Robust Statistical Methods for Automated Outlier Detection

J. R. Jee
Navigation Systems Section

The computational challenge of automating outlier, or blunder point, detection in radio metric data requires the use of nonstandard statistical methods because the outliers have a deleterious effect upon standard least squares methods. The particular nonstandard methods most applicable to the task are the robust statistical techniques that have undergone intense development since the 1960s. These new methods are by design more resistant to the effects of outliers than standard methods. Because the topic may be unfamiliar, a brief introduction to the philosophy and methods of robust statistics is presented. Then the application of these methods to the automated outlier detection problem is detailed for some specific examples encountered in practice.

I. Introduction: A Specific Problem and a Solution Strategy

Radio metric data must routinely be screened for spurious values which may result, for example, from a temporary malfunction in the instruments gathering the data or from a human blunder in the data collection process. These spurious values typically reveal themselves as outliers, or points which lie outside the normal range of the good data. An efficient way for a data analyst to detect these outliers is to view a computer-generated plot of the radio metric data (or, as in actual practice, a transformed version of them) against their respective time coordinates. The outliers show up in the plot as unusually large deviations from a mean curve followed by the good data. After examining such a plot, the data analyst can delete the outliers from the data set. While this mode of operation for outlier detection has been an acceptable method in the past, it is realized that the expected large increase in the amount of radio metric data to be processed will require excessive amounts of manpower unless steps are taken to automate some portion of the process. The consequent mathematical challenge is that of developing computational algorithms which can perform the job of outlier detection for radio metric data.

Since radio metric data are directly derived from physical quantities such as velocity and range, the mean curve traced by a plot of the good data should be smooth and continuous (in the absence of abrupt dynamic changes); thus, function fitting techniques seem to constitute a natural mathematical tool to use for initial analytical estimation of the mean curve. Then, assuming that the good data fall within a distinct nominal noise band about the mean curve, one can use the analytic function estimate to characterize the outliers as those data which lie outside the band. The special difficulty of this function fitting problem is that the very presence of the out-

liers foils classical function fitting attempts. Indeed, the very reason that outliers need to be removed is that least squares estimation algorithms can be adversely affected by them. More specifically, a function fit by least squares to outlier-contaminated data may not follow the good data well. A common consequence is that the good data deviate as much from the function estimate as the outliers do. To salvage this strategy based on function estimation, it is necessary to employ methods which are robust against the presence of outliers. Fortunately, outlier-robust methods have been a major area of research and development in the statistical community since the 1960s. A major goal of this article is to provide the reader with a brief introduction to these modern statistical methods by way of application to the specific problem described.

A summary of the rest of this article is now given. Section II provides an introduction to some of the philosophy and methods of robust statistics. Sources of information for this section include the seminal article by Huber [1] and his follow-up textbook [2]. Two less theoretical treatments of this topic are given in [3] and [4]. Section III details the application of these methods to the outlier detection problem. The description is devoted more to providing the rationale behind automating outlier detection than to specifying fine algorithmic details. The methods are applied to two sets of radio metric data from the International Cometary Explorer (ICE) spacecraft. Section III also outlines the application of two additional statistical tests which can be found in more standard texts such as [5].

II. Robust Statistical Methods

The operative mode for much of classical statistics is to assume an appropriate probability model and then to employ the optimal procedure for the model. For the purposes of this discussion, the efficiency of a procedure is measured by its theoretical variance. In contrast to classical procedures, robust statistical methods by design rely less heavily upon an appropriate choice of the probability model. Considerations of robustness lead to the development of methods which compromise the requirement of optimal efficiency for the ability to accommodate a range of deviations from a specific assumed model with reasonable, rather than optimal, efficiency. For example, a very common model is the Gaussian density denoted by $\phi(\cdot;\mu,\sigma)$. A class of deviations from this model is given by the mixture densities $(1-\epsilon)\,\phi(\cdot;\mu,\sigma) + \epsilon\phi(\cdot;\mu,N\sigma)$, where N is a large number and ϵ ranges from 0 for no deviation to 1/2 for large deviations. These particular mixture densities with ϵ between 0.01 and 0.1 often provide a more realistic model of real data which tend to be contaminated with outliers. The main fault of classical procedures is that they perform optimally for their intended case $\epsilon = 0$ but often lose efficiency quite drastically as ϵ increases. Robust alternatives remedy this fault by insuring against unacceptably poor performance while not necessarily providing optimal efficiency for any one case.

A. Robust Location and Scale Estimation

The problems of locating the center of a distribution and of determining its scale are the simplest cases for illustration of these concepts. As a specific example, suppose a set of data $\{x_i\}_{i=1}^{n}$ is assumed to come from a Gaussian distribution for which the mean and variance are to be estimated. The classical location estimator is the minimizer of the least squares error criterion $\rho_2(T) = \Sigma(x_i - T)^2$. This estimator is, of course, the sample mean denoted by $\bar{x}_n$. The classical estimate of scale is derived by taking the square root of the sample variance, $\sqrt{s_n^2}$. Suppose that the particular data set has $\bar{x}_n = 0$ and $s_n^2 = 1$, and consider the effect of an additional datum on these estimates. Then the sample mean $\bar{x}_{n+1} = x_{n+1}/(n+1)$ while the sample variance $s_{n+1}^2 = [(n-1)/n] + x_{n+1}^2/(n+1)$. These equations show that a single outlier x_{n+1} can cause the estimates to be arbitrarily large. This is one sense in which the classical estimators of location and scale are not robust against the presence of outliers.

The extreme sensitivity of the sample mean to outliers can be traced to the error criterion from which it is derived. In an effort to balance the total squared error, the sample mean is forced to overcompensate for the one outlier. Consider now changing the form of the error criterion to $\rho_1(T) = \Sigma|x_i - T|$. This criterion is minimized by the median M_n = median (x_i). It is easy to verify that one outlier, or even a small percentage of outliers, has a limited effect upon the location estimate based upon minimizing ρ_1. Thus, the median offers some robustness against outliers and consequently is receiving renewed attention as a location estimator. Similarly, a preferred robust estimate of scale is the Median Absolute Deviation from the sample median, MAD_n = median $\{|x_i - M_n|\}$. For a Gaussian data set, $E[s_n] \approx \sigma$, but $E[\mathrm{MAD}_n] \approx (2/3)\sigma$. Thus, a factor of 3/2 is required to make the MAD_n directly comparable with s_n as an estimator.

B. Robust Linear Regression

In brief review, linear regression analysis is concerned with equations of the form $\mathbf{y} = \mathbf{X}\beta + \epsilon$, where $\mathbf{X}$ is a matrix of known quantities, $\mathbf{y}$ is also a known vector often called the vector of observations, β is a vector of unknown parameters to be estimated, and ϵ is a vector of random errors with covariance $E[\epsilon\epsilon^T] \equiv \mathbf{W}^{-1}$. For example, least squares function fitting is one of the problems which may be formulated in the linear regression framework. Classical regression proceeds to a solution by finding the β which minimizes the summed weighted-squared errors $(\mathbf{y} - \mathbf{X}\beta)^T\mathbf{W}(\mathbf{y} - \mathbf{X}\beta)$. The extremal condition

obtained by differential calculus is $\mathbf{X}^T\mathbf{W}(\mathbf{y} - \mathbf{X}\beta) = 0$, from which the solution is quickly obtained. The sensitivity of this least squares solution to outliers is seen by examining the predicted observation vector $\mathbf{y}_{LS} = \mathbf{X}\beta_{LS} = \mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y}$. This equation shows that if all y-values are held fixed except for y_i, then a change in y_i produces a proportional change in the least squares fitted value $(y_{LS})_i$. Thus, a deviant value can have an arbitrarily large effect upon least squares estimates.

The technique of modifying the error criterion to obtain a robust location estimator is now applied to the regression problem. The least squares criterion may be written as

$$\sum_i (y_i - \mathbf{x}_i\beta)^2 w_i$$

if $\mathbf{x}_i$ denotes a row of $\mathbf{X}$ and $\mathbf{W}$ is a diagonal matrix with nonzero elements w_i. Another standard notation is that for the residual $r_i = y_i - \mathbf{x}_i\beta$. The particular robust error criteria considered here are of the form $\Sigma_i \rho_c(r_i)$, where

$$\rho_c(r) = \begin{cases} c|r| - (1/2)c^2 & \text{if } |r| \geqslant c \\ (1/2)|r|^2 & \text{if } |r| < c \end{cases}$$

where each positive value of c defines a candidate error measure. As c is chosen to be arbitrarily large, the criterion approaches the least squares measure; as c approaches 0, the absolute value criterion is produced. At a minimizer of $\Sigma_i \rho_c(r_i)$, the β must satisfy

$$\sum_{|r_i| \leqslant c} (y_i - \mathbf{x}_i\beta)\mathbf{x}_i^T + \sum_{|r_i| > c} \frac{c}{|r_i|}(y_i - \mathbf{x}_i\beta)\mathbf{x}_i^T = 0$$

This equation reveals that the optimality condition resembles something arising from a weighted least squares problem with two types of weights: weights of 1 for residuals smaller than c and weights of $\sqrt{c/|r_i|}$ for larger residuals. Unfortunately, the proper weighting cannot in general be known before the problem is solved, and the equation as it stands is nonlinear in the unknowns. As for the constant c, setting $c = 1\sigma$ yields an estimator that has good efficiency for independent, identically distributed Gaussian errors while providing robustness against mild contamination. However, the variance of the errors is not always known beforehand, so c may also be another unknown parameter to be determined.

One popular technique of solving the extremal condition equations is called Iteratively Reweighted Least Squares (IRLS). This technique is favored because the algorithm (1) is easy to understand, (2) is easy to implement if a common weighted least squares routine is available, (3) works well in practice, and (4) has fairly well understood convergence properties. For simplicity of exposition, the regression problem is now assumed to have $\mathbf{W}^{-1} = \sigma^2\mathbf{I}$. Then the IRLS algorithm works as follows: (1) the regression problem is solved by the standard least squares method, and residuals from this regression are calculated; (2) the MAD is used to estimate σ, the scale of the residuals; (3) the standard weighted regression equations are solved, where weights of 1 are assigned to residuals less than σ and weights of $\sqrt{\sigma/|r_i|}$ are used otherwise; and (4) steps 2 and 3 are iterated until convergence is achieved. While the theoretical convergence properties of the IRLS algorithm are not completely known, partial results indicate that it may be globally convergent, as is the case in computational practice. Furthermore, it often converges at a linear rate to a minimum of the error criterion.

III. Application to Automated Outlier Detection

The statistically robust estimation methods previously described are now applied to the automated outlier detection problem initially presented. More specifically, the problem is that a set of sequential observations $\{(t_i, y_i)\}_{i=1}^n$ are to be screened for deviant y-values. The observations are quantities derived from radio metric data such as range and doppler. In the particular case to be studied, the derived data are doppler pseudoresiduals generated by the Deep Space Stations while tracking the International Cometary Explorer (ICE) spacecraft. Pseudoresiduals are essentially numerical differences between the observed values and the predicted values.

A. Model Selection

As in any applied mathematical problem, one assumes, either explicitly or tacitly, some model for the physical situation which is to be studied. The two main model assumptions for this application are stated in this section.

Dynamic considerations can be used to derive analytic descriptions of the observations as functions of time. For a single pass of data of several hours' length, polynomials may be used to approximate these functions. From empirical studies, it was found that polynomials of degree $2n$ usually suffice to accurately approximate the mean curve of a set of data of length n hours from a spacecraft in interplanetary cruise. In choosing the degree of the polynomial, it is more desirable to underestimate the required degree than to overestimate it because an overly high order polynomial can fit both good data and outliers. While splines and trigonometric series are other possible candidates for the function approximation, time has not permitted an investigation into their use for this application.

Extensive experience has shown that the nominal noise about the mean curve is well modeled by independent, identically distributed Gaussian random variables. Thus, if good estimates of the variance of the Gaussian noise are obtained, then a 3σ rejection rule should falsely remove only about 1 out of 500 good values. On this basis, an outlier may be defined as a value that is deviant by more than 3σ, where σ is a robust estimate of the scale of the nominal Gaussian noise. From experience, the portion of outliers in a pass is typically between 1 and 10 percent, and they may deviate from the mean curve by as much as a few orders of magnitude above the nominal σ.

B. Algorithm Selection

The mathematical tools required for automated outlier detection have been presented in Section II. In particular, the iteratively reweighted least squares, or IRLS, algorithm is used to fit a polynomial to the outlier-contaminated data according to the error criterion ρ_c. The choice of $c = 1\sigma$ gives an estimator with good statistical efficiency for pure Gaussian errors and resistance to approximately 10 percent outlier contamination. The MAD is used in the IRLS algorithm to estimate the scale. After the IRLS fit is completed, the MAD can be used once more on the residuals from the fit to obtain a final scale estimate upon which the 3σ rejection rule is based.

Figures 1 and 2 illustrate the application of the IRLS algorithm to a specific problem. The data are 2-way doppler pseudoresiduals from the ICE spacecraft. Figure 1 shows a plot of the data along with a standard least squares fourth degree polynomial fit to the entire data set (recall that this is the first step in the IRLS algorithm). Also in the plot are 3σ bands about the least squares fit calculated by the usual standard deviation formula. The two main effects of the outliers on these classical estimators are clearly shown in Fig. 1. First, the extreme outlier near the end of the time segment pulls the least squares polynomial away from the good data. Then the outliers and the oversized residuals near the end of the segment combine to inflate the standard deviation estimate. Figure 2 shows the final IRLS polynomial fit to the data with robust 3σ bands. In this plot, the ordinate has a different scale and the extreme outlier is marked by an arrow at its abscissa. As can be seen, the robust methods give a more agreeable estimate of both the mean curve and the spread of the nominal noise about the curve.

C. Model Verification

If the function estimation and the outlier rejection steps are performed correctly, then according to the assumptions for the model, the remaining residuals should be distributed as Gaussian noise. As is often the case, however, models are not always as accurate as they need to be for the mathematical methods to perform as hoped. Consequently, it is essential that measures for verifying model adequacy are included in this automatic data screening algorithm. The major assumptions which must be checked are that the polynomial does give a good approximation to the function followed by the good data, that the portion of outliers detected is less than 10 percent, and that the data remaining after the editing are Gaussian.

As a specific case, Fig. 3 shows a plot of 1-way doppler pseudoresiduals, an IRLS fifth degree polynomial fit to these data, and a robustly determined 3σ band. Again, arrows at the borders of the plot denote outliers that are out of range. This is an example of model underfitting, as the polynomial does not follow some distinct features of the data. Consequently, the basis for the automated outlier detection program is undermined, and the user of the algorithms should be warned of the unreliability of results obtained in this case.

Underfitting can often be characterized by a tendency for strings of data to lie on one side of the polynomial rather than being more randomly strewn about the fitted curve. Fortunately, there exist standard statistical procedures for verifying the randomness of data based on this idea. A statistical test which considers only the signs of the residuals is called the runs test. The statistic upon which it is based is the total number of runs, denoted by R, of both consecutive positive signs and consecutive negative signs. The exact theoretical mean and variance of R under the hypothesis of randomness are given by

$$E[R] = \frac{2NP}{N+P} + 1$$

and

$$\mathrm{Var}[R] = \frac{2NP(2NP - N - P)}{(N+P)^2(N+P-1)}$$

where N is the number of negative signs and P is the number of positive signs. For data sets of size 50 or more, the random variable R standardized by its mean and variance is approximately a zero mean unit variance Gaussian random variable; thus, the measure of randomness given by R can be calculated easily.

If the retained residuals test negatively for underfitting, then an additional chi-square test for Gaussian behavior can be applied. In a nutshell, this test is based upon measuring the amount of agreement between a histogram of the data and a theoretical histogram determined by a "perfect" Gaussian

sample. If $\hat{\eta}_i, \ldots, \hat{\eta}_k$ denote the number of points in each histogram bin and $\eta_i, \ldots, \eta_k$ the number of points in a "perfect" Gaussian histogram, then the chi-squared test statistic C is given by

$$C = \sum_{i=1}^{k} \frac{(\hat{\eta}_i - \eta_i)^2}{\eta_i}$$

Finally, if the polynomial does not seem to underfit the data and if the data seem to be distributed as Gaussian noise about the polynomial, then a simple count of the percentage of outliers should be performed. Since the algorithms are geared to handle data with 10 percent or less outliers, cases which contain greater than 10 percent detected outliers should be flagged as potential problem sets requiring manual inspection.

IV. Discussion

A prime motivation for employing function fitting in the outlier detection problem for radio metric data is that this seems to mimic the operation of a human data analyst. As stated in the introduction, a data analyst usually bases much of outlier screening upon visual inspection of a plot. The human eye "smooths" the data while ignoring outliers to extract the underlying curve. Function fitting attempts to imitate this process, and robust function fitting in particular is required to accurately produce the underlying curve. After the curve is estimated, the robust 3σ rule provides some analytical basis for outlier detection comparable to an analyst's decision to remove data separated by a "gap" from the bulk of the data spread about the underlying curve.

Another reason for choosing robust statistical methods is that they are fairly easy to understand at the conceptual level. There are more classical statistical procedures for outlier detection in the linear regression framework, but their application in this setting is less straightforward. Typically, classical procedures require more involved statistical reasoning than do robust methods while providing comparable performance; hence the choice for robust methods. For a more detailed comparative discussion, the interested reader can consult [6], which is an extensive survey of methodologies (classical, Bayesian, and robust) for handling outliers in various contexts.

To think that robust statistical methods were first invented in the 1960s is incorrect, since the median and other robust statistics were in use long before then. However, it was not until the 1960s that a unifying framework was established for considerations of robustness. This development prompted considerable theoretical and computational investigation into the subject. Also, while the emphasis in this article has been on robustness against outliers, it should be known that the goals of robust statistics include protection against more than just outliers. A more complete description of this modern statistical methodology can be found in the references. As a note of caution, robust statistical methods are not a new class of foolproof methods which will replace classical methods based on least squares and maximum likelihood. Instead, they constitute another class of methods with its own domain of application alongside those of other statistical methods. One appropriate domain of application is the outlier detection problem of this article. For this case, robust methods should seem not only reasonable but also ideal for the problem.

References

[1] P. J. Huber, "Robust Estimation of a Location Parameter," *Annals of Mathematical Statistics*, vol. 35, pp. 73–101, 1964.

[2] P. J. Huber, *Robust Statistics*, New York: John Wiley, 1981.

[3] D. C. Hoaglin, F. Mosteller, and J. W. Tukey, *Understanding Robust and Exploratory Data Analysis*, New York: John Wiley, 1983.

[4] D. C. Hoaglin, F. Mosteller, and J. W. Tukey, *Exploring Data Tables, Trends, and Shapes*, New York: John Wiley, 1985.

[5] I. M. Chakravarti, R. G. Laha, and J. Roy, *Handbook of Methods of Applied Statistics*, New York: John Wiley, 1967.

[6] R. J. Beckman and R. D. Cook, "Outliers," *Technometrics*, vol. 25, pp. 119–163, 1983.

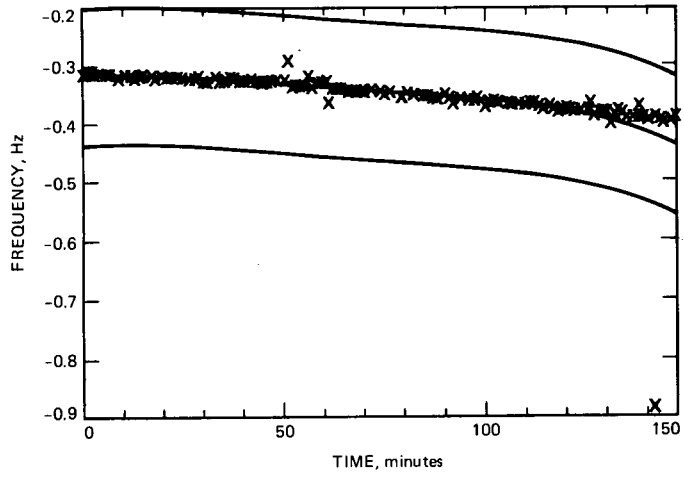

Fig. 1. Fourth degree least squares polynomial and 3 standard deviations based on 2-way doppler pseudoresiduals from ICE. Extreme outlier at time 141 pulls the polynomial away from the rest of data near the end of the time segment and inflates the variance estimate. Time segment begins at 13:18, February 5, 1987.

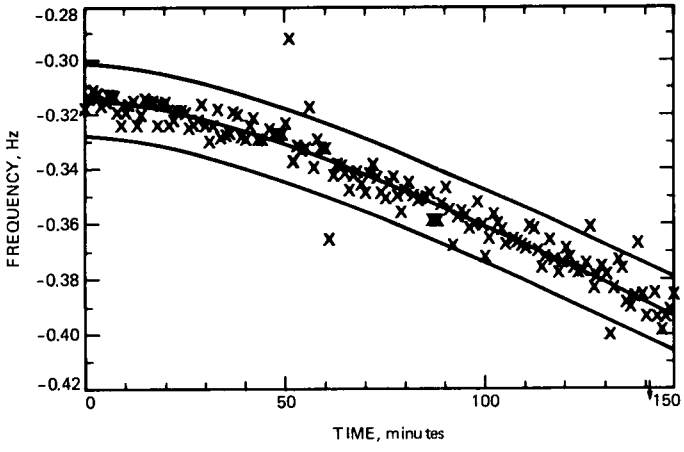

Fig. 2. Fourth degree IRLS polynomial and MAD-based 3σ for 2-way doppler data of Fig. 1. Out-of-range extreme outlier at time 141 is marked by the arrow at the bottom of the graph. Time segment begins at 13:18, February 5, 1987.

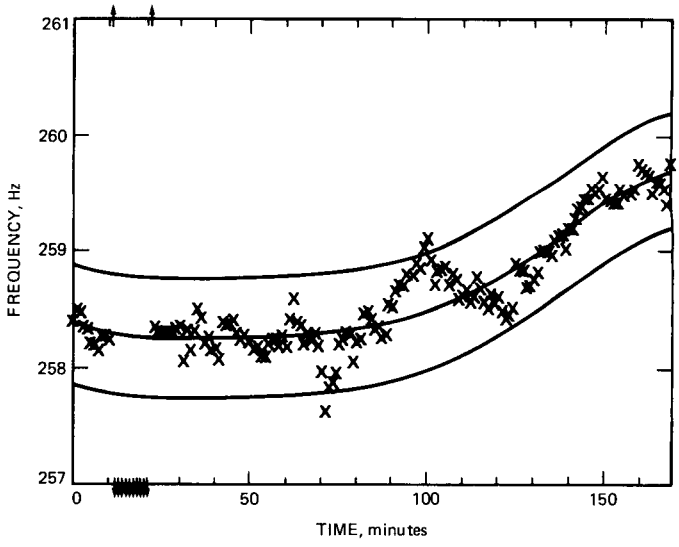

Fig. 3. Fifth degree IRLS polynomial and MAD-based 3σ for 1-way doppler pseudoresiduals from ICE. Out-of-range outliers are marked by arrows. Polynomial underfitting is characterized by the number of runs of deviations of the same sign from the polynomial. Time segment begins at 18:23, February 7, 1987.