On the Robust Estimation of Power Spectra, Coherences, and Transfer Functions

ALAN D. CHAVE[1]

Earth and Space Sciences Division, Los Alamos National Laboratory, Los Alamos, New Mexico

DAVID J. THOMSON

AT&T Bell Laboratories, Murray Hill, New Jersey

MARK E. ANDER

Earth and Space Sciences Division, Los Alamos National Laboratory, Los Alamos, New Mexico

Robust estimation of power spectra, coherences, and transfer functions is investigated in the context of geophysical data processing. The methods described are frequency-domain extensions of current techniques from the statistical literature and are applicable in cases where section-averaging methods would be used with data that are contaminated by local nonstationarity or isolated outliers. The paper begins with a review of robust estimation theory, emphasizing statistical principles and the maximum likelihood or M-estimators. These are combined with section-averaging spectral techniques to obtain robust estimates of power spectra, coherences, and transfer functions in an automatic, data-adaptive fashion. Because robust methods implicitly identify abnormal data, methods for monitoring the statistical behavior of the estimation process using quantile-quantile plots are also discussed. The results are illustrated using a variety of examples from electromagnetic geophysics.

INTRODUCTION

Reliable estimation of power spectra for single data sequences or of transfer functions and coherences between multiple time series is of central importance in many areas of geophysics and engineering. While the effects of the underlying Gaussian distributional assumptions on such estimates are generally understood, the ability of a small fraction of non-Gaussian noise or localized nonstationarity to affect them is not. These phenomena can destroy conventional estimates, often in a manner that is difficult to detect.

Problems with conventional (i.e., nonrobust) time series procedures arise because they are essentially copies of classical statistical procedures parameterized by frequency. Once Fourier transforms are taken, estimating a spectrum is the same process as computing a variance, and estimating a transfer function is a similar procedure to linear regression. Because these methods are based on the least squares or Gaussian maximum likelihood approaches to statistical inference, their advantages include simplicity and the optimality properties established by the Gauss-Markov theorem [e.g., *Kendall and Stuart*, 1977, chapter 19]. For example, linear regression yields the best linear unbiased estimate when the errors are uncorrelated and share a common variance; this holds independent of any distributional assumptions about them. If, in addition, the residuals are drawn from a multivariate normal probability distribution, then the least squares result is also a maximum likelihood, fully efficient, minimum variance estimate. In practice, the regression model is rarely an accurate description due to departures of the data from the model requirements. Most data contain a small fraction of unusual observations or "outliers" that do not fit the model distribution or share the characteristics of the bulk of the sample. These can often be described by a probability distribution which has a nearly Gaussian shape in the center and tails which are heavier than would be expected for a normal one, or by mixtures of Gaussian distributions with different variances.

Two forms of data outliers are common: point defects and local nonstationarity. Point defects are isolated outliers that exist independent of the structure of the process under study. Typical examples include dropped bits in digital data, transient instrument failures, and spike noise due to natural phenomena (e.g., lightning). Local nonstationarity means a departure from a stationary base state that is of finite duration and must be differentiated from complete nonstationarity, in which the concept of a spectrum must be reformulated [e.g., *Priestley*, 1965; *Martin and Flandrin*, 1985]. A geophysical example of local nonstationarity is seen in observations of the time-varying geomagnetic field: most of the time the data statistics are approximately constant, but this stationary process is interrupted sporadically by brief but intense disturbances such as magnetic storms with markedly different characteristics. In some studies these events are regarded as contaminating noise, and they must be removed to study the underlying process. The influence of these types of outliers on regression problems can be complicated, as aberrant data in the dependent and independent variables produce quite

[1] On leave from Institute of Geophysics and Planetary Physics, University of California, San Diego, La Jolla; now at AT&T Bell Laboratories, Murray Hill, New Jersey.

Paper number 5B5911
0148-0227/87/005B-5911$05.00

different changes in the output, and correlations between apparent data outliers in both variables can still yield reasonable regression parameters. In addition, new classes of outliers can occur in linear regression. In any case, it is a serious statistical error to blindly accept mixture situations of these types and analyze the combination as a unit. With such data, conventional least squares-based techniques will give inefficient and often seriously misleading estimates. This breakdown of the least squares model, while sometimes spectacular in form, is more typically insidious in that a seemingly reasonable answer is obtained, and considerable effort has gone into devising diagnostics to detect this sort of problem [e.g., *Belsley et al.*, 1980; *Cook and Weisberg*, 1982].

Because the Fourier transform of even moderately long-tailed data tends to be Gaussian as the length of the series increases (essentially by the central limit theorem, but see *Brillinger* [1981, chapter 5] for details), it is sometimes claimed that outliers in time series are not a serious problem. However, this is often not true, as shown by the power spectrum examples of *Thomson* [1977*b*], *Kleiner et al.*, [1979], and *Martin and Thomson* [1982]. Furthermore, coherences and transfer functions are substantially more sensitive to the presence of outliers, since multiple time series and ratios of spectra are involved. It is frequently argued that a careful analyst will examine a data set and use ad hoc remedies to avoid outlier difficulties. While this may work for obvious discordancies in small samples, it is impractical for large data sets or when, as often happens in time series, the outliers have a scale comparable to or smaller than that of the process under study. It is preferable to use statistical procedures that are robust, in the sense of being relatively insensitive to the presence of a moderate amount of bad data or to inadequacies in the model, and that react gradually rather than abruptly to perturbations of either. Such methods have been developed over the past two decades and are reviewed by *Huber* [1981], *Hoaglin et al.* [1983], and *Hampel et al.* [1986].

In this paper the principles of robust statistics are adapted to univariate and multivariate spectral analysis within a geophysical context. The treatment begins with a review of some critical statistical concepts, especially robustness in the estimation of location and scale. This is followed by the introduction of the maximum likelihood or M-estimators for computing robust averages and solving robust regression problems. After considering numerical implementation of the M-estimators, some diagnostic plotting methods to help elucidate the extent of outlier contamination or nonstationarity in data are discussed. These tools are then combined with the section-averaging method of spectral analysis and applied to the estimation of power spectra, transfer functions, and coherences. The results are illustrated with a variety of examples from natural source electromagnetic geophysics. The paper closes with a discussion of distributional aspects and some suggestions for further work.

STATISTICAL PARAMETERS AND ROBUSTNESS

Given a continuous probability density function (pdf) $f(x)$, the cumulative distribution function (cdf) is denoted by $F(x)$, where

$$F(x)=\int_{-\infty}^{x} du\, f(u) \tag{1}$$

To make the correspondence between theoretical and sample entities clearer, write $dF(x)$ for $dx\, f(x)$ and indicate the empirical or sample cdf by $d\hat{F}(x)$. For a set of N data samples $\{x_i\}$, this is given by

$$d\hat{F}(x)=\frac{1}{N}\sum_{i=1}^{N}\delta(x-x_i)\,dx \tag{2}$$

where $\delta(x)$ is the Dirac delta function. Substitution of (2) into (1) yields the usual empirical cdf in which each data point corresponds to a step in probability of $1/N$.

A set of N real samples $\{x_i\}$ may be sorted into the ascending order

$$x_{(1)}\leqslant x_{(2)}\leqslant \ldots \leqslant x_{(N)}$$

where $x_{(j)}$ is called the jth order statistic. Note that the probability distribution of the ordered data is different from that of the original data; clearly $x_{(i)}$ depends on $x_{(i-1)}$ and $x_{(i+1)}$ even when the $\{x_i\}$ are independent. Assuming these samples to be independent and characterized by the pdf $f(x)$, the corresponding theoretical entities are the set of N quantiles Q_j. These are found from the inverse cdf or quantile function $F^{-1}(\alpha)$, defined for $0\leqslant\alpha\leqslant 1$ as the solution of

$$\int_{-\infty}^{Q_\alpha} dx\, f(x)=\alpha \tag{3}$$

with $\alpha=(j-½)/N$ for $j=1,2,\ldots,N$. The Q_j divide the area under the distribution into $N+1$ probability intervals, with the first and last points assigned cumulative probabilities of $1/2N$ and $1-1/2N$, respectively, and with a probability step of $1/N$ occurring at each of the intermediate points.

In the following, it will usually be assumed that the data $\{x_i\}$ are independent, or at least uncorrelated. While this may appear strange in a time series setting, the $\{x_i\}$ should be thought of as Fourier transforms of a windowed section of data at some frequency, not as the original data in the time domain. Explicitly, consider N segments of data from a nominally stationary process $y(t)$, with each segment consisting of T discrete samples spaced Δ time units apart and offset by an amount b from the preceding one, and compute

$$x_k(\omega)=\sum_{t=0}^{T-1} v_t\, y(t\Delta+(k-1)b)\, e^{i\omega t\Delta} \qquad k=1,\ldots,N$$

where v_t is the data window. The quantity $x_k(\omega)$ is the windowed Fourier transform of the kth data segment at radian frequency ω. Even if closely spaced samples of the original process $y(t)$ are highly correlated, the $\{x_k\}$ will be uncorrelated at a given frequency for reasonable values of the base offset b. Similarly, information in a single data section at frequencies spaced at least the window bandwidth apart is uncorrelated.

One of the most useful procedures in statistics is the summary characterization of a distribution or sample using various types of averages. Of these, the most common one is location, specifying (in a loose sense) the center or peak value of a distribution or sample; examples include the mean and median. It is less commonly realized that such averages are the result of minimizing various norms. For example, the distribution mean μ and the sample mean, or average, $\bar{x}$ are obtained by minimizing the L_2 or least squares norms of the residuals about the distributions $[\int |x-\mu|^2 dF(x)]^{1/2}$ and $[\int |x-\bar{x}|^2 d\hat{F}(x)]^{1/2}$ with respect to μ and $\bar{x}$, respectively. Performing the minimization gives the familiar expressions

$$\mu = \int_{-\infty}^{\infty} x\, dF(x) = \int_{-\infty}^{\infty} dx\, x\, f(x) \tag{4}$$

$$\bar{x} = \int_{-\infty}^{\infty} x\, d\hat{F}(x) = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{5}$$

The sample median $\tilde{x}$ is obtained by minimizing the L_1 norm of the residuals $\int |x-\tilde{x}|\, d\hat{F}(x)$. This is the same as minimizing the summed absolute deviations $\sum_{i=1}^{N} |x_i - \tilde{x}|$ and reduces to the middle order statistic for N odd, $\tilde{x} = x_{(\lfloor N/2 \rfloor + 1)}$, where $\lfloor\ \rfloor$ denotes the integer part. The sample median is ambiguous for N even but is typically chosen as $(x_{(\lfloor N/2 \rfloor)} + x_{(\lfloor N/2 \rfloor + 1)})/2$. This is the simplest example of an order statistic. The population median is $\tilde{\mu} = Q_{1/2}$ from (3).

The L_1 and L_2 norms have the advantages of being well known and, for the L_2 norm in particular, of resulting in algebraically simple equations. However, there are few good reasons to consider them exclusively or to believe that they are the best choices except in restricted circumstances. Much of the work in robustness has effectively been on finding more appropriate norms for actual data, rather than applying what J. W. Tukey describes as "over-utopian" assumptions.

The second most common characterization of a distribution or sample is a measure of its width, dispersion, spread, variability, or standard deviation, which is included under the generic term scale. There are numerous available estimates of scale; *Gross* [1976] compares 25 variants of 12 distinct forms, and many others have been suggested. Among these are the minimum value of the norms achieved around the corresponding location estimates. This class includes the theoretical standard deviation

$$\sigma = \left[\int_{-\infty}^{\infty} |x-\mu|^2\, dF(x) \right]^{1/2} \tag{6}$$

and the sample standard deviation

$$s = \left[\int_{-\infty}^{\infty} |x-\bar{x}|^2\, d\hat{F}(x) \right]^{1/2} = \left[\frac{1}{N}\sum_{n=1}^{N} |x_n - \bar{x}|^2 \right]^{1/2} \tag{7}$$

for the L_2 norm, and the average absolute deviation

$$\tilde{\sigma} = \int_{-\infty}^{\infty} |x-\tilde{\mu}|\, dF(x) \tag{8}$$

or its sample counterpart

$$\tilde{s} = \int_{-\infty}^{\infty} |x-\tilde{x}|\, d\hat{F}(x) = \frac{1}{N}\sum_{n=1}^{N} |x_n - \tilde{x}| \tag{9}$$

for the L_1 norm.

A second class of scale estimate is obtained by reapplying the same estimator used for location to its absolute residuals. The median absolute deviation (MAD) is obtained by taking the median value of the absolute residuals about the L_1 location estimate

$$s_{\mathrm{MAD}} = \mathrm{median}\{\, |x_i - \tilde{x}|\,\} \tag{10}$$

The expected value of the MAD is the solution σ_{MAD} of

$$F(\tilde{\mu} + \sigma_{\mathrm{MAD}}) - F(\tilde{\mu} - \sigma_{\mathrm{MAD}}) = 1/2 \tag{11}$$

The interquartile distance is a related scale estimate given by

$$s_{\mathrm{IQ}} = x_{(3N/4)} - x_{(N/4)} \tag{12}$$

and is the spacing between the 75% and 25% points of the sample distribution, or the center range containing half of the probability. The corresponding theoretical value is

$$\sigma_{\mathrm{IQ}} = Q_{3/4} - Q_{1/4} \tag{13}$$

using (3), and is just twice the MAD for symmetric distributions. Both the average absolute deviation $\tilde{s}$ and the standard deviation s are sensitive to outliers, and $\tilde{s}$ is also inefficient. Further information on the MAD, interquartile distance, and other robust estimates of scale is available from *Mosteller and Tukey* [1977].

The simplest extension of these ideas to the general linear regression model is obtained by minimizing a norm of the residuals in

$$x_i = \sum_{j=1}^{p} u_{ij}\beta_j + r_i \qquad i = 1,\ldots,N \tag{14}$$

where $\{x_i\}$ is an N vector of data or observations, $\{u_{ij}\}$ is an $N \times p$ matrix of known coefficients, $\{\beta_j\}$ is a p vector of model parameters, and $\{r_i\}$ is an N vector of residuals. The solution of (14) depends on the norm that is chosen, and will not be the same even for the L_1 and L_2 cases, as discussed by *Claerbout and Muir* [1973].

It is well-known that least squares estimates are notoriously lacking in robustness, and both the sample mean and standard deviation may be strongly influenced by a single discordant data point. The resistance of the sample median and the scale estimates s_{IQ} and s_{MAD} to outliers is also well-established, and there are many applications where L_1 norm methods yield dramatic improvements over their L_2 norm counterparts. This has led to suggestions to substitute the minimum absolute deviation for the least squares method for many geophysical problems [*Claerbout and Muir*, 1973]. There are at least three reasons why this course of action is imprudent. First, the Gauss-Markov theorem [e.g., *Kendall and Stuart*, 1977, chapter 19] establishes the optimality of the L_2 estimator under general conditions on the structure of the regression residuals. No comparable result is known for the L_1 estimator, and it is not difficult to show, for example, that with Gaussian data, $\mathrm{var}(\tilde{x}) = \pi/2\, \mathrm{var}(\bar{x})$. This reduction in efficiency for the L_1 estimator means that about 60% more data is required to achieve equivalent parameter uncertainties to the L_2 estimator. Second, the natural probability distribution for L_1 estimates is the Laplace or double exponential type, whereas L_2 estimates involve the familiar normal distribution. This makes statistical inferences based on L_1 results more complex and less intuitively appealing. Finally, for most data the residuals from a least squares procedure are in large part Gaussian, with the addition of a small fraction of outliers having different statistics. This suggests that some method for treating the contamination within the framework of a Gaussian model, rather than outright abandonment of that model, is the logical course of action [e.g., *Tukey*, 1975; *Mallows*, 1983].

One obvious way to achieve robustness is by the rejection of outliers on the basis of some type of statistical test, followed by the use of least squares on the remaining and presumably good data. While a substantial amount of effort has gone into the development of outlier tests [*Barnett and Lewis*, 1978; *Hawkins*, 1980; *Barnett*, 1983; *Beckman and Cook*, 1983], the bulk of the results apply only to single, isolated outliers, usually in a parent normal population. The dual phenomena of masking and swamping, in which one bad value may hide the presence of others, has been widely documented for the more common multiple outlier situation [*Barnett*, 1983]. Outlier detection becomes even more complicated in time series because of correlations between the data [*Fox*, 1972; *Abraham and Box*, 1979]. As a consequence, methods are sought that accommodate outliers and minimize their influence in a semi-automatic fashion.

ROBUST ESTIMATION

In addition to heuristic methods, two major classes of robust procedures are the L-estimators, based on combinations of the order statistics and the M-estimators, a variant of maximum likelihood. L-estimates are especially useful for location problems with nonsymmetric distributions. In a time series context, *Thomson* [1977*a*] applied a maximum likelihood L-estimator proposed by *Mehrotra and Nanda* [1974] for censored, exponentially distributed populations to get robust, section-averaged power spectra of contaminated data. While simpler than an M-estimator, this approach suffers from a loss of efficiency because the truncation point must be fixed in advance, so that the result is not data adaptive. L-estimators do not generalize readily to regression problems and are not considered further in this paper.

To motivate the concept of an M-estimator, consider again the problem of determining the location parameter θ given independent samples $\{x_i\}$ drawn from a common pdf $f(x-\theta)$ by maximum likelihood. The logarithm of the likelihood function $L(\theta)$ is obtained in the usual way by inserting the data into the sampling pdf, yielding

$$\log L(\theta) = \sum_{i=1}^{N} \log f(r_i) \qquad (15)$$

where $r_i = x_i - \theta$ is the ith residual. The strict maximum likelihood solution $\hat{\theta}$ for θ comes from maximizing $L(\theta)$, or its logarithm, and obviously depends on knowing the pdf $f(x)$ exactly. A generalization of (15) based on a quantity $\rho(x)$, which is called a loss function, can be written

$$M(\theta) = \int_{-\infty}^{\infty} \rho(x-\theta)\, d\hat{F}(x) = \sum_{i=1}^{N} \rho(r_i) \qquad (16)$$

Clearly, if $\rho = -\log f$, minimizing M yields the ordinary maximum likelihood result (15). Furthermore, the estimate is identical to that obtained by the norm minimization methods discussed earlier. This equivalence is the basis for identifying L_1 and L_2 as the natural norms for the Laplace and Gaussian distributions, respectively. Performing the minimization of (16) gives the equation

$$\sum_{i=1}^{N} \psi(r_i) = 0 \qquad (17)$$

where $\psi(x) = \partial_x \rho(x)$ is called an influence function. Equations (16) and (17) reduce to the least squares or least absolute deviations forms when $\rho(x) = x^2/2 + c$, $\psi(x) = x$ or $\rho(x) = |x| + c$, $\psi(x) = \mathrm{sgn}(x)$, respectively, yielding the sample mean or the sample median as solutions. The formulation in (16) or (17) is equally valid for more complicated distributions, and the solution $\hat{\theta}$ is called an M-estimate.

If enough data are available, the sampling pdf or its logarithm can be estimated directly from it, resulting in a data adaptive, maximum likelihood estimate of location. In practice, it is rarely feasible to characterize the distribution tails with the available data, and the loss function is usually chosen on theoretical grounds to retain high efficiency over a family of expected distributions in preference to yielding the maximum likelihood estimate for a single distribution. Since most data yield largely Gaussian residuals, with a small outlier fraction, it is customary to use loss functions giving high efficiency (say, >95%) for normal data but which still provide reasonable protection for contaminated data. This slight loss in nominal efficiency is the inevitable penalty that must be paid to get a stable estimate in the general case. Since typical data contain 1–20% outliers, the increase in sample size needed to compensate for this rejection is small compared to the extra ≈60% needed for the L_1 estimator. Some commonly used loss functions are described by *Holland and Welsch* [1977] and *Hampel et al.* [1986].

There is an additional complication that must be considered with M-estimators: the solution $\hat{\theta}$ of (16) or (17) will not be scale invariant in the sense that multiplying the data $\{x_i\}$ by a constant will not necessarily result in a similar change in $\hat{\theta}$. To correct for this, it is necessary to replace (16) and (17) by

$$\min \sum_{i=1}^{N} \rho\left(\frac{r_i}{d}\right) \qquad (18)$$

and

$$\sum_{i=1}^{N} \psi\left(\frac{r_i}{d}\right) = 0 \qquad (19)$$

where d is a robust estimate of scale. While joint optimization with respect to both location and scale is described by *Huber* [1981, chapter 6] and *Hampel et al.* [1986, chapter 4], two practical but lower efficiency choices are

$$d_1 = \frac{s_{\mathrm{MAD}}}{\sigma_{\mathrm{MAD}}} \qquad (20)$$

and

$$d_2 = \frac{s_{\mathrm{IQ}}}{\sigma_{\mathrm{IQ}}} \tag{21}$$

where s_{MAD} and s_{IQ} are the sample MAD and interquartile distance (10) and (12), and σ_{MAD} and σ_{IQ} are their theoretical counterparts for the appropriate standard pdf from (11) and (13). Because of the extreme sensitivity of the L_2 estimate to outliers, use of the standard deviation (6) and (7) is not recommended.

The concept of an M-estimator may be extended to the general linear regression model (14) by identifying the error r_i as the regression residual [*Andrews*, 1974]. While (18) is not changed in form, (19) must be rewritten as

$$\sum_{i=1}^{N} \psi\left(\frac{r_i}{d}\right) x_{ij}^{*} = 0 \qquad j=1,\ldots,p \tag{22}$$

where the superscript asterisk denotes the complex conjugate.

A variety of numerical procedures to solve the generally nonlinear forms of (18), (19), and (22) exist, but it is easiest to rewrite the problem as a weighted least squares one and iterate to linearize it. This allows the use of fast, accurate, and stable matrix algorithms. The weighted least squares forms of (18) and (22) are, respectively,

$$\min \sum_{i=1}^{N} \tilde{w}_i^{\,2} r_i^{\,2} \tag{23}$$

where $\tilde{w}^2 = \rho(r/d)/r^2$ and

$$\sum_{i=1}^{N} w_i\, r_i\, x_{ij}^{*} = 0 \tag{24}$$

where $w = \psi(r/d)/r$. Note that this procedure gives two different weights: the $\tilde{w}_i$ from treating the loss function formulation as an equivalent weighted least squares problem, and w_i from its equivalence with the influence function. While distinct, they are not always clearly separated in the literature. In addition, w or ψ are often chosen a priori, implicitly defining $\tilde{w}$ and ρ. To linearize the weighted least squares problem, an initial solution is obtained using ordinary (unweighted) least squares, and both the residuals and a scale estimate like (20) or (21) are computed. The weights are calculated from these, and the solution to (23) or (24) is found. This procedure is repeated using the residuals and scale estimate from the previous iteration at each stage until convergence is achieved. Note that the weights are data adaptive; the robust formulation ensures that data corresponding to residuals which are large compared to the scale will be downweighted.

There are several forms of the weights in (23) that work well in spectral problems. The first of these was introduced by *Huber* [1964] on theoretical grounds and is based on a density function with a Gaussian center and Laplacian tails, yielding

$$\begin{aligned} \rho(x) &= \frac{x^2}{2} & |x| &\leqslant k \\ &= k|x| - \frac{k^2}{2} & |x| &> k \end{aligned} \tag{25}$$

and a weight function

$$\begin{aligned} \tilde{w}(x) &= 1 & |x| &\leqslant k \\ &= \sqrt{\frac{2k}{|x|} - \frac{k^2}{x^2}} & |x| &> k \end{aligned} \tag{26}$$

where a value of $k=1.5$ gives better than 95% efficiency for outlier-free normal data. The corresponding Huber influence function $\psi(x)=x$ for $|x|<k$ and $k\,\mathrm{sgn}(x)$ otherwise has weights which never descend to zero. Because it has a discontinuous derivative, the Huber function may introduce slight distortions in time series work if used exclusively. Since the weights (26) fall off slowly for large residuals, they provide inadequate protection against severe outliers. However, (25) is a convex function, so that convergence to a local, as opposed to a global, minimum cannot occur in the iterative weighted least squares solution. Use of the Huber weights gives a good starting point for the application of more severe types of weight functions.

Another class of influence functions is called redescending because the influence function and weights approach zero in the presence of large outliers. The biweight of *Mosteller and Tukey* [1977, chapter 10] is a typical redescending influence function. However, for time series work it has too much curvature near the origin and can introduce serious distortion [*Thomson*, 1977*b*]. To reduce this effect, Thomson [1977a] proposed a new weight function

$$w(x) = \exp\{-e^{\beta(|x|-\beta)}\} \tag{27}$$

from a heuristic extension of the extreme value distribution. The parameter β determines the scale at which downweighting begins. While strictly empirical, the Nth quantile of the appropriate probability distribution from (3) is an excellent choice for β. This form has the advantages of being smooth, close to unity near the origin, and including an implicit dependency on the number of data in the weighting procedure; as the number of samples rises, extreme values become more common, and β must increase to avoid affecting valid data. For example, $\beta = Q_N$ corresponding to $N=10^3$, 10^4, 10^5, and 10^6 for a normal distribution are 3.09, 3.72, 4.26, and 4.75 standard deviations, respectively. However, as with other redescending influence functions, the solution of (23) or (24) with (27) is not unique, so this weight should only be used after a good starting value has been found.

NUMERICAL CONSIDERATIONS

Equations (18) and (23) are generalizations of the usual linear regression statement (14), while (19), (22), and (24) are generalizations of the familiar normal equations. Since the numerical ill-conditioning of the normal equations is well-known, it is best to solve the matrix form (23) by a numerically stable method such as QR or singular value decomposition [e.g., *Lawson and Hanson*, 1974]. The QR decomposition method of solving (23) is used throughout this work. Since spectral analysis involves the Fourier transform, the matrices in (23) are complex. Complex least squares procedures are reviewed by *K. S. Miller* [1974] and are available in standard packages (e.g., CQRDC and CQRSL in LINPACK [*Dongarra et al.*, 1979]). Alternately, the problem may be broken down into real and imaginary parts and solved using standard real algorithms. To avoid problems with numerical round-off, it is recommended that 64-bit arithmetic be utilized.

Additional modes of failure in the solution of regression equations from collinearity of the coefficient matrix and similar effects are possible [*Belsley et al.*, 1980], and a method for checking for these based on the condition number of u_{ij} is given by *Lanzerotti et al.* [1986]. To illustrate, consider a geomagnetic example where x_i is a Fourier-transformed electric field from the ith data segment. As independent variables in (14), take the corresponding transforms of the $p=3$ magnetic field variables $u_{i1}=H_i$, $u_{i2}=D_i$, and $u_{i3}=Z_i$ and solve for the $\{\beta_j\}$, or impedances. If H, D, and Z are linearly independent, a solution will exist, and collinearity is not a problem. However, in the presence of complex geologic structure or source field inhomogeneity, Z may depend on H and D through the tipper functions T_H and T_D defined by

$$Z = T_H H + T_D D$$

in which case the coefficient matrix would be theoretically of rank 2 (or collinear), and the numerical solution of (14) would be unstable. This problem is flagged by large values of the condition number, and a better solution is obtained by taking $p=2$ and, as is the usual practice, analyzing the vertical magnetic field separately. Note also that the presence of outliers in data can induce collinearity into an otherwise stable system of equations [*Mason and Gunst*, 1985]. For methods to help decide when (and how) to truncate near-collinear matrix systems, see *Lawson and Hanson* [1974, chapter 26] or *Vogel* [1986].

The matrix elements in the normal equations in a time series setting may be identified as the auto-spectra and cross-spectra of the data and coefficient variables. This may lead to the temptation to estimate these quantities independently and robustly, followed by solution of the normal equations for the transfer functions, rather than treating the entire problem as a unit using a common set of weights as described in the last section. While such a procedure will result in good estimates of the individual matrix elements, the resulting matrix structure may be seriously in error because the normal equations encountered in spectral applications are often marginally conditioned even with ideal data, and small changes in the matrix elements induced by the individual robust procedures can easily result in nonphysical solutions. Such an approach is discouraged even more than use of the ordinary normal equations.

DIAGNOSTICS

It is useful to develop diagnostic procedures that disclose the extent of outlier contamination and give a visual indication of goodness of fit to a specified probability distribution. These find application both in the early stages of the analysis, when decisions about the suitability of particular statistical models must be made, and as the final step in robust estimation, when the efficacy of the method in reducing outlier influence must be assessed. While a myriad of techniques based on residual plotting have been proposed [e.g., *Belsley et al.*, 1980; *Cook and Weisberg*, 1982], a simple and effective method is based on the order statistics. The quantile-quantile (q-q) plot serves to simultaneously yield indications of goodness of fit, sketch the extent of outlier contamination, and provide the key location and scale parameters. An important property of the N order statistics of a data sample $\{x_i\}$ is that they divide the area under a pdf into $N+1$ parts of not necessarily equal size. The q-q plot is obtained by comparing the quantiles of a specified distribution, Q_j, which do divide the area under the pdf into equal-sized pieces, to the order statistics $x_{(j)}$. If the latter are drawn from the assumed distribution, they will be similar to the quantiles, and the q-q plot will be an approximately straight line. Systematic departures from a straight line show that the model is inconsistent and can be used to guide the search for a better one. Outliers usually appear as deviations from a line at the extreme quantiles. Finally, the slope and intercept of the line give the scale and location parameters for the experimental distribution. In the present work, q-q plots of the residuals from a robust procedure are examined as a function of frequency. However, instrumentation problems, such as stuck bits, often appear as plateaus or staircases in q-q plots of the raw data. For a lucid discussion of q-q plots and their use, see *Kleiner and Graedel* [1980] or *Lewis and Fisher* [1982].

Other regression diagnostics, such as plots of the residual against the input or output power, become unmanageable if applied directly at all frequencies but can be tried at specific frequencies of interest. Another worthwhile check requiring only one additional plot is investigation of the condition number of the coefficient matrix as a function of frequency. The condition number is just the ratio of the largest to smallest singular values at each frequency and is available trivially if (23) is solved with the SVD method. Estimates of the condition number are also available from QR solutions.

One important aspect of robust techniques is their ability to identify unusual data. Typically, anomalous effects are either localized in frequency or in time and may occasionally be localized simultaneously in both. A common type of outlier is local nonstationarity and may be detected by the stationarity test given by *Thomson* [1977*a,b*]. This is Bartlett's M-test (no relation to M-estimates) for variance homogeneity [*Bartlett*, 1937] computed as a function of frequency

$$M(\omega) = N\nu \log\left[\frac{1}{N}\sum_{j=1}^{N}\hat{S}_j(\omega)\right] - \nu\sum_{j=1}^{N}\log\hat{S}_j(\omega) \qquad (28)$$

where $\hat{S}_j(\omega)$ is the spectral estimate for the jth data section at radian frequency ω and there are N individual estimates, each having ν degrees of freedom, typically, 2 for raw spectra. If the series is stationary, M will be distributed approximately as χ^2_{N-1}. Excessive variability between sets appears as large values of M, while narrowband processes produce values that are smaller than expected. Unusual sections may be identified by observing the change in M when a given $\hat{S}_k(\omega)$ is deleted. In this case, it may prove simpler to examine the variance or total power in the raw spectral estimates (i.e., the integral over frequency of the spectrum) against the subset index to find anomalous sections.

Another useful procedure to detect sections of a data series which are different is based on the innovations variance. This is just the variance or power associated with the difference between an observation of a process and a linear prediction of it based on all of the earlier data. The one-step prediction variance is treated in detail by *Priestley* [1981, chapter 10]. In the present context, an estimate of the innovations variance is

$$\hat{\sigma}^2_{1,k} = \exp\left[\int_{-\omega_N}^{\omega_N} d\omega \log \hat{S}_k(\omega)\right] \tag{29}$$

where ω_N is the Nyquist frequency. Higher than normal values for the innovations variance suggest a change in the structure of the underlying process because it cannot be predicted adequately from earlier data. In addition, the ratio of the innovations and ordinary variances is useful as a measure of the relative predictability of the process and is also scale invariant.

SPECTRUM ESTIMATION

A major problem in time series analysis is the choice of an algorithm that yields a spectral estimate, given a finite observation of the process of interest, such that the result is not badly biased, yet remains statistically consistent and efficient. That these requirements are usually in conflict is attested to by the plethora of techniques in the literature. For the purposes of this paper, only nonparametric estimates will be considered, eliminating those in which a specific functional form for the spectrum is assumed (e.g., maximum entropy). In addition, only direct estimates based on the discrete Fourier transform are of immediate interest.

Computation of a direct estimate consists of the following steps: (1) tapering a data sequence or a subset of a data sequence (either the raw data or the residuals from a prewhitening operation) with a data window, (2) taking the discrete Fourier transform, (3) converting the result to a cross-spectrum or auto-spectrum by a suitable multiplication, (4) smoothing the result to achieve statistical consistency, and (5) correcting for any prewhitening. The smoothing operation may be done by some combination of convolution with a second type of data window (band-averaging) and combining a set of independent raw estimates computed from a longer data sequence (section-averaging). It is usually assumed that the data window primarily controls bias, while the smoothing operation primarily controls variance, but interactions between the two procedures do exist.

In applying robust M-estimation to spectral problems, only the section-averaging approach will be used. The data are assumed to consist of long sequences of contiguous values that may be subdivided into smaller pieces of equal length. A data window with good bias characteristics is then applied to the subsets with enough overlap between them to yield high efficiency, yet ensure approximate independence of the raw spectra. For this purpose, the superiority of the prolate spheroidal sequences as data windows is well-documented [*Thomson*, 1977*a*; *Slepian*, 1978]. A prolate data window with a time-bandwidth product of 4 and 70% overlap between estimates is used throughout this paper; this gives over 100 dB of bias protection outside an inner domain of full width $8/(T\Delta)$, where T is the number of samples in the data section and Δ is their spacing, but shows partial correlation of the result inside that band. The correlation is the value of the equivalent lag window at the subset offset; see *Thomson* [1977*a*, section 3.3] for details. The raw spectral estimates obtained in this way serve as the input to an M-estimator, as described in the next two sections.

Cases where only a few disjoint sections of data are available, or where the whole time series is short (in the sense that the frequencies of interest are comparable to the reciprocal series length), have traditionally been treated by band averaging with variable width smoothers and are not amenable to the robust procedures of this paper. In any case, band-averaged spectra have a low variance efficiency if a low-bias data window is employed, and the multiple prolate window method of *Thomson* [1982] offers vastly superior performance without a significant loss of bias protection. Robustification of the latter is feasible under some circumstances and will be treated in a future paper.

Nonparametric spectral estimates formally characterize only purely nondeterministic, stochastic processes. In many practical cases, additional deterministic signals, or components with a bandwidth comparable to the reciprocal series length, and so apparently deterministic, are present in a time series and can complicate the spectrum estimation problem considerably. Special methods are required to handle these even for contamination-free data [*Thomson*, 1977*a,b*, 1982], and attention to the presence of line terms (periodic signals) is required when using robust techniques to avoid treating them incorrectly as outliers.

ROBUST ESTIMATION OF POWER SPECTRA

Robust computation of power spectra is a special case of the simple location parameter problem discussed earlier. At each frequency a set of independent raw power spectra are to be averaged together using an M-estimator; for slowly varying spectra, additional band averaging can be incorporated by combining several adjacent frequencies from each raw spectrum. It should be remembered that outlier contamination of some of the spectral subsets can only result in a power spectrum that is biased upwards since it is not possible to subtract from the power in the background process. This means that only individual estimates deviating in a positive sense from the current robust average are downweighted during the iterative processing.

The robust algorithm employed for univariate power spectra is as follows: given a set of spectral sections to be averaged on a frequency-by-frequency basis, an initial robust solution is obtained from the sample median and used to find both the residuals and a scale estimate. Either the MAD or interquartile distance scaling (20) or (21) yields satisfactory results, although the latter offers a slight computational advantage. An iterative solution to the weighted location problem (23) is then sought using the weight function (27) modified to affect only data whose scaled residuals r_i/d exceed the robust average by a critical amount (i.e., the absolute value operation in the

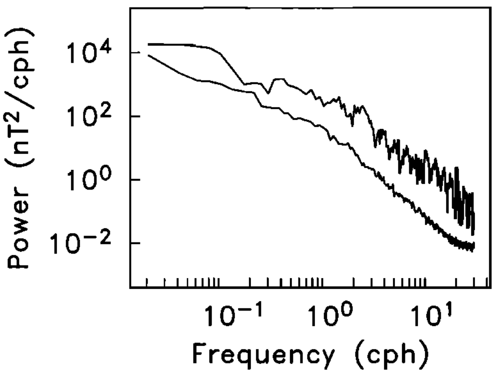

Fig. 1. Conventional (top curve) and robust (bottom curve) power spectra for the time variations of the north magnetic field at Victoria, British Columbia, Canada, during July 1982. The nonrobust spectrum is the arithmetic average of 42 independent pieces of data, with additional band-averaging yielding 84 estimates over 0.1–0.5 cph and 168 estimates over 0.5–30 cph. Because of severe nonstationarity it actually possesses only 10–20 degrees of freedom. The bottom line is the robust average of the same raw spectra, such that the high-amplitude magnetic storms are eliminated, yielding a much higher equivalent degrees of freedom (see text).

exponent is removed). Convergence is achieved when the answer does not change substantially, typically after 3–5 iterations.

If outliers have been eliminated, power spectra are the sums of squares of almost normally distributed variates, and hence distributed as chi-square (χ^2). It is crucial to the correct operation of the nonlinear M-estimator that the proper distribution be used in getting the theoretical MAD and interquartile distance in (20) and (21). Using the discrete Fourier transform, each raw power estimate possesses 2 degrees of freedom at each frequency, neglecting the DC and Nyquist components, which have only 1 degree of freedom. The χ^2 distribution with 2 degrees of freedom (χ_2^2) is equivalent to the exponential distribution and has a pdf given by the standard form [*Johnson and Kotz*, 1970, chapter 18]

$$f(x) = \tfrac{1}{2} e^{-x/2} \quad (x \geqslant 0) \qquad (30)$$

for which the median $\tilde{\mu}$ is $2 \log 2$, the MAD μ_{MAD} is $2 \sinh^{-1}(\tfrac{1}{2}) \approx 0.9624$, and the interquartile distance is $\sigma_{\text{IQ}} = 2 \log 3 \approx 2.1972$. The quantiles of the exponential distribution are given by

$$Q_j = 2 \log \left[\frac{N}{N-j+\tfrac{1}{2}} \right] \qquad j = 1, \ldots, N \qquad (31)$$

Using the Nth quantile of the χ_2^2 distribution as the weight parameter β in (27), robust averaging becomes an adaptive procedure that operates automatically in all save exceptional cases. Detection of such exceptional cases is facilitated by examination of the q-q plot of the final, weighted data. If the weighting has been performed correctly, then the final q-q plot of the weighted spectral estimates (after discarding values with zero weight) against the χ_2^2 quantiles (31) should approximate a straight line. If it remains long-tailed, then the outlier fraction exceeds a typical value of 10–20%, and the weight parameter β must be reduced. Since the scale parameters (20) and (21) are chosen to be consistent with a χ_2^2 distribution, the new weight parameter may be taken directly from the ordinate of the q-q plot as the point where the distribution tails begin to be evident.

The robust power spectral method is best illustrated with some examples. Figure 1 compares robust and conventional spectra of the north magnetic field time variations at Victoria Observatory for the month of July 1982. The data are typical of the mid- to high-latitude geomagnetic field, consisting of a quiet background component interspersed with violent, short-duration storm activity. The latter comprise only 10–20% of the data but exhibit power levels that are easily a decade above the background. The spectra in Figure 1 were computed by averaging, both arithmetically and robustly, raw estimates of 2 days duration. The conventional, arithmetic-averaged spectrum is about a factor of 10 larger than the robust spectrum and displays much greater point-to-point variability. Owing to the data adaptive weighting, the nominal number of estimates per frequency for the robust spectrum is variable but lower by a factor of about 0.9 compared to the conventional type. An argument ignoring robustness would imply that the equivalent degrees of freedom per frequency (about 90 in this example) is higher for the straight arithmetic-averaged spectrum, but this is contradicted by the higher variability seen in Figure 1. In fact, the conventional estimate is dominated by only a fraction of the data, so that it has only 10–20 degrees of freedom at most, accounting for the larger uncertainty. The robust spectrum is a much better measure of the time-averaged behavior of the geomagnetic field, while the conventional spectrum is little different from that given by analysis of only the storm time data.

The q-q plots in Figure 2 illustrate the statistical nature of the robust averaging procedure. For convenience in plotting, the raw and weighted estimates in each frequency bin have been scaled so that their sum of squares is 8, the value expected for the second moment of a χ_2^2 variate. The top plot show the original (unweighted) q-q plot for a few frequencies in the range 0.3–0.4 cph which are typical of the entire spectrum. The infrequent but intense storm activity gives a typical long-tailed distribution. The shape of the weight function causes the bottom q-q plot of the final, weighted power estimates to be slightly short-tailed. Increasing the weight parameter β would bring this closer to a true χ_2^2 result. However, the short-tailed nature of the result does not appreciably alter the robust spectrum of Figure 1, especially on the logarithmic scale over which significant power changes occur.

The stationarity test (28) was computed using the 2-day-long raw section estimates of Figure 1. The value of M was about 130 from the lowest frequency to about 0.2 cph, decreased slowly to a value of about 70 at the Nyquist frequency, and did not display any structure that was localized in frequency. The expected value of M is 64.3, indicating strong nonstationarity at low frequencies. The reduction of M at high frequencies is caused by a rise in the noise level as the spectrum decreases; instrument noise is generally quite stationary. Further examination of the series shows that the detailed form of the nonstationarity is complicated. The ordinary variance in each section of data normalized to its average over all of the sections is plotted against time in Figure 3 (top). The subset variance exceeds twice the average value at five

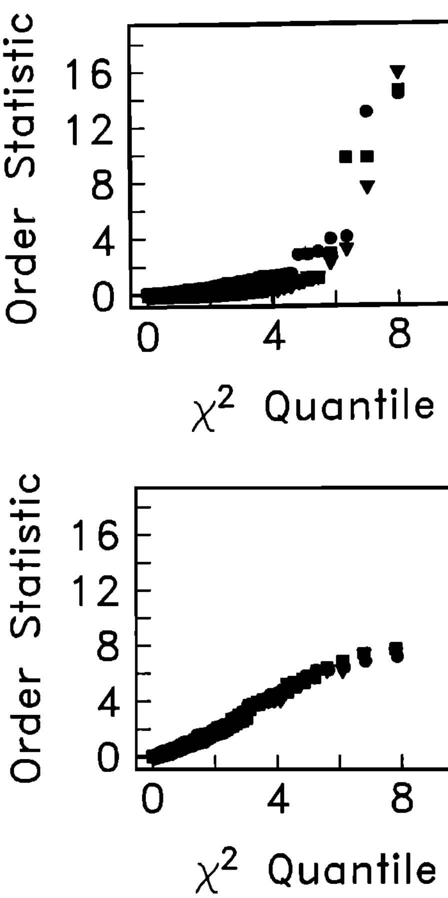

Fig. 2. Quantile-quantile plots for the conventional (top) and robust (bottom) spectra of Figure 1 in the frequency range 0.3–0.4 cph. The plots show the N ranked raw power estimates in a given frequency bin and scaled so that their sums of squares is 8 against the N quantiles of the χ_2^2 distribution; each plotted symbol corresponds to a single data section, and different symbols correspond to different frequency bands. The nonrobust q-q plot at the top shows a largely χ_2^2 population, with an additional long-tailed component that is attributed to brief but intense magnetic storms. The q-q plot at the bottom shows the effect of adaptive weighting of the data, eliminating the storms and yielding a slightly short-tailed χ_2^2 result.

separate and isolated places; these are associated with small magnetic storms. The innovations variance from (29) normalized by its average over all of the data sections is shown against time in Figure 3 (bottom). This was much higher than the mean for extended periods after the five events of Figure 3 (top), suggesting complex and long-term changes in the underlying process. The innovations variance is also higher between days 14 and 20 without any obvious association with the power, implying a different type of change in the structure. The ratio of the innovations and ordinary variances stayed near its median value during the high power event at day 3, implying that the structure of the process did not change much even though the ordinary variance increased dramatically. However, the process was altered after this event, as evidenced by an increase in the ratio, and returned slowly to normal with time. The largest overall change in the structure of the process occurs during a 4-day interval near the center of the record where the relative prediction variance decreases by a factor of 4 from its median value; this is also apparent in Figure 3 (bottom).

Figure 4 shows both conventional and robust spectra of the electric field variations in the frequency band 10^{-4} to 1 Hz collected on Adams Mesa, Arizona, in 1979. The differences between the robust and nonrobust results are more substantial than for Figure 1, especially at the highest frequencies, where the conventional spectrum is dominated by outliers. Note in particular the oscillatory nature of the conventional estimate, remembering that it is plotted on a logarithmic frequency scale. This behavior is typical of a few large outliers in a single subset. Figure 5 shows original and final q-q plots for two frequency intervals, 0.010–0.013 Hz and 0.200–0.204 Hz. The former band shows typical long-tailed behavior caused by a small fraction of outliers that is easily corrected by the robust method. The higher-frequency interval also shows the effects of a few extremely large outliers; these are again removed by the robust averaging procedure with a dramatic effect on the spectrum.

ROBUST ESTIMATION OF TRANSFER FUNCTIONS AND COHERENCES

The robust computation of transfer functions is the frequency-domain equivalent of multivariate robust linear regression, while the coherences are similar to the correlation coefficients of statistical inference theory, and are derived from the output and residual powers obtained during the M-estimation procedure. The spectral problem differs from more conventional ones because the data are complex rather than real numbers, and there are at least two frameworks within which determination of the robust scale and iterative reweighting may be performed. In the

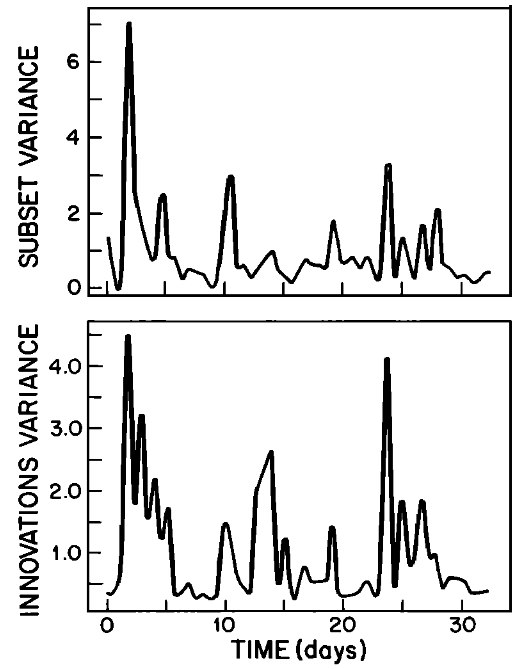

Fig. 3. The top panel shows the total power or variance in each windowed data subset used in obtaining Figure 1 as a function of the center time of the subset. These values have been normalized by the average power for the entire data series. The bottom panel shows the innovations variance for the different subsets against the center time of the subset. The innovations variance has been normalized by its mean over all of the subsets. See text for details.

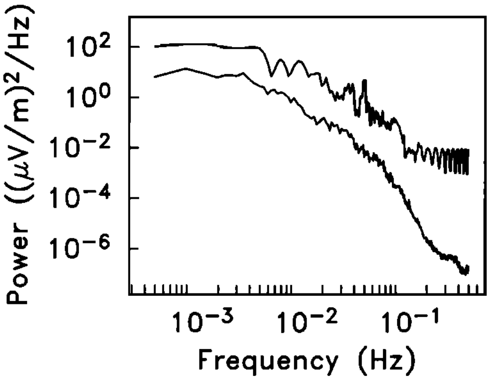

Fig. 4. Conventional (top curve) and robust (bottom curve) power spectra of the north electric field collected on Adams Mesa, Arizona. The nonrobust spectrum is the arithmetic average of 43 independent raw estimates, with additional band-averaging increasing this to 86 estimates over 0.01–0.1 Hz and 172 estimates over 0.1–0.5 Hz. The robust spectrum has ≈70 degrees of freedom over 0–0.01 Hz, ≈130 degrees of freedom over 0.01–0.1 Hz, and ≈250 degrees of freedom over 0.1–0.5 Hz. The severe effect of outliers is apparent on comparing the two results, especially at the highest frequencies.

first instance, the data (i.e., the raw Fourier transforms) may be regarded as having independent Gaussian real and imaginary parts, so that separate weights are applied to them. In the second case, only the magnitudes of the complex numbers are considered, and identical weights are applied to the real and imaginary parts of the data. *Zeger* [1985] has shown that the latter choice, which has a Rayleigh distribution, is preferable because it is rotationally (phase) invariant. It is also more conservative in that an outlier in either the real or the imaginary part of the data will result in downweighting. Extensive practical experience reinforces this: in severely contaminated data, treating the complex magnitude, rather than the independent real and imaginary parts, yields more consistent transfer functions and better convergence. Only the Rayleigh method is used in this paper.

The robust M-estimator used for transfer function computation is similar to the standard procedures described earlier. The data input to the M-estimator are a set of k raw section Fourier transforms for a single output data series $\{x_k\}$ and similar sets for p input data series $\{u_{k_j}\}$; both are also parameterized by frequency. The solution is initialized by computing an unweighted least squares result using QR decomposition on (14), from which residuals and a scale estimate (20) or (21) are obtained. The Huber weights (26) computed using the scaled residuals are then applied to the rows of the matrix regression problem (23). The weighted regression problem is solved, and the entire process is repeated until the total residual power does not change below a threshold value. This procedure gives a final scale estimate and an initial set of residuals for use in the last part of the algorithm. This is also based on iteratively reweighted least squares but uses the weight function (27) and a fixed scale estimate derived from the final Huber iteration, again terminating when the residual power does not vary. A modification of this procedure which substitutes an L_1 simplex programming algorithm for the iterative Huber one has also been used successfully. The residuals from an L_1 solution are found and their MAD is used to get a final scale estimate. Iteratively reweighted least squares using the weight function (27) is then used until the residual power does not change substantially.

The standard probability distribution for a variate which is the square root of the sum of the squares of two

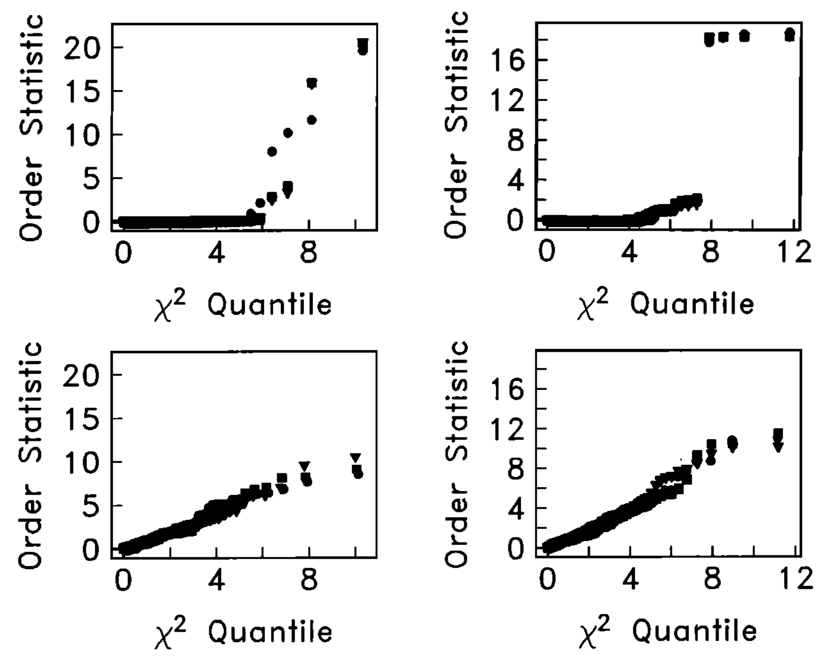

Fig. 5. Quantile-quantile plots for the nonrobust (top panels) and robust (bottom panels) power spectra of Figure 4. The left-hand side shows the results in the 0.010–0.013 Hz band, while the right-hand side covers the range 0.200–0.204 Hz. In both cases the original distribution is a long-tailed χ_2^2 type and is changed to a slightly short-tailed one by the robust averaging. The effect of only a few severe outliers on the spectrum is seen at the higher frequencies and by examining Figure 4.

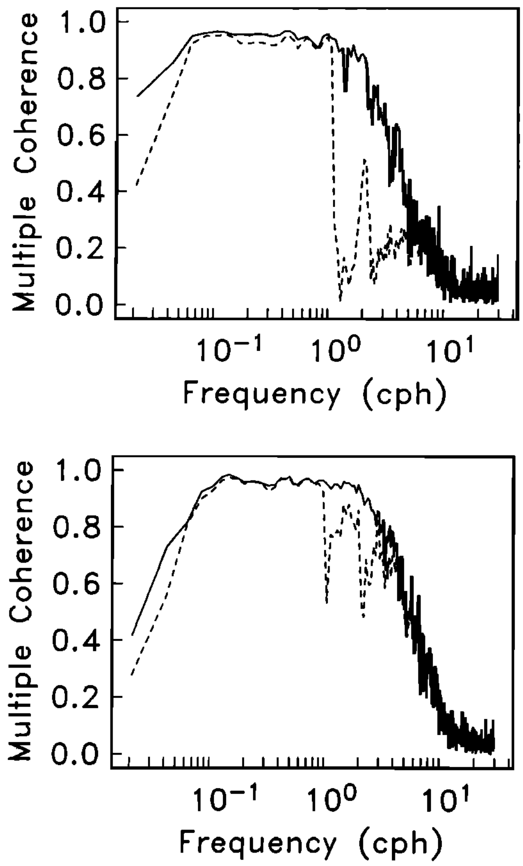

Fig. 6. Conventional (dashed lines) and robust (solid lines) squared multiple coherence plots for the north magnetic field at two seafloor stations off Vancouver Island, site A (bottom) and site B (top), with all three magnetic field components at Victoria Observatory. The data have been contaminated by an instrument fault, accounting for the low conventional coherence between 1 and 5 cph. Robust regression completely removes this effect. See text for details.

independent Gaussian variables is Rayleigh, a transformed version of the χ_2^2 distribution. The pdf is given by [*Johnson and Kotz*, 1970, p. 197]

$$f(x) = xe^{-x^2/2} \quad (x \geqslant 0) \tag{32}$$

Using (32), the median and interquartile distance are $\tilde{\mu} = \sqrt{2 \log 2}$ and $\sigma_{\mathrm{IQ}} = \sqrt{2 \log 4} - \sqrt{2 \log 4/3}$. The MAD is a solution of the transcendental equation

$$2 \sinh (\tilde{\mu}\, \sigma_{\mathrm{MAD}}) = e^{(\sigma_{\mathrm{MAD}})^2/2}$$

yielding a value of ≈ 0.44845. The quantiles are

$$Q_j = \sqrt{2 \log \left[\frac{N}{N-j+1/2}\right]} \quad j = 1,\ldots,N \tag{33}$$

As for the other robust methods, the Nth quantile from (33) serves as the weight parameter β in (27).

The robust squared multiple coherence of the output with the p input time series is computed on a frequency-by-frequency basis from

$$\hat{\gamma}^2 = \frac{S_{xx} - S_{rr}}{S_{xx}} \tag{34}$$

where S_{xx} and S_{rr} are the weighted output and residual powers

$$S_{xx} = \frac{\sum_{i=1}^{N} w_i^2 |x_i|^2}{\sum_{i=1}^{N} w_i^2} \tag{35}$$

$$S_{rr} = \frac{\sum_{i=1}^{N} w_i^2 |r_i|^2}{\sum_{i=1}^{N} w_i^2} \tag{36}$$

and (34) is understood to be zero if $S_{rr} > S_{xx}$. The weight w_i is computed using (27) and the scaled residuals from the last iteration in the weighted linear regression, and setting $w_i = 1$ yields the conventional, nonrobust coherence. Note that (34) reduces to the amplitude of the ordinary magnitude-squared coherence function when there is only one input data sequence ($p=1$). The weighted power spectrum in (35) is not the same as the robust result that is found using the methods of the last section because the weights are derived on the basis of regression rather than location residuals. The processes producing large regression outliers may be different from those that generate anomalous power spectra; in particular, the local nonstationarity seen in Figure 1 may not lead to regression problems if the data and coefficient variables change in similar ways. In addition, new classes of regression outliers can occur that do not affect power spectra, and the origin of these outliers can be difficult to determine.

Figure 6 compares the robust and nonrobust squared multiple coherence (34) between the north magnetic field at two seafloor sites off Vancouver Island and all three magnetic field components from the standard observatory at Victoria. The seafloor data were contaminated to varying degrees by a nearly sinusoidal component of about 1-hour period that was later attributed to magnetized tape cassettes in the instrument data recorders. Both the period and the amplitude of the offending signal decreased continuously over the 1-month-long record, making it a frequency-localized outlier, rather than simply a deterministic component. Site B (top) was more heavily affected than site A (bottom), and the noise component was readily visible as a large amplitude sinusoid in the former case but only produced a slight fuzziness in the site A record. In either case, the conventional coherence (dashed lines) is reduced substantially by the contamination, while the robust algorithm has readily eliminated its effects. The conventional coherence possesses about 100 nominal degrees of freedom per frequency, while the robust coherence has 85–90 degrees of freedom outside the zone of contamination and somewhat fewer inside it. Note also the presence of harmonics of the ≈ 1 cph fundamental. Because of instrument drift, a reduced signal level, and the electromagnetic effects of internal waves the coherence falls off at low and high frequencies in both examples: the robust method does not artificially increase the coherence when the underlying process is itself incoherent. This example illustrates the ability of the robust algorithm to handle both gross (site B) and subtle (site A) outliers in an automatic fashion.

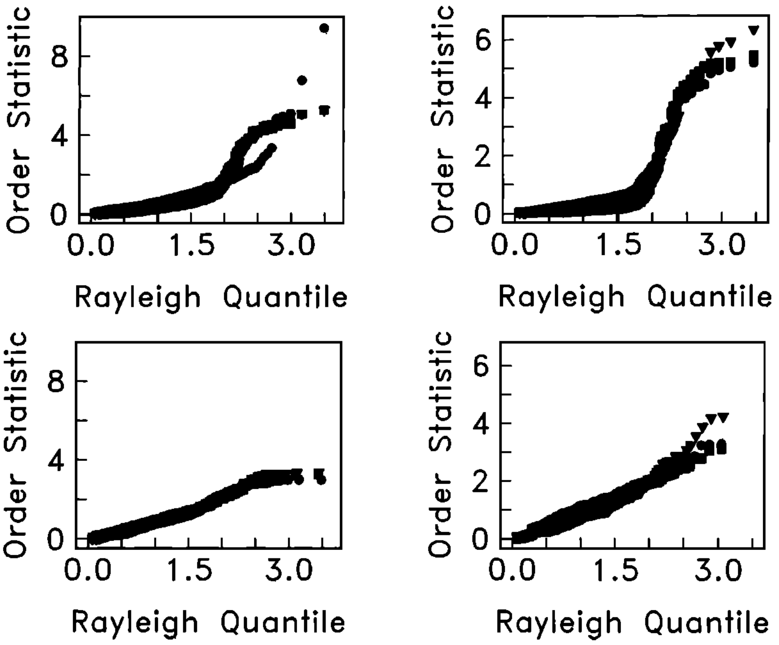

Fig. 7. Unweighted (top) and weighted (bottom) residual q-q plots for the results in Figure 5 computed using the Rayleigh residual method described in the text. The left-hand side shows the initial long-tailed and final, slightly short-tailed distributions for the site A data over 0.97–1.14 cph. The right-hand side shows equivalent values over 1.2–1.4 cph for the more severely contaminated site B data.

Figure 7 shows sample Rayleigh q-q plots for the site B data over 1.2–1.4 cph (right panels) and for the site A data over 0.97–1.14 cph (left panels). For convenience in plotting, the residuals in each frequency band have been scaled so that the sum of squares is 2, as expected for the second moment of a Rayleigh-distributed variate. In both cases the contamination causes the original, unweighted q-q plots to be extremely long-tailed; this is especially evident for the site B data. The final, weighted q-q plot shows a slightly short-tailed Rayleigh distribution for the residuals. There is a suggestion that a smaller value of the weight parameter β would improve the result at site B; this is corroborated by the dips in the robust coherence between 1 and 3 cph in Figure 6 (top). Such tuning is normally not required, but this is an exceptional case because the interfering signal is both strong and persistent.

Figure 8 shows robust and conventional coherences between the north electric field collected on the bed of Lake Washington near Seattle in 1981 (A. Schultz, private communication, 1985) and two horizontal magnetic components from a nearby land site computed using 223 raw estimates. The roll-off at low and high frequencies is caused by the filters applied during data collection, but outlier effects seriously degrade the conventional coherence estimate between 0.001 and 0.03 Hz. The robust method yields a coherence of better than 0.99 over the same frequency band. Figure 9 compares the nonrobust and robust transfer functions for the same north electric and east magnetic components; this is just an element of the magnetotelluric response function. Outlier noise in the electric field causes the conventional transfer function to be excessively variable and consistently biased downward. By contrast, the robust transfer function is much smoother over the frequency interval of Figure 8 where the coherence is high.

Figure 10 compares the original and final q-q plots for two frequency bands obtained using the Rayleigh residual method to get Figures 8 and 9. At the lower frequency (0.0013–0.0017 Hz) the contamination consists of a few large outliers, and the conventional coherence is degraded only a little. The final q-q plot shows a typical short-tailed Rayleigh result. In the higher-frequency band of 0.0063–0.0067 Hz the nonrobust coherence is substantially lower and the outlier contamination much worse, but the robust algorithm is still successful in removing them with a concomitant improvement in the coherence and transfer functions.

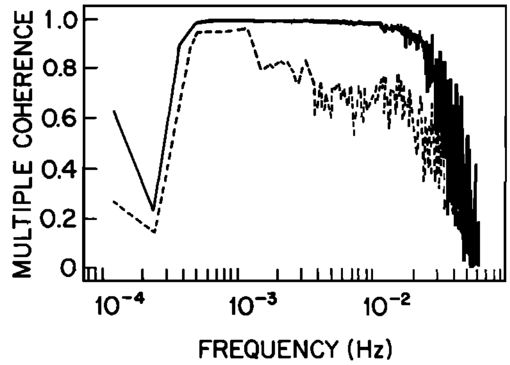

Fig. 8. The conventional (dashed line) and robust (solid line) squared multiple coherence between the north electric field at the bed of Lake Washington and the two horizontal magnetic field components from a nearby land site. There are 223 raw estimates used in both cases, but the robust coherence has fewer than 446 degrees of freedom because outliers are downweighted. The robust squared coherence exceeds 0.99 over most of the range 0.0005–0.02 Hz.

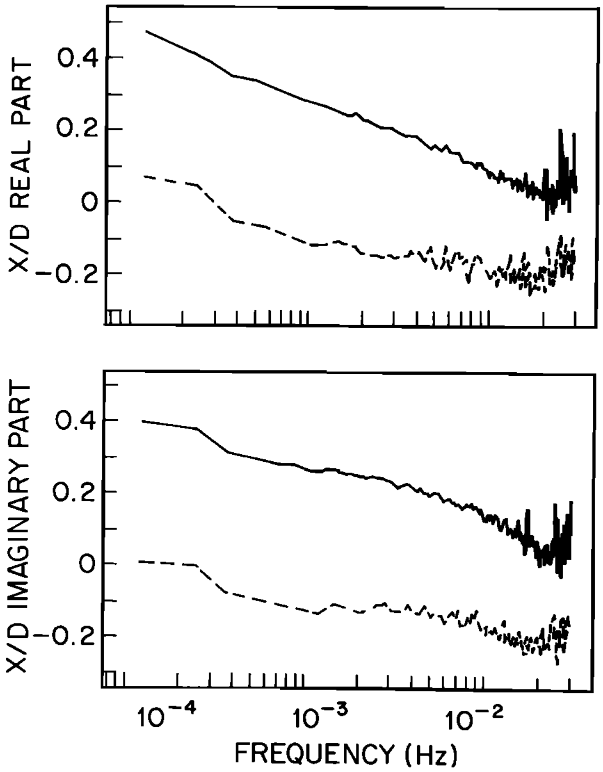

Fig. 9. The real (top) and imaginary (bottom) parts of the conventional (dashed lines) and robust (solid lines) transfer functions between the north electric and east magnetic field data of Figure 8. The nonrobust and robust curves have been offset by 0.4 for clarity. The robust result is smoother than its conventional counterpart, while the latter is also biased downward by outlier noise.

DISTRIBUTIONS AND RELIABILITY OF ESTIMATES

A detailed discussion of the distributions for estimates of power spectra, coherences, and transfer functions is beyond the scope of this paper, and the subject is covered in detail by *Brillinger* [1981]. Traditionally, tests of hypotheses or the placement of confidence intervals on spectral parameters have required distributional assumptions, with the complexity of the exact distributions necessitating further simplification and the use of asymptotic forms to give a tractable result. However, the standard assumption of independence and identical statistics for either the data or the regression residuals implies freedom from outliers. This suggests that the robust spectral estimates will match the statistical model more correctly, so that more realistic hypothesis testing can be performed using them. For the robust estimators of this paper, reasonable approximations are obtained using Gaussian-based statistics if

$$N_{\mathrm{eff}}(\omega) = \sum_{i=1}^{N} w_i^2(\omega)$$

is used as the effective or nominal number of estimates, where $w_i(\omega)$ is the robust weight for the univariate case. In the multivariate case this number is averaged over the p inputs, and N_{eff} is reduced by a factor of $p-1$. This result is usually not as reliable as the jackknifed value mentioned below. In either case, variance calculations must include the correlation caused by the overlap of the windowed subsections, about 25% for the prolate taper with a time-bandwidth product of 4.

In keeping with the nonparametric philosophy adopted

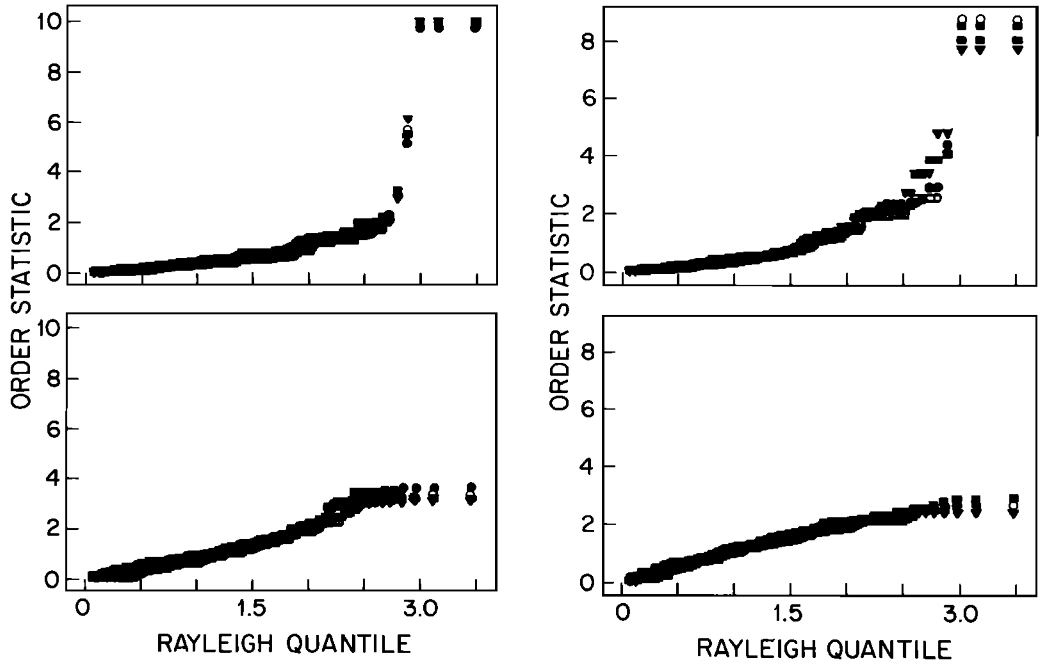

Fig. 10. Unweighted (top) and weighted (bottom) residual q-q plots for the results in Figures 8 and 9 using the Rayleigh residual method described in the text. Two frequency bands are shown, 0.0013–0.0017 Hz (left panels) and 0.0063–0.0067 Hz (right panels). The initial result is typically long-tailed, and the outlier fraction is larger at the higher frequency. The adaptive weighting makes the final residuals slightly short-tailed.

in this work, it is possible to compute error and bias estimates by use of the jackknife method [*R. G. Miller*, 1974; *Efron*, 1982]. The jackknife is based on N repeated analyses of the data using $N-1$ subsets each time, in addition to the original estimate using all N subsets. Appropriate combinations of these $N+1$ estimates give values for both the bias and the variance that are valid under a wide range of parent distributions and estimation procedures, and are generally considered to be more reliable than error estimates based on Gaussian statistics. Such a procedure clearly involves more computing than that for a single estimate, but the amount is not excessive, as it is not necessary to recompute the Fourier transforms.

DISCUSSION

The methods for computing robust power spectra, coherences, and transfer functions presented in this paper operate in an automatic, data adaptive fashion that breaks down only under unusual circumstances. Detection of their failure may be obvious on inspection and is eased by q-q diagnostic plotting. In addition, there are a few tools that can prove useful either in preventing problems or in correcting them when they appear. As has been noted, outliers in power spectra are usually completely different from those in the transfer functions and coherences. The former are generally associated with nonstationarity or obvious contamination of the data (e.g., spike noise) and are not difficult to detect or eliminate. Problems with the robust regression methods used to find transfer functions and coherences are more complicated and may be due to failures of either the data or the model. The following ideas are aimed primarily at improving the robust transfer function and coherence computations.

Under some physical conditions there may be a correlation of the rate of occurrence or the size of outliers with the power in either the input or output data sequences. One obvious example of this is an instrument problem where clipping, saturation, or distortion at high signal levels on one or more data channels will result in poor correlations with other, clean, time series. There is also some evidence for an increase in the frequency of outlier appearance during large magnetic storms at mid- to high-latitudes that can affect electromagnetic induction data. These problems may be detected by plotting the power in the unweighted regression residuals against the power in the individual data series and looking for correlations. While the standard robust algorithms usually work with such data, a substantial improvement may result if the rows of (23) are weighted by the inverse of the power in the pertinent section of the offending time series at all stages in the iterative procedure. Weighting that is proportional to the data power is appropriate if the outliers occur during intervals of low activity.

In both the robust power spectrum and transfer function algorithms the residuals and a scale estimate are computed independently using the solution from an earlier iteration. At first glance this seems to be unnecessary since, for example, a Gaussian model for the data implies a scaled χ^2 distribution for the power spectrum, yielding coupled estimates of location (i.e., the residuals) and scale. However, if periodic or other deterministic background signals are present the Fourier transform has a nonzero expected value and the distribution becomes noncentral χ^2 [*Johnson and Kotz*, 1970, chapter 28], and location and scale estimates must be made separately. Similar arguments apply to the transfer functions and coherences, although their distributions are substantially more complicated in the noncentral case. In other instances, deterministic components may be known to be present (e.g., tidal signals) and are removed by a preliminary least squares procedure. Such approaches must be used with caution: first, by using a robust fitting procedure; second, by being careful to remove the right frequencies.

The effect of noise in the data channels in introducing bias and erratic behavior into computed magnetotelluric response functions is well-known, and many techniques to circumvent this problem have been proposed. Numerical approaches include the selection or weighting of different data sections using coherence [*Kao and Rankin*, 1977; *Jones and Jödicke*, 1984] or iterative refinement of the response function [*Larsen*, 1975, 1980]. The most widely used and generally effective method is employment of a remote reference to minimize the uncorrelated noise between several distinct time series [*Gamble et al.*, 1978]. These techniques are aimed at the removal of outliers from the data (especially the electric field), since it is a few discordant data values rather than persistent background noise components that produce most of the erratic response function behavior. The effectiveness of a robust estimation approach in substantially reducing bias and variability is seen in Figure 9 and has been verified on many other magnetotelluric time series. The use of robust spectral analysis to reduce bias and related problems in magnetotelluric processing appears to be quite promising.

The methods presented in this paper are based on conventional M-estimators and provide good protection against outliers in the data $\{x_i\}$ of (14). They are less effective when outliers occur only in the coefficient variables $\{u_{ij}\}$ and cannot cope with grossly aberrant values in them. This leads to the concept of a breakdown point ϵ^* which is the smallest percentage of contaminated data that will cause the estimator to yield incorrect estimates. The breakdown point is zero for ordinary least squares in the presence of outliers in either the data vector or coefficient matrix and is also zero for outliers in the coefficient matrix with L_1 or M-estimators. M-estimators are preferred over L_1 types on statistical and efficiency grounds, not because of improved outlier resistance.

This observation has led to the introduction of different robustification schemes for linear regression. Generalized M or G-M estimates were proposed with the goal of bounding the influence of outlying coefficient variables and are discussed by *Mallows* [1975]. The breakdown point for G-M estimates is at most $1/(p+1)$, or about 30% for simple linear regression. *Rousseeuw* [1984] suggested an alternate method based on replacing the sum in (14) with the median operator, yielding a breakdown point which approaches 50%. Larger values for ϵ^* are meaningless, since discrimination of good from bad data becomes a matter of semantics. Neither of these approaches to robust regression have been applied to time series, but both are promising candidates to further improve robust spectral analysis.

While the examples of this paper have been drawn from electromagnetic geophysics, the applications of robust spectral estimation are by no means limited to that field and will find uses in many branches of geophysics and oceanography. Since spectral analysis is fundamentally a statistical tool, emphasis has been placed on the importance of getting consistency between data and model, as well as on verifying the underlying statistical assumptions. The techniques presented here are capable of achieving this by giving substantial, automatic, and data adaptive protection from at least two classes of outliers and should be used whenever the section-averaging spectral methods would be appropriate. Use of robust spectrum estimates can also substantially reduce data editing chores; within reason, the effects of isolated, bad data are virtually eliminated without user intervention. In conclusion, three points should be reemphasized. First, both spectrum estimation and robust statistics are fields of active research, so that the methods presented here are unlikely to be the last word on the subject. Second, while these methods can help to identify subtle outliers, this does not mean that they should always be summarily rejected: in some instances, the outliers may be the important part of the data. Finally, while there are many arguments about which robust methods are best, there is general agreement among statisticians that almost any robust method is better than none at all.

Acknowledgments. This work was supported by National Science Foundation grants OCE83-08595 and OCE84-20578 and by funding from the Institute of Geophysics and Planetary Physics branch at Los Alamos National Laboratory (LANL). It was completed during A.D.C.'s tenure as J. Robert Oppenheimer Fellow at LANL. We would like to thank L. K. Law for permission to use the seafloor data for Figures 6 and 7 and A. Schultz for providing the data for Figures 8–10.

REFERENCES

Abraham, B., and G. E. P. Box, Bayesian analysis of some outlier problems in time series, *Biometrika, 66*, 229–236, 1979.

Andrews, D. F., A robust method for multiple linear regression, *Technometrics, 16*, 523–531, 1974.

Barnett, V., Principles and methods for handling outliers in data sets, in *Statistical Methods and the Improvement of Data Quality*, edited by T. Wright, pp. 131–166, Academic, Orlando, Fla., 1983.

Barnett, V., and T. Lewis, *Outliers in Statistical Data*, John Wiley, New York, 1978.

Bartlett, M. S., Properties of sufficiency and statistical tests, *Proc. R. Soc. London, Ser. A, 160*, 268–282, 1937.

Beckman, R. J., and R. D. Cook, Outlier....s, *Technometrics, 25*, 119–163, 1983.

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

Brillinger, D. R., *Time Series: Data Analysis and Theory*, Holden-Day, San Francisco, Calif., 1981.

Claerbout, J. F., and F. Muir, Robust modeling with erratic data, *Geophysics, 38*, 826–844, 1973.

Cook, R. D., and S. Weisberg, *Residuals and Influence in Regression*, Chapman and Hall, New York, 1982.

Dongarra, J. J., C. B. Moler, J. R. Bonch, and G. W. Stewart, *LINPACK User's Guide*, SIAM, Philadelphia, Pa., 1979.

Efron, B., *The Jackknife, the Bootstrap, and Other Resampling Plans*, SIAM, Philadelphia, Pa., 1982.

Fox, A. J., Outliers in time series, *J. R. Stat. Soc. Ser. B, 34*, 340–363, 1972.

Gamble, T. D., W. M. Goubau, and J. Clarke, Magnetotellurics with a remote reference, *Geophysics, 44*, 53–68, 1978.

Gross, A. M., Confidence interval robustness with long-tailed symmetric distributions, *J. Am. Stat. Assoc., 71*, 409–416, 1976.

Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel, *Robust Statistics*, John Wiley, New York, 1986.

Hawkins, D. M., *Identification of Outliers*, Chapman and Hall, London, 1980.

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

Holland, P. W., and R. E. Welsch, Robust regression using iteratively reweighted least squares, *Commun. Stat., A6*, 813–827, 1977.

Huber, P. J., Robust estimation of a location parameter, *Ann. Math. Stat., 35*, 73–101, 1964.

Huber, P. J., *Robust Statistics*, John Wiley, New York, 1981.

Johnson, N. L., and S. Kotz, *Continuous Univariate Distributions-1*, Houghton Mifflin, Boston, 1970.

Jones, A. G., and H. Jodicke, Magnetotelluric transfer function improvement by a coherence-based rejection technique, paper presented at 1984 Society of Exploration Geophysics meeting, Atlanta, Ga., 1984.

Kao, D., and D. Rankin, Enhancement of signal-to-noise ratio in magnetotelluric data, *Geophysics, 42*, 103–110, 1977.

Kendall, M., and A. Stuart, *The Advanced Theory of Statistics*, 2 vols., MacMillan, New York, 1977.

Kleiner, B., and T. E. Graedel, Exploratory data analysis in the geophysical sciences, *Rev. Geophys., 18*, 699–717, 1980.

Kleiner, B., R. D. Martin, and D. J. Thomson, Robust estimation of power spectra, *J. R. Stat. Soc. Ser. B, 41*, 313–351, 1979.

Lanzerotti, L. J., D. J. Thomson, A. Meloni, L. V. Medford, and C. G. Maclennan, Electromagnetic study of the Atlantic continental margin using a section of a transatlantic cable, *J. Geophys. Res., 91*, 7417–7428, 1986.

Larsen, J. C., Low frequency (0.1–6.0 cpd) electromagnetic study of deep mantle electrical conductivity beneath the Hawaiian Islands, *Geophys. J. R. Astron. Soc., 43*, 17–46, 1975.

Larsen, J. C., Electromagnetic response functions from interrupted and noisy data, *J. Geomagn. Geoelectr., 32, Supp. 1*, SI89–SI103, 1980.

Lawson, C. L., and R. J. Hanson, *Solving Least Squares Problems*, Prentice-Hall, Englewood Cliffs, N.J., 1974.

Lewis, T., and N. I. Fisher, Graphical methods for investigating the fit of a Fisher distribution to spherical data, *Geophys. J. R. Astron. Soc., 69*, 1–13, 1982.

Mallows, C. L., On some topics in robustness, technical memorandum, Bell Telephone Lab., Murray Hill, N.J., 1975.

Mallows, C. L., Robust methods, *Proc. Symp. Appl. Math, 28*, 49–74, 1983.

Martin, R. D., and D. J. Thomson, Robust-resistant spectrum estimation, *Proc. IEEE, 70*, 1097–1115, 1982.

Martin, W., and P. Flandrin, Wigner-Ville spectral analysis of nonstationary processes, *IEEE Trans. Acoust. Speech Signal Process., ASSP-33*, 1461–1470, 1985.

Mason, R. L., and R. F. Gunst, Outlier-induced collinearities, *Technometrics, 27*, 401–407, 1985.

Mehrotra, K. G., and P. Nanda, Unbiased estimation of parameters by order statistics in the case of censored samples, *Biometrika, 61*, 601–606, 1974.

Miller, K. S., *Complex Stochastic Processes*, Addison-Wesley, Reading, Mass., 1974.

Miller, R. G., The jackknife-a review, *Biometrika, 61*, 1–15, 1974.

Mosteller, F., and J. W. Tukey, *Data analysis and linear regression*, Addison-Wesley, Reading, Mass., 1977.

Priestley, M. B., Evolutionary spectra and non-stationary processes, *J. R. Stat. Soc. Ser. B, 27*, 204–229, 1965.

Priestley, M. B., *Spectral Analysis and Time Series*, Academic, Orlando, Fla., 1981.

Rousseeuw, P. J., Least median of squares regression, *J. Am. Stat. Assoc., 79*, 871–880, 1984.

Slepian, D., Prolate spheroidal wavefunctions, Fourier analysis, and uncertainty, V, The discrete case, *Bell Syst. Tech. J., 57*, 1371–1429, 1978.

Thomson, D. J., Spectrum estimation techniques for characterization and development of WT4 waveguide, I, *Bell Syst. Tech. J., 56*, 1769–1815, 1977*a*.

Thomson, D. J., Spectrum estimation techniques for characterization and development of WT4 waveguide, II, *Bell Syst. Tech. J., 56*, 1983–2005, 1977*b*.

Thomson, D. J., Spectrum estimation and harmonic analysis, *Proc. IEEE, 70*, 1055–1096, 1982.

Tukey, J. W., Instead of Gauss-Markov, what?, in *Applied Statistics*, edited by R.P. Gupta, pp. 351–372, Amsterdam, North-Holland, 1975.

Vogel, C. R., Optimal choice of a truncation level for the truncated SVD solution of linear first kind integral equations when the data are noisy, *SIAM J. Numer. Anal., 23*, 109–117, 1986.

Zeger, S. L., Exploring an ozone spatial time series in the frequency domain, *J. Am. Stat. Assoc., 80*, 323–331, 1985.

M. E. Ander, Earth and Space Sciences Division, Los Alamos National Laboratory, Los Alamos, NM 87545.

A. D. Chave, AT&T Bell Laboratories, 600 Mountain Ave., Murray Hill, NJ 07974.

D. J. Thomson, AT&T Bell Laboratories, 600 Mountain Ave., Murray Hill, NJ 07974.

(Received December 26, 1985;
revised September 30, 1986;
accepted October 6, 1986.)