct a root $\mathbf{S}$ of $\mathbf{C}$:

$$\mathbf{S} = \begin{bmatrix} \sqrt{1-\varepsilon} & 0 \\ 0 & \sqrt{(1-\varepsilon)(1+\varepsilon u^2)} \end{bmatrix}.$$

For u tending to ∞, the transformed $(\tilde{x}_1, \tilde{x}_2) = (x_1, x_2)\mathbf{S}^{-1}$ are situated as in Figure 5*b*: The $(1-\varepsilon)$-fraction gets concentrated on the $\tilde{x}_1$-axis with

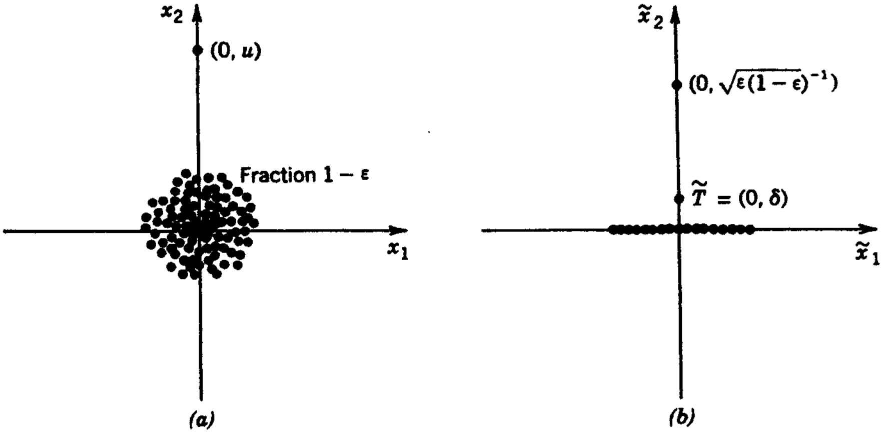


Figure 5. Sketch of the example: (*a*) original data, (*b*) after scaling by means of the classical covariance matrix.

univariate normal distribution $N(0, (1-\varepsilon)^{-1})$, and the ε-fraction lands at the point $(0, (\varepsilon(1-\varepsilon))^{-1/2})$.

For any finite value of u, the L_1 estimate of the transformed data lies on the vertical axis by symmetry, so we may denote it by $\tilde{T} = (0, \delta(u))$. Let us now look at the limit $\delta = \lim_{u\to\infty} \delta(u)$. By definition, δ minimizes the expected value of $\|(\tilde{x}_1, \tilde{x}_2) - (0, \delta)\|$, which is denoted by

$$D := \varepsilon((\varepsilon(1-\varepsilon))^{-1/2} - \delta) + (1-\varepsilon)\int_{-\infty}^{+\infty} [\delta^2 + \tilde{x}_1^2]^{1/2}\, dF(\tilde{x}_1)\,,$$

where F is the normal distribution $N(0, (1-\varepsilon)^{-1})$. By the substitution $z = (1-\varepsilon)^{1/2}\tilde{x}_1$, the average distance D equals

$$D = \varepsilon((\varepsilon(1-\varepsilon))^{-1/2} - \delta) + (1-\varepsilon)^{1/2}\int_{-\infty}^{+\infty} [(1-\varepsilon)\delta^2 + z^2]^{1/2}\, d\Phi(z)\,,$$

where Φ is the standard normal cdf. Because δ minimizes D, it follows that

$$0 = \partial D/\partial\delta = -\varepsilon + (1-\varepsilon)\int_{-\infty}^{+\infty} \{(1-\varepsilon)\delta^2/[(1-\varepsilon)\delta^2 + z^2]\}^{1/2}\, d\Phi(z)\,;$$

hence

$$\int_{-\infty}^{+\infty} \{(1-\varepsilon)\delta^2/[(1-\varepsilon)\delta^2 + z^2]\}^{1/2}\, d\Phi(z)$$

equals the positive constant $\varepsilon/(1-\varepsilon)$. This cannot happen for $\delta=0$, hence $\delta>0$. As the final estimate T equals $\tilde{T}\mathbf{S}$, we conclude that

$$\|T\|=\|\tilde{T}\mathbf{S}\|=\|(0,\delta(u))\mathbf{S}\|=\{(1-\varepsilon)(1+\varepsilon u^2)\}^{1/2}\delta(u)$$

tends to infinity for $u\rightarrow\infty$. (Note that this effect is caused by the explosion of $\mathbf{S}$!) This means that any fraction $\varepsilon>0$ can make T break down, so $\varepsilon^*(T)=0\%$.

Therefore, rescaling does not solve our problem unless we could start with a high-breakdown covariance estimator, but that is precisely what we were looking for in the first place.

2. ROBUST TIME SERIES ANALYSIS

Many people are involved in time series analysis and forecasting (e.g., in economics, engineering, and physical sciences). Software packages based on the work of Box and Jenkins (1976) are widely available, but unfortunately they are restricted to the least squares approach and do not provide for handling outliers. Indeed, the field of robust time series analysis has come into existence only fairly recently and has seen most of its activity during the last decade. (For an up-to-date survey, see Martin and Yohai 1984a.) This is partly because one had to wait for the development of robust regression techniques (of which extensive use is made) and also because of the increased difficulty inherent in dealing with dependencies between the observations. In this section, we shall first briefly describe the most commonly used time series models together with two main types of outliers and then consider robust estimation in this framework.

a. Autoregressive Moving Average Models and Types of Outliers

A time series is a sequence of n consecutive univariate observations

$$Y_1, Y_2, \ldots, Y_n$$

measured at regular intervals. Often the observations are fitted by a model

$$Y_t=\mu+X_t \qquad (t=1,\ldots,n), \tag{2.1}$$

where μ is a location parameter, and X_t follows a zero-mean *autoregressive moving average* model, abbreviated ARMA(p,q). This means that

$$X_t - \alpha_1 X_{t-1} - \cdots - \alpha_p X_{t-p} = e_t + \beta_1 e_{t-1} + \cdots + \beta_q e_{t-q}\,, \tag{2.2}$$

where the e_t are distributed like $N(0, \sigma^2)$. The unknown parameters are therefore

$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_p \end{bmatrix}, \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_q \end{bmatrix}, \qquad \mu, \quad \text{and} \quad \sigma\,.$$

The left-hand side of (2.2) is called the *autoregressive part*, because X_t depends on its previous (lagged) values $X_{t-1}, \ldots, X_{t-p}$. The right-hand side is referred to as the *moving average part*, because the actual error term at time t is a linear combination of the original $e_t, e_{t-1}, \ldots, e_{t-q}$ which cannot be observed directly.

The general ARMA(p, q) model is rather difficult to deal with numerically because of the moving average part. Fortunately, it is often sufficient to use an AR(p) model, which is much simpler because it contains only the autoregressive part. This model is obtained by putting $q = 0$ in (2.2), which reduces to

$$X_t = \alpha_1 X_{t-1} + \cdots + \alpha_p X_{t-p} + e_t\,. \tag{2.3}$$

(Such AR(p) models always provide a good approximation to ARMA data if one is willing to switch to a larger value of p.) In terms of the actually observed values $Y_t = \mu + X_t$, (2.3) becomes

$$(Y_t - \mu) = \alpha_1 (Y_{t-1} - \mu) + \cdots + \alpha_p (Y_{t-p} - \mu) + e_t\,,$$

which can be rewritten as

$$Y_t = \alpha_1 Y_{t-1} + \cdots + \alpha_p Y_{t-p} + \gamma + e_t\,, \tag{2.4}$$

where the intercept term γ equals $\mu(1 - \alpha_1 - \cdots - \alpha_p)$. There are $n - p$ complete sets $(Y_t, Y_{t-1}, \ldots, Y_{t-p})$ as t ranges from $p + 1$ to n, which can be used to estimate $\alpha_1, \ldots, \alpha_p, \gamma$ in (2.4) by means of some regression estimator. Indeed, we now have the linear model

$$\begin{bmatrix} Y_{p+1} \\ Y_{p+2} \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} Y_p & Y_{p-1} & \cdots & Y_1 & 1 \\ Y_{p+1} & Y_p & \cdots & Y_2 & 1 \\ \vdots & \vdots & & \vdots & \vdots \\ Y_{n-1} & Y_{n-2} & \cdots & Y_{n-p} & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_p \\ \gamma \end{bmatrix} + \begin{bmatrix} e_{p+1} \\ e_{p+2} \\ \vdots \\ e_n \end{bmatrix}.$$

We do not believe it is useful to add rows with the unobserved

$Y_0, Y_{-1}, \ldots$ put equal to zero, because this only leads to outliers as illustrated by Martin (1980, example 2).

A very simple (but frequently used) example is the AR(1) model

$$Y_t = \alpha_1 Y_{t-1} + \gamma + e_t . \tag{2.5}$$

In order for the Y_t to be stationary, the absolute value of α_1 must be less than 1. There are $n-1$ complete pairs (Y_t, Y_{t-1}), which may conveniently be plotted in a scatterplot of Y_t (vertically) versus Y_{t-1} (horizontally). Figure 6*a* is a graph of a well-behaved AR(1) time series without outliers, together with the corresponding scatterplot. In such an uncontaminated situation, it is not difficult to estimate the slope α_1 and the intercept γ adequately.

Fox (1972) and Martin (1981) considered two types of outliers that may occur in time series data. The first (and relatively innocent) class is that of *innovation outliers*, which may be modeled using the original ARMA framework (2.1)–(2.2), except that the distribution of the e_t is no longer normal but has heavy tails. This implies that when a certain e_{t_0} is outlying, it immediately affects X_{t_0}. Because the model (2.2) still holds, this exceptional value of X_{t_0} will then influence X_{t_0+1}, and then X_{t_0+2}, and so on. However, after a while this effect dies out. For real-world examples of innovation outliers in the area of speech recognition, see Lee and Martin (1987). Figure 6*b* shows a sample path similar to Figure 6*a*, but with one innovation outlier at time t_0.

Let us now look at the corresponding scatterplot in Figure 6*b*. The innovation outlier first appears at the point $A = (Y_{t_0-1}, Y_{t_0})$ which is an outlier in the vertical direction. But afterwards it yields some large (Y_{t-1}, Y_t) that do lie close to the original line of slope α_1. In the scatterplot, the innovation outlier therefore results in one outlier in the response variable and a certain number of "good" leverage points, which have the potential of improving the accuracy of the regression estimate of α_1. When least squares is used, these good points tend to compensate for the effect of the one outlier (Whittle 1953). In fact, when the heavy-tailed distribution of e_t is symmetric, one may estimate α_1 with a better precision than in the case of Gaussian e_t.

Let us now look at the second type of outlier. *Additive outliers* occur when contamination is added to the Y_t themselves, so the resulting observations no longer obey the ARMA model. To formalize this, we must extend (2.1) to

$$Y_t = \mu + X_t + V_t . \tag{2.6}$$

For most observations V_t is zero, but we assume that $P(V_t \neq 0) = \varepsilon$ where

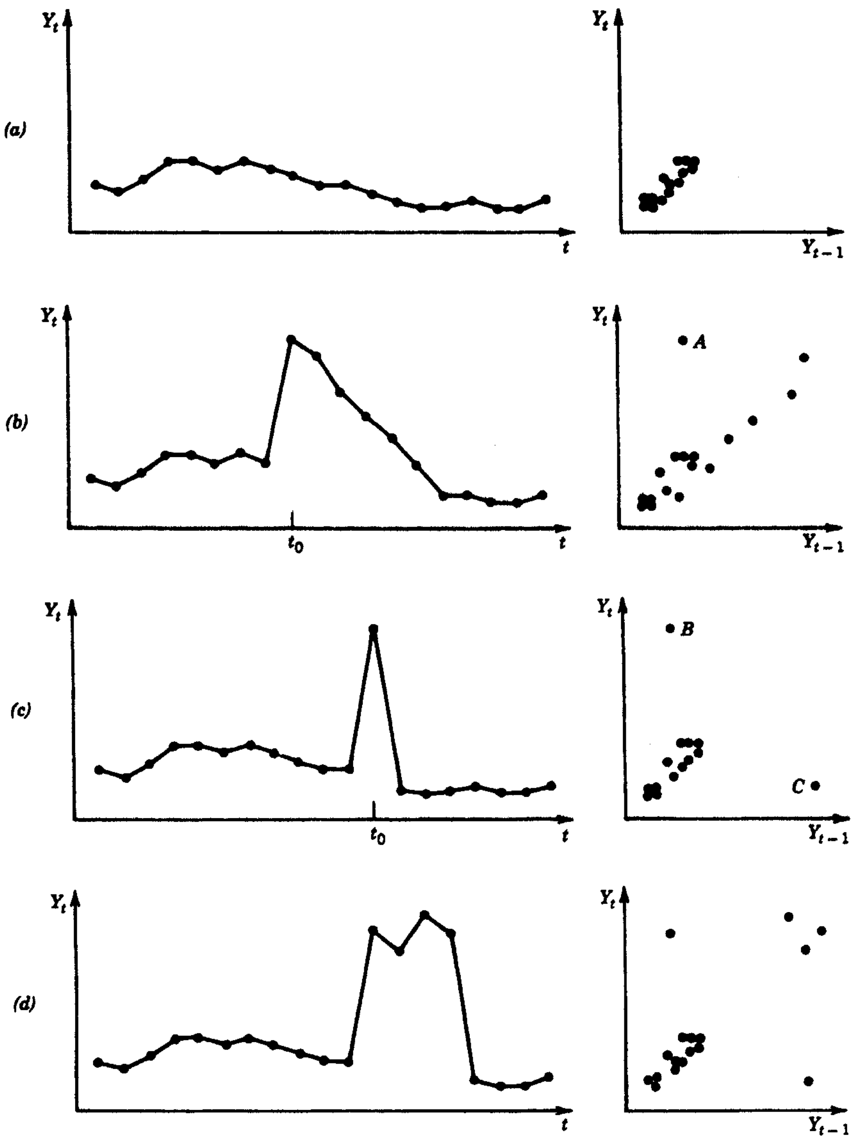


Figure 6. Sketch of time series with corresponding plot of Y_t versus Y_{t-1} for (*a*) no outliers, (*b*) innovation outlier, (*c*) isolated additive outlier, and (*d*) patch of additive outliers.

this fraction of additive outliers is positive and not too large. (This is similar to replacement contamination, which is the basis of the finite-sample breakdown point approach adopted in this book.) According to Martin (1981), many time series with outliers arising in practice have ε between 1% and 25%. Letting the V_t be i.i.d. random variables yields a model for "isolated" or "scattered" outliers, occurring independently of each other. On the other hand, certain dependency structures for the V_t could be used to describe outliers coming in "patches," "clumps," or "bursts."

Figure 6*c* provides an example with an isolated additive outlier, as might be caused by a keypunch or transcription error (e.g., a misplaced decimal point) or someone accidentally touching a recording instrument. This one outlying Y_{t_0} gives rise to *two* exceptional points in the scatterplot, namely $B = (Y_{t_0-1}, Y_{t_0})$, which is an outlier in the response variable, and $C = (Y_{t_0}, Y_{t_0+1})$, which is a bad leverage point. [In an AR(p) model, we obtain one outlier in the vertical direction and p leverage points.] Additive outliers are therefore a cause for much greater concern than innovative ones, because the usual least squares (LS) method cannot cope with leverage points and will yield biased parameter estimates. In Figure 6*c*, the LS estimate of the slope will be biased toward 0.

Figure 6*d* shows a whole patch of additive outliers, as is not uncommon in real data. For instance, weekly sales may be positively affected by a bargain month. Another example is that of industrial production that temporarily decreases because of a strike. The resulting effects in the scatterplot are equally unpleasant, because the leverage points will tend to make the estimated slope biased toward 1.

Additive outliers appear to occur more often than innovation outliers in actual applications, and they are more difficult to deal with because leverage points pose bigger problems than vertical outliers. This indicates the need for reliable robust estimators for time series parameters.

A research area that has recently received a lot of attention is the qualitative robustness of such estimators, generalizing Hampel's (1971) i.i.d. definition. We shall not treat this topic here, but refer the reader to Martin (1980, p. 237), Papantoni-Kazakos and Gray (1979), Bustos (1981), Cox (1981), Boente et al. (1982), and Papantoni-Kazakos (1984).

b. *M*- and GM-Estimators

If an ARMA process contains only innovation outliers, then the usual *M*-estimators for regression (which bound the influence of vertical outliers) are appropriate. Let us first restrict our attention to AR(p) models (2.4). In this case, classical least squares corresponds to

$$\underset{\hat{\gamma},\hat{\alpha}}{\text{Minimize}} \sum_t (Y_t - \hat{\gamma} - \hat{\alpha}_1 Y_{t-1} - \cdots - \hat{\alpha}_p Y_{t-p})^2 . \tag{2.7}$$

This can easily be generalized to M-estimation:

$$\underset{\hat{\gamma},\hat{\alpha}}{\text{Minimize}} \sum_t \rho\left(\frac{Y_t - \hat{\gamma} - \hat{\alpha}_1 Y_{t-1} - \cdots - \hat{\alpha}_p Y_{t-p}}{\hat{\sigma}}\right), \tag{2.8}$$

where $\hat{\sigma}$ is a robust estimate of the error scale. This means that any computer program for regression M-estimation may be applied to such data. Finally, a natural estimate of the location parameter μ in (2.1) is given by

$$\hat{\mu} = \frac{\hat{\gamma}}{1 - \hat{\alpha}_1 - \cdots - \hat{\alpha}_p} . \tag{2.9}$$

Under regularity conditions, $\hat{\gamma}$, $\hat{\alpha}$, and $\hat{\mu}$ are consistent and asymptotically normal (Lee and Martin 1982). Their Cramer–Rao bound and efficiency robustness can be found in Martin (1982).

In the general ARMA(p, q) framework, the least squares estimator is given by

$$\underset{\hat{\gamma},\hat{\alpha},\hat{\beta}}{\text{Minimize}} \sum_t r_t(\hat{\gamma}, \hat{\alpha}, \hat{\beta})^2 . \tag{2.10}$$

Since the residual r_t is now a nonlinear function of the parameters, we already face a nonlinear estimation problem in the classical approach (Box and Jenkins 1976). In order to provide for innovation outliers, Martin (1981) proposed to replace (2.10) by an M-estimator

$$\underset{\hat{\gamma},\hat{\alpha},\hat{\beta}}{\text{Minimize}} \sum_t \rho\left(\frac{r_t(\hat{\gamma}, \hat{\alpha}, \hat{\beta})}{\hat{\sigma}}\right) \tag{2.11}$$

and then to estimate μ as in (2.9). This yields essentially the same asymptotic behavior as in the autoregression case, as was proved by Lee and Martin (1982).

However, ordinary M-estimators are not robust for additive outliers, as was pointed out by Martin (1979) in the context of autoregression. Therefore, it seems natural to apply GM-estimators [formulas (2.12) and (2.13) of Chapter 1] instead, because they can also bound the influence of leverage outliers. Some computational aspects and Monte Carlo results concerning the application of GM-estimators to autoregression models can be found in Martin and Zeh (1978), Denby and Martin (1979), and Zeh (1979), and some actual analyses were performed by Martin (1980).

Bustos (1982) proved consistency and asymptotic normality. Influence functions of GM-estimators in autoregression models were defined by Martin and Yohai (1984b, 1986); see also Künsch (1984). For the computation of GM-estimators in general ARMA models, Stockinger (1985) constructed an algorithm based on nonlinear regression and carried out a small Monte Carlo study with it.

Other methods of estimation involve robust instrumental variables (Martin 1981, Section VI), approximate non-Gaussian maximum likelihood (Martin 1981, Section VII; Pham Dinh Tuan 1984), and residual autocovariance estimators (Bustos and Yohai 1986, Bustos et al. 1984). Bruce and Martin (1987) considered multiple outlier diagnostics for time series. Robust tests for time series were considered by Basawa et al. (1985). A summary of robust filtering, smoothing, and spectrum estimation can be found in Martin and Thomson (1982), and Kassam and Poor (1985) presented a survey of robust techniques for signal processing.

c. Use of the LMS

Any attempt to define breakdown points for time series parameter estimators must take the nature of the failure mechanism into account, because breakdown can occur in different ways. Martin (1985) gave breakdown points for some estimators that yield to analytical calculations. Moreover, he constructed maximal bias curves that display the global bias–robustness of an estimator.

At any rate, it appears extremely useful to have a high-breakdown regression method when fitting autoregressive models to time series with outliers. Therefore, we propose to apply the LMS to the AR(p) model (2.4). Indeed, as we saw above, an isolated additive outlier occurs in $p+1$ subsequent regression cases $(Y_t, Y_{t-1}, \ldots, Y_{t-p})$, yielding one vertical outlier and p leverage points. This means that the fraction of additive outliers in the original time series may give rise to a much higher fraction of contaminated data when it comes to fitting the autoregression, which is all the more reason for wanting to use a very robust regression estimator. The LMS can be applied to such situations [with $n-p$ cases and $p+1$ coefficients, as in (2.4)] as easily as to any other data set. However, further research is necessary before the LMS may be used in general ARMA(p, q) models, because of the nonlinearity caused by the moving average part.

Example

Table 1, part (a) lists the monthly time series RESX which originated at Bell Canada. It describes the installation of residential telephone exten-

Table 1. Residential Extensions Data

Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sep.	Oct.	Nov.	Dec.
(*a*) *Original Series* RESX$_t$ *from January 1966 to May 1973*											
10165	9279	10930	15876	16485	14075	14168	14535	15367	13396	12606	12932
10545	10120	11877	14752	16932	14123	14777	14943	16573	15548	15838	14159
12689	11791	12771	16952	21854	17028	16988	18797	18026	18045	16518	14425
13335	12395	15450	19092	22301	18260	19427	18974	20180	18395	15596	14778
13453	13086	14340	19714	20796	18183	17981	17706	20923	18380	17343	15416
12465	12442	15448	21402	25437	20814	22066	21528	24418	20853	20673	18746
15637	16074	18422	27326	32883	24309	24998	25996	27583	22068	75344	47365
18115	15184	19832	27597	34256							
(*b*) *Seasonal Differences* $Y_t = \text{RESX}_{t+12} - \text{RESX}_t$ *for* $t = 1, \ldots, 77$											
380	841	947	−1124	447	48	609	408	1206	2152	3232	1227
2144	1671	894	2200	4922	2905	2211	3854	1453	2497	680	266
646	604	2679	2140	447	1232	2439	177	2154	350	−922	353
118	691	−1110	622	−1505	−77	−1446	−1268	743	−15	1747	638
−988	−644	1108	1688	4641	2631	4085	3822	3495	2473	3330	3330
3172	3632	2974	5924	7446	3495	2932	4468	3165	1215	54671	28619
2478	−890	1410	271	1373							

Source: Martin et al. (1983).

sions in a fixed geographic area from January 1966 to May 1973, so it contains 89 observations. Looking at these data, it appears that the values for November and December 1972 are extremely large. These outliers have a known cause, namely a bargain month (November) with free installation of residence extensions, and a spillover effect in December because all of November's orders could not be filled in that month.

When analyzing these data, Brubacher (1974) first performed seasonal differencing, that is, he constructed the series

$$Y_t = \mathrm{RESX}_{t+12} - \mathrm{RESX}_t\,, \qquad t = 1, \ldots, 77\,, \tag{2.12}$$

which is listed in Table 1, part (b) and shown in Figure 7. The outliers are now the observations 71 and 72. Brubacher found that this new series was stationary with zero mean (except for the outliers) and that it could adequately be described by an autoregressive model without intercept. In what follows, we shall first fit an AR(1) model and then an AR(2).

The AR(1) model is given by

$$Y_t = \alpha_1 Y_{t-1} + e_t\,. \tag{2.13}$$

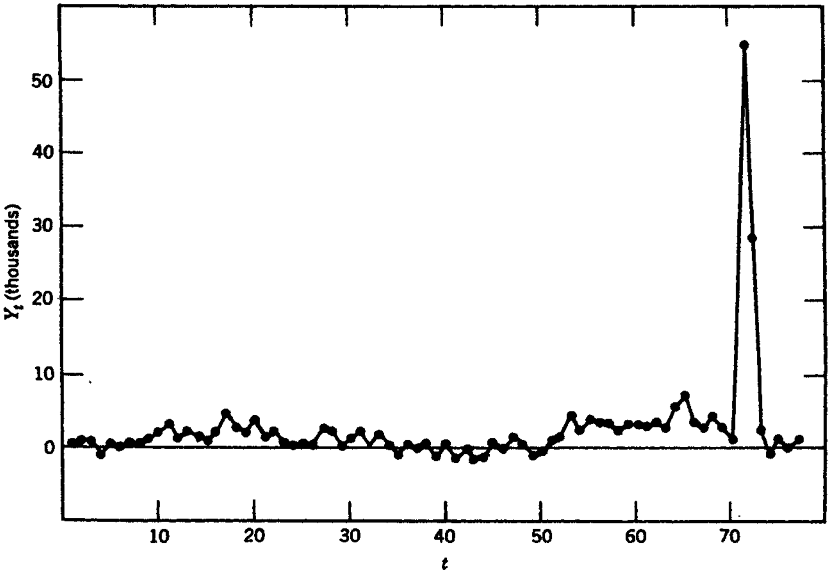


Figure 7. Plot of the series Y_t, $t = 1, \ldots, 77$.

This model does not contain an intercept term because the expected value of Y_t appears to be about zero, so $\mu = 0$ in (2.1) and hence $\gamma = \mu(1 - \alpha_1) = 0$ in (2.4). The lag-one scatterplot (i.e., the plot of Y_t versus Y_{t-1}) is presented in Figure 8 and displays 76 pairs (Y_1, Y_2), $(Y_2, Y_3), \ldots, (Y_{76}, Y_{77})$. It looks a little bit like Figure 6*d*, indicating that a (small) patch of two additive outliers would provide a plausible description of the situation. The pair (Y_{70}, Y_{71}) is a regression outlier in the vertical direction, (Y_{71}, Y_{72}) is a leverage point that does not lie far away from the linear pattern formed by the majority, and (Y_{72}, Y_{73}) is a bad leverage point. Applying least squares to this data set yields the upper portion of column 1 in Table 2. The outliers have blown up $\hat{\sigma}$, so the only case with large standardized residual $|r_i/\hat{\sigma}|$ is (Y_{70}, Y_{71}). This means that the December 1972 value is masked in a routine LS analysis. On the other hand, the LMS obtains a higher slope $\hat{\alpha}_1$ because it is not dragged down by the bad leverage point, and its $\hat{\sigma}$ is four times smaller. Consequently, both (Y_{70}, Y_{71}) and (Y_{72}, Y_{73}) are now identified by their large $|r_i/\hat{\sigma}|$, whereas (Y_{71}, Y_{72}) has a small LMS residual, which confirms its classification as a good leverage point. The next column of Table 2 shows the LMS-based reweighted least squares (RLS) estimates, which are quite similar.

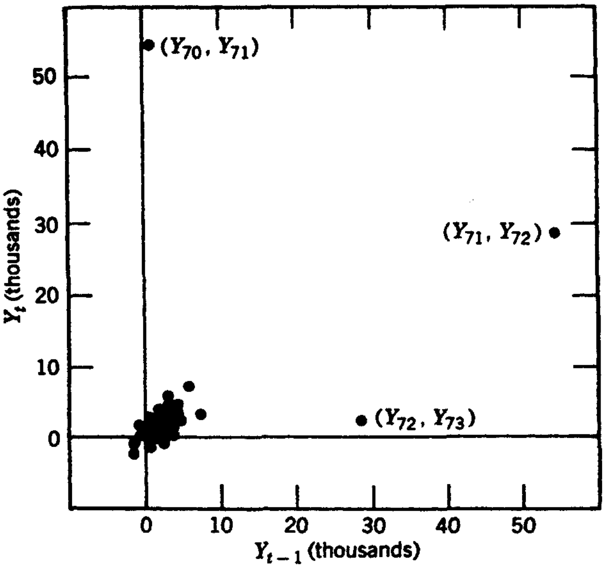


Figure 8. Plot of Y_t versus Y_{t-1}.

Table 2. Parameter Estimates for Y_t Series

Estimates	LS	LMS	RLS
AR(1) $\hat{\alpha}_1$	0.482	0.535	0.546
$\hat{\sigma}$	6585	1501	1422
AR(2) $\hat{\alpha}_1$	0.533	0.393	0.412
$\hat{\alpha}_2$	−0.106	0.674	0.501
$\hat{\sigma}$	6636	1162	1111

Let us now consider an AR(2) model

$$Y_t = \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + e_t\,, \tag{2.14}$$

which does not have an intercept term either (for the same reason as before). We try to fit the model (2.14) to the 75 triples (Y_1, Y_2, Y_3), $(Y_2, Y_3, Y_4), \ldots, (Y_{75}, Y_{76}, Y_{77})$. Now the two outliers Y_{71} and Y_{72} affect four regression cases:

Case no. 69: (Y_{69}, Y_{70}, Y_{71}) is a vertical regression outlier

$$\left.\begin{array}{ll}\text{Case no. 70:} & (Y_{70}, Y_{71}, Y_{72})\\ \text{Case no. 71:} & (Y_{71}, Y_{72}, Y_{73})\\ \text{Case no. 72:} & (Y_{72}, Y_{73}, Y_{74})\end{array}\right\}\ \text{are leverage points}\,.$$

Applying LS to all 75 cases yields the lower portion of column 1 in Table 2. Because the LS fit is attracted by the leverage points and $\hat{\sigma}$ is blown up, only case 69 may be identified through its large $|r_i/\hat{\sigma}|$, whereas for all other cases, $|r_i/\hat{\sigma}|$ is very small. The LMS fit in the next column is totally different and possesses a much lower value of $\hat{\sigma}$. The observations with large $|r_i/\hat{\sigma}|$ are now 69, 70, 71, and 72. Reweighting on the basis of the LMS yields comparable estimates.

Martin (1980) applied a GM-estimator to these data, yielding $\hat{\alpha}_1 = 0.51$ and $\hat{\alpha}_2 = 0.38$, and the robust filter used by Martin et al. (1983) gives a similar result. Both solutions are close to the estimates obtained by Brubacher (1974) by means of an interpolation technique described in Brubacher and Wilson (1976), which, however, requires that one specify the outlier positions in advance.

For the AR(2) model, all these robust estimators yield a positive $\hat{\alpha}_2$ and a small $\hat{\sigma}$, which makes them rather distinct from the naive LS fit. Nevertheless, it is somewhat discomforting to see relatively large differences between the various robust estimates (although they all identify the same outliers). A possible explanation may be the high correlation between Y_{t-1} and Y_{t-2}, at least when the outliers are taken out first (this can be seen from the lag-one plot in Figure 8), so we are almost in a collinearity situation: Whenever $\hat{\alpha}_1$ increases, $\hat{\alpha}_2$ will have to decrease. A way to avoid this problem would be to use only one explanatory variable, by sticking to the AR(1) model. However, Martin (1980) found that the AR(2) model is more suitable in this example, by means of a robustified version of Akaike's AIC function for estimating the order of the autoregression.

Also other people have used the LMS in a time series context. Heiler (1984) gave an example concerning the interest rates of a certain type of bonds. The data were recorded monthly from January 1965 to December 1982, so $n = 216$. He fitted AR(1), AR(2), and AR(3) models by means of both LS and LMS, and showed that the LMS coefficients were much easier to interpret. He further replaced the usual R^2 and partial correlation coefficients by means of Kendall's tau, which he then used to select the AR(2) fit. In his diploma thesis, Stahlhut (1985) analyzed the same example, as well as the residential extensions data and one other time series, and found that the LMS and the corresponding RLS yielded more reliable estimates and forecasts than did LS.

3. OTHER TECHNIQUES

In this final section we skim some other statistical topics in which robustification is emerging, and we provide a few supplementary references. Our selection is, of course, subjective and largely incomplete.

To begin with, there is the area of *orthogonal regression*, in which the residuals are no longer measured vertically, but instead are measured perpendicular to the regression line or hyperplane. This means that the actual (Euclidean) distances from the observations to the regression hyperplane are computed. This approach is more symmetric, in the sense that it does not matter which variable is selected as the response. On the other hand, standardization of the variables does not preserve orthogonality, so everything depends very much on the measurement units that are chosen by the user. In classical theory, the sum of the squared orthogonal distances is minimized. It is trivial to generalize the LMS by minimizing the *median* of the squared orthogonal distances. (Note that the LMS then really corresponds to the narrowest band covering half of the points, because the thickness of the band is now measured in the usual way.) Also from the computational point of view this presents no problems, because we only have to divide the original objective $\text{med}_i\, r_i^2$ (with "vertical" r_i^2) by the factor $\theta_1^2 + \cdots + \theta_{p-1}^2 + 1$, at each trial estimate. By adding a couple of lines of code, we made orthogonal versions of both PROGRESS (described in Section 1 of Chapter 5) and the simple regression program (see Section 2 of Chapter 5). The more general *errors in variables* model can be dealt with in the same way, as far as LMS is concerned. Recently, Brown (1982) constructed w-estimators in this context, whereas Zamar (1985) extended M-estimators. Kelly (1984) defined influence functions for the errors in variables situation.

Another important field is that of *nonlinear regression*, which is also useful in time series analysis and generalized linear models. The classical approach is based on the LS criterion and makes use of several iterative methods such as the Gauss–Newton, steepest descent, and Marquardt algorithms. Dutter and Huber (1981) generalized the nonlinear LS algorithm of Nagel and Wolff (1974) to nonlinear robust regression. Ways in which the LMS estimator may be applied to nonlinear problems are presently being investigated.

Nowadays there is a lot of interest in the robustification of *analysis of variance*. For the approach based on M-estimators we refer to Huber (1981, Section 7.10). Robust tests in general linear models (making use of GM-estimators) can be found in Hampel et al. (1986, Chapter 7). In two-way tables, the median polish provides a robust estimator of row and column effects (for a recent survey, see Hoaglin et al. 1983, 1985). By means of zero–one variables, a two-way table can also be written as a linear model, but it is not always possible to apply the LMS to it because the number of cases (cells) must be larger than twice the number of parameters. On the other hand, the LMS becomes more useful when also interval-scaled covariates occur, which may contain outliers.

Another topic is estimation on the circle and the sphere, corresponding to *directional data* (Watson 1983). On these spaces the worst kind of contamination is of an asymmetric type, when the outliers are 90° away from the "true" direction. Ko (1985) used the LMS in this framework and found that it could withstand a high percentage of contamination. For instance, he applied the LMS (which amounts to the "shortest arc" on the circle) to the well-known frog data (Ferguson et al. 1967, Collett 1980), in which the sun compass orientation of 14 frogs was investigated. In this example, the LMS comes close to the true home direction. Recently, Ko and Guttorp (1986) considered a standardized version of the gross-error-sensitivity that is tailored to estimators for directional data.

The currently used algorithms for *multidimensional scaling* are very susceptible to outlying dissimilarities, partly because these methods employ some variant of the classical Young–Householder–Torgerson scaling procedure for their initial estimator. There is a growing awareness of this problem, causing a recent interest in robustness aspects. A type of influence function was defined in this setting by de Leeuw and Meulman (1986), following a suggestion of Kruskal and Wish (1978, pp. 58–60).

Spence and Lewandowsky (1985) proposed a resistant multidimensional scaling method and implemented it in a program called TUFSCAL. Their algorithm is a median-type modification of Newton's method, and to avoid iteration toward a wrong solution they use a very robust starting configuration based on ranks. The new method was compared to some of the currently employed alternatives in a simulation study, and its breakdown value turned out to be quite high (although it is probably hard to summarize in a single number because it depends on the particular pattern of outliers). Influence functions for the *principal components* problem have been computed recently by Critchley (1985), and influence functions for various forms of *correspondence analysis* and *canonical analysis* can be deduced from Gifi (1981) and de Leeuw (1984).

In biological studies of morphological change, one makes use of *shape comparison* techniques. Suppose that the first shape is characterized by the points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ in the plane, and the second is characterized by $\mathbf{y}_1, \ldots, \mathbf{y}_n$. The points $\mathbf{x}_i$ now have to go through a rigid transformation f (composed of rotation, translation, and magnification) such that the images $f(\mathbf{x}_i)$ come as close as possible to the $\mathbf{y}_i$. The classical LS criterion is

$$\underset{f}{\text{Minimize}} \sum_{i=1}^{n} \|f(\mathbf{x}_i) - \mathbf{y}_i\|^2 . \tag{3.1}$$

Its nonrobustness causes the effects of large deformations to become smeared out and therefore obscured. Siegel and Benson (1982) use repeated medians to obtain a robust comparison, which succeeds in detecting localized shape differences (e.g., in an example on primate skulls). The method was generalized to three dimensions by Siegel and Pinkerton (1982) and used for protein molecules. By replacing the sum in (3.1) by a median, the LMS could also be applied.

In astronomy, one often has to *estimate shifts* between stellar spectra. For instance, this happens when studying the orbits of binary star systems, in which the same star may be moving toward us at one point in time, but away from us at another time. This change of velocity may be measured by means of the Doppler effect, which causes the star spectrum to shift to the red or to the blue. To estimate the actual shift between two observed spectra in an automatic way, astronomers have typically used LS or product cross-correlation methods. However, it turns out that these techniques run into trouble in the case of "noisy" spectra, which occur frequently for hot and luminous stars. For this reason, Rousseeuw (1987a) proposed to apply the L_1 approach, that is, to choose the shift that minimizes the sum of absolute values of the differences between the intensities g and h:

$$\underset{\Delta\lambda}{\text{Minimize}} \sum_i |g(\lambda_i) - h(\lambda_i + \Delta\lambda)| . \qquad (3.2)$$

This method is more robust against inaccuracies in the data, and it appears to be very powerful for analyzing complicated spectra, where asymmetric peaks, as well as a combination of absorption and emission features, may occur. The method was tried out on artificial profiles and on the spectrum of Beta Arietes. In all cases it compared favorably with the classical approach. The reason why L_1 does so well here is that there can be no leverage points in this one-dimensional framework. Therefore, it does not appear necessary to replace the sum in (3.2) by a median, which would be a lot more expensive in view of the size of astronomical data sets.

A similar situation occurs in cluster analysis, at least in those cases where orthogonal equivariance is sufficient (things become more difficult when affine equivariance is imposed). Indeed, in the social sciences many data sets consist of a collection of subjective distances (or dissimilarities) between entities for which no measurements or coordinates are available. Such entities are not easily represented as points in a linear space, so affine equivariance is not applicable, whereas orthogonal transformations on the original points (when the latter exist) do preserve the collection of distances. In this context, Kaufman and Rousseeuw (1988) apply L_1-type methods. For instance, the k-median approach searches for k representative objects (called "medoids") such that

$$\frac{1}{n} \sum_{i=1}^{n} d(i, m(i)) \qquad (3.3)$$

is minimized, where $d(i, m(i))$ is the distance (or dissimilarity) of object i to the closest medoid, denoted by $m(i)$. Then a partitioning into k clusters is constructed by assigning each object to the nearest medoid. This method is a robust alternative to the classical k-means technique, in which sums of *squared* distances are minimized. Again, the L_1 approach is suitable because there exist no leverage points in this situation. Also, graphical methods can be used to discover outliers in such data sets (see, e.g., Rousseeuw 1987b and the references cited therein). Recently, Cooper and Milligan (1985) examined the effect of outliers on hierarchical clustering procedures.

EXERCISES AND PROBLEMS

Section 1

1. Show that the L_1 estimator is orthogonal equivariant, and construct a small two-dimensional example to illustrate that it is not affine equivariant for $p \geq 2$. What happens for $p = 1$?
2. Show that the coordinatewise median lies in the convex hull of the sample when $p \leq 2$.
3. Suppose that T is affine equivariant, $\mathbf{A}$ is a nonsingular matrix, $\mathbf{b}$ is any vector, and X is any sample of size n. Then show that $\varepsilon_n^*(T, X\mathbf{A} + \mathbf{b}) = \varepsilon_n^*(T, X)$ by using the fact that affine transformations map bounded sets on bounded sets.
4. (From Donoho 1982.) Show that convex peeling is affine equivariant, and draw the subsequent convex hulls by hand in a two-dimensional example. What is the most harmful configuration of outliers?
5. Show that the outlyingness-weighted mean given by (1.19) and (1.20) is affine equivariant.
6. Construct an approximate algorithm for the outlyingness-weighted mean, by either (a) projecting on lines through pairs of points, as done by Donoho or (b) projecting on lines orthogonal to hyperplanes through p points, as proposed by Stahel. Which algorithm is faster? Show that (b) is affine equivariant but that (a) is not.
7. What happens to the estimators (1.4), (1.8), (1.9), (1.11), (1.18), (1.20), (1.22), and (1.31) if $p = 1$? Compare these special cases by means of one of the univariate data sets of Chapter 4.
8. Show that the breakdown point of the MVE estimator becomes exactly (1.36) when $h = [(n + p + 1)/2]$.
9. Use some standard statistical program to compute the principal components of the $\mathbf{x}$-part of the modified wood gravity data, before and after deletion of the outliers identified by the MVE. Repeat this for the $\mathbf{x}$-part of the Hawkins–Bradu–Kass data.
10. Show that the breakdown point of any translation equivariant estimator of location is at most (1.38), and give an example to show that this bound is sharp.
11. Show that the breakdown point of any affine equivariant covariance estimator is at most $[(n - p + 1)/2]/n$. (Solution in Davies 1987.)

Section 2

12. (From Heiler 1984 and Stahlhut 1985.) Use PROGRESS to fit AR(1), AR(2), and AR(3) models to the time series of Table 3.

Table 3. Monthly Interest Rates of Bonds, from January 1965 to December 1982

Year	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sep.	Oct.	Nov.	Dec.
1965	6.3	6.2	6.4	6.6	6.9	7.1	7.2	7.4	7.4	7.4	7.5	7.6
1966	7.6	7.6	7.6	7.7	7.8	8.0	8.3	8.6	8.5	8.1	7.9	7.6
1967	7.6	7.5	7.3	7.0	6.8	6.7	6.8	6.8	6.8	6.8	6.7	6.8
1968	6.8	6.8	6.8	6.8	6.6	6.5	6.5	6.3	6.3	6.3	6.4	6.4
1969	6.2	6.2	6.3	6.5	6.6	6.8	6.9	7.1	7.1	7.3	7.2	7.1
1970	7.4	7.8	8.2	8.2	8.1	8.4	8.6	8.5	8.5	8.6	8.6	8.3
1971	7.9	7.7	7.7	7.8	7.9	8.1	8.3	8.3	8.2	8.0	7.9	7.8
1972	7.8	7.6	7.4	7.6	8.0	8.1	8.2	8.1	8.1	8.1	8.3	8.4
1973	8.6	8.5	8.5	8.6	9.2	9.8	9.8	9.9	9.6	9.8	9.4	9.5
1974	9.5	9.6	10.3	10.6	10.6	10.5	10.6	10.6	10.3	10.3	10.2	9.7
1975	9.4	8.8	8.7	8.6	8.3	8.2	8.3	8.3	8.7	8.7	8.5	8.2
1976	8.2	8.0	7.8	7.7	7.9	8.1	8.2	8.2	8.0	7.9	7.5	7.3
1977	7.1	6.9	6.6	6.6	6.3	6.3	6.3	6.0	6.0	5.9	5.9	5.9
1978	5.8	5.7	5.5	5.4	5.8	6.0	6.2	6.5	6.3	6.3	6.4	6.5
1979	6.6	6.8	6.9	7.1	7.4	7.9	7.9	7.6	7.6	7.7	8.1	7.9
1980	7.9	8.2	9.1	9.5	8.7	8.2	8.0	7.8	8.1	8.4	8.8	8.8
1981	9.0	9.6	10.1	10.0	10.2	10.9	10.5	11.0	11.2	10.4	10.0	9.7
1982	9.8	9.7	9.5	8.9	8.7	9.1	9.3	9.0	8.7	8.3	8.2	7.9

Source: Statistische Beihefte zu den Monatsberichten der Deutschen Bundesbank: Renditen von festverzinslicher Wertpapiere, Neuemissionen.

Compare the LS and LMS coefficients. Can you identify outliers? (Also look at a plot of the data.) Which p would you select?

13. Take a time series (from your own experience) that is adequately described by an AR(p) model and analyze it by means of PROGRESS to compare the LS and LMS estimates. Then throw in a few additive outliers and run the program again to investigate their effect on both estimators.

Section 3

14. Take a simple regression data set (e.g., from Chapter 2). Is the LMS regression of y on x the same as that of x on y? What happens when the orthogonal variant of LMS is used?

15. (From Ko 1985.) Ferguson et al. (1967) described an experiment in which 14 frogs were captured, transported to another place, and then released to see if they would find their way back. Taking 0° to be due north, the directions taken by these frogs were

104°, 110°, 117°, 121°, 127°, 130°, 136°,
145°, 152°, 178°, 184°, 192°, 200°, 316°.

(a) Plot these data on a circle with center 0 and radius 1. Indicate the approximate position of the "mode" (in a loose sense).

(b) Compute the population mean, given by the point

$$T = \frac{\sum_{i=1}^{14} \mathbf{x}_i}{\left\| \sum_{i=1}^{14} \mathbf{x}_i \right\|},$$

where the $\mathbf{x}_i$ are the observations on the unit circle.

(c) Collett (1980) found that there is some evidence to indicate that 316° is an outlier. Does it make much difference to the population mean whether this point is deleted? Explain.

(d) Compute (approximately) the spherical median of these data, given by

$$\underset{\hat{\boldsymbol{\theta}}}{\text{Minimize}} \sum_{i=1}^{14} d(\mathbf{x}_i, \hat{\boldsymbol{\theta}}),$$

where the distance d is measured along the circle.

(e) Compute the LMS estimator, corresponding to the midpoint of the shortest arc containing (at least) eight points. Which points do now appear to be outliers?

(f) Which of these estimates comes closest to the true home direction 122°? Compare the various estimates with the "mode" (a). Which outliers are influential for (b)?

16. How would you generalize the LMS to data on a sphere?

17. Apply a standard multidimensional scaling program to a data set in which you replace a few dissimilarities by large values.