ELSEVIER

Computational Statistics & Data Analysis 34 (2000) 279–298

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

BACON: blocked adaptive computationally efficient outlier nominators

Nedret Billor[a], Ali S. Hadi[b,*], Paul F. Velleman[b]

[a] *Department of Mathematics, Cukorova University, Turkey*
[b] *Department of Statistical Sciences, Cornell University, USA*

Received 1 March 1999; received in revised form 1 November 1999

Abstract

Although it is customary to assume that data are homogeneous, in fact, they often contain outliers or subgroups. Methods for identifying multiple outliers and subgroups must deal with the challenge of establishing a metric that is not itself contaminated by inhomogeneities by which to measure how extraordinary a data point is. For samples of a sufficient size to support sophisticated methods, the computation cost often makes outlier detection unattractive. All multiple outlier detection methods have suffered in the past from a computational cost that escalated rapidly with the sample size. We propose a new general approach, based on the methods of Hadi (1992a,1994) and Hadi and Simonoff (1993) that can be computed quickly — often requiring less than five evaluations of the model being fit to the data, regardless of the sample size. Two cases of this approach are presented in this paper (algorithms for the detection of outliers in multivariate and regression data). The algorithms, however, can be applied more broadly than to these two cases. We show that the proposed methods match the performance of more computationally expensive methods on standard test problems and demonstrate their superior performance on large simulated challenges.

Keywords: Data mining; Mahalanobis distance; Multivariate outliers; Outlier detection; Prediction error; Regression outliers; Residuals; Robust distance; Robust statistics

> Whoever knows the ways of Nature will more easily notice her deviations; and, on the other hand, whoever knows her deviations will more accurately describe her ways. *Francis Bacon (1620)*, *Novum Organum II* 29.

* Corresponding author. Tel.: +1-607-255-2748; fax: +1-607-255-8484.
E-mail address: ali-hadi@cornell.edu (A.S. Hadi).

0167-9473/00/$ - see front matter
PII: S 0167-9473(99)00101-2

1. Introduction

Data often contain outliers. Most statistics methods assume homogeneous data in which all data points satisfy the same model. However, as the aphorism above illustrates, scientists and philosophers have recognized for at least 380 years that real data are not homogeneous and that the identification of outliers is an important step in the progress of scientific understanding.

Robust methods relax the homogeneity assumption, but they have not been widely adopted, partly because they hide the identification of outliers within the black box of the estimation method, but mainly because they are often computationally infeasible for moderate to large size data. Several books have been devoted either entirely or in large part to robust methods; see, for example, Huber (1981), Hampel et al. (1986), Rousseeuw and Leroy (1987), and Staudte and Sheather (1990).

Outlier detection methods provide the analyst with a set of proposed outliers. These can then be corrected (if identifiable errors are the cause) or separated from the body of the data for separate analysis. The remaining data then more nearly satisfy homogeneity assumptions and can be safely analyzed with standard methods. There is a large literature on outlier detection; see, for example, the books by Hawkins (1980), Belsley et al. (1980), Cook and Weisberg (1982), Atkinson (1985), Chatterjee and Hadi (1988), and Barnett and Lewis (1994), and the articles by Gray and Ling (1984), Gray (1986), Kianifard and Swallow (1989), Rousseeuw and van Zomeren (1990), Paul and Fung (1991), Simonoff (1991), Hadi (1992b), Hadi and Simonoff (1993,1994), Atkinson (1994), Woodruff and Rocke (1994), Rocke and Woodruff (1996), Barrett and Gray (1997), and Mayo and Gray (1997).

A good outlier detection method defines a robust method that works simply by omitting identified outliers and computing a standard nonrobust measure on the remaining points. Conversely, each robust method defines an outlier detection method by looking at the deviation from the robust fit (robust residuals or robust distances). Often outlier detection and robust estimation are discussed together, as we do here.

Although the detection of a single outlier is now relatively standard, the more realistic situation in which there may be multiple outliers poses greater challenges. Indeed, a number of leading researchers have opined that outlier detection is inherently computationally expensive.

Outlier detection requires a metric with which to measure the "outlyingness" of a data point. Typically, the metric arises from some *model* for the data (for example, a center or a fitted equation) and some measure of *discrepancy* from that model. Multiple outliers threaten the possibility that the metric itself may be contaminated by an unidentified outlier. The *breakdown point* of an estimator is commonly defined as the smallest fraction of the data whose arbitrary modification can carry estimator beyond all bounds (Donoho and Huber, 1983). Contamination of the outlier metric breaks down an outlier detector and, of course, any robust estimator based on that outlier detector. Attempts in the literature to solve this problem are summarized in Section 2.

2. Optimality, breakdown, equivariance, and cost of outlier detection

Suppose that the data set at hand consists of n observations on p variables and contains $k < n/2$ outliers. In practice, the number k and the outliers themselves are usually unknown. One method for the detection of these outliers is the brute force search. This method checks all possible subsets of size $k = 1,\ldots,n/2$ and for each subset determines whether the subset is outlying relative to the remaining observations in the data. The number of all possible subsets, $\sum_{k=1}^{n/2} \binom{n}{k}$, is so huge that brute force is clearly not a computationally feasible approach for even modest amounts of data. Nevertheless, it is guaranteed to find subsets of the data that are compact and that exclude multiple outlying points.

Alternative outlier detection methods try to form a clean subset of the data that can safely be presumed to be free of outliers, and test the outlyingness of the remaining points relative to the clean subset. For example, for regression data, Rousseeuw (1984) and Rousseeuw and Leroy (1987) propose to minimize the median of the squared (or absolute) residuals, yielding the *least median of squares* (LMS) method. For multivariate data, Rousseeuw and van Zomeren (1990) propose finding the subset of $h = [(n + p + 1)/2]$ observations within a *minimum volume ellipsoid* (MVE). More recently, Rousseeuw and van Driessen (1999) propose finding the subsets of h observations with the *minimum covariance determinant* (MCD). The observations obtained by the MVE or MCD can then be used to define a metric for nominating outliers. Finding the MVE or MCD requires computing the volumes of $\binom{n}{h}$ ellipsoids and choosing the subset which gives the minimum volume or minimum determinant. Although $\binom{n}{h}$ is much smaller than $\sum_{k=1}^{n/2} \binom{n}{k}$, MVE and MCD are still computationally infeasible. For this reason several researchers have proposed algorithms to approximate MVE and LMS. Rousseeuw and Leroy (1987, p. 259), discussing MVE, propose drawing random elemental subsets of p different observations, where p is the dimension of the data. They suggest a minimum number of samples based on a probabilistic argument on the likelihood of drawing a subset with truly minimum volume. They propose a similar sampling rule for estimating the LMS solution.

Cook and Hawkins (1990) question these rules, offering their own computations on the 20-point "Wood Gravity" example used by Rousseeuw and Leroy (1987). They found that 57,000 samples were required before the outliers generally recognized by researchers for these data were identified. Hawkins and Simonoff (1993) concur, recommending that all subsets of size p be examined if possible (which still would not guarantee the exact solution), or that at least 10,000 subsets be sampled if complete enumeration is not feasible. Ruppert and Simpson (1990) and Portnoy (1987) reach similar conclusions. The same criticism can be applied to the MCD method.

In one special case, Souvaine and Steele (1987) give an algorithm for LMS regression on a single dependent variable that is $O(n^2 \log n)$. Stromberg (1993) describes an algorithm for exact LMS for any number of predictors, which involves looking at all subsets of size $p + 1$. Hawkins et al. (1994) give details, including Fortran implementations of both serial and distributed versions of the algorithm. This is still

computationally intensive, of course, although the calculations are highly amenable to a parallel or distributed implementation.

Many estimators proposed for outlier detection and robust regression (such as the MVE, MCD, and LMS) satisfy an optimality condition. Optimality conditions have advantages for understanding the properties of the estimators, but they impose some costs as well. For example, Steele and Steiger (1986) show that the LMS objective function has on the order of n^2 local minima, making the true minimum hard to find by any systematic search.

An estimator, T, is *affine equivariant* if and only if

$$T(\boldsymbol{XA}+\boldsymbol{b}) = T(\boldsymbol{X})\boldsymbol{A}+\boldsymbol{b} \tag{1}$$

for any vector $\boldsymbol{b}$ and nonsingular matrix $\boldsymbol{A}$. For example, the brute force method is affine equivariant because both the multivariate mean and the covariance matrix are themselves affine equivariant. Clearly, affine equivariance is a desirable property both for the model and for discrepancy measures; one would not want a robust regression or the nomination of outliers to depend on the location, scale, or orientation of the data.

It is difficult to find affine equivariant methods with high-breakdown points. Rousseeuw and Leroy (1987, p. 253) report that Donoho (1982) studied many affine equivariant methods and found that they had breakdown points of at most $1/(p+1)$. Siegel's (1982) repeated median estimators have 50% breakdown point, but are not affine equivariant, and are quite computationally expensive. The LMS, MVE, and MCD have high-breakdown points and are affine equivariant methods. However, they are also too computationally expensive to be practical for large data sets.

Rousseeuw and Leroy (1987, p. 145) note that many affine equivariant high-breakdown regression methods are related to Projection Pursuit (see, e.g. Friedman and Stuetzle, 1981) because their breakdown properties are determined by behavior in certain special projections. In principle, a full solution is equivalent to checking all possible projections, so Rousseeuw and Leroy (1987) consign them to "the highly computer-intensive part of statistics". Ruppert and Simpson (1990, p. 646) agree, saying that "High-breakdown point regression appears to be unavoidably computer-intensive..."

If affine equivariant high-breakdown estimation is inherently computationally intensive, robust estimation and outlier detection must either approximate the solution or sacrifice affine equivariance.

Hadi (1992a,1994) and Hadi and Simonoff (1993) propose outlier detection methods that are location and scale invariant but are not affine equivariant. These methods identify a clean subset of the data that can be presumed to be free of outliers, and then perform a "forward search". They test the remaining points relative to the clean subset and allow the subset to grow one observation at a time as long as the new subset remains clean of outliers. They thus require ordering of the observations at each step of a process that can have $n-p$ steps. Although $n-p$ represents significant reduction in computing expense when compared to earlier methods, these methods too can be prohibitively expensive to compute for large data sets.

This article presents computationally efficient, high-performance multiple outlier detection methods (and corresponding robust methods) for any situation in which the data analyst can specify a model for the data capable of generating predicted values for all observations from a subset of observations. We then apply the general method to two special cases: (a) multivariate data and (b) regression data. Two versions of the methods are proposed. One version is nearly affine equivariant, has high breakdown points (upwards of 40%), and nevertheless is computationally efficient even for very large data sets. The other version is affine equivariant at the expense of a somewhat lower breakdown point (about 20%), but with the advantage of even lower computational cost.

Our methods require very few steps regardless of the sample size. The simulation results of Section 7 and the examples of Section 8 show outlier detection capabilities comparable to previously published high-breakdown outlier detection methods, but obtained in 4 or 5 iterations for data sets of sizes from 100 to 10,000 and dimensions from 5 to 20.

Section 3 presents the algorithm in a general form not given in previous work. Sections 4 and 5 give the details of the general algorithm as applied to multivariate and regression data and specify modifications to improve computing efficiency. Section 6 discusses the assumptions and the role of the data analyst in outlier detection. Section 7 reports on a simulation experiment that shows the computational savings and demonstrates that there is no loss of performance relative to previous methods. Section 8 gives illustrative examples. Section 9 discusses a potential application in very large data sets such as those encountered in data mining. Section 10 summarizes and offers concluding remarks and recommendations.

3. The general BACON algorithm

To obtain computationally efficient robust point estimators and multiple outlier detection methods, we propose to abandon optimality conditions and work with iterative estimates. Experiments and experience have shown that the results of the iteration are relatively insensitive to the starting point. Nevertheless, a robust starting point offers greater assurance of high breakdown and, in simulation trials, a breakdown point in excess of 40%. However, the robust starting point is not affine equivariant, and thus we cannot claim affine equivariance for iterations that start from it. We offer an affine equivariant start that is not, however, robust. Our simulations show that it has a lower breakdown point near 20%, but that when fewer than 20% of the points are outliers, it performs as well as the robust start. It is also computationally less expensive than the robust start.

The methods we offer are so computationally efficient that they can easily be applied to data sets of hundreds of thousands of points or more — something not imaginable for previous methods. Moreover, these methods match the performance of MVE, MCD, and LMS on all published test problems, and match the performance of Hadi's (1994) method in extensive simulation studies.

We base our proposal on the methods of Hadi (1992a,1994). He finds a small subset of the data that can safely be presumed free of outliers, then allows this clean subset to grow gradually until it includes all the data values not nominated as outliers. Because the basic subset of "clean" values increases by one at each step, Hadi's method requires at most n–p covariance matrix computations and inversions. Hadi and Simonoff (1993) give related methods for linear regression. These methods have been shown to perform well in many real-life and simulated data sets and have been implemented in statistics packages such as Data Desk (Velleman, 1998) and Stata (Gould and Hadi, 1993). Atkinson (1994) bases a graphical outlier detection method on Hadi's (1992) method, and Sullivan and Barrett (1997) improve further on his approach.

The principal improvements proposed here over these forward selection methods derive from allowing the subset of outlier-free points to grow rapidly, testing against a criterion and incorporating blocks of observations at each step. This saves computations both by reducing the number of covariance matrices computed and inverted, and by eliminating the need to sort potentially long arrays of discrepancies. We call this class of algorithms blocked adaptive computationally efficient outlier nominators or by its acronym, BACON, after the author of the aphorism (given at the beginning of the article) that the approach embodies.

The General BACON Algorithm consists of the following steps:

Algorithm 1: the general BACON algorithm

Step 1: Identify an initial *basic subset* of $m > p$ observations that can safely be assumed free of outliers, where p is the dimension of the data and m is an integer chosen by the data analyst.

Step 2: Fit an appropriate *model* to the basic subset, and from that model compute *discrepancies* for each of the observations.

Step 3: Find a larger basic subset consisting of observations known (by their discrepancies) to be homogeneous with the basic subset. Generally, these are the observations with smallest discrepancies. This new basic subset may omit some of the previous basic subset observations, but it must be as large as the previous basic subset.

Step 4: Iterate Steps 2 and 3 to refine the basic subset, using a *stopping rule* that determines when the basic subset can no longer grow safely.

Step 5: Nominate the observations excluded by the final basic subset as outliers.

The discrepancies can be displayed to check for gaps and to identify points that just barely were nominated as outliers or just barely failed to be so nominated.

Hadi (1992, 1994) and Hadi and Simonoff (1993,1997) give methods for identifying initial basic subsets for multivariate and regression situations, respectively. We use these methods here for Step 1 (after some modifications that make them even more computationally efficient), in part because extensive experience has shown that they work well.

The iterations in Steps 2 to 4 increase the basic subset size, but restrict membership to observations consistent with the current basic subset, and thus reliably not

outliers. The larger subset size yields more reliable estimates of the model and the corresponding discrepancies, refining the definition of the basic subset as it grows. The details of Steps 2–4 as applied to multivariate and regression data are given in Sections 4 and 5, respectively.

4. BACON algorithm for multivariate data

Given a matrix $\boldsymbol{X}$ of n rows (observations) and of p columns (variables), Step 1 of Algorithm 1 requires finding an initial basic subset of size $m > p$. This subset can either be specified by the data analyst or obtained by an algorithm. The analyst may have reasons to believe that a certain subset of observations is "clean". In this case, the number m and/or the observations themselves can be chosen by the analyst. There is some tension between the assurance that a small initial basic subset will be outlier-free and the need for a sufficiently large basic subset to make stable estimates of the model. If the desired basic subset size is $m = cp$, where c is a small integer chosen by the data analyst, then the estimation of parameters is based on at least c observations per parameter. The simulation results of Section 7 shows that $c = 4$ or 5 perform quite well.

The initial basic subset can also be found algorithmically in one of two ways as given in Algorithm 2 below.

Algorithm 2: initial basic subset in multivariate data

Input: An $n \times p$ matrix $\boldsymbol{X}$ of multivariate data and a number, m, of observations to include in the initial basic subset.

Output: An initial basic subset of at least m observations.

Version 1 ($V1$): *Initial subset selected based on Mahalanobis distances*

For $i = 1, \ldots, n$, compute the Mahalanobis distances

$$d_i(\bar{\boldsymbol{x}}, \boldsymbol{S}) = \sqrt{(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^{\mathrm{T}} \boldsymbol{S}^{-1} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})}, \quad i = 1, \ldots, n, \tag{2}$$

where $\bar{\boldsymbol{x}}$ and $\boldsymbol{S}$ are the mean and covariance matrix of the n observations. Identify the $m = cp$ observations with the smallest values of $d_i(\bar{\boldsymbol{x}}, \boldsymbol{S})$. Nominate these as a potential basic subset.

This start is not robust, but it is affine equivariant. The simulations of Section 7 show that the subsequent iterations tend to make up for the non-robustness of the start as long as the fraction of outliers is relatively small (20% in five dimensions, 10% in 20 dimensions). The advantages of this start are its affine equivariance (and thus the affine equivariance of the entire method) and its low computational cost.

Version 2 ($V2$): *Initial subset selected based on distances from the medians*

For $i = 1, \ldots, n$, compute $||\boldsymbol{x}_i - \boldsymbol{m}||$, where $\boldsymbol{m}$ is a vector containing the coordinatewise median, $\boldsymbol{x}_i$ is the *ith* row of $\boldsymbol{X}$, and $||\cdot||$ is the vector norm. Identify the m observations with the smallest values of $||\boldsymbol{x}_i - \boldsymbol{m}||$. Nominate these as a potential basic subset.

This start is robust, but it is not affine equivariant because the coordinatewise median is not affine equivariant. Because the subsequent iterations are robust, the

entire procedure is robust, with a high breakdown points (about 40%). Because the subsequent iterations are affine equivariant, the overall algorithm tends to be nearly affine equivariant. This start requires more computations than Version 1 because of the computational cost of finding medians in all dimensions.

In both versions, let $\bar{\boldsymbol{x}}_{\mathrm{b}}$ and $\boldsymbol{S}_{\mathrm{b}}$ be the mean and covariance matrix of the potential basic subset. If $\boldsymbol{S}_{\mathrm{b}}$ is not of full rank, then increase the basic subset by adding observations with smallest distances until it has full rank, and increase m by the number of observations added to make the subset full-rank.

Algorithm 3: the BACON algorithm for identifying outliers in multivariate data

Input: An $n \times p$ matrix $\boldsymbol{X}$ of multivariate data.

Output: A set of observations nominated as outliers and the discrepancies for all observations based on (3) relative to the final basic subset.

Step 1: Select an initial basic subset of size m using either V1 or V2 of Algorithm 2.

Step 2: Compute the *discrepancies*

$$d_i(\bar{\boldsymbol{x}}_{\mathrm{b}}, \boldsymbol{S}_{\mathrm{b}}) = \sqrt{(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{\mathrm{b}})^{\mathrm{T}} \boldsymbol{S}_{\mathrm{b}}^{-1} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{\mathrm{b}})}, \quad i = 1, \ldots, n, \tag{3}$$

where $\bar{\boldsymbol{x}}_{\mathrm{b}}$ and $\boldsymbol{S}_{\mathrm{b}}$ are the mean and covariance matrix of the observations in the basic subset.

Step 3: Set the new basic subset to all points with discrepancy less than $c_{npr}\chi_{p,\alpha/n}$, where $\chi^2_{p,\alpha}$ is the $1-\alpha$ percentile of the chi square distribution with p degrees of freedom, $c_{npr} = c_{np} + c_{hr}$ is a correction factor, $c_{hr} = \max\{0, (h-r)/(h+r)\}, h = [(n+p+1)/2], r$ is the size of the current basic subset, and

$$c_{np} = 1 + \frac{p+1}{n-p} + \frac{1}{n-h-p} = 1 + \frac{p+1}{n-p} + \frac{2}{n-1-3p}. \tag{4}$$

(When the size of the basic subset r is much smaller than h, the elements of the covariance matrix tend to be smaller than they should be. Thus, one can think of c_{hr} as a variance inflation factor that is used to inflate the variance when r is much smaller than h. Note also that when $r = h$, c_{npr} reduces to c_{np}.)

Step 4: The *stopping rule:* Iterate Steps 2 and 3 until the size of the basic subset no longer changes.

Step 5: Nominate the observations excluded by the final basic subset as outliers.

Hadi's (1994) method starts from a basic subset of size $p+1$, then increases the basic subset one observation at a time until it reaches $m = (n+p+1)/2$. The observation with the smallest distance $d_i(\bar{\boldsymbol{x}}_{\mathrm{b}}, \boldsymbol{S}_{\mathrm{b}})$ in (3) is then tested. If it is an outlier, the method stops declaring all observations in the nonbasic subset as outliers. Otherwise this observation is added to the current basic subset to form a new basic subset and the testing step is repeated. If there are k outliers in the data, Hadi's (1994) method requires about $n-k-p-1$ iterations. In each of these iterations an ordering of the distances is required. This can be computationally burdensome for large n.

Previously published forward selection methods such as this one increased the basic subset by a single observation at each step. But in early steps of the algorithm, most

of the non-outlying observations can easily be seen to be consistent with the basic subset. BACON treats these "easy" observations in a block, adding all observations that are clearly consistent at once, and concentrating effort on the cases near the boundaries.

Algorithm 3 offers substantial computional efficiencies by blocking the addition of observations to the basic subset, substantially reducing the number of iterations. Each of these iterations requires computing and inverting a covariance matrix, but the number of iterations does not grow with the sample size n. In addition, the BACON algorithm does not require the ordering of the n discrepancies, but rather just compares them to a standard value, resulting in further savings.

BACON uses the Mahalanobis distances (V1) or the distances from the coordinate-wise medians (V2) only to nominate a small subset of observations. These are then used to find a mean and covariance matrix which, in turn, nominate a new set of central observations. If the first subset of observations is not near enough to the center of the (non-outlying) data, the successive provisional basic subsets identified by the algorithm tend to drift toward the center. As the basic subset grows in size, its mean and covariance matrix become more stable.

5. BACON algorithm for regression data

Consider the standard linear model $\boldsymbol{y}=\boldsymbol{X}\boldsymbol{\beta}+\varepsilon$, where $\boldsymbol{y}$ is an n-vector of responses, $\boldsymbol{X}$ is an $n \times p$ matrix representing p explanatory variables with rank $p < n$, β is a p-vector of unknown parameters, and ε is an n-vector of random disturbances (errors) whose conditional mean and variance are given by $E(\varepsilon \mid X)=0$ and $\mathrm{Var}(\varepsilon \mid X)=\sigma^2\boldsymbol{I}_n$, where σ^2 is an unknown parameter and $\boldsymbol{I}_n$ is the identity matrix of order n.

The least-squares estimates of β and σ^2 are given by $\hat{\beta}=(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$ and the residual mean square, $\hat{\sigma}^2=SSE/(n-p)$, respectively, where $\boldsymbol{e}=(\boldsymbol{I}_n-\boldsymbol{P})\boldsymbol{y}$ is the vector of ordinary residuals, $SSE=\boldsymbol{e}^{\mathrm{T}}\boldsymbol{e}$ is the residual sum of squares, and $\boldsymbol{P}=\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}$. Let b be the set of indices of the observations in the basic subset and $\boldsymbol{y}_{\mathrm{b}}$ and $\boldsymbol{X}_{\mathrm{b}}$ be the subsets of observations indexed by b. Let $\hat{\beta}_{\mathrm{b}}$ be the estimated regression coefficients computed from fitting the model to the subset b and let SSE_{b} be the corresponding residual sum of squares and $\hat{\sigma}^2_{\mathrm{b}}$ be the corresponding residual mean square.

The method of Hadi and Simonoff (1993,1997) fits within the general BACON algorithm. For the initial basic subset Hadi and Simonoff (1993,1997) propose several alternatives. We suggest the following algorithm to find an initial subset of size $m = cp$:

Algorithm 4: initial basic subset in regression data

Input: An $n \times 1$ vector holding the response variable $\boldsymbol{y}$, an $n \times p$ matrix $\boldsymbol{X}$ of covariate data, and a number, m, of observations to include in the initial basic subset.

Output: An initial basic subset of at least m observations that is free of outliers.

Step 0: Apply Algorithm 3 to the $\boldsymbol{X}$ data (after removing the constant column, if any). Let $\boldsymbol{y}_m$ and $\boldsymbol{X}_m$ be the set of m observations with the smallest values of the

distances $d_i(\bar{\boldsymbol{x}}_b, \boldsymbol{S}_b)$ computed in the final iteration of Algorithm 3. If $\boldsymbol{X}_m$ is not of full rank, then increase the basic subset by adding observations with smallest values of $d_i(\bar{\boldsymbol{x}}_b, \boldsymbol{S}_b)$, until it has full rank. For $i = 1, \ldots, n$, compute

$$t_i(\boldsymbol{y}_m, \boldsymbol{X}_m) = \begin{cases} \dfrac{y_i - \boldsymbol{x}_i^T \hat{\beta}_m}{\hat{\sigma}_m \sqrt{1 - \boldsymbol{x}_i^T (\boldsymbol{X}_m^T \boldsymbol{X}_m)^{-1} \boldsymbol{x}_i}} & \text{if } \boldsymbol{x}_i \in \boldsymbol{X}_m, \\ \dfrac{y_i - \boldsymbol{x}_i^T \hat{\beta}_m}{\hat{\sigma}_m \sqrt{1 + \boldsymbol{x}_i^T (\boldsymbol{X}_m^T \boldsymbol{X}_m)^{-1} \boldsymbol{x}_i}} & \text{if } \boldsymbol{x}_i \notin \boldsymbol{X}_m, \end{cases} \tag{5}$$

where $\hat{\beta}_m$ and $\hat{\sigma}_m^2$ are the least-squares estimates of β and σ^2 based on the observations in the subset $\boldsymbol{y}_m$ and $\boldsymbol{X}_m$. Identify the $p+1$ observations with smallest $|t_i(\boldsymbol{y}_m, \boldsymbol{X}_m)|$, and declare them to be the initial basic subset $\boldsymbol{y}_b$ and $\boldsymbol{X}_b$.

Step 1. If this new $\boldsymbol{X}_b$ is not of full rank, increase the subset by as many observations as needed for it to become full rank, adding observations with smallest $|t_i(\boldsymbol{y}_b, \boldsymbol{X}_b)|$ first, and increase m by the number of observations added to make the subset full-rank. For $i = 1, \ldots, n$, compute

$$t_i(\boldsymbol{y}_b, \boldsymbol{X}_b) = \begin{cases} \dfrac{y_i - \boldsymbol{x}_i^T \hat{\beta}_b}{\hat{\sigma}_b \sqrt{1 - \boldsymbol{x}_i^T (\boldsymbol{X}_b^T \boldsymbol{X}_b)^{-1} \boldsymbol{x}_i}} & \text{if } \boldsymbol{x}_i \in \boldsymbol{X}_b, \\ \dfrac{y_i - \boldsymbol{x}_i^T \hat{\beta}_b}{\hat{\sigma}_b \sqrt{1 + \boldsymbol{x}_i^T (\boldsymbol{X}_b^T \boldsymbol{X}_b)^{-1} \boldsymbol{x}_i}} & \text{if } \boldsymbol{x}_i \notin \boldsymbol{X}_b, \end{cases} \tag{6}$$

where $\hat{\beta}_b$ and $\hat{\sigma}_b^2$ are the least-squares estimates of β and σ^2 based on the observations in the basic subset $\boldsymbol{y}_b$ and $\boldsymbol{X}_b$.

Step 2. Let r be the size of the current basic subset. Identify the $r+1$ observations with smallest $|t_i(\boldsymbol{y}_b, \boldsymbol{X}_b)|$, and declare them to be the new basic subset. (Note that these observations need not include all of the previous basic subset observations.)

Step 3. Repeat Steps 1 and 2 until the basic subset contains m observations.

Note that Step 1 of Algorithm 4 applies Algorithm 3 to the $\boldsymbol{X}$ matrix because multivariate outliers in the $\boldsymbol{X}$ space are points with high leverage. Thus the distance computed in the last iteration of Algorithm 3 can be used as a measure of the leverage of the ith observation. This measure is not as affected by the masking problem as the traditional measure of leverage, p_{ii}, the diagonal elements of the projection matrix $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$.

Note also that when $\boldsymbol{x}_i \in \boldsymbol{X}_b$, $t_i(\boldsymbol{y}_b, \boldsymbol{X}_b)$ in (6) is simply the scaled ordinary least squares residual obtained from the regression of $\boldsymbol{y}_b$ on $\boldsymbol{X}_b$; whereas when $\boldsymbol{x}_i \notin \boldsymbol{X}_b$, $t_i(\boldsymbol{y}_b, \boldsymbol{X}_b)$ is the scaled prediction error.

Hadi and Simonoff's (1993) method starts from the basic subset of Algorithm 4 with $m = (n + p + 1)/2$, uses the distances, $t_i(\boldsymbol{X}_b, \boldsymbol{y}_b)$ from (6) to define *discrepancies*. The method starts with an initial subset of size $p+1$ and increases the basic subset by one observation at each iteration until it reaches $(n + p + 1)/2$ observations. The method continues increasing the subset size beyond $(n + p + 1)/2$ but it stops when the $(r + 1)$ smallest absolute discrepancy exceeds $t_{(\alpha/2(r+1), r-p)}$, where $t_{(\alpha, r-p)}$ is the $1 - \alpha$ percentile of the t-distribution with $r - p$ degrees of freedom, where r is the size of the current basic subset at each step.

This method fits within the general Algorithm 1, but the repeated fitting of the regression model, and the computing and sorting of discrepancies at each step is computationally expensive. We propose to grow the basic subset in blocks, retaining

its ability to adapt to the data, but avoiding unneeded intermediate calculations. Thus the BACON Algorithm for robust regression is as follows.

Algorithm 5: the BACON algorithm for robust regression.

Input: An $n \times 1$ vector holding the response variable $\boldsymbol{y}$, and an $n \times p$ matrix $\boldsymbol{X}$ of covariate data.

Output: A regression model fit to non-outlying observations, the set of observations identified as outliers, and the distances found by (6) based on the final basic subset.

Step 1: Use Algorithm 4 to select an initial basic subset of size $m = cp$. The value of c is either selected by the data analyst or set by default to a small number (usually 4 or 5).

Step 2: Find *discrepancies*, $t_i(\boldsymbol{X}_b, \boldsymbol{y}_b)$ as in (6).

Step 3: The new basic subset consists of all points with distances less than $t_{(\alpha/2(r+1),r-p)}$, where r is the size of the current basic subset.

Step 4: The *stopping rule:* Iterate Steps 2 and 3 until the size of the basic subset no longer changes.

Step 5: Nominate the observations excluded by the final basic subset as outliers.

The computational efficiency of Algorithm 5 arises from the rapid expansion of the basic subset, resulting in far fewer regression calculations and evaluations of discrepancies, and from the fact that there is no longer any need to order the discrepancies, but rather only a need to check them against a constant, which requires only n operations.

As a byproduct of Algorithms 4 and 5, a useful diagnostic plot can be obtained by plotting the predicted errors $t_i(\boldsymbol{y}_b, \boldsymbol{X}_b)$ obtained at the final iteration of Algorithm 5 versus the distances $d_i(\bar{\boldsymbol{x}}_b, \boldsymbol{S}_b)$ obtained in Step 1 of Algorithm 4. For further discussion of this plot, see Rousseeuw and van Zomeren (1990) and Hadi and Simonoff (1997).

6. Assumptions and the role of the data analyst

All outlier nomination and robust methods must assume some simple structure for the non-outlying points — otherwise one cannot know what it means for an observation to be discrepant. The BACON algorithms assume that the model used to define the basic subsets is a good description of the non-outlying data. In the regression version, there must in fact be an underlying linear model for the non-outlying data. In the multivariate outlier nominator, the non-outlying data should be roughly elliptically symmetric. Although the algorithms will often do something reasonable even when these assumptions are violated, it is hard to say what the results mean.

It is possible to construct data where the BACON algorithms may not seem to work. For example, in the multivariate case, consider a hollow sphere with observations uniformly distributed over its surface, along with a small cluster of at least m but fewer than $n/2$ observations at its center. The data satisfy the elliptical symmetry assumption. The coordinatewise median will be in the central cluster. If that cluster is large enough to hold the initial basic subset, then BACON will nominate the

central cluster as the basic subset and the observations on the surface of the sphere as outliers. Of course, this is a reasonable response to these data because it helps to diagnose the structure of the data. Indeed, it is more useful than the MVE solution, which would not nominate any point as an outlier.

As an example of regression data, consider a data set where nearly all of the points are generated by the equation $y_i = \beta x_i^2 + \varepsilon_i$ with ε_i normally distributed and the x_i's symmetric around 0. One or a few outlying points relative to this model are planted at $x_i = 0$ and $y_i = \bar{y}$, where $\bar{y}$ is the mean of the y_i's. If we erroneously assume a linear model of the form $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ and apply the BACON Algorithms 4 and 5, the planted outliers would not be detected because a linear model is inappropriate for the non-outlying cases (the line that "fits" the parabola is forced to pass through $(0, \bar{y})$). However, using the correct linear model $y_i = \beta_0 + \beta_1 w_i + \varepsilon$, where $w_i = x_i^2$, allows the BACON algorithm to find the outliers easily.

Although such anomalous data sets help us to understand the limitations of these algorithms, they do not represent realistic challenges to the methods. They do, however, alert us to the need for intelligent participation in all data analyses.

We believe that data analysis is not a process to be performed by computers alone. No algorithm should make decisions about data without oversight by a responsible analyst. One of the advantages of the BACON algorithms is that they provide the data analyst with discrepancy measures. We strongly recommend that these be displayed and that other evidence of the structure of the data be examined to try to ascertain whether the assumptions made about the non-outlying data are reasonable. Our choice of the term "outlier nomination" reflects our conviction that an algorithm should not go beyond nominating observations as potential outliers. It is left to the data analyst to ultimately label outlying points as such.

7. Simulations

Hadi (1994) showed that his method matched or surpasses the performance of other published methods. Therefore, we performed simulation experiments to (a) compare the BACON method with the Hadi's (1994) method (H94) with regard to both performance and computational expense and (b) assess the performance of the BACON method for large data. The experiment considers outlier detection in multivariate data.

The H94 method is computationally expensive for large data sets. Therefore for comparison purposes we set $p = 5$, the contamination level $\phi = 0.05$, and selected three sample sizes, n (100, 500, and 1000). For assessing the performance of the BACON method the number of variables p was set to 5 (low dimension) or 20 (high dimension), and we took $n = 500$, 5000, and 10,000 observations.

For each value of n and p, we generated $N = 100$ data sets each of size $n \times p$. The first $k = \phi n$ of the observations were generated as outliers. The outliers were generated from a mean slippage model with a multivariate mean of moved 4 standard deviations from the mean of the remaining data. Thus, for each data set, $n - k$ observations are generated from the multivariate normal distribution, $\mathrm{N}(\mathbf{0}, \boldsymbol{I}_p)$, where $\boldsymbol{I}_p$ is the $p \times p$

Table 1
The simulation results for Hadi's (1994) method (H94) and the two versions of BACON (V1) and (V2), for the case $p = 5$ and $\phi = 0.05$ contamination level. The performance criteria A, B, and C are defined in (7)

		A			B			C		
n	m	H94	V1	V2	H94	V1	V2	H94	V1	V2
100	20	1.0060	1.0040	1.0060	1.00	0.998	1.00	89	3	4
	25	1.0060	1.0040	1.0060	1.00	0.998	1.00	89	3	4
500	20	1.0032	1.0032	1.0032	1.00	1.000	1.00	467	4	5
	25	1.0032	1.0032	1.0032	1.00	1.000	1.00	467	4	5
1000	20	1.0014	1.0014	1.0014	1.00	1.000	1.00	944	4	5
	25	1.0014	1.0014	1.0014	1.00	1.000	1.00	944	4	5

identity matrix. The k planted outliers are generated from $\mathrm{N}(4\mathbf{1}, \boldsymbol{I}_p)$, where $\mathbf{1}$ is a vector of p ones.

For the contamination parameter ϕ, we selected four values: 0.1, 0.2, 0.3, and 0.4, representing 10%, 20%, 30%, and 40% contamination. The simulation experiment thus consisted of 24 configurations: 2 dimensions (p), 3 sample sizes (n), and 4 contamination levels (ϕ). For each of these configurations, we tested two values for the size of the initial basic subset, $m = cp$, with c set to 4 or 5.

Finally, we set the significance level $\alpha = 0.05$ for our outlier tests and considered the following measures of performance:

$$A = \frac{\sum_{i=1}^{N} Out_i}{Nk},$$
$$B = \frac{\sum_{i=1}^{N} TrueOut_i}{Nk},$$
$$C = \text{Average number of iterations}, \tag{7}$$

where Out_i is the number of observations declared as outliers in the ith simulation run and $TrueOut_i$ is the number of observations correctly identified as outliers in the ith simulation run. Thus, perfect performance occurs to when $A = B = 1$, indicating that all of the planted outliers have been found, and no non-outlying observations have been declared to be outliers.

Table 1 gives the simulation results for $p = 5$ and $\phi = 0.05$. All three methods identified all of the planted outliers reliably in virtually all trials and occasionally identified a non-planted observation as an outlier, but in all cases, only very rarely. The substantial computational savings are clear. The number of iterations, C, for Hadi's (1994) method is expected to be $C = n - p - \phi n$. Thus for $n = 100$, 500, and 1000, the average number of iterations should be approximately 90, 470, and 945, respectively. As Table 1 shows, the observed average numbers of iterations C were 89, 467, and 944, as expected.

By contrast, the BACON method required only four or five iterations for all sample sizes. This is extraordinary because the number of iterations does not grow with the sample size.

Table 2
The simulation results for version 1 (V1) and version 2 (V2) of the BACON algorithm for the null case. The performance criteria A is the average percentage of observations declared as outliers and C is the average number of iterations

			A		C	
p	n	m	V1	V2	V1	V2
5	500	20	0.068	0.068	3	4
		25	0.068	0.068	3	4
	5000	20	0.054	0.054	4	5
		25	0.054	0.054	4	5
	10000	20	0.056	0.056	4	5
		25	0.056	0.056	4	5
20	500	80	0.010	0.014	2	3
		100	0.012	0.016	2	3
	5000	80	0.018	0.028	2	3
		100	0.006	0.022	2	3
	10,000	80	0.042	0.050	2	4
		100	0.024	0.046	2	3

To assess the performance of V1 and V2, we first investigate their performance under the null case, where there are no planted outliers in the data ($\phi = 0$). Table 2 holds the simulation results. In most of the tested cases, the average number of values identified (incorrectly) as outliers is close to the nominal 5% that should be expected. That value is, of course, under the user's control and can be set to a lower number to reduce the number of false positives, at the usual corresponding risk of increasing the number of marginal outliers not identified. In the 20 dimension trials, the rate of false positives is lower than 5% rising to the nominal value only for larger sample sizes.

In all trials the number of iterations remains small, showing no tendency to increase with increasing sample size. There seems to be some tendency toward fewer iterations for higher dimensional data. Overall, the computing efficiency of the algorithm is clear. Even for 10,000 cases in 20 dimensions, the method required on average only four evaluations and inversions of a covariance matrix.

Now we assess the performance of V1 and V2 in the presence of outliers. Tables 3 and 4 give the simulation results for $p = 5$ and 20, respectively. The simulation results show that V2 is very reliable even when the contamination is as large as 40%. V1 breaks down at about 30% contamination for both $p = 5$ and $p = 20$. Both V1 and V2 require at most six iterations, as compared to $n - p - \phi n = 4745$ expected iteration for the H94 method for $n = 5000$, $p = 5$, and $\phi = 0.05$, for example. This is a substantial saving in computation especially in view of the fact that the BACON algorithm does not require sorting the data at each iteration.

For a mean shift of 4 used in our simulation, some of the planted outliers can be ambiguously interspersed with "real" non-outlying data, posing a greater challenge to the methods. We have also repeated the simulation for a mean shift of 10, following

Table 3
The simulation results for versions 1 (V1) and 2 (V2) of the BACON algorithm for $p = 5$. The performance criteria A, B, and C are defined in (7) and ϕ is the contamination level

			A		B		C	
n	m	ϕ	V1	V2	V1	V2	V1	V2
500	20	0.1	1.0008	1.0010	0.9998	0.9998	4	5
		0.2	0.7888	1.0004	0.7885	0.9999	4	5
		0.3	0.5801	1.0002	0.5799	0.9999	4	5
		0.4	0.0302	1.0000	0.0301	0.9999	3	5
	25	0.1	1.0008	1.0010	0.9998	0.9998	4	5
		0.2	0.7880	1.0004	0.7877	0.9999	4	5
		0.3	0.5499	1.0002	0.5497	0.9999	4	5
		0.4	0.0302	1.0000	0.0301	0.9999	3	5
5000	20	0.1	1.0001	1.0001	0.9999	0.9999	5	6
		0.2	0.9900	1.0000	0.9899	0.9999	5	6
		0.3	0.7099	0.9999	0.7099	0.9999	5	5
		0.4	0.2900	0.9999	0.2900	0.9999	4	5
	25	0.1	1.0001	1.0001	0.9999	0.9999	5	6
		0.2	0.9400	1.0000	0.9399	0.9999	5	6
		0.3	0.7100	0.9999	0.7099	0.9999	5	5
		0.4	0.2100	0.9999	0.2100	0.9999	4	5
10,000	20	0.1	0.9998	0.9998	0.9997	0.9997	5	6
		0.2	0.9398	0.9998	0.9398	0.9998	5	6
		0.3	0.6698	0.9998	0.6698	0.9998	5	6
		0.4	0.3997	0.9998	0.3997	0.9998	4	5
	25	0.1	0.9998	0.9998	0.9997	0.9997	5	6
		0.2	0.9198	0.9998	0.9198	0.9998	5	6
		0.3	0.6299	0.9998	0.6299	0.9998	5	6
		0.4	0.3199	0.9998	0.3199	0.9998	4	5

Rousseeuw and van Driessen (1999). The results do not require a table; the BACON method has a perfect performance, that is $A = B = 1.0$ in all cases.

8. Examples

We illustrate the computational efficiency of the proposed methods using two data sets: The Wood Gravity data and the Philips data. Rousseeuw and Leroy (1987) use the wood gravity data (originally given by Draper and Smith, 1966) to illustrate the performance of LMS. The data consist of 20 observations on six variables. Observations 4, 6, and 19 are known to be outliers. Cook and Hawkins (1990) apply MVE with sampling to these data and they report that they needed over 57,000 samples to find the correct solution. The BACON method finds the outliers in these data in one step beyond the initial basic subset of size $2p = 12$. However, the wood gravity data are not a particularly good test problem for outlier detection because of its small sample size.

Table 4
The simulation results for versions 1 (V1) and 2 (V2) of the BACON algorithm for $p = 20$. The performance criteria A, B, and C are defined in (7) and ϕ is the contamination level

			A		B		C	
n	m	ϕ	V1	V2	V1	V2	V1	V2
500	80	0.1	0.9804	1.0008	0.9800	1.0000	3	4
		0.2	0.0101	1.0002	0.0101	1.0000	2	4
		0.3	0.0000	1.0001	0.0000	1.0000	2	4
		0.4	0.0000	1.0001	0.0000	1.0000	2	4
	100	0.1	0.9702	1.0006	0.9700	1.0000	3	4
		0.2	0.0001	1.0003	0.0001	1.0000	2	4
		0.3	0.0000	1.0001	0.0000	1.0000	2	4
		0.4	0.0000	1.0001	0.0000	1.0000	2	4
5000	80	0.1	1.0000	1.0000	1.0000	1.0000	3	4
		0.2	0.5400	1.0000	0.5400	1.0000	3	4
		0.3	0.0000	1.0000	0.0000	1.0000	2	4
		0.4	0.0000	1.0000	0.0000	1.0000	2	4
	100	0.1	1.0000	1.0000	1.0000	1.0000	3	4
		0.2	0.3400	1.0000	0.3400	1.0000	2	4
		0.3	0.0000	1.0000	0.0000	1.0000	2	4
		0.4	0.0000	1.0000	0.0000	1.0000	2	4
10,000	80	0.1	1.0000	1.0000	1.0000	1.0000	3	5
		0.2	0.7600	1.0000	0.7600	1.0000	3	4
		0.3	0.0000	1.0000	0.0000	1.0000	2	4
		0.4	0.0000	1.0000	0.0000	1.0000	2	4
	100	0.1	1.0000	1.0000	1.0000	1.0000	3	4
		0.2	0.6500	1.0000	0.6500	1.0000	3	4
		0.3	0.0000	1.0000	0.0000	1.0000	2	4
		0.4	0.0000	1.0000	0.0000	1.0000	2	4

The Philips data consist of 677 observations on nine variables (diaphragm parts for TV sets which are thin metal plates, molded by press). Because of masking, the classical Mahalanobis distance classifies only a few points as outliers, but both versions of the BACON methods V1 and V2 identified the same 92 observations (see Fig. 1) as outliers. Of these, 75 observations are contiguous (observations number 491–565). Rousseeuw and van Driessen (1999), who kindly sent us the data, used it to test their MCD method, which has nearly a 50% breakdown point, but requires substantially more computational time because it depends on resampling.

9. Large data sets

Remarkably, the computing cost of the BACON algorithms for multivariate outliers and for regression is low. The major costs are the computing of a covariance matrix and the computing of the distances themselves. Because the number of iterations is

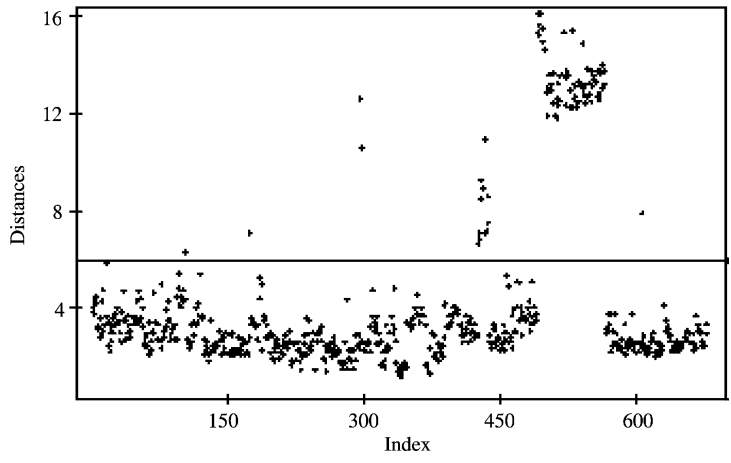

Fig. 1. The Philips Data: The index plot of robust distances $d_i(\bar{x}_b, S_b)$.

small, none of these costs grows out of bounds. In practical terms, current desktop computers can find Mahalanobis distances for a million cases in about ten seconds. It is thus practical to apply BACON algorithms to data sets of millions of cases on desktop computers and to expect virtually instant results for data sets of only 10,000 cases.

The extraordinarily small computing effort required by BACON algorithms, and in particular the fact that this effort grows slowly with increasing sample size, makes these methods particularly well suited for large data sets.

Several steps help extend the application of BACON algorithms to very large data sets. The initial basic subset can be constructed from a representative sample of the data, so that medians need not be computed for very large samples. If there is doubt about generating a representative sample, the algorithm can be computed several times from different starts. In particularly large data sets, it may be necessary to eliminate the calculation of the model for the final basic subset simply because that subset may be too large for practical computing. However, a model that is already based on millions of cases (practical for today's desktop computers) is unlikely to change very much when computed on hundreds of millions of cases.

One potential application of BACON algorithms is in *data mining*. Data mining (see, for example, Glymour et al., 1997) ordinarily deals with large, multivariate data sets, but has thus far employed methods that are not resistant to outliers. The ability to apply general multivariate outlier detection to such large data sets before fitting data mining models can significantly improve the performance and predictive ability of those methods.

10. Summary and recommendations

Outlier detection methods have suffered in the past from a lack of generality and a computational cost that escalated rapidly with the sample size. Small samples

provide too small a base for reliable detection of multiple outliers, so suitable graphics are often the detection method of choice. Samples of a size sufficient to support sophisticated methods rapidly grow too large for previously published outlier detection methods to be practical. The BACON algorithms given here reliably detect multiple outliers at a cost that can be as low as four repetitions of the underlying fitting method. They are thus practical for data sets of even millions of cases.

The BACON algorithms balance between affine equivariance and robustness. Versions that start from an affine equivariant subset of the data are themselves affine equivariant, but are generally less robust, although the robustness of the subsequent steps often adjusts for any bias in the initial basic subset due to outliers. Versions of the algorithm that start from a more robust start are not affine equivariant, although the affine equivariance of the subsequent steps often adjusts for any sensitivity to rotation in the initial basic subset. Future research may identify a computationally efficient way to identify an outlier-free robust starting subset of the data with an affine equivariant algorithm.

We define BACON algorithms for two models in this paper. However, the algorithms can be applied more broadly. To apply the BACON approach, one must be able to identify an initial basic subset clean of outliers. One must then be able to fit a model for data that generates discrepancy measures for all data values in a data set from a model fit to a subset of the data, and a suitable cutoff value for those discrepancies. The model can be arbitrarily complex and the fitting method may be iterative. For example, BACON algorithms can be applied to non-linear models provided the analyst is willing to assume an error distribution to use as a basis for determining a cutoff value for discrepancies. Researchers trying to establish general methods should take care in defining algorithms for automatically determining the initial basic subset. Data analysts may be more comfortable identifying an initial basic subset from additional knowledge they may have, or simply by examining displays of the data.

BACON methods are easy to implement in statistics packages that have programming or macro languages. Templates for Data Desk 6.0 (Velleman, 1998) are currently in preparation.

References

Atkinson, A.C., 1985. In: Plots, Transformations, and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis. Clarendon Press, Oxford.

Atkinson, A.C., 1994. Fast very robust methods for the detection of multiple outliers. Journal of the American Statistical Association 89, 1329–1339.

Bacon, F., 1620. In: Urbach, P., Gibson, J. (Translators, Eds.), Novum Organum. Open Court Publishing Co, Chicago, 1994.

Barrett, B.E., Gray, J.B., 1997. On the use of robust diagnostics in least squares regression analysis. Proceedings of the Statistical Computing Section, the American Statistical Association, pp. 130–135.

Barnett, V., Lewis, T., 1994. In: Outliers in Statistical Data. Wiley, New York.

Belsley, D.A., Kuh, E., Welsch, R.E., 1980. In: Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. Wiley, New York.

Chatterjee, S., Hadi, A.S., 1988. In: Sensitivity Analysis in Linear Regression. Wiley, New York.

Cook, R.D., Hawkins, D.M., 1990. In: Comment on unmasking multivariate outliers and leverage points. Journal of the American Statistical Association 85, 640–644.

Cook, R.D., Weisberg, S., 1982. In: Residuals and Influence in Regression. Chapman & Hall, London.
Donoho, D.L., 1982. In: Breakdown properties of multivariate location estimators. Qualifying Paper. Harvard University, Boston, MA.
Donoho, D.L., Huber, P.J., 1983. The notion of breakdown point. In: Bickel, P., Doksum, K., Hodges J.L. Jr. (Eds.), a Festschrift for Erich Lehmann. Wadsworth, Belmont, CA.
Draper, N., Smith, H., 1966. Applied Regression Analysis. John Wiley and Sons, New York.
Friedman, J.H., Stuetzle, W., 1981. Projection pursuit regression, Journal of the American Statistical Association 76, 817–823.
Glymour, C., Madigan, D., Pregibon, D., Smyth, P., 1997. Statistical themes and lessons for data mining. Data Mining and Knowledge Discovery, 1:1, http://www.research.microsoft.com/research/datamine/vol1-1.
Gould, W., Hadi, A.S., 1993. Identifying multivariate outliers. Stata Technical Bulletin 11, 2–5.
Gray, J.B., 1986. A simple graphic for assessing influence in regression. Journal of Statistical Computation and Simulation 24, 121–134.
Gray, J.B., Ling, R.F., 1984. K-clustering as a detection tool for influential subsets in regression. Technometrics 26, 305–330.
Hadi, A.S., 1992a. Identifying multiple outliers in multivariate data. Journal of the Royal Statistical Society Series (B) 54 (3), 761–771.
Hadi, A.S., 1992b. A new measure of overall potential influence in linear regression. Computational Statistics and Data Analysis 14, 1–27.
Hadi, A.S., 1994. A modification of a method for the detection of outliers in multivariate samples. Journal of the Royal Statistical Society Series (B) 56, 393–396.
Hadi, A.S., Simonoff, J.S., 1993. Procedures for the identification of multiple outliers in linear models. Journal of the American Statistical Association 88, 1264–1272.
Hadi, A.S., Simonoff, J.S., 1994. Improving the estimation and outlier identification properties of the least median of squares and minimum volume ellipsoid estimators. Parisankhyan Sammikkha 1, 61–70.
Hadi, A.S., Simonoff, J.S., 1997. A more robust outlier identifier for regression data. Bulletin of the International Statistical Institute 281–282.
Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A., 1986. In: Robust Statistics: The Approach Based on Influence Functions. Wiley, New York.
Hawkins, D.M., 1980. In: Identification of outliers. Chapman & Hall, London.
Hawkins, D.M., Simonoff, J.S., 1993. High breakdown regression and multivariate estimation. Applied Statistics 42, 423–432.
Hawkins, D.M., Simonoff, J.S., Stromberg, A.J., 1994. Distributing a computationally intensive estimator: the case of exact LMS regression. Computational Statistics 9, 83–95.
Huber, P.J., 1981. In: Robust Statistics. Wiley, New York.
Kianifard, F., Swallow, W.H., 1989. Using recursive residuals, calculated on adaptively-ordered observations, to identify outliers in linear regression. Biometrics 45, 571–585.
Mayo, M.S., Gray, J.B., 1997. Elemental subsets: the building blocks of regression. Journal of the American Statistical Association 51, 122–129.
Paul, S.R., Fung, K.Y., 1991. A generalized extreme studentized residual multiple-outlier-detection procedure in linear regression. Technometrics 33, 339–348.
Portnoy, S., 1987. Using regression Fractiles to identify outliers. In: Dodge, Y. (Ed.), Statistical Data Analysis Based on the L1-norm and Related Methods. North-Holland, Amsterdam, pp. 345–356.
Rocke, D.M., Woodruff, D.L., 1996. Identification of outliers in multivariate data. Journal of the American Statistical Association 91, 1047–1071.
Rousseeuw, P.J., 1984. Least median of squares regression. Journal of the American Statistical Association 79, 871–880.
Rousseeuw, P.J., Leroy, A., 1987. Robust regression and outlier detection. Wiley, New York.
Rousseeuw, P.J., van Driessen, K., 1999. A fast algorithm for the minimum covariance determinant estimator. Technometrics 41, 212–223.
Rousseeuw, P.J., van Zomeren, B., 1990. Unmasking multivariate outliers and leverage points (with discussion). Journal of the American Statistical Association 85, 633–639.

Ruppert, D., Simpson, D.G., 1990. Comment on unmasking multivariate outliers and leverage points. Journal of the American Statistical Association 85, 644–646.

Siegel, A.F., 1982. Robust regression using repeated medians. Biometrika 69, 242–244.

Simonoff, J.S., 1991. In: General approaches to stepwise identification of unusual values in data analysis. In: Stahel, W., Weisberg, S. (Eds.), Directions in Robust Statistics and Diagnostics: Part II. Springer, New York, 223–242.

Souvaine, D.L., Steele, J.M., 1987. Time- and space-efficient algorithms for least median of squares regression. Journal of the American Statistical Association 82, 794–801.

Staudte, R.G., Sheather, S.J., 1990. In: Robust Estimation and Testing. Wiley, New York.

Steele, J.M., Steiger, W.L., 1986. Algorithms and complexity for least median of squares regression. Discrete Applied Mathematics 13, 509–517.

Stromberg, A.J., 1993. Computing the exact least median of squares estimate and stability diagnostics in multiple linear regression. SIAM Journal on Scientific Computing 14, 1289–1299.

Sullivan, J.H., Barrett, B.E., 1997. Multivariate outlier detection using an extended stalactite plot. Proceedings of the Statistical Computing Section, the American Statistical Association, pp. 120–123.

Velleman, P.F., 1998. In: Data Desk. Data Description Inc, Ithaca, NY.

Woodruff, D.L., Rocke, D.M., 1994. Computable robust estimation of multivariate location and shape in high dimension using compound estimators. Journal of the American Statistical Association 89, 888–896.