In: DIRECTIONS IN ROBUST STATISTICS AND DIAGNOSTICS, PART II,
eds. W. Stahel and S. Weisberg. The IMA Volumes in Mathematics
and its Applications, Vol. 34, Springer-Verlag, New York (1991).

ROBUST DISTANCES: SIMULATIONS AND CUTOFF VALUES

PETER J. ROUSSEEUW* AND BERT C. VAN ZOMEREN**

Abstract. It is possible to detect outliers in multivariate point clouds by computing distances based on robust estimates of location and scale. It has been suggested to use the Minimum Volume Ellipsoid estimator, which can be computed using a resampling algorithm. In this paper the small sample behavior of the robust distances is studied by means of simulation. We obtain a correction factor yielding approximately correct coverage percentages for the corresponding ellipsoids. In addition, a projection-type algorithm is considered to overcome the computational complexity of the resampling algorithm. Advantages and disadvantages of the second algorithm are discussed.

1. Motivation. It can be very hard to detect outliers in multivariate point clouds. The classical Mahalanobis distances based on the sample mean and the sample covariance matrix often fail in the presence of groups of outliers. This is called the masking effect. In Rousseeuw and Leroy (1987) it was suggested to calculate robust distances based on estimators of location and scale with high breakdown point. By using the Minimum Volume Ellipsoid estimator (MVE) it is indeed possible to detect *multiple* outliers.

Unfortunately it is difficult to find an exact solution to the minimization problem corresponding to the MVE. Instead we approximate the MVE by a resampling algorithm, which will be described in section 2. The robust distances produced by this algorithm roughly have a χ^2 distribution. However, a correction factor is needed to prevent that in small samples too many good observations will be flagged as outliers. We performed a simulation study to investigate the finite sample behavior of the distances. The results, which will be presented in section 3, show that approximately correct coverage percentages can be achieved for clean data as well as for contaminated data.

The resampling algorithm for the MVE can be efficiently combined with a similar algorithm for regression with high breakdown point. This will be briefly discussed in section 4.

The computational complexity of the resampling algorithm is high. Yet, for not too large a sample size and not too many variables an analysis on a personal computer is still feasible. The third column of table 1 shows some computation times for analyses performed on a fast (20 MHz) PC with mathematical coprocessor. Although the speed of the hardware is rapidly improving, computing the MVE by means of the resampling algorithm will still be tedious in high dimensions. Implementing accelerations in the algorithm helps. Using parallel computers or vector computers helps even more, because the resampling algorithm can be vectorized very easily. In section 5 we will describe a different algorithm to compute robust distances. It is a variant of an idea of Stahel (1981) and Gasko and Donoho (1982) and related to projection pursuit. In this paper it will be called "projection algorithm". The projection algorithm is not affine equivariant, but it is much faster

*UIA, Vesaliuslaan 24, B-2650 Edegem, Belgium
**Delft University of Technology, Delft, The Netherlands

TABLE 1. APPROXIMATE COMPUTATION TIMES FOR THE MVE (IN SECONDS) USING THE RESAMPLING METHOD OR PROJECTION METHOD ON A FAST PC.

n	p	time (resampling)	time (projection)
50	3	32	1
50	5	78	2
50	6	99	2.5
100	10	259	6
100	20	—	7.5

than the resampling algorithm, as is shown in the fourth column of table 1. In section 6 we will report the results of a simulation study, similar to the one described in section 3, and present some cutoff values for these robust distances as well. Some conclusions will be formulated in section 7.

2. MVE and resampling. The minimum volume ellipsoid estimator was introduced in Rousseeuw (1983). It is defined as the pair (T, C) where $T(\mathrm{X})$ is a p-vector, and $\mathrm{C}(\mathrm{X})$ is a positive semi-definite p-by-p matrix, such that the determinant of C is minimized subject to

$$\#\{i;\ (\mathbf{x}_i - T)\mathrm{C}^{-1}(\mathbf{x}_i - T)^t \leq a^2\} \geq h. \tag{1}$$

Here n is the number of observations, p is the number of variables, $h = [(n+p+1)/2]$ and X is the data set. The constant a^2 will be discussed later. The MVE has a breakdown point of nearly 50%, which means that $T(\mathrm{X})$ will remain bounded and the eigenvalues of $\mathrm{C}(\mathrm{X})$ will stay away from zero and infinity when less than half of the data are replaced by arbitrary values (see, e.g. Lopuhaä and Rousseeuw (1989)). The robust distances RD_i are defined relative to the MVE:

$$RD_i = \sqrt{(\mathbf{x}_i - T(\mathrm{X}))\mathrm{C}(\mathrm{X})^{-1}(\mathbf{x}_i - T(\mathrm{X}))^t} \tag{2}$$

The exact solution to the minimization problem corresponding to the MVE is difficult to find. This is why we use the following resampling algorithm.

Repeat the following steps m times:

- Draw a random subsample $J = \{i_1, \ldots, i_{p+1}\}$ of size $p+1$.
- Compute the mean and covariance matrix of this sample:

$$T_J = \frac{1}{p+1}\sum_J \mathbf{x}_i$$

$$\mathrm{C}_J = \frac{1}{p}\sum_J (\mathbf{x}_i - T_J)^t(\mathbf{x}_i - T_J).$$

- For all observations compute squared distances using T_J and C_J: $D^2_{J_i} = (\mathbf{x}_i - T_J)\mathrm{C}_J^{-1}(\mathbf{x}_i - T_J)^t$.

- Find m_J^2, the h-th order statistic of the D_{Ji}^2,
  and let $V_J = m_J^{2p} \det(\mathbf{C}_J)$.

Keep the J for which V_J is minimal across all m replications. Finally, the MVE estimates for location and scale are given by

$$T(\mathbf{X}) = T_J$$
$$\mathbf{C}(\mathbf{X}) = c_{n,p}^2 (\chi_{p,0.50}^2)^{-1} m_J^2 \mathbf{C}_J$$

where $c_{n,p}^2$ is a correction factor for small samples.

Remark 1. The number m of subsamples to be drawn is determined by a probability argument. It should satisfy

$$1 - (1 - (1 - \epsilon)^{p+1})^m \geq p_0, \tag{3}$$

where ϵ is taken to be 0.5 and p_0 is near unity. Table 2 shows in the middle column some values of m using $p_0 = 0.95$ in (3). The right column shows the values of m in the actual implementation of the MVE, which was used to obtain table 1.

TABLE 2. NUMBER OF SUBSAMPLES THEORETICALLY NEEDED (m_1) AND ACTUALLY USED (m_2) FOR DATASETS IN DIFFERENT DIMENSIONS (p).

p	m_1	m_2
2	23	1500
3	47	2000
4	95	2500
5	191	3000
6	382	3000

Remark 2. An acceleration of the resampling algorithm is possible. Let V_k be the value of the objective function after the k-th subsample is drawn. In order to decrease the objective function further in the $(k+1)$-th replication we need

$$m_{k+1}^2 < \left(\frac{V_k}{\det(\mathbf{C}_{k+1})}\right)^{1/p} = M.$$

So if in step 3 of the algorithm at least $n - h + 1$ of the distances exceed M, the objective function cannot improve for the $(k+1)$-th subsample, and the remaining distances need not be computed. Because an analysis of the algorithm using a profiler showed that a very large part of the total computation time is spent calculating such distances, this speeds up the algorithm considerably (by about 25%).

3. Cutoff values for the resampling method. When do we consider a robust distance RD_i to be suspiciously large? In order to flag outliers we need a rough cutoff value. The classical Mahalanobis distances of a data set in p dimensions, based on the sample mean and the sample covariance matrix, approximately have a $\chi^2(p)$-distribution. Therefore it was originally suggested to compare the robust distances

based on the MVE with a quantile of the χ^2 distribution with p degrees of freedom. For instance, by choosing the 97.5% quantile, one would expect that only about 2.5% of a clean dataset drawn from a multivariate normal distribution would fall outside the 97.5% tolerance ellipsoid.

Some simulation results are presented in table 3, which gives *median* coverage percentages for samples of different size and dimension. The third column indicates that too many observations are outside the naive tolerance ellipse, something which was also noticed by the authors when using the MVE with real data.

TABLE 3. MEDIAN COVERAGE PERCENTAGES FOR THE RESAMPLING ALGORITHM WITHOUT AND WITH THE SMALL SAMPLE CORRECTION FACTOR (50 REPLICATIONS).

n	p	without correction	with correction
20	2	70.0	95.0
50	2	88.0	96.0
100	2	92.5	97.0
20	3	72.5	95.0
50	3	85.0	96.0
100	3	91.5	96.0
20	4	70.0	95.0
50	4	84.0	96.0
100	4	91.5	96.0

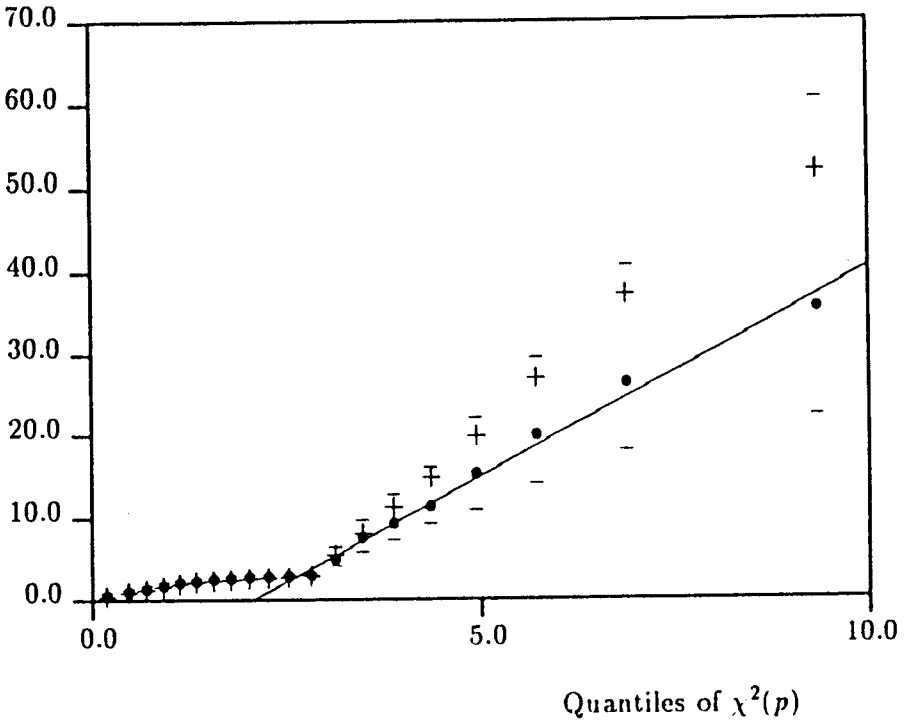
A further illustration of the distribution of the robust distances is given by the Q-Q plots in figures 1 and 2. The first shows the results of 50 replications of a data set of 20 observations in 3 dimensions. Data were generated from a multivariate normal distribution with zero expectation and unit covariance matrix, and contained no outliers. The robust squared distances in this plot are from the resampling algorithm *without* correction factor. For each replication the distances were sorted, and means (+), medians (•) and the interquartile range (-) of the simulated empirical quantiles were plotted against the corresponding quantiles of the $\chi^2(p)$ distribution.

A remarkable feature of figure 1 is the sudden change in pattern which occurs at the h-th distance. One could argue that this is caused by the nature of the estimator, which essentially is an ellipsoid forced to cover at least h observations out of n. From the interquartile ranges in figure 1 it is also clear that the variability of the first h distances is very small, whereas it increases considerably for the larger distances.

Furthermore it appears that the robust distances have a distribution with a much longer tail than $\chi^2(p)$, which explains the bad coverage percentages in the third column of table 3.

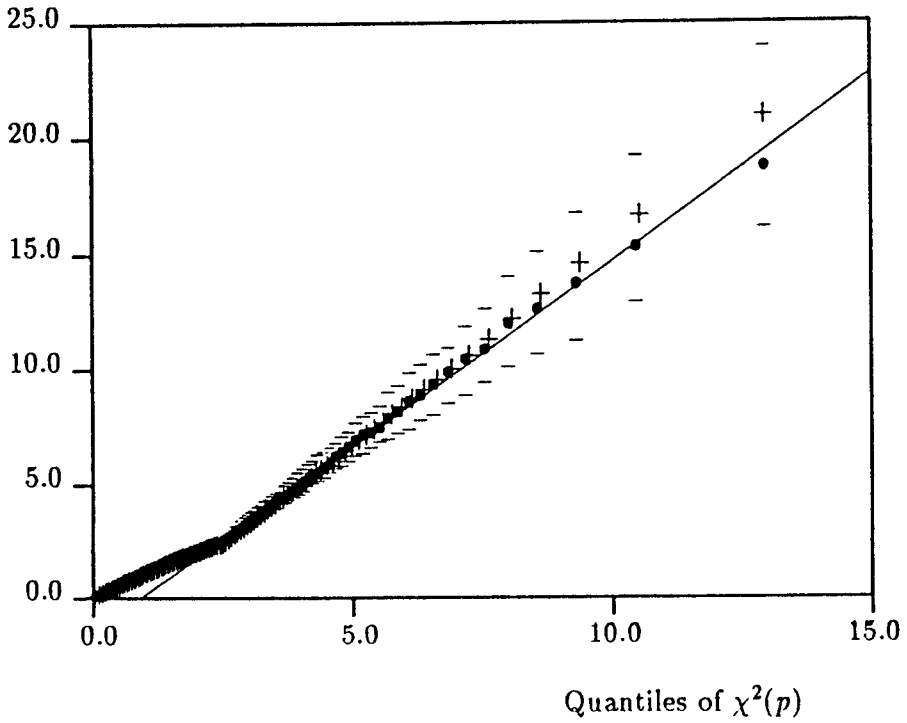
Figure 2 shows a Q-Q plot based on 100 observations in 3 dimensions. Here the change in pattern is much less pronounced, and the tail behaves much more like $\chi^2(p)$, which suggests that for larger data sets the χ^2 approximation will produce

FIGURE 1: Q-Q PLOT OF (SQUARED) DISTANCES OBTAINED BY RESAMPLING, FOR $n = 20$ AND $p = 3$, BASED ON 50 REPLICATIONS: MEANS (+), MEDIANS (•) AND QUANTILES (−) OF SIMULATED EMPIRICAL QUANTILES OF THE DISTANCES $(\mathbf{x}_i - T)\mathbf{C}^{-1}(\mathbf{x}_i - T)^t$



better results. Table 3 confirms this, but the low coverage percentages for $n = 100$ indicate that the convergence is rather slow.

FIGURE 2: Q-Q PLOT OF (SQUARED) DISTANCES AS IN FIGURE 1, FOR N=100 AND P=3



For many combinations of n and p we generated similar Q-Q plots and fitted the median quantiles of the right-hand parts by straight lines (as in figures 1 and 2), which were then used to compute actual cutoff values corresponding to the 97.5% quantile of $\chi^2(p)$. A plot of the square roots of these values against the factor $1/(n-p)$ showed a linear relationship. The coverage percentages in the fourth column of table 3 were obtained by inflating the tolerance ellipsoids by the factor

$$c^2_{n,p} = (1 + 15/(n-p))^2 \tag{4.}$$

They are good approximations to the ideal value of 97.5%.

We also checked the performance of this correction factor in the presence of different amounts of contamination, and found that the ability of the MVE to detect far outliers was not impaired.

Summarizing, we proceed as follows. First we apply the resampling algorithm, in which C(X) is computed taking into account the correction factor $c^2_{n,p}$ of (4). Then the resulting RD_i of (2) are compared to the cutoff value $\sqrt{\chi^2_{p,0.975}}$. Note however, that we do not just flag points as "outlying" according to this rule, but that we do look at the RD_i themselves (preferably in graphical displays) to distinguish far outliers from boundary cases.

4. An application of resampling. The resampling method for computing the MVE can also be applied in the context of robust regression with high breakdown point, such as Least Trimmed Squares (LTS) (Rousseeuw 1984). Suppose one wants to fit a regression model with p parameters ($p-1$ coefficients of predictor variables and a constant). The aim of the analysis would not only be to compute robust estimates of the regression coefficients and robust residuals, which is done by LTS, but also to detect leverage points, which can be done using the MVE. The resampling algorithm for LTS requires subsamples of size p, as does the resampling algorithm for the MVE applied to the $p-1$ predictor variables. This way the same subsample is used for LTS as well as for MVE and one obtains robust LTS-residuals and robust MVE-distances relatively cheaply. The plot of standardized robust residuals versus robust distances is a very powerful diagnostic tool (Rousseeuw and van Zomeren (1990)).

5. The Projection Method. As an alternative to the resampling algorithm for computing robust distances in high-dimensional data we consider a variant of a procedure introduced independently by Stahel (1981) and Gasko and Donoho (1982). For each point $\mathbf{x}_i$ we consider

$$u_i = \max_{\mathbf{v}} \frac{|\mathbf{x}_i\mathbf{v}^t - L(\mathbf{x}_1\mathbf{v}^t, \ldots, \mathbf{x}_n\mathbf{v}^t)|}{S(\mathbf{x}_1\mathbf{v}^t, \ldots, \mathbf{x}_n\mathbf{v}^t)}, \tag{5}$$

in which L and S are estimates of location and scale, applied on the projections $\mathbf{x}_i\mathbf{v}^t$ of the observations on selected directions $\mathbf{v}$. For L and S we use the midpoint and the length of the shortest half (Rousseeuw and Leroy (1988)). Again, the determination of the true maxima of (5) is not feasible, and we maximize over a finite set

of directions. A possible selection of directions $\mathbf{v}_l$ is generated by considering only $\mathbf{v}_l = \mathbf{x}_l - \mathbf{m}$, where $\mathbf{m}$ is the vector of coordinatewise medians of the observations. This leads to the following algorithm to compute the u_i .

First compute $\mathbf{m}$ and for all n directions $\mathbf{v}_l = \mathbf{x}_l - \mathbf{m}$ repeat:

- Project all observations on $\mathbf{v}_l$, producing projections $y_i = \mathbf{x}_i \mathbf{v}_l^t$, $i = 1, \ldots, n$.
- Calculate univariate location and scale estimates of the projections, giving $L = L(y_1, \ldots, y_n)$ and $S = S(y_1, \ldots, y_n)$.
- Standardize the projections: $z_i = \frac{|y_i - L|}{S}$, $i = 1, \ldots, n$.
- Calculate $u_i \leftarrow \max(u_i, z_i)$, $i = 1, \ldots, n$.

The final u_i are approximations of RD_i which can be plotted, or used for reweighting by comparing them to some quantile of an appropriate distribution.

This algorithm is no longer affine equivariant. In the first place the coordinatewise median itself is not affine equivariant. Furthermore, there exist affine transformations which can mask the outliers, for instance standardizing the data using the arithmetic mean and covariance matrix. (In subsequent work we will also consider Stahel's original suggestion to project on directions $\mathbf{v}$ orthogonal to hyperplanes through random subsets of p points, which gives an affine equivariant algorithm.)

On the other hand, the projection algorithm has the same high breakdown point. Also, it is permutation invariant, whereas resampling algorithms are not. The loss of affine equivariance is the greatest disadvantage, but sometimes one could give up this property when the observations have "natural units", and do not have to be transformed or scaled. Finally, the algorithm is much faster than resampling. This is demonstrated in the fourth column of table 1, which shows that computing robust distances using the projection method is feasible even for high-dimensional data.

6. Cutoff values for the projection method. When using the robust distances computed with the projection method for identification of outliers, there is also a need for cutoff values. A simulation study similar to the one discussed in section 3 was performed to get more insight into the distribution of these distances. Figure 3 shows the Q-Q plot of projection distances versus quantiles of the $\chi^2(p)$ distribution, obtained from 50 replications of a data set of 100 observations in 15 dimensions. The data were drawn from a $N(0, \mathrm{I})$ distribution and contained no outliers. Similar plots were constructed for many combinations of n and p. They all showed a nearly linear pattern. Again, straight lines were fitted through the right-hand part of the Q-Q plots, and then used to construct cutoff values corresponding to a coverage percentage of 97.5%. When these values were plotted against n or p no clear pattern emerged, so instead we calculated different cutoffs for several values of n and p. Table 4 gives the median coverage percentages from 100 replications of data sets of different size and dimension, in the case of no contamination (column 4) and 25% contamination (column 5). The regular data were generated from a $N(0, \mathrm{I})$ distribution, whereas outliers were created by shifting regular observations 10 units on the first variable. Table 4 also shows which cutoff values were used to

FIGURE 3: Q-Q PLOT OF (SQUARED) DISTANCES, OBTAINED BY PROJECTION, FOR N=100 AND P=15, AS IN FIGURE 1

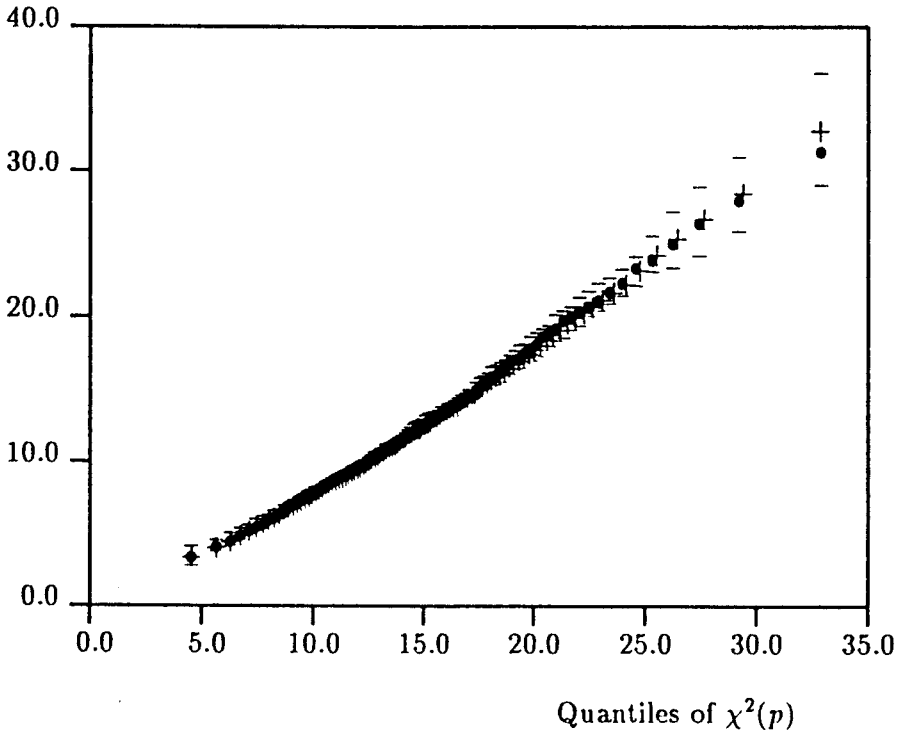


TABLE 4. MEDIAN COVERAGE PERCENTAGES FOR THE PROJECTION ALGORITHM FOR 100 REPLICATIONS OF CLEAN AS WELL AS CONTAMINATED DATA SETS.

		u_i^2	contamination	
n	p	cutoff	0%	25%
50	10	18.57	96	74
100	10	22.57	98	75
50	15	21.02	98	76
100	15	27.52	98	75
50	20	22.56	98	89
100	20	29.56	97	76

identify outliers. In general the results in the table are satisfactory, only in the case of 50 observations in 20 dimensions the coverage for the contaminated data is too high. This is in agreement with our rule of thumb that there should be *at least five* observations per dimension.

7. Conclusions. When faced with the difficult task of detecting outliers in a multivariate data set, one would like to compute robust distances. Therefore there is a need for a high-breakdown-point estimator of multivariate location and scale, on which robust distances can be based. The Minimum Volume Ellipsoid estimator is such an estimator. In data sets of small dimension, the resampling algorithm can be used to approximate the MVE. By using the small sample correction factor (4) proposed in section 3, the adequacy of the MVE (using the resampling method) for outlier detection has been greatly improved. Depending on the availability of a fast

computer, this approximation to the MVE can be computed for data in not too many dimensions.

In cases where affine equivariance is not essential, or in many dimensions, we propose to use the projection method which is much faster than the resampling algorithm. The resulting robust distances seem to be very useful to expose far outliers, although their distributional properties are still not fully understood. These distances can also be used to identify leverage points in regression with many predictor variables.

REFERENCES

GASKO, M. AND DONOHO, D., *Influential Observation in Data Analysis*, in *American Statistical Association Proceedings of the Business and Economic Statistics Section*, 1982, pp. 104–109.

LOPUHAÄ, H.P. AND ROUSSEEUW, P.J., *Breakdown Points of Affine Equivariant Estimators of Multivariate Location and Covariance Matrices*, Annals of Statistics (1989) (to appear).

ROUSSEEUW, P.J., *Multivariate Estimation with High Breakdown Point*, paper presented at Fourth Pannonian Symposium on Mathematical Statistics and Probability, Bad Tatzmannsdorf, Austria, September 4-9, 1983. Abstract in *IMS Bulletin*, 1983. Appeared in *Mathematical Statistics and Applications, Volume B*, eds. W.Grossmann, G.Pflug, I.Vincze, and W.Wertz, Dordrecht: Reidel Publishing Company, (1985).

ROUSSEEUW, P.J. AND LEROY, A., *Robust Regression and Outlier Detection*, New York: John Wiley, 1987.

ROUSSEEUW, P.J. AND LEROY, A., *A Robust Scale Estimator based on the Shortest Half*, Statistica Neerlandica, 42 (1988), pp. 103–116.

ROUSSEEUW, P.J. AND VAN ZOMEREN, B.C., *Unmasking multivariate outliers and leverage points*, Journal of the American Statistical Association, (to appear in the September issue) (1990).

STAHEL, W.A., *Robuste Schätzungen: Infinitesimale Optimalität und Schätzungen von Kovarianzmatrizen*, Ph.D.Thesis, ETH Zürich (1981).