Appl. Statist. (2003)
52, *Part* 3, *pp.* 307–322

A bounded influence regression estimator based on the statistics of the hat matrix

Alan D. Chave

Woods Hole Oceanographic Institution, USA

and David J. Thomson

Queens University, Kingston, Canada

[Received November 2000. Final revision December 2002]

Summary. Many geophysical regression problems require the analysis of large (more than 10^4 values) data sets, and, because the data may represent mixtures of concurrent natural processes with widely varying statistical properties, contamination of both response and predictor variables is common. Existing bounded influence or high breakdown point estimators frequently lack the ability to eliminate extremely influential data and/or the computational efficiency to handle large data sets. A new bounded influence estimator is proposed that combines high asymptotic efficiency for normal data, high breakdown point behaviour with contaminated data and computational simplicity for large data sets. The algorithm combines a standard M-estimator to downweight data corresponding to extreme regression residuals and removal of overly influential predictor values (leverage points) on the basis of the statistics of the hat matrix diagonal elements. For this, the exact distribution of the hat matrix diagonal elements p_{ij} for complex multivariate Gaussian predictor data is shown to be $\beta(p_{ij}, m, N - m)$, where N is the number of data and m is the number of parameters. Real geophysical data from an auroral zone magnetotelluric study which exhibit severe outlier and leverage point contamination are used to illustrate the estimator's performance. The examples also demonstrate the utility of looking at both the residual and the hat matrix distributions through quantile–quantile plots to diagnose robust regression problems.

Keywords: Bounded influence estimator; Hat matrix; Projection matrix; Robust regression; Time series analysis; Transfer function estimation

1. Introduction

The application of regression methods to large (more than 10^4 values) data sets is increasingly common in the geophysical sciences. In many of these problems, the data are complex in the mathematical sense of having real and imaginary parts. Complex data arise either because the original data consist of time series measurements, but the data used in the regression problem consist of windowed Fourier transforms of them, or when quantities such as wind velocity that are a dominantly two-dimensional vector are best described and analysed as complex numbers (e.g. Calman (1978)). Complex data are also common in other areas of science and technology, notably communications and electronic engineering. Further, geophysical data are typically direct measurements of natural processes collected with limited control of environmental systematics, in contrast with the standard situation in the laboratory sciences. As a result, the measurable quantities may be the result of mixtures of distinct but concurrent processes with

Address for correspondence: Alan D. Chave, Deep Submergence Laboratory, Department of Applied Ocean Physics and Engineering, Woods Hole Oceanographic Institution, Woods Hole, MA 02543, USA.
E-mail: alan@whoi.edu

 0035–9254/03/52307

widely varying statistical properties, some of which may be non-stationary, so that violations of the standard regression conditions are commonplace. In many cases, standard robust estimators fail on geophysical data, and new approaches are necessary which are both computationally efficient and robust to a large fraction of outliers in all the variables.

This situation is typified by magnetotelluric studies of the electrical structure of Earth through measurements of its response to variations of natural electromagnetic sources in the ionosphere and magnetosphere (e.g. Vozoff (1972)). Magnetotelluric data consist of time series of the variations of the horizontal electric and magnetic fields at Earth's surface at a variety of locations. They are usually collected for days to months at sample rates of 1 Hz or more and hence constitute large data sets. In the regression context, the magnetic fields are the input, or predictor, and the electric fields are the output, or response, variables. Further, the input–output relationship is derived from the Maxwell equations and so is not in doubt. Data are typically analysed site by site, although they are interpreted collectively. The diffusive form of electromagnetic waves in a conductive medium results in the information contained in them being spread out smoothly over a wide range of frequencies, and hence data are typically treated in the frequency domain rather than in the time domain.

The impulsive and non-stationary behaviour of natural source electromagnetic fields is epitomized by (although not limited to) familiar auroral and geomagnetic storm phenomena. The latter can at times interfere with power and telecommunications systems (e.g. Lanzerotti *et al.* (1999)) and hence constitute physical as well as statistical outliers. Human activity such as the transients from switching direct current trains also produces impulsive and non-stationary electromagnetic fields (e.g. Egbert *et al.* (2000)); since this type of source is at ground level, it does not give the same field configuration as the ionospheric sources assumed in magnetotellurics. These can lead to serious outlier and, in the most severe cases, leverage problems. Over the past decade, conventional robust methods have revolutionized the use of magnetotellurics in geophysics (Chave and Thomson, 1989; Jones *et al.*, 1989; Egbert, 1997) and are now applied routinely and automatically, producing reliable magnetotelluric responses in most instances. However, at sites within the auroral zone at high latitudes where source field non-stationarity is especially severe, robust methods frequently break down (often spectacularly) owing to extremely influential predictor data (Garcia *et al.*, 1997). Other situations where leverage problems are important include geomagnetic storms at mid-latitudes and any point within 100 km of a direct current train track (i.e. most of western Europe). The interpretation of data from these regions is problematical unless reliable and automatic methods to remove influential data effects from large data sets can be devised.

The bounded influence (BI) estimator that is presented in this paper provides the required reliability for extreme magnetotelluric data and has proven to be more generally applicable to a variety of other situations. This estimator combines high asymptotic efficiency for normal data, high breakdown point performance with contaminated data and computational simplicity that is suitable for large data sets. It is based on the combination of a standard M-estimator to remove data corresponding to large regression residuals with leverage point removal based on the statistics of the hat matrix diagonal, for which the exact distribution given complex multivariate Gaussian data is derived. Its performance is illustrated with both standard data and a large, severely contaminated magnetotelluric data set from central British Columbia.

2. Regression and robustness

The standard linear regression model relates the $N \times 1$ vector $\mathbf{y}$ of observations of the response (sometimes called dependent) variable to the $N \times m$, rank m matrix $\mathbf{X}$ of predictors (some-

times called explanatory, regressor or carrier variables) through the $m \times 1$ vector of unknown parameters β:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon \tag{1}$$

where ε is an $N \times 1$ vector of random errors. For complex response and predictor data, the least squares solution for model (1) is

$$\hat{\beta} = (\mathbf{X}^{\mathrm{H}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{H}}\mathbf{y} \tag{2}$$

where the superscript H denotes the Hermitian transpose. The predicted values of the response variable from the regression are derived from the observed values by

$$\hat{\mathbf{y}} = \mathbf{P}\mathbf{y} \tag{3}$$

where the $N \times N$ prediction or hat matrix is given by

$$\mathbf{P} = \mathbf{X}(\mathbf{X}^{\mathrm{H}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{H}}. \tag{4}$$

The regression residuals $\mathbf{r}$ are the differences between the observed $\mathbf{y}$ and predicted $\hat{\mathbf{y}}$ response variables.

The classical Gauss–Markov theorem gives the conditions on the response, predictor and residual variables and their moments under which the least squares estimator will be the best unbiased linear estimator, and the high efficiency of least squares when these are met is well known. In geophysical applications, the predictor as well as the response variables are typically random rather than fixed. Shaffer (1991) has shown that the Gauss–Markov theorem also holds when the joint distribution of $\mathbf{X}$ and $\mathbf{y}$ is multivariate normal with unknown parameters, the distribution of $\mathbf{X}$ is continuous and non-degenerate but otherwise unknown or, under mild conditions, if $\mathbf{X}$ is a realization of a random sample from a finite population. It is in this sense that model (1)–(2) will be considered in this paper.

Regardless of whether $\mathbf{X}$ is fixed or random, problems with the least squares estimator when the regression residuals are markedly heteroscedastic and heavy tailed and/or when some of the points in $\mathbf{X}$ are unduly influential have been extensively documented. Typically, these result in both a loss of efficiency and substantial errors in statistical inference about the parameters in $\hat{\beta}$. Over the past three decades, this has led to the development of robust regression methods which, in varying ways and to different degrees, automatically reduce the influence of a small fraction of data which cause problems for least squares. Recent reviews of this topic appear in Ryan (1997) and Wilcox (1997).

A simple robust regression method is the M-estimator which systematically reduces the effect of data on the basis of the size of the elements of $\mathbf{r}$. M-estimators may not be sensitive to overly influential predictors (usually termed leverage points), depending on whether they produce unusual residuals or not. In fact, M-estimators have a breakdown point of only $1/N$ so a single leverage point can completely dominate the ensuing estimate, and their influence functions as defined by Hampel (1974) are unbounded. These limitations have led to the development of estimators that bound the influence of any single element or row of $\mathbf{X}$, so that they guard against leverage points as well as regression outliers. These will be called BI estimators, although the term generalized M- or GM estimator is often used as well. Indeed, a good GM estimator has BI, but a poor choice of leverage downweighting strategy can result in GM estimates that are not bounded. A standard statistical measure of leverage is the size of the diagonal elements of the hat matrix, and many estimators use this quantity to detect and downweight leverage values (e.g. Mallows (1975), Handschin *et al.* (1975) and Krasker and Welsch (1982)). These BI estimators have a breakdown point of at most $1/m$ (Ryan, 1997). In practice, standard

BI estimators have proven to be less than satisfactory in the face of multiple, large leverage values, as is typical of geophysical data, in part because downweighting of bad data typically is mild and limited rather than aggressive. More capable, high breakdown point (up to 0.5), estimators such as the least median of squares and least trimmed sum of squares (Rousseeuw, 1984) have been developed, and hybrid combinations with BI estimators have also been proposed (e.g. Coakley and Hettmansperger (1993)). However, both of these approaches entail a substantial increase in computational complexity that limits their applicability to large data sets and would have to be adapted to complex data and the method of instrumental variables to be widely useful in the earth sciences. Moreover, although advances in computing have been substantial, we concur with the conclusion of Hawkins and Olive (2002) that high breakdown estimators are impractical to compute exactly for large samples, and hence we regard computational efficiency to be an important consideration.

3. Statistics of the hat matrix diagonal

The hat matrix is an important auxiliary quantity in regression theory and is a standard measure of predictor influence (e.g. Hoaglin and Welsch (1978), Belsley *et al.* (1980) and Chatterjee and Hadi (1988)). The ith diagonal element of $\mathbf{P}$ (denoted by p_{ii}) is a measure of the potential influence or leverage of the ith predictor observation. Because it is a projection matrix, $\mathbf{P}$ is symmetric and idempotent, so that $0 \leqslant p_{ii} \leqslant 1$. The eigenvalues of a projection matrix are either 0 or 1, and the number of non-zero eigenvalues equals its rank; hence the trace of $\mathbf{P}$ is m and the expected value of p_{ii} is m/N. The factor by which the hat matrix diagonal estimate must exceed the expected value to be considered a leverage point is not well defined, but statistical lore suggests that values which are more than 2 or 3 times m/N are problematic.

It is often reasonable to assume that the rows of $\mathbf{X}$ are multivariate normal, although this is not required in least squares theory. Further, the normal assumption serves as a convenient base-line for evaluating statistical entities (e.g. p_{ii}) derived from measured predictors which often contain influential points, as will be demonstrated later. A derivation of the exact distribution of p_{ii} for Gaussian predictors does not appear to have been published, although asymptotic forms for real data are available (e.g. Rao (1973), Belsley *et al.* (1980) and Chatterjee and Hadi (1988)). In Appendix A, the exact distribution of the diagonal elements of the hat matrix for complex multivariate Gaussian predictors is shown to be $\beta(p_{ii}, m, N-m)$. This relationship has been tested by extensive numerical simulation and holds for real data as a special case. Interestingly, the exact derivation proves to be simple for complex data whereas it is not tenable for real data; see Appendix A for details.

From the beta distribution, the expected value of p_{ii} is m/N, in agreement with the heuristic value given above. The cumulative distribution function is the incomplete beta function ratio $I_x(m, N-m)$. Appendix B gives an exact series expression for the cumulative distribution function for integer m and N which is useful for computation.

Fig. 1 shows the percentile of the beta distribution scaled by N/m at the 0.90, 0.95 and 0.99 probability levels as a function of the regression order m in the limit $N \gg m$. If a 5% penalty for Gaussian data is acceptable, the 0.95-line is the approximate factor by which the expected value m/N should be scaled to obtain a threshold value for p_{ii} to define leverage points. Note the rapid decrease from a value of nearly 3 for simple regression ($m = 1$) and the nearly constant value of about 1.2 for $m > 50$. Defining leverage points to be those corresponding to $p_{ii} > 2m/N$ would carry a 5% penalty for Gaussian data only for $m = 3$, with a larger penalty obtaining for smaller m and a rapidly decreasing value applying to larger regression problems. This would inevitably leave an increasingly large fraction of potential leverage points in place as m increases.

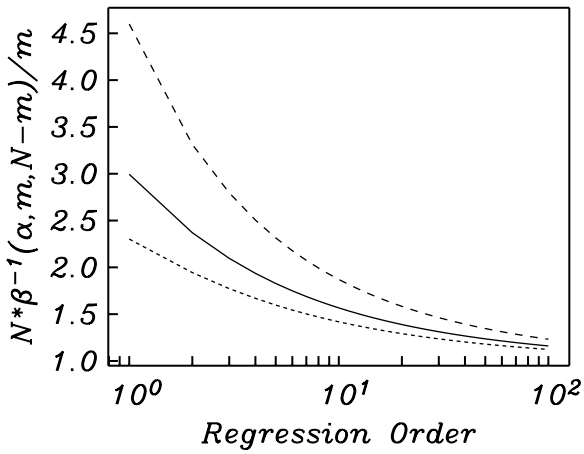

Fig. 1. Percentile of the beta distribution for $(m, N-m)$ degrees of freedom divided by its expected value m/N as a function of log(m) for the limit $N \gg m$ at the 0.99- (-------), 0.95- (———) and 0.90- (·····) levels: see the text for discussion

For this reason, it is recommended that the common statistical practice of designating data with $p_{ii} > \alpha m/N$, where α is 2 or 3, as leverage points be discontinued. A better value for α is the percentile of the beta distribution scaled by N/m at a suitable probability threshold. This ensures that, on average, the same proportion of the population is flagged for any size regression problem. However, note that such statistical rules of thumb may be inappropriate for non-Gaussian predictors which lead to a hat matrix diagonal which is systematically longer tailed than beta, as is illustrated in Section 5.

4. The estimator

Consider the class of estimators defined by

$$\sum_{i=1}^{N} w_i \, \Psi\left\{ \frac{r_i}{\tau(\mathbf{x}_i) d} \right\} x_{ij} = 0 \tag{5}$$

for $j = 1, \ldots, m$, where r_i is the ith regression residual, Ψ is the derivative of the loss function (or the influence function), d is a robust estimate of the scale of the residuals, w_i is a weight which depends on a measure of leverage and $\tau(\mathbf{x}_i)$ is a function of the ith row of the predictor matrix. Two versions of equation (5) are in common use as BI estimators. If $\tau(\mathbf{x}_i) = 1$ and $w_i = \sqrt{(1 - p_{ii})}$, then the form is that originally suggested by Mallows (1975), in which leverage points are gently downweighted according to the size of the hat matrix diagonal, whereas residual outliers are dealt with through a standard M-estimator. If $\tau(\mathbf{x}_i) = w_i$, then equation (5) uses Schweppe weights as originally suggested by Handschin *et al.* (1975). The Schweppe approach is more efficient than the Mallows approach since large leverage points corresponding to small residuals are not heavily penalized, but Carroll and Welsh (1988) have shown under general conditions that it can lead to parameter estimates which are not consistent.

In the authors' experience with severely contaminated geophysical data, neither the Mallows nor the Schweppe approach is adequate in the presence of strong leverage because they do not eliminate influential data sufficiently aggressively. A similar comment pertains to M-estimators

using non-descending loss functions such as the Huber type. Rather, it is essential to identify and remove the most severe outliers and leverage points. For large data sets (10^4 values or more per regression problem, with many such problems per application), it is also essential that this be accomplished with high computational efficiency.

An algorithm which accomplishes these goals has been developed. As with most robust estimators, the algorithm uses a base-line Gaussian model contaminated by a fraction of outlying values. There are at least three reasons for using this approach, especially in a time series context. First, windowed Fourier estimates are a linear transformation of time series data. The narrow effective bandwidths employed conform to the requirements of the theorems in Mallows (1967), and hence Fourier transforms must be approximately complex Gaussian in the absence of outlying data. Second, the effectiveness of the Gaussian mixture model as a basis for robust estimation on time series data in both the time and the frequency domains has been repeatedly demonstrated in a wide range of situations (Thomson, 1977; Kleiner *et al.*, 1979; Chave *et al.*, 1987; Chave and Thomson, 1989, 2003). Third, non-stationarity and contamination effects in time series data imply that the variance will be time dependent, suggesting that the appropriate model is a Gaussian mixture.

For simplicity, reduced concern about high statistical efficiency given large data sets, and the consistency issues raised by Carroll and Welsh (1988), only the Mallows (1975) approach in equation (5) has been used, so that $\tau(\mathbf{x}_i) = 1$. This leads to the iteratively reweighted least squares form

$$\sum_{i=1}^{N} w_i^{[k]} v_i^{[k]} r_i^{[k]} x_{ij} = 0 \tag{6}$$

for $j = 1, \ldots, m$, where the superscript $[k]$ is the iteration number, $r_i^{[0]}$ is a residual from the ordinary least squares solution (2),

$$v_i^{[k]} = \frac{\Psi(r_i^{[k-1]}/d^{[k-1]})}{r_i^{[k-1]}/d^{[k-1]}}$$

and $d^{[k-1]}$ is a robust estimate of scale computed from the $[k-1]$ order residuals. The $\{w_i^{[k]}\}$ and $\{v_i^{[k]}\}$ will be termed leverage and residual weights as they downweight leverage points and residual outliers respectively. The two sets of weights in equation (6) are decoupled and will be considered separately.

Leverage weights are required which rapidly remove the effects of severe leverage points in a smooth manner to maintain computational stability, which obviates against simple hard limiting. A variant on the weight motivated by the form of the Gumbel extreme value distribution that was originally suggested by Thomson (1977) takes the form

$$w_i^{[k]} = w_i^{[k-1]} \exp\{\exp(-\chi^2)\} \exp[-\exp\{\chi(t_i - \chi)\}] \tag{7}$$

where $w_i^{[0]} = 1$. This weight is especially suitable because the parameterization is both continuous and continuously differentiable. In addition, the exponential terms in equation (7) have a maximum value of 1, so the iteratively multiplicative nature of the procedure implies that, once a given observation has been downweighted, the weight can only decrease on subsequent iterations. For leverage point weighting, the statistic t_i in equation (7) is the normalized hat matrix diagonal element $M^{[k]} p_{ii}^{[k]}/m$ with

$$p_{ii}^{[k]} = u_i^{[k-1]} \mathbf{x}_i (\mathbf{X}^{\mathrm{H}} \mathbf{U}^{[k-1]} \mathbf{U}^{[k-1]} \mathbf{X})^{-1} \mathbf{x}_i^{\mathrm{H}} \mathbf{u}_i^{[k-1]} \tag{8}$$

where $\mathbf{U}^{[k]} = \mathbf{W}^{[k]}\mathbf{V}^{[k]}$, $\mathbf{W}^{[k]}$ and $\mathbf{V}^{[k]}$ are diagonal leverage and residual weight matrices, $\mathbf{W}^{[0]}$ and $\mathbf{V}^{[0]}$ are identity matrices and $M^{[k]}$ is the trace of $\mathbf{U}^{[k]}$ which is initially the number of data points N. The free parameter χ in equation (7) determines the point where leverage point downweighting begins. Following on the discussion of Section 2, an empirical choice for χ is the percentile of the beta distribution normalized by m/M at some specified probability level, typically lying between 2 and 4. Using a 0.95-criterion results in a 5% leverage penalty with Gaussian predictors, but less severe criteria may prove suitable for many data sets, especially when the actual predictor distribution is longer tailed than Gaussian.

The residual weighting in equation (6) is based on the standard Huber approach parameterized to a 5% penalty for Gaussian data for the initial few iterations, followed by more severe outlier removal in the final iterations by using the weight function

$$v_i^{[k]} = \exp\{\exp(-\xi^2)\}\ \exp[-\exp\{\xi(|r_i^{[k-1]}|/d^{[k-1]} - \xi)\}] \tag{9}$$

where ξ defines the point at which downweighting begins. Chave *et al.* (1987) suggested using the Mth quantile of the target distribution for the residuals as an empirical choice for ξ to allow implicitly for increasing departure of some residuals from the population centre as the number of data increases. They also noted that, although a Gaussian residual model is appropriate for real data, the target distribution for complex data should be Rayleigh because the absolute value of the complex residual is the most appropriate residual statistic. The robust scale $d^{[k-1]}$ may be obtained for either a Gaussian or Rayleigh residual model in the standard way using the residuals from the $(k-1)$th iteration. For severely contaminated data, the median absolute deviation offers good performance in estimating $d^{[k-1]}$. The iterative solution of equation (6) ends when the weighted sum of squared residuals does not change beyond a threshold value.

The difference between the forms of the weights in equations (7) and (9) is required because of the well-known tendency for M-estimator residual-based weighting to increase leverage significantly. If the leverage weights in equation (7) are recomputed at each iteration, instability usually results due to interaction between the residual and leverage weights, especially in severely contaminated data. Because the weights in a given iteration multiply those from previous iterations in equation (7), this instability is eliminated with only a slight decrease in efficiency.

As a heuristic demonstration that the estimator proposed here is bounded, consider the simple case where the ith row of $\mathbf{X}$ contains one or more outliers but all the remaining rows are uncontaminated. As in Appendix A, define $\mathbf{X}_*$ to be $\mathbf{X}$ with the ith row removed and assume that $\mathbf{X}_*$ contains independent standardized complex random variables, so that $\mathbf{X}_*^{\mathrm{H}}\mathbf{X}_* \cong (N-1)\mathbf{I}$ where $\mathbf{I}$ is the $m \times m$ identity matrix. Further assume that $\mathbf{X}_*$ is such that all the weights in $\mathbf{U}$ except u_i are 1. Let the ith row of $\mathbf{X}$ be $\mathbf{x}_i = \alpha\tilde{\mathbf{x}}_i$ where $\tilde{\mathbf{x}}_i$ is uncontaminated so that $\|\tilde{\mathbf{x}}_i\|^2 = m$ and α is a measure of the degree of contamination. Then, the norm of the scaled and weighted ith row of $\mathbf{X}$ is $\|\mathbf{x}_i\|^2 = \alpha^2(u_i)^2 m$. Scaling p_{ii} by M/m, where $M = \mathrm{tr}(\mathbf{U}) = N - 1 + u_i$, yields

$$\frac{Mp_{ii}}{m} \cong \frac{\alpha^2(u_i)^2}{1 + \alpha^2(u_i)^2 m/(N-1)}. \tag{10}$$

Because u_i always lies between 0 and 1, for small values of α, $Mp_{ii}/m \approx \alpha^2(u_i)^2$, whereas for very large values $Mp_{ii}/m \approx (N-1)/m$. Since typical values of χ range from 2 to 4, the leverage weight in equation (7) decreases at an exponential rate with α, and hence the influence is bounded. Further, the weights in equation (7) are the product of the present value of the exponential term and the previous weight, so once a point has been downweighted because of excess leverage it stays that way for subsequent iterations.

Diagnostic plotting in regression has been extensively described in Belsley *et al.* (1980) and Chatterjee and Hadi (1988), and an extension of some standard approaches to robust regression is treated in McKean *et al.* (1993), who focused primarily on inferences about model order from residual plots. In most geophysical situations, the form of the model is prescribed by the relevant physics, and hence greater attention should be paid to the estimators' statistical performance, and especially to whether outliers and leverage points have been adequately removed through the iterative solution of equation (6). For this, the most useful diagnostic plots are quantile–quantile (q–q-) plots of both the weighted regression residuals against the appropriate target distribution quantiles (truncated Gaussian for real and truncated Rayleigh for complex data) and of the weighted hat matrix diagonal (8) against the truncated beta distribution quantiles. The residual q–q-plot should be reasonably straight and free of extreme values when using the final weights from equation (6). The hat matrix diagonal q–q-plot should be consistent with the conditions defined by Shaffer (1991); note that these do not require that the hat matrix diagonal be beta distributed unless the predictors are actually Gaussian. However, even when the predictor distribution is markedly non-Gaussian, a plot of the hat matrix diagonal against the beta distribution quantiles (or, in some cases, log-beta quantiles) is effective in detecting extreme outliers, as shown in the next section. Note also that it is essential to use the quantiles from the truncated form of the original residual or hat matrix distribution, or else the result will inevitably appear short tailed; see Appendix C for details.

5. Examples

In this section, the performance of the BI estimator described in Section 4 will be illustrated by using two physical data sets. The first of these is the star data set from Rousseeuw and Leroy (1987), Table 3. Although not complex, this data set is intended to be illustrative and has become a bench-mark for testing robust regression methods. The second data set consists of several components from the frequency domain analysis of a very large (more than 700 000 points each) group of time series of the electric and magnetic field variations from western Canada reported by Jones (1993). These data show especially severe leverage and outlier contamination.

The star data consist of 47 measurements of the logarithms of the effective surface temperature and light intensity of stars from the cluster Cygnus OB1. A plot of the data shows a direct relationship between the two variables except for four red giant stars (points 11, 20, 30 and 34) which are outlying with low temperatures and a high output of light (Fig. 2); these represent a different population rather than bad data. An ordinary least squares (OLS) estimator (equation (2)) including an intercept applied to the data (the broken line in Fig. 2; $\hat{\beta} = [-0.4133, 6.7935]$) is badly pulled by the four red giant stars, crossing the bulk of the population obliquely and fitting nothing very well, as was previously shown by Rousseeuw and Leroy (1987).

Fig. 3 shows a q–q-plot of the diagonal elements of equation (4) from the OLS solution. The result is long tailed, and the five most extreme values are, in decreasing order, data points 30, 34, 20, 11 and 7. A q–q-plot of the regression residuals (not shown) is somewhat short tailed and does not suggest the presence of serious outliers. The first four extreme values in Fig. 3 correspond to the red giant stars in Fig. 2, whereas datum 7 has an intermediate temperature between this group and the main population. Fig. 3 graphically illustrates the ability of a hat matrix q–q-plot to identify multiple leverage points, although the well-known masking and swamping phenomena do occur and hence there is no guarantee that all leverage points can be identified at one time with this diagnostic. However, an iterative approach, in which the most extreme leverage values are removed, the least squares line and hat matrix are recomputed

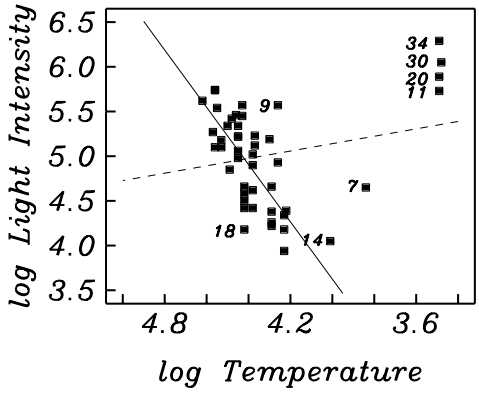

Fig. 2. Star data taken from Rousseeuw and LeRoy (1987) with selected data points highlighted, together with an OLS fit to the data (-------) and a BI fit (———) by using the estimator proposed in this paper

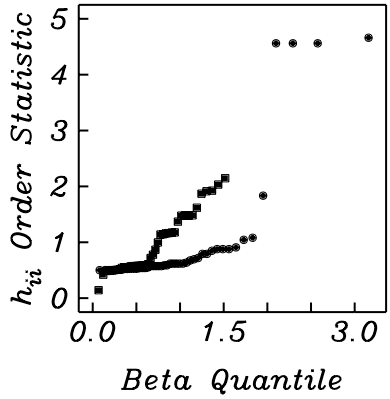

Fig. 3. Beta quantiles (scaled by $M/2$) plotted against the ranked hat matrix diagonal (also scaled by $M/2$) for the star data of Fig. 2 (see the text for discussion): ●, OLS hat matrix for which $M = N$ and the target distribution is $\beta(2, N-2)$; ■, BI hat matrix for which M is the number of data points after censoring and the quantiles are those of the truncated $\beta(2, N-2)$ distribution

and a new q–q-plot is generated, has proven very effective in the face of multiple leverage points.

The BI estimator was applied to the star data set with the leverage point rejection threshold set to the percentile of the beta distribution with parameters (2,45) at the 0.99-level. The Huber stage converged to within a 1% change in the sum of squared residuals after five iterations, and one additional iteration using equation (9) did not change the fit significantly, reflecting the weakness of outlying response data. The final BI estimate is $\hat{\beta} = (3.2038, -9.1905)$ and is shown by the full line in Fig. 2. The 95% confidence intervals on this estimate easily intersect the least median of squares result given in Rousseeuw and Leroy (1987). The leverage weights have eliminated points 7, 11, 20, 30 and 34 (see Fig. 2) and downweighted point 14 ($w_{14}^{[6]} = 0.14$). The final hat matrix q–q-plot is shown in Fig. 3 and is reasonably free of extreme values.

The second data set consists of a 732160-point magnetotelluric time series of the vector horizontal electric and magnetic field variations recorded at a sampling rate of 12 Hz at site 006 in central British Columbia, as described by Jones (1993). The magnetotelluric statistic of interest is the second rank tensor **Z** relating the horizontal electric and magnetic fields (**E** and **B**) as a

function of frequency given by

$$\mathbf{E} = \mathbf{ZB}. \tag{11}$$

At a given frequency, this is equivalent to estimating the rows of $\mathbf{Z}$ by solving two independent regressions relating either the north or east electric field response variable simultaneously to the north and east magnetic field predictor variables.

The data series were prewhitened by using a short robust autoregressive filter, subdivided into sections whose length is of the order of the inverse of the frequency of interest, windowed by using a Slepian sequence data window of unity time bandwidth, and Fourier transformed by using an overlapped section averaging approach (Percival and Walden (1993), section 6.17). After correction for prewhitening, the frequencies of interest are obtained from each section and become complex data to which the BI estimator is applied. Because noise in the magnetic field data produces downward bias in the tensor elements, a geophysical adaptation of the method of instrumental variables called the remote reference method (Gamble *et al.*, 1979) was used effectively to replace all autocovariance terms in equation (2) or (6) involving the local magnetic field with cross-covariance terms involving a reference magnetic field. Finally, the standard error on the response tensor elements was estimated by using the jackknife, as described by Thomson and Chave (1991).

Fig. 4 compares the magnitude and phase of the response tensor element between the north electric and east magnetic field (Z_{xy}), which is expected to be the dominant component involving this electric field element according to magnetotelluric theory. This quantity has been computed by using an M-estimator (the estimator of Section 4 with the leverage weights always fixed to 1, hereafter called the ordinary robust (OR) estimator) and the BI estimator. In conformity with geophysical practice, the magnitude has been expressed as apparent resistivity in ohm-metres, which is 2×10^{-4} times the period in seconds times the absolute square of a given response tensor element when the electric and magnetic fields are in Système International units; this corresponds to the true subsurface resistivity if it is depth independent. The BI result has been computed with the cut-off parameter in equation (7) taken as the 0.99999-point of the appropriate beta distribution; this choice reflects the inherent very long-tailed nature of the predictor distribution and, as will be demonstrated, is not especially critical. In general, the OR and BI results are similar, although there are subtle but significant differences at short (below 1 s) periods and substantial differences between 2 and 20 s. The latter are marked by substantial heteroscedasticity of the OR result due to leverage which is reflected in a much larger confidence limit estimate.

Fig. 5 is a complex plane view of the response tensor element at a period of 5.3 s. The OLS estimate has 8694 degrees of freedom. Both the OLS and the OR estimates display large uncertainties, reflecting heteroscedasticity that is not removed by using weights based entirely on the size of the residuals despite the elimination of about 7.2% of the data, or about 313 values. This heteroscedasticity is reflected in both the residual and the hat matrix diagonal q–q-plots from the OLS solution (Fig. 6) which suggest the presence of severe multiple outliers and extreme leverage. The most serious outliers are about 50 standard deviations from the Rayleigh mean, and the most serious leverage points are over 1000 times the expected value of the hat matrix diagonal for Gaussian predictors. Note that the hat matrix distribution is approximately a long-tailed version of a log-beta rather than beta distribution. Ionospheric and magnetospheric processes are highly non-linear, and hence their electromagnetic effects are the result of many multiplicative steps. This means that the statistical distribution of the magnetic variations will tend towards log-normal rather than normal (Lanzerotti *et al.*, 1991), and hence the resulting hat matrix diagonal will tend towards log-beta rather than beta. The q–q-plot for the hat matrix

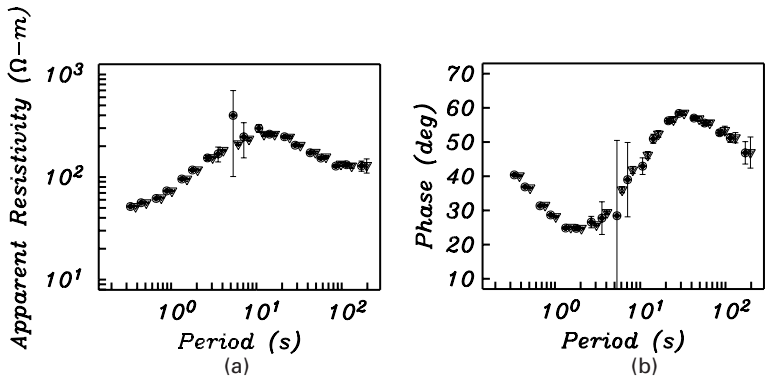

Fig. 4. (a) Apparent resistivity and (b) phase as a function of period in seconds for the magnetotelluric response tensor element between the north electric and east magnetic field (Z_{xy}) for the BC87 data described in the text; ⌶, double-sided 95% confidence limits computed by using the jackknife; ●, *M*-estimator; ▼, BI results described in the text (the BI estimates have been offset to slightly longer periods for clarity of presentation)

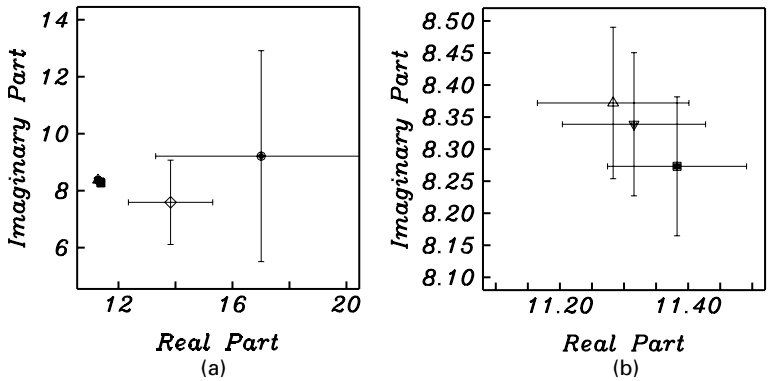

Fig. 5. (a) Complex plane view of the estimate in Fig. 4 at a period of 5.3 s (each symbol is plotted with the jackknife standard error) for the $\beta(2, N-2)$ distribution, where N is the number of estimates, and (b) a magnified view with different x- and y-axis limits showing the three BI estimates at the left-hand side of (a): ◇, OLS estimate ($N = 4342$); ●, OR estimate; ■, BI estimate with cut-off parameter χ at the 0.999 99-percentile; ▼, BI estimate with cut-off parameter χ at the 0.9999-percentile; △, BI estimate with cut-off parameter χ at the 0.999-percentile

diagonal is virtually unchanged for the OR solution, although the residual q–q-plot is slightly long tailed, reflecting pervasive though weak residual heteroscedasticity. The BI estimates for different values of the cut-off parameter in equation (7) are indistinguishable in Fig. 5; these are based on the 0.999 99-, 0.9999- and 0.999-points on the beta distribution, resulting in the rejection of 14 % (608 values), 15.7% (682 values) and 18.7% (813 values) of the data respectively. The three BI estimates are statistically identical despite the wide range in data elimination, and hence the choice of the cut-off point in equation (7) is not critical. A value of 0.999 99 was selected for production because a smaller value results only in reduced efficiency with no improvement in performance. The standard error on the 0.999 99 BI result is reduced by more than a factor of 34 compared with that from the OR estimator. In fact, the strong bias in the regression solution and its error estimate is caused by approximately 300 data points, amounting to about 7% of the total. The final residual q–q-plot for the BI estimate is slightly long tailed and upwardly concave

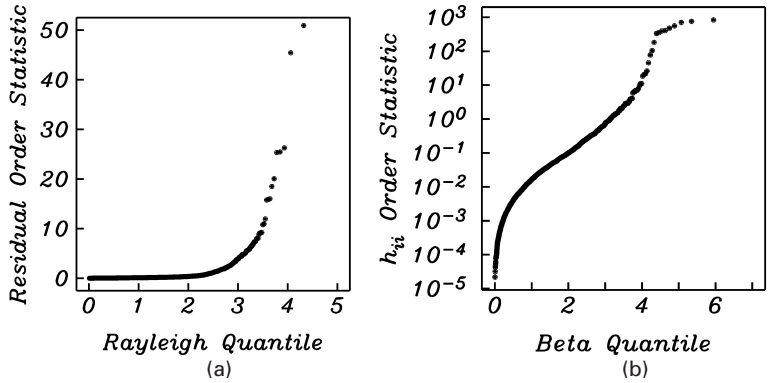

Fig. 6. q–q-plots for (a) the regression residuals and (b) the hat matrix diagonal for the OLS solution of Fig. 5: the absolute value of the regression residuals are scaled to have a variance of 2 and plotted against the Rayleigh quantiles in (a), whereas the hat matrix diagonal elements are scaled by $N/2$, converted to logarithms and plotted against $\beta(2, N-2)$ quantiles, also scaled by $N/2$, in (b); note the extremely long-tailed form of both quantiles

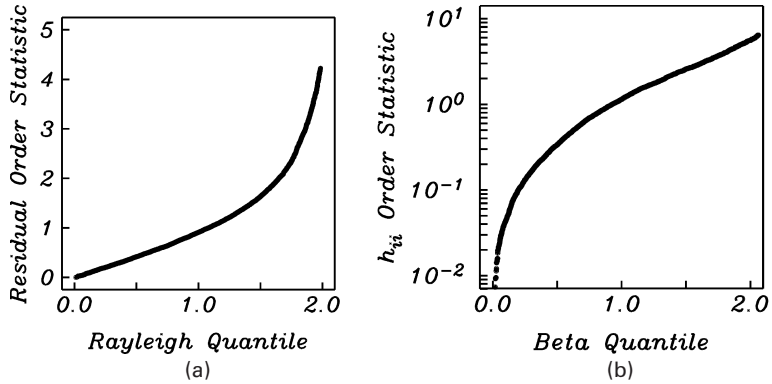

Fig. 7. q–q-plots for (a) the regression residuals and (b) the hat matrix diagonal for the 0.999 99 BI solution in Fig. 5 (see the caption to Fig. 6 for the plotting details): the target distributions are the truncated forms of the Rayleigh and β-distributions; the residual distribution is slightly longer tailed than the Rayleigh distribution, whereas the hat matrix diagonal is approximately log-β with some extreme values at the lower end

compared with the Rayleigh distribution (Fig. 7), and the upward concavity of the result is indicative of a residual distribution which is slightly longer tailed than the Rayleigh rather than the presence of significant outliers. The final hat matrix q–q-plot is approximately log-beta with a larger fraction of lower end values than would be expected. It is easy to modify the estimator of Section 4 to eliminate unusual data at both the lower and the upper ends of the distribution, but this has no significant effect on the result.

Fig. 8 shows the solutions for the same data components at a period of 0.89 s, for which the OLS solution has 69 577 degrees of freedom. In this instance, the OR solution displays a substantially smaller standard error than the OLS value but is also markedly offset compared with the BI result. The OR solution differs from the BI result by more than 6 standard errors of the latter. A comparison of the OLS and OR q–q-plots for the hat matrix diagonal (Fig. 9) shows that robust weighting has not improved things, and in fact the most extreme OR values are increased. The BI method controls the leverage and eliminates this problem. Residual q–q-plots for these data are qualitatively like those in Figs 6 and 7, and are not shown.

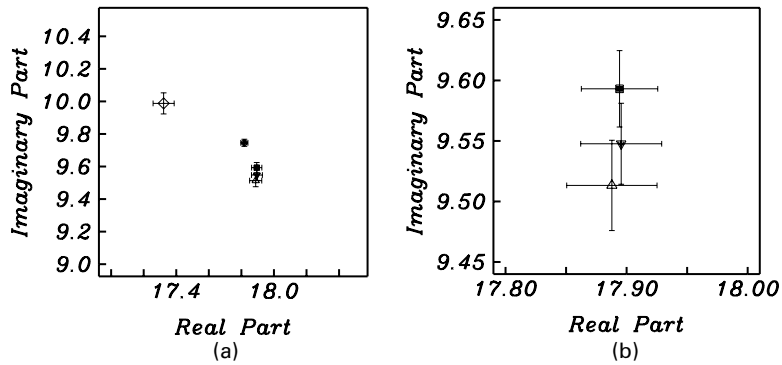

Fig. 8. (a) Complex plane view of the estimate in Fig. 4 at a period of 0.89 s (each symbol is plotted with the jackknife standard error) for the $\beta(2, N-2)$ distribution, where N is the number of estimates, and (b) a magnified view with different x- and y-axis scales showing the three BI estimates at the bottom of (a): ◇, OLS estimate (N = 34 788); ●, OR estimate; ■, BI estimate with cut-off parameter χ at the 0.999 99-percentile; ▼, BI estimate with cut-off parameter χ at the 0.9999-percentile; △, BI estimate with cut-off parameter χ at the 0.999-percentile

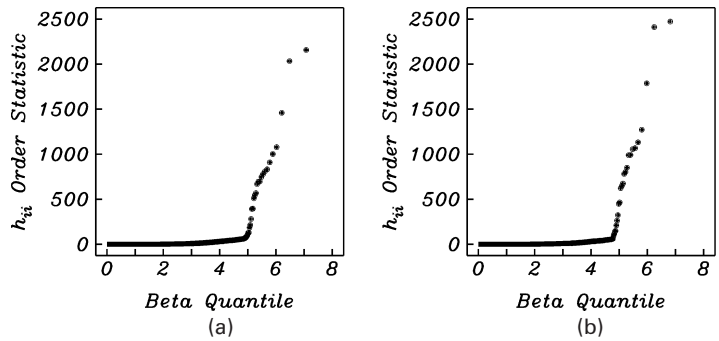

Fig. 9. q–q-plots for the hat matrix diagonal elements for (a) the OLS and (b) the OR estimates of Fig. 5: note that robust weighting in (b) has increased the size of the leverage points

The sort of behaviour that is seen in Figs 4, 5 and 8 is not unusual with magnetotelluric data, especially when they come from a high geomagnetic latitude where auroral effects can be severe or when cultural noise is a problem. Further examples showing the importance of robustness and leverage control in magnetotellurics may be found in Jones *et al.* (1989), Garcia *et al.* (1997) and Chave and Thomson (2003). In fact, part of the motivation for developing this algorithm was that earlier approaches performed unsatisfactorily. One of the advantages of working in the frequency domain is that regression estimation is done independently at many frequencies. For stationary processes, the data at different frequencies are strictly uncorrelated. Empirically, this still holds approximately in the observed mixtures; the magnitudes of the data at different frequencies are partially correlated, but the phases remain approximately independent. Consequently, if regression estimates at some frequencies follow the outliers rather than the data, the result can be physically impossible changes in slope for the apparent resistivity and phase curves.

The authors have not seen any magnetotelluric situation where BI estimation using the approach in this paper causes problems, and the computational overhead is only slightly larger than for ordinary M-estimation. With magnetotelluric data, an empirical breakdown point for the

method approaching 0.5 is routinely achieved. However, as with other high breakdown estimators, when the fraction of unusual data exceeds 0.5, the ensuing estimate can reflect the extreme data rather than the underlying, presumably good, data; this is illustrated in a magnetotelluric context for some extremely energetic auroral events by Garcia *et al.* (1997).

Acknowledgements

The Woods Hole Oceanographic Institution component of this work was supported by Department of Energy contract DE-FG02-94ER114435. This is Woods Hole Oceanographic Institution contribution 9932.

Appendix A: Derivation of the hat matrix diagonal distribution

An approximate distribution for the diagonal elements of the hat matrix for real Gaussian predictor data is given in Chatterjee and Hadi (1988), section 2.3.7. The derivation for complex Gaussian data follows similar reasoning but yields a simpler and exact result. A major complication with real data is the ubiquitous column of 1s in most predictor matrices; this has no equivalent in the complex data matrices that are used here. The diagonal elements of equation (4) may be written

$$p_{ii} = \mathbf{x}_i(\mathbf{X}^{\mathrm{H}}\mathbf{X})^{-1}\mathbf{x}_i^{\mathrm{H}} \tag{12}$$

where $\mathbf{x}_i$ denotes the ith row of $\mathbf{X}$. Because $\mathbf{X}$ contains $\mathbf{x}_i$, a direct statistical characterization of equation (12) is difficult. Chatterjee and Hadi (1988) suggested simplifying it by replacing $\mathbf{X}$ by the $(N-1)\times m$ matrix $\mathbf{X}_*$ which has the ith row deleted. Under this assumption, $\mathbf{X}_*^{\mathrm{H}}\mathbf{X}_*$ will have a complex Wishart distribution with $2(N-1)$ degrees of freedom (Srivastava, 1965). Further, because $\mathbf{x}_i$ is independent of $\mathbf{X}_*^{\mathrm{H}}\mathbf{X}_*$, the Mahalanobis distance from $\mathbf{x}_i$ to the remainder of the sample, as described by $\mathbf{X}_*$, is given by

$$p_{ii}^* = \mathbf{x}_i(\mathbf{X}_*^{\mathrm{H}}\mathbf{X}_*)^{-1}\mathbf{x}_i^{\mathrm{H}} \tag{13}$$

where the notation p_{ii}^* denotes the corresponding hat matrix diagonal elements. These will have a complex Hotelling T^2-distribution. Expressing the inverse of $\mathbf{X}^{\mathrm{H}}\mathbf{X}$ by using the Sherman–Morrison formula to update that of $\mathbf{X}_*^{\mathrm{H}}\mathbf{X}_*$ gives

$$p_{ii} = \frac{p_{ii}^*}{1+p_{ii}^*} \tag{14}$$

and hence the distribution of equation (12) follows from that of equation (13). The complex Hotelling T^2-distribution is given by Giri (1965) and, together with an expression for the distribution of T^2 in terms of the central F-distribution (Muirhead (1982), page 98), it can be shown that the statistic

$$f = \frac{N-m}{m}\frac{p_{ii}}{1-p_{ii}} \tag{15}$$

will be distributed as $F_{2m,2(N-m)}$. Further simplification is possible through the usual transformation from an F- to a beta distribution (Johnson *et al.* (1995), page 327) by letting $n = N-m$ and $n/(n+mf) = 1-p_{ii}$. From this, it is easy to show that the hat matrix diagonal elements have a beta distribution with parameters m and $N-m$, which has probability density function

$$\rho(p_{ii}) = \frac{1}{B(m,N-m)}p_{ii}^{m-1}(1-p_{ii})^{N-m-1} \tag{16}$$

where $B(a,b)$ is the beta function. The cumulative distribution function is the incomplete beta function ratio $I_x(m,N-m)$ obtained by integration of equation (16) and given in closed form in Appendix B.

Appendix B: Closed form expression for the beta cumulative distribution function

A closed form series expression for the incomplete beta function ratio $I_x(a,b)$ may be derived directly from the integral definition when a and b are integers. The definition is

$$I_x(a,b) = \frac{1}{B(a,b)} \int_0^x t^{a-1}(1-t)^{b-1}\,\mathrm{d}t. \tag{17}$$

Integrating by parts a times and substituting $b = N - a$ yields the series expansion

$$I_x(a, N-a) = 1 - \frac{1}{B(a, N-a)} \sum_{i=1}^{a} \frac{A_i}{B_i} x^{a-i}(1-x)^{N-a+i-1} \tag{18}$$

where

$$A_i = \prod_{j=1}^{i-1} (a-j),$$

$$B_i = \prod_{j=0}^{i-1} (N-a+j)$$

and $A_1 = B_1 = 0$. The beta function is given by

$$B(a, N-a) = (a-1)! \Big/ \prod_{i=1}^{a} (N-i). \tag{19}$$

The quantiles may be obtained numerically by solving $I_x(a, N-a) = (j - \frac{1}{2})/N$ for x, and other statistical quantities of interest follow directly.

Appendix C: Quantiles of a truncated distribution

Suppose that data are censored in the process of robust or BI weighting by using an estimator such as that described in Section 4. The target distribution for the relevant statistics must then be truncated to reflect data censoring. Let $f_X(x)$ be the probability density function of a random variable X before censoring; in the present context, this may be the Gaussian or Rayleigh distribution for the residuals or the beta distribution for the hat matrix diagonal. After truncation, the probability density function of the censored random variable X' is

$$f_{X'}(x') = \frac{f_X(x')}{F_X(d) - F_X(c)} \tag{20}$$

where $c \leqslant x' \leqslant d$ and $F_X(x)$ is the cumulative distribution function for $f_X(x)$. Let N be the original and M be the final number of data, so that $M = N - m_1 - m_2$, where m_1 and m_2 are the numbers of data censored from below and above respectively. Suitable choices for c and d are the m_1th and $(N - m_2)$th quantiles of the original distribution $f_X(x)$. The M-quantiles of the truncated distribution can then be computed from that of the original distribution by using

$$\int_{-\infty}^{Q_j} f_X(x)\,\mathrm{d}x = \{F_X(d) - F_X(c)\}\frac{j - \frac{1}{2}}{M} + F_X(c). \tag{21}$$

References

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

Calman, J. (1978) On the interpretation of ocean current spectra. *J. Phys. Oceanogr.*, **8**, 627–652.

Carroll, R. J. and Welsh, A. H. (1988) A note on asymmetry and robustness in linear regression. *Am. Statistn*, **42**, 285–287.

Chatterjee, S. and Hadi, A. S. (1988) *Sensitivity Analysis in Linear Regression*. New York: Wiley.

Chave, A. D. and Thomson, D. J. (1989) Some comments on magnetotelluric response function estimation. *J. Geophys. Res.*, **94**, 14215–14225.

Chave, A. D. and Thomson, D. J. (2003) Bounded influence magnetotelluric response function estimation. *Geophys. J. Int.*, to be published.

Chave, A. D., Thomson, D. J. and Ander, M. E. (1987) On the robust estimation of power spectra, coherences, and transfer functions. *J. Geophys. Res.*, **92**, 633–648.

Coakley, C. W. and Hettmansperger, T. P. (1993) A bounded influence, high breakdown, efficient regression estimator. *J. Am. Statist. Ass.*, **88**, 872–880.
Egbert, G. D. (1997) Robust multiple-station magnetotelluric data processing. *Geophys. J. Int.*, **130**, 475–496.
Egbert, G. D., Eisel, M., Boyd, O. S. and Morrison, H. F. (2000) DC trains and PC3s: source effects in mid-latitude geomagnetic transfer functions. *Geophys. Res. Lett.*, **124**, 25–28.
Gamble, T. D., Goubau, W. M. and Clarke, J. (1979) Magnetotellurics with a remote magnetic reference. *Geophysics*, **44**, 53–68.
Garcia, X., Chave, A. D. and Jones, A. G. (1997) Robust processing of magnetotelluric data from the auroral zone. *J. Geomagn. Geoelectr.*, **49**, 1451–1468.
Giri, N. (1965) On complex analogues of T^2 and R^2 tests. *Ann. Math. Statist.*, **36**, 664–670.
Hampel, F. R. (1974) The influence curve and its role in robust estimation. *J. Am. Statist. Ass.*, **69**, 383–393.
Handschin, E., Schweppe, F. C., Kohlas, J. and Fiechter, A. (1975) Bad data analysis for power system state analysis. *IEEE Trans. Pwr Appar. Syst.*, **94**, 329–337.
Hawkins, D. M. and Olive, D. J. (2002) Inconsistency of resampling algorithms for high-breakdown regression estimators and a new algorithm (with discussion). *J. Am. Statist. Ass.*, **97**, 136–159.
Hoaglin, D. C. and Welsch, R. E. (1978) The hat matrix in regression and ANOVA. *Am. Statistn*, **32**, 17–22.
Johnson, N. L., Kotz, S. and Balakrishnan, N. (1995) *Continuous Univariate Distributions*, vol. 2, New York: Wiley.
Jones, A. G. (1993) The BC87 dataset: tectonic setting, previous EM results, and recorded MT data. *J. Geomagn. Geoelectr.*, **45**, 1089–1105.
Jones, A. G., Chave, A. D., Egbert, G. D., Auld, D. and Bahr, K. (1989) A comparison of techniques for magnetotelluric response function estimation. *J. Geophys. Res.*, **94**, 14201–14214.
Kleiner, B., Martin, R. D. and Thomson, D. J. (1979) Robust estimation of power spectra (with discussion). *J. R. Statist. Soc.* B, **41**, 313–351.
Krasker, W. S. and Welsch, R. E. (1982) Efficient bounded-influence regression estimation. *J. Am. Statist. Ass.*, **77**, 595–604.
Lanzerotti, L. J., Gold, R. E., Thomson, D. J., Decker, R. E., Maclennan, C. G. and Krimigis, S. M. (1991) Statistical properties of shock-accelerated ions in the outer heliosphere. *Astrophys. J.*, **380**, L93–L96.
Lanzerotti, L. J., Thomson, D. J. and Maclennan, C. G. (1999) Engineering issues in space weather. In *Modern Radio Science 1999* (ed. M. A. Stuchly), pp. 35–50. Oxford: Oxford University Press.
Mallows, C. L. (1967) Linear processes are nearly Gaussian. *J. Appl. Probab.*, **4**, 313–329.
Mallows, C. L. (1975) On some topics in robustness. *Technical Memorandum*. Bell Telephone Laboratories, Murray Hill.
McKean, J. W., Sheather, S. J. and Hettmansperger, T. P. (1993) The use and interpretation of residuals based on robust estimation. *J. Am. Statist. Ass.*, **88**, 1254–1263.
Muirhead, R. T. (1982) *Aspects of Multivariate Theory*. New York: Wiley.
Percival, D. and Walden, A. (1993) *Spectral Analysis for Physical Applications*. Cambridge: Cambridge University Press.
Rao, C. R. (1973) *Linear Statistical Inference and Its Applications*. New York: Wiley.
Rousseeuw, P. J. (1984) Least median of squares regression. *J. Am. Statist. Ass.*, **79**, 871–880.
Rousseeuw, P. J. and Leroy, A. M. (1987) *Robust Regression and Outlier Detection*. New York: Wiley.
Ryan, T. P. (1997) *Modern Regression Methods*. New York: Wiley.
Shaffer, J. P. (1991) The Gauss-Markov theorem and random regressors. *Am. Statistn*, **45**, 269–273.
Srivastava, M. S. (1965) On the complex Wishart distribution. *Ann. Math. Statist.*, **36**, 313–315.
Thomson, D. J. (1977) Spectrum estimation techniques for characterization and development of WT4 waveguide, I. *Bell Syst. Tech. J.*, **56**, 1769–1815.
Thomson, D. J. and Chave, A. D. (1991) Jackknifed error estimates for spectra, coherences, and transfer functions. In *Advances in Spectrum Analysis and Array Processing*, vol. 1 (ed. S. Haykin), ch. 2, pp. 58–113. Englewood Cliffs: Prentice Hall.
Vozoff, K. (1972) The magnetotelluric method in the exploration of sedimentary basins. *Geophysics*, **37**, 98–141.
Wilcox, R. R. (1997) *Introduction to Robust Estimation and Hypothesis Testing*. New York: Academic Press.