

Detection of Multivariate Outliers by a Simulated Epidemic *

Beat HULLIGER and Cédric BEGUIN

*Swiss Federal Statistical Office
Espace de l'Europe 10
CH-2010 Neuchâtel*

e-mail: beat.hulliger@bfs.admin.ch / cedric.beguin@bfs.admin.ch

Abstract: A new method for the detection of multivariate outliers is proposed. It is based on the simulation of an epidemic in a point cloud in p -dimensional space. The epidemic starts on a well chosen point and then spreads through the point cloud with probabilities that decrease with the distances between points. Outliers typically have a high probability of being infected late. Outlying infection times are therefore used to detect outliers. The method is flexible and compares well with methods based on robust Mahalanobis distances.

Keywords: Multivariate outlier, stochastic algorithm, non-elliptical data.

1 Introduction

We want to detect outliers in a population of n points in p -dimensional space. The idea is to start an epidemic from a well chosen point. The epidemic will spread through the population and eventually all points will be infected. In this process the outliers should either not be infected or be infected late due to their isolation. We use the infection time to judge on the outlyingness of a point. In other words the epidemic defines a random mapping from the population into the time axes which should give high values for outliers. The algorithm will be described in the Section 2. In Section 3 the method will be evaluated on both synthetic and real datasets. The results will be compared with those obtained by other methods using a robust Mahalanobis distance to detect outliers.

*This work is done for the EUREDIT project under the Information Society Technology Programme of Framework Programme 5 of the European Union. It is financed by the Swiss Federal Office of Education and Science.

Conclusions on the evaluation and remarks on some theoretical aspects as well as on some future developments will be made in the last Section.

2 The Epidemic algorithm

The transmission probability of the epidemic depends on the distance and decreases with it. The transmissions are independent. The time is discrete. An infected point can transmit the epidemic as long as the epidemic lasts. Denote the population with U . The points are described by the vector valued variable $x_i \in \mathbb{R}^p$, ($i = 1, \dots, n$). The distance of points i and j is the Euclidean distance: $d_{ij} = \|x_i - x_j\|_2$. To avoid unbalanced effects of the different variables, their variances shall be standardised before calculating the distances, e.g. by

$$\tilde{x}_{ik} = \text{med}(x_{ik}) + \frac{x_{ik} - \text{med}(x_{ik})}{\text{mad}(x_{ik} - \text{med}(x_{ik}))}.$$

Alternatively one may weight the contribution of each variable to the distance by the inverse of a robust measure of scale:

$$d_{ij} = \left(\sum_{k=1}^p q_k (x_{ik} - x_{jk})^2 \right)^{1/2},$$

where e.g. $q_k = (\text{mad}(x_{ik}))^{-2}$. In Section 3 the first solution is implemented.

The starting point of the epidemic shall be the "sample spatial median" (ssm), namely the sample point that has the characterizing minimal property of the usual spatial median :

$$ssm = x_i \text{ with } i \text{ such that } \sum_{j \in U} d_{ij} = \min_{k \in U} \left\{ \sum_{j \in U} d_{kj} \right\}.$$

Note that the sample spatial median is not necessarily close to the real spatial median. E.g. for a uniform distribution on a circle the spatial median will be near the center and the ssm will be on the circle. However the ssm will be in the bulk of the data. Moreover as all the distances d_{ij} will be needed anyway for the Epidemic algorithm, the computation of ssm is cheap.

Given a point i that is infected, the probability that a non-infected point j is infected by i at any time t is $P[j|i] = h(d_{ij}) = P[i|j]$, where the function h is monotone decreasing for growing d and $0 \leq h(d_{ij}) \leq 1$. We write $h_{ij} = h(d_{ij})$ for brevity. There are many possible choices for the transmission function h . Two examples are:

a) A simple linear transmission function

$$h(d) = \begin{cases} 1 - \beta d & \text{if } d \leq 1/\beta \\ 0 & \text{if not} \end{cases}$$

This function becomes exactly 0 at $d_{ij} = 1/\beta$ and thus no transmission is possible beyond this distance. The parameter β may be chosen in the following way. Calculate the maximum distance to a nearest neighbor $d_0 = \max_i(\min_j(d_{ij}))$. Then $\beta = (1 - 1/n) \min\{d_0, 2\sqrt{p}\}$. Thus β is chosen such that the transmission probability is $1/n$ at d_0 or at $2\sqrt{p}$ if d_0 is inflated by one or several single outliers.

- b) The logistic function: $h_{ij} = \frac{\exp(\alpha + \beta d_{ij})}{(1 + \exp(\alpha + \beta d_{ij}))}$ with $\alpha > 0$ and $\beta < 0$. The transmission probability is close to 1 for $d_{ij} = 0$ and $= 0.5$ at $d_{ij} = -\alpha/\beta$. The slope at this latter distance is $\beta/4$. We propose to choose the parameters α and β in such a way that the transmission probability is 0.5 at the median of the interpoint distances and $1/n$ at the maximal distance d_0 .

In the examples of the next Section, transmission function a) is used. The choice of the transmission function and its parameters is crucial for the detection capability of the algorithm and for its speed.

2.1 The steps of the Epidemic algorithm

1. Set the infection time of all points to zero: $t_j = 0, \forall j \in U$.
2. Set the time to one: $t = 1$. Choose the sample spatial median as a starting point. Set its infection time to one: $t_{sm} = 1$.
3. Increase the infection time by 1: $t = t + 1$.
4. Denote by I the subset of all the points infected before time t : $I = \{i : 0 < t_i \leq t - 1\}$. Calculate the total infection probability $P[j|I]$ for all non-infected points $j \notin I$:

$$P[j|I] = 1 - \prod_{i \in I} (1 - P[i|j]) = 1 - \prod_{i \in I} (1 - h_{ij}), \forall j \notin I.$$

5. Independent Bernoulli trials with success probability $P[j|I]$ decide on whether the points $j \notin I$ are infected or not at time t . If a point is infected, its infection time t_j is set to t : $t_j = t$. Update the set of infected points I .
6. If $|I| = n$ or $t - \max\{t_i : i \in I\} > l$ then stop. Otherwise go to step 3.

The algorithm stops if all points are infected or if no infection occurs during a period of length l . The non-infected points will keep infection time $t_j = 0$. The integer number l is chosen by the statistician. In the next Section it is set to 10. Alternatively the choice of l may be guided by an upper bound on the probability of no infection in l trials: $(1 - h(d_0))^l$. In the following we sometimes abbreviate Epidemic algorithm to EA.

2.2 Computational complexity

In the beginning we have to calculate the $(n(n-1))/2$ distances, each involving $p+1$ operations. For one epidemic there are at each time-point at most $|I| = k$ ($0 < k < n$) points that are infected and $n-k$ points that are not yet infected. Thus $(n-k)k \leq \lceil \frac{n}{2} \rceil^2$ possible transmissions have to be checked. We will stop the algorithm if there is no

infection after a fixed number of l trials (see Step 6). Thus at each stage of the EA we will have to perform at most $l \lceil \frac{n}{2} \rceil^2$ trials. Since there are at most n sizes for I there are at most $ln \lceil \frac{n}{2} \rceil^2$ trials in an epidemic. In other words the number of trials is polynomial in n , and, which is more important, is independent of the dimension p . The dimension of the space only affects the initial calculation of the distances.

3 Application to synthetic and real datasets

The algorithm has been implemented in S-Plus 2000, on a PC with a 600 MHz Intel Pentium Processor and 128 Mb RAM. The S -language is not efficient for the EA as any use of loops should be avoided in S . It was mainly motivated by the fact that the other methods (MCD, Stahel-Donoho, BACON) were available in that language. Therefore one should not consider the comparison of computing times as totally relevant. Moreover memory problems were quickly encountered when dealing with the $n \times n$ distance matrix: the 128 Mb RAM were not enough as soon as $n = 2000$ and memory swapping made the computing time explode.

3.1 Behavior of the Epidemic algorithm with normally distributed data

To analyse the behavior of the algorithm in the absence of outliers several datasets were simulated with a multivariate normal distribution in \mathbb{R}^p , with mean at the origin and covariance matrix equal to \mathbf{I}_p (identity matrix). The following table gives the total number of infected points at each infection time for 10 different datasets with n ranging from 100 to 2000 and p from 2 to 100.

Data sets	n	100	100	500	500	1000	1000	1000	2000	2000	2000
	p	2	10	10	20	10	20	50	20	50	100
Infection time (t)	1	1	1	1	1	1	1	1	1	1	1
	2	13	15	53	81	78	79	75	199	96	136
	3	52	61	369	435	715	665	516	1758	1027	1335
	4	78	89	477	489	948	943	900	1981	1815	1887
	5	89	95	490	495	980	965	950	1990	1909	1963
	6	95	97	494	497	989	976	970	1996	1938	1975
	7	97	97	494	498	992	987	980	1998	1952	1982
	8	99	97	496	499	992	991	985		1962	1984
	9		98	497		994	992	989		1972	1987
	10			497		994	992	990		1976	1987
Largest inf. time		8	9	11	8	15	14	25	7	47	34
Non-infected		1	2	2	1	3	4	2	2	3	2
Comp. time		0.7	0.8	3.4	3.4	9.2	10.4	15.0	388.5	776.1	252.3

This table shows that under normal distribution the median infection time is always 3 and that after $t = 7$ more than 95% of the population has been infected in all cases for any values of n and p (the worst case occurred when $n = 100$ where only 97% is infected at $t = 7$). We therefore use $t = 7$ as critical time under normal distribution. The number of non-infected points does not seem to depend on n or p , in all simulations it has never exceeded 5. In contrast the length of the epidemic does vary very much, even if half of the population has been infected after $t = 3$ in all cases! It seems that for fixed n the largest infection time increases with p . The computing time for $n = 2000$ is not too relevant because a large part of it is due to memory swapping.

3.2 Competing methods

The results obtained by the EA will be compared with three other methods, all using a robust estimation of location and scatter to compute a robust Mahalanobis distances. Thus these three methods need the assumption that the "good part" of the data is uni-modal and has some elliptical shape.

MCD The Minimum Covariance Determinant: see [13] and [14].

MSD The Modified Stahel-Donoho method: see [16], [7], [10] and [8].

BACON The "Blocked Adaptative Computationally-Efficient Outlier Nominators": see [5].

The results obtained by these methods will be illustrated by Q-Q plots of transforms of Mahalanobis distances (MD_i) using the following approximation for normal data :

$$D_i = F(0.5, p, n - p) \frac{MD_i}{\text{med}(MD_i)} \approx f_i = F\left(\frac{i}{n+1}, p, n - p\right)$$

where $F(\alpha, k, l)$ is the α -quantile of the F distribution with k and l degrees of freedom.

3.2.1 The Bushfire data

The first real dataset has 38 observations in dimension 5.

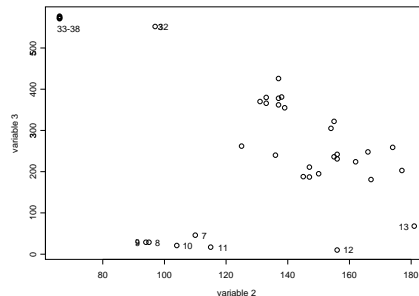


Figure 1: Bushfire dataset

It was used by Campbell in 1989 [6] to locate bushfire scars. It contains satellite measurements on five different frequency bands corresponding to each of 38 pixels. It has the advantage of having been well studied [10] and of allowing a two dimensional plot (in variable 2 and 3) that reveals almost all the outliers (see Figure 1). The data contains an outlying cluster of observations 32 to 38 and a few other outlying values 32 and 7 to 11, eventually also 12 and 13. A classical multivariate analysis using the sample mean and covariance estimate would not detect anything. Figure 2 shows that the results obtained from the three comparative methods are quite similar (MSD used 100 different projections). The table below gives the observations with the largest MD_i in decreasing order for the three methods. All of them detect the above mentioned outliers.

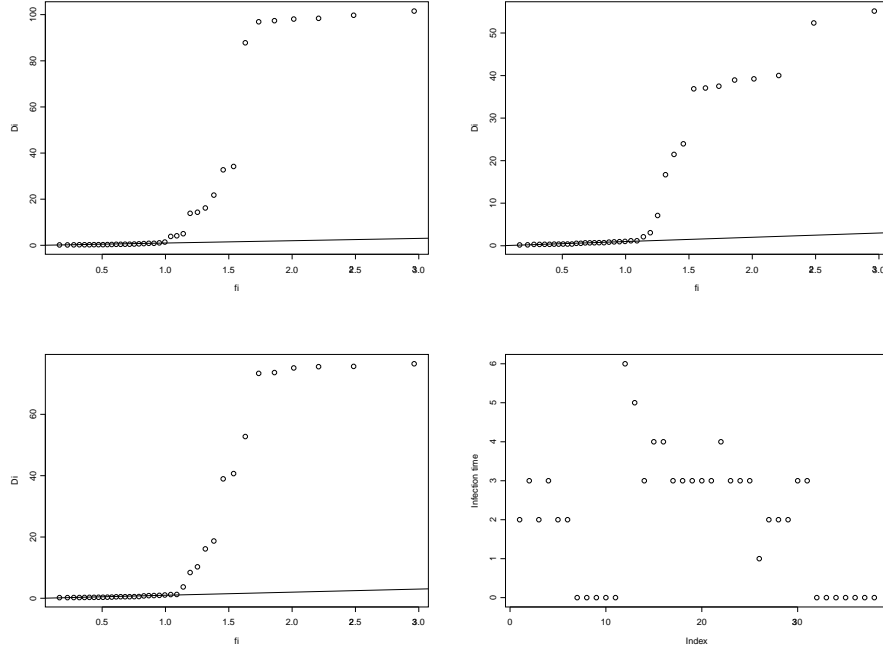


Figure 2: Q-Q plots of $M D_i$ for MCD, MSD and BACON and Epidemic history

MCD	33	35	34	38	37	36	32	9	8	31	10	11	7
MSD	9	8	38	37	35	33	36	34	32	10	11	7	
BACON	38	35	37	33	34	36	32	9	8	10	11	7	

MSD considers 8 and 9 as more outlying than the 32 – 38 group and MCD detects also 31 as an outlier. The EA applied to the Bushfire data did not infect any points after time $t = 6$ (see Figure 2). Only non-infected observations will therefore be declared as outliers, namely points 7 to 11 and 32 to 38. Clearly in that case all methods are equivalent. Finally, due to the small size of the dataset all computing times are moderate : MCD 0.23s, MSD 1.6s, BACON 0.08s and Epidemic 0.68s.

3.2.2 The Ionosphere data

The second dataset was taken from the UCI Machine Learning Database Repository [2] and was suggested to us by Ricardo Maronna [11]. This dataset was part of a study of the Ionosphere carried out by the Space Physics Group of the Applied Physics Laboratory of the Johns Hopkins University [15]. Radar data were collected by a system in Goose Bay, Labrador. The targets were free electrons in the ionosphere. "Good" radar returns were those showing evidence of some type of structure in the ionosphere.

These good radar measurements form the dataset which is studied here: there are 225 observations in 32 dimensions (two variables with no variance were eliminated).

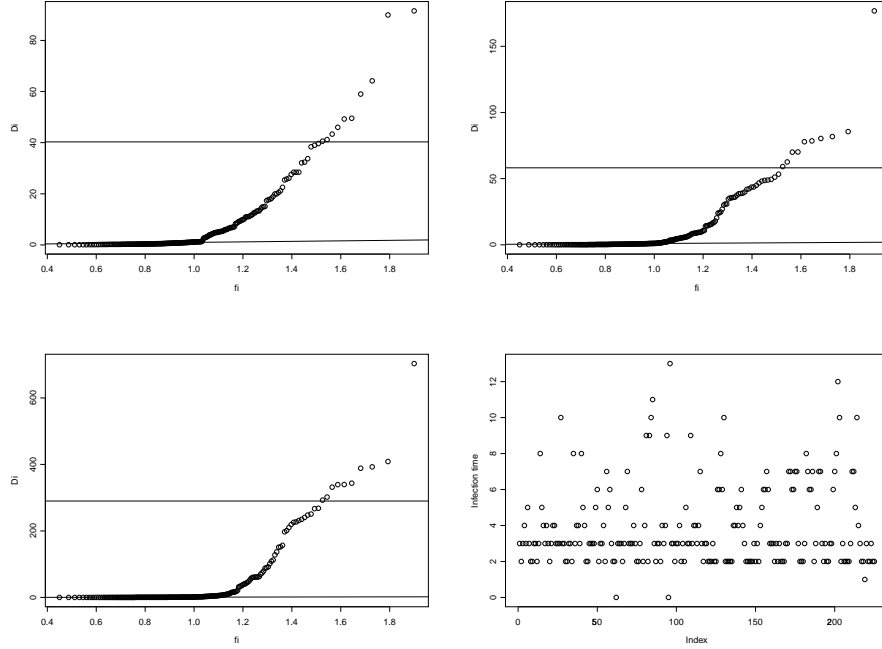


Figure 3: Q-Q plots of MD_i for MCD, MSD and BACON and Epidemic history

The EA was run first and gave the following results (Figure 3): Two observations were not infected (62 and 95) and 18 others were infected after time $t = 7$. To compare these results with the other methods, the Q-Q plots are given in Figure 3 with a horizontal line just below the 10 largest MD_i for each method. MSD used 1000 different directions. These plots show that about 60% (= 135 observations) of the data behave like normally distributed. Note that after time $t = 3$ the EA had infected 140 observations! Clearly something is happening for the remaining data. Choosing a value where to cut for outlyingness would require more knowledge of the data. To compare all the results we give two tables with the number of common points in the "central part" of each method and in the "extreme part". The central part of a method consists of the 140 observations which are least outlying (lowest MD_i or infection time ≤ 3) while the extreme part consists of the 10 most outlying observations (highest MD_i or infection time > 9 or 0). EA, MSD and BACON do agree well on the central part (actually the three of them shared 112 common points in their respective central parts), while MCD seems to react somewhat different (86 points are shared by the four methods). For the outlying part there is no consensus, but if we look closer at the Q-Q plots, MCD has two clear outliers (96 and 95), MSD has one (27) and BACON also has one (27) and these outliers are all detected by the EA with very high infection time: $t = 10$ for 27,

$t = 13$ for 96 and not infected for 95!

Central part					Extreme part				
	MCD	MSD	BAC	EA		MCD	MSD	BAC	EA
MCD	140	111	98	87	MCD	10	3	2	5
MSD	111	140	125	113	MSD	3	10	7	5
BAC	98	125	140	125	BAC	2	7	10	2
EA	87	113	125	140	EA	5	5	2	10

The computing times diverge. EA took 2.3s, MCD 21.8s, MSD 73.9s and BACON 0.41s. Note that our implementation of MSD is not optimized. When the dimension of the data grows, the computing time of MCD and MSD grows too. This was expected as well as the fact that the computing time of EA is not much affected by the growth of dimension (remember that the dimension appears in the algorithm only in the distance computation). BACON remains by far the fastest (see [3]).

3.3 Behavior of the Epidemic algorithm with concentrated contamination

In [12] Rocke and Woodruff made two observations: 1) it is very hard to detect outliers in data with a contamination fraction of 35% or higher; 2) compactly spaced outliers are harder to find. To test the quality of Epidemic algorithm we combined here the two difficulties: we generated a dataset with 500 observations in \mathbb{R}^{10} with observations 1 to 300 that followed a multinormal distribution centered at the origin with a covariance matrix set to $10 * \mathbf{1}_{10}$ and two contaminations formed by two other clouds centered at two randomly chosen points in \mathbb{R}^{10} , one at distance 70 (observations 301 to 400) and one at distance 100 (observations 401 to 500), both with multinormal distribution with covariance matrix of $\mathbf{1}_{10}$.

Here, as we know the indices of the outliers, the results of all methods are just plotted with infection time or MD versus index (Figure 4). All possible cases do occur here:

- EA** The 300 outliers have not been infected and they are therefore perfectly detected. Three other points are infected after time 7 and are therefore suspicious. The algorithm did not make any difference between the two clouds.
- MCD** The 300 outliers have the smallest Mahalanobis distances and were not detected. The Q-Q plot looks very strange but can only tell that there is a problem.
- MSD** The more distant outlier cloud was perfectly detected with high Mahalanobis distances, but the closer outlier cloud was not detected.
- BACON** The detection is perfect. It even distinguishes between the two outlier clouds. This is no surprise since BACON is perfect in such cases (see [3]).

For that particular example EA was slightly slower than the other algorithm: MCD took 5.28s, MSD 6.0s (with 200 directions, no improvement with 500), BACON 0.28s and EA 10.1s.

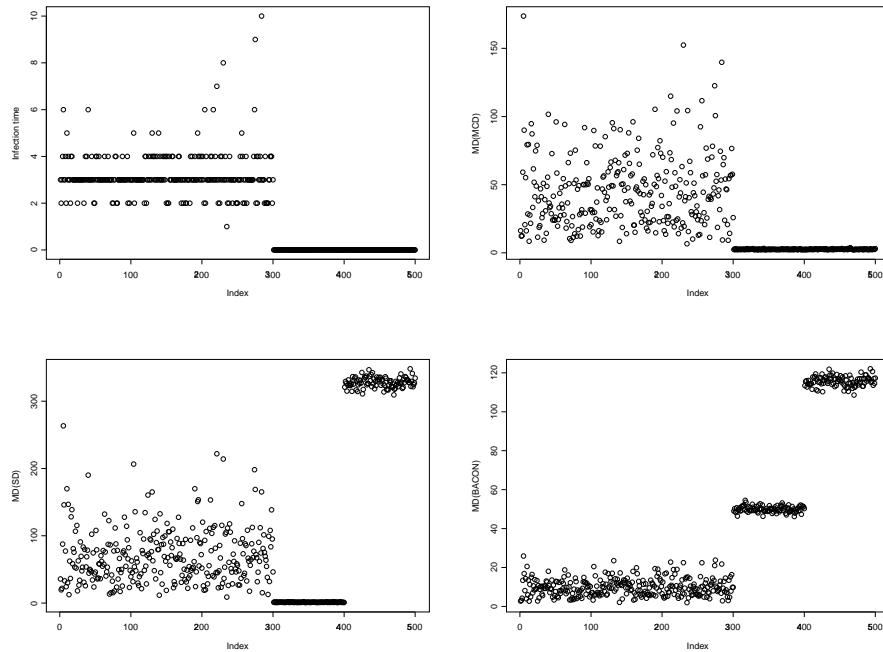


Figure 4: Time resp. MD_i for EA, MCD, MSD and BACON

4 Conclusions and Outlook

The Epidemic algorithm is computationally feasible. It is somewhat slower than the most efficient algorithms. However its computing time does not grow exponentially with the number of dimensions. It does not need any assumption on the data except that the good data is not divided into well separated clusters. No transformation is necessary to apply EA. It is based on the intuitive notion of an outlier as an isolated point or group of points. The starting point of a sample spatial median seems to be very fruitful. The Bushfire example shows that EA has good detection capabilities for outliers in moderately high dimensions. The Ionospheric data shows that it also works in higher dimensions well. Though the different methods disagree on the outliers (except the most outlying ones) the "good" part of the data is very similar for EA, MSD and BACON, while it is quite different for MCD. The last, synthetic example shows that EA seems to be unaffected by highly concentrated outliers. Furthermore, examples which are not reported here show that EA has good detection capabilities in situations where the bulk of the data is far from elliptically shaped. The situation where EA may give comparatively bad results is when the bulk of the data follow a nice model like a regression model which can be detected and used by other methods.

However when the bulk of the data is multivariate normal we did not find a considerable disadvantage of the EA. The discussion of the theory for the EA has to be taken further. The EA has connections to clustering algorithms and to nearest neighbor methods. However, by exploiting the dynamics of the epidemic, it takes into account local and global properties at the same time. The adaption of the Epidemic algorithm to missing values is straightforward. Also categorical variables and sampling weights can be taken into account. However, these extensions have to be investigated further along with a thorough discussion of the choice of the transmission function and its parameters.

References

- [1] A.C. Atkinson. Stalactite plots and robust estimation for the detection of multivariate outliers. In *Data Analysis and Robustness*. Morgenthaler, S., Ronchetti, E. and Stahel, W. (Ed.), Birkhäuser, 1993.
- [2] S. D. Bay. The UCI KDD archive [<http://kdd.ics.uci.edu>], 1999.
- [3] C. Béguin. Outlier detection in multivariate data. Master's thesis, Université de Neuchâtel, 2001. Preprint.
- [4] C. Béguin and B. Hulliger. Develop and evaluate new methods for statistical outlier detection and outlier robust multivariate imputation. Workplan for EUREDIT workpackage x.2, EUREDIT, November 2000.
- [5] N. Billor, A. S. Hadi, and P. F. Velleman. BACON: Blocked Adaptative Computationally-efficient Outlier Nominators. To be published in CS DA, 2000.
- [6] N.A. Campbell. Bushfire mapping using noaa avhrr data. Technical report, CSIRO, 1989.
- [7] D.L. Donoho. *Breakdown Properties of Multivariate Location Estimators*. Ph.d. qualifying paper, Department of Statistics, Harvard University, 1982.
- [8] S. Franklin, S. Thomas, and M. Brodeur. Robust multivariate outlier detection using Mahalanobis' distance and a modified Stahel-Donoho estimator. Technical report, Statistics Canada, 2000.
- [9] A. S. Hadi. Identifying multiple outliers in multivariate data. *J. R. Statist. Soc. B*, 54(3):761–771, 1992.
- [10] R. A. Maronna and V. J. Yohai. The behaviour of the Stahel-Donoho robust multivariate estimator. *Journal of the American Statistical Association*, 90(429):330–341, 1995.
- [11] R.A. Maronna and R.H. Zamar. Robust multivariate estimates for high dimensional data sets. Preprint, 2001.
- [12] D.M. Rocke and D.L. Woodruff. Identification of outlier in multivariate data. *Journal of the American Statistical Association*, 91(435):1047–1061, 1996.
- [13] P. J. Rousseeuw and A. M. Leroy. *Robust Regression and Outlier Detection*. John Wiley & Sons, 1987.
- [14] P.J. Rousseeuw and K. van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41:212–223, 1999.
- [15] V. G. Sigillito, S. P. Wing, L. V. Hutton, and K. B. Baker. Classification of radar returns from the ionosphere using neural networks. *Johns Hopkins APL Technical Digest*, 10:262–266, 1989.
- [16] W.A. Stahel. *Robuste Schätzungen: infinitesimale optimalität und Schätzungen von Kovarianzmatrizen*. Ph.d. thesis, Swiss Federal Institute of Technology, 1981.