Geophysical Journal (1988) **94,** 131–142

Parameter estimation in non-Gaussian noise

C. G. Constable

Institute of Geophysics and Planetary Physics, Scripps Institution of Oceanography, University of California, San Diego, La Jolla, Ca 92093, *USA*

Accepted 1988 January 22. Received 1988 January 21; in original form 1986 June 25

SUMMARY

Least squares (LS) estimation of model parameters is widely used in geophysics. If the data errors are Gaussian and independent the LS estimators will be maximum likelihood (ML) estimators and will be unbiased and of minimum variance. However, if the noise is not Gaussian, e.g. if the data are contaminated by extreme outliers, LS fitting will result in parameter estimates which may be biased or grossly inaccurate. When the probability distribution of the errors is known it is possible, using the maximum likelihood method, to obtain consistent and efficient (minimum variance) estimates of parameters. In some cases the distribution of the noise may be determined empirically, and the resulting distribution used in the ML estimation. A procedure for doing this is described here. Hourly values of geomagnetic observatory data are used to illustrate the technique. These data sets contain a number of periodic components, whose amplitudes and phases are geophysically interesting. Geomagnetic storms and other phenomena in the record make the noise distribution long-tailed, asymmetric and variable with location. Using an iterative procedure, one can model the form of these distributions using smoothing splines. For these data ML estimation yields quite different results from standard robust and LS procedures. The technique has the potential for widespread application to other problems involving the recovery of a known form of signal from non-Gaussian noise.

Key words: non-Gaussian noise, maximum likelihood

INTRODUCTION

Parameter estimation is a problem that almost always arises in geophysics when one wishes to draw any inferences from data. A variety of estimators is available and the one chosen should obviously depend on the nature of the problem to be solved. In choosing an appropriate technique it is important to decide what properties are required of the *best* method. Two desirable properties are that an estimator be *consistent* and *unbiased.* An estimator t_n of a parameter θ, computed from a sample of n values is said to be *consistent* if t converges to θ in probability; i.e. if for any δ, $\eta > 0$ there exists N such that

$$P\{|t_n - \theta| < \delta\} > 1 - \eta, \qquad n > N$$

The requirement that an estimator be unbiased is simply that its expected value should be the true value of the parameter

$$E(t) = \theta.$$

Thus consistency includes the property of being unbiased in the limit of large data samples. In general there is more than one consistent estimator of a parameter, even if one only considers those that are unbiased (e.g. for the normal distribution both the sample mean and median are consistent and unbiased estimators of the mean). Clearly, one needs further criteria for discriminating between them; an obvious choice is to make use of the estimator which has a smaller variance, since on the average it will deviate less from the true value than one with a large variance, and may thus be regarded as better. Under fairly general conditions, (Kendall & Stuart 1979, chapter 17) it may be shown that there exists a bound below which the variance of an unbiased estimator cannot fall (the minimum variance bound) although this bound is not necessarily attained. Since most estimators are asymptotically normally distributed (by virtue of the Central Limit Theorem), the distribution of estimators will depend for large samples only on the mean and variance. The estimator with minimum variance in large samples is said to be efficient. Maximum likelihood (ML) estimators can be shown to be consistent, asymptotically normal and efficient (Kendall & Stuart 1979, chapter 18); we can therefore regard them as *best* in situations involving large samples.

Least-squares (LS) estimation is widely used, and in some cases will be the optimum method. In particular, if the data errors are Gaussian, the Gauss–Markov theorem (see e.g. Kendall & Stuart 1979; Priestley 1981) shows that the LS estimators will be unbiased and of minimum variance, and coincide with the ML estimators. However, although data errors are often assumed to be Gaussian, an *a priori* knowledge of the distribution is comparatively rare in geophysical problems. Often the best that we can hope to do is look at the residuals after model fitting in order to determine whether our initial assumptions are justified.

Even when the majority of the data errors are

approximately Gaussian, there may be contamination of the data by a small number of outlying points. LS estimation relies on the minimization of a *loss* (or penalty) function, in this case the sum of the squares of the residuals from the model fit to the data. This form for the loss function means that LS estimation is not robust in the presence of outliers; their residual contribution to the loss function is squared, giving them what might be regarded as an excessive influence on the resulting model. The presence of outliers in many data sets has led to the development of robust methods for performing regression estimates (see e.g. Barnett & Lewis 1984; Huber 1981; Montgomery & Peck 1982, chapter 9; Hampel *et al.* 1986). These methods include M-type (or maximum likelihood type) estimation and R- and L- estimation. M-type estimation is similar to LS estimation in that it involves the minimization of a loss function; however, for outlying points the loss function usually invokes a lesser penalty than the square of the residual. The outcome from M-type estimation will thus depend on the choice of loss function and the distribution of the residuals. R-estimation is a procedure based on the ranks of the residuals and L-stimators are based on order statistics (e.g. the sample median is an order statistic). M-type regression estimates are probably the most widely used, partly because of their simplicity and partly because their effects may be more readily determined than those of R- or L-estimates.

There are some situations, however, in which the data errors cannot even be regarded as approximately normally distributed. Then the LS parameter estimates cannot be expected to be minimum variance, and the results of robust estimation procedures may no longer be optimum either. A true ML estimate is desirable, because biased or inconsistent estimates may give us the wrong answer, while those which are not of minimum variance will have larger confidence intervals than necessary. I present here a method for obtaining ML estimates of parameters in cases where there are sufficient data available to make an estimate of the actual error distribution. The method is illustrated by estimating the amplitudes and phases of a number of sinusoidal signals of known period that occur in geomagnetic observatory data.

MAXIMUM LIKELIHOOD ESTIMATION

Let us assume that we have n data y_i which may be written as a linear combination of p known basis functions, $c_j(x)$, plus an error or noise term ϵ_i, i.e.

$$y_i = g(x_i) + \epsilon_i \qquad i = 1, 2, \ldots, n, \tag{1}$$

where

$$g(x_i) = \sum_{j=1}^{p} \beta_j c_j(x_i) \tag{2}$$

or in matrix notation

$$\mathbf{y} = \mathbf{g} + \boldsymbol{\epsilon} = \boldsymbol{\beta}\mathbf{C} + \boldsymbol{\epsilon}. \tag{3}$$

The errors are independent random variables with probability density function (p.d.f.) $f(\epsilon)$, $-\infty < \epsilon < \infty$, and we will assume for the time being that f is of known functional form. Determining the model that best fits the data thus involves the estimation of the parameter vector β.

The ML method finds the parameter estimates, $\hat{\beta}$, which maximize the probability of getting the data that were actually observed. The likelihood function $L(\hat{\beta})$ is the joint p.d.f., F, of the sample errors at y, i.e.,

$$L(\hat{\beta}) = F(\epsilon, \hat{\beta}). \tag{4}$$

Assuming the errors in the y_i to be independent and identically distributed (i.i.d.), this is the product of the individual density functions for each measurement y_i

$$L(\hat{\beta}) = \prod_{i=1}^{n} f(\epsilon_i, \hat{\beta}). \tag{5}$$

The maxima of the likelihood and of the log of the likelihood functions coincide so we may rewrite the problem as the minimization of a loss function $\rho(\epsilon) = -\ln[f(\epsilon, \hat{\beta})]$

$$\max_{\hat{\beta}} \ln L(\hat{\beta}) = \min_{\hat{\beta}} \sum_{i=1}^{n} \rho(\epsilon_i), \tag{6}$$

where there is an implicit dependence of ρ on the parameter vector β. The necessary condition for the minimum is thus

$$\sum_{i=1}^{n} c_j(x_i)\psi(\epsilon_i) = 0 \qquad j = 1, 2, \ldots, p, \tag{7}$$

where

$$\psi(\epsilon) = \rho'(\epsilon)$$

and is sometimes called the influence function (apart from a constant of proportionality). The equations (7) are known as the likelihood equations and their solution yields the ML estimate for β, provided that it corresponds to a global minimum in the loss function ρ; a sufficient condition is that ρ be convex.

It is evident that the ML estimate depends entirely on the form assumed for $f(\epsilon)$, the p.d.f. for the errors. For a normal distribution the loss function $\rho(\epsilon) = \epsilon^2$ and (7) are simply the well-known normal equations; then the ML solution for β is identical to that obtained by least squares. The fact that it is minimum variance and unbiased for normal distributions is part of the reason for the widespread popularity of LS estimation. However, if the errors are non-Gaussian we should use a different loss function. As another example let us consider errors with a Laplacian distribution, whose tails fall off exponentially as $e^{-|\epsilon|}$. ML estimation of the parameters yields a loss function of $\rho(\epsilon) = |\epsilon|$, and therefore will involve minimization of the one norm of the residuals.

Robust M-type estimation techniques make use of the above loss function minimization formulation. For data which have a normal error distribution contaminated by a few gross outliers it is possible to reduce the influence of these outliers on the LS parameter estimates. This is done by minimizing a loss function $\rho(\epsilon)$ which increases less rapidly than the residual sum of squares. Examples of the influence functions associated with some commonly used robust loss functions are shown in Fig. 1, along with the LS influence function. The M-estimator obtained using Huber's influence function is the maximum likelihood estimator corresponding to a density with a normal centre and double-exponential tails. This could be viewed as the sort of distribution that one obtains when the noise arises from two different sources, one with a normal distribution and the

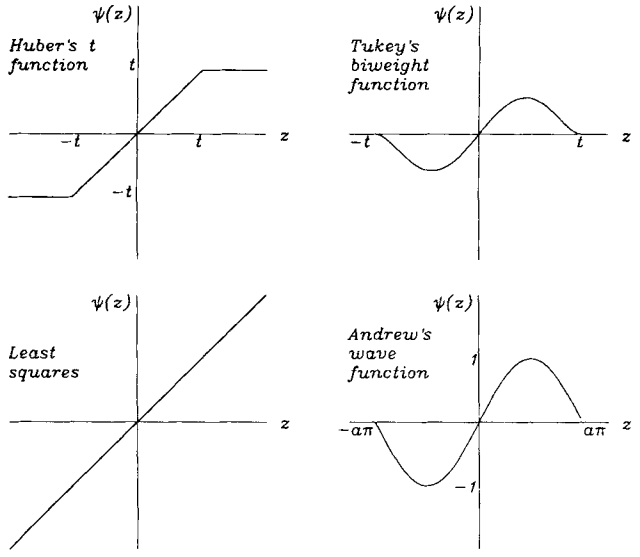

Figure 1. LS influence function and examples of some commonly used robust influence functions.

other contributing outliers. Such a combination of noise sources is referred to as a finite mixture distribution. Mixture distributions are discussed extensively by Titterington, Smith & Makov (1985), along with means of estimating their parameters. Problems involving outliers are often discussed in terms of mixture distributions as even when underlying categories for the noise processes cannot be identified they can provide a convenient means for modelling the overall p.d.f. for the noise. Some of the loss functions of Fig. 1 are somewhat *ad hoc* in that they require the data analyst to specify certain tuning constants in order to optimize their performance. Ideally, one would hope that the solution would not depend too critically on the choice of these constants. The density estimation discussed in the next section gets around this problem by enabling us to tailor the loss function used to the underlying noise distribution.

The solution of equations (7) is in general a non-linear problem and may not be readily obtainable in exact form. Newton's method may be used by making an initial guess, $\hat{\beta}_0$, for the solution and then iterating to obtain an improved solution

$$\frac{\partial(\ln L)}{\partial\hat{\beta}} \approx \frac{\partial(\ln L)}{\partial\hat{\beta}}\bigg|_{\beta_0} + \frac{\partial^2(\ln L)}{\partial\hat{\beta}^2}\bigg|_{\beta_0}(\hat{\beta}-\hat{\beta}_0) = 0.$$

Alternatively, an approximate solution may be obtained using an iteratively reweighted LS method. Let us suppose once again that we have an initial estimate $\hat{\beta}_0$ and write (7) as

$$\sum_{i=1}^{n} c_j(x_i)\frac{\psi(\epsilon_i)}{\epsilon_i}\epsilon_i = 0 \qquad j = 1, 2, \ldots, p. \tag{8}$$

Now let g_0 be our estimate of g computed from $\hat{\beta}_0$ and

$$w_{i0} = \frac{\psi(\epsilon_i)}{\epsilon_i} = \frac{\psi[y_i - g_0(x_i)]}{[y_i - g_0(x_i)]} \quad \text{if } y_i \neq g_0(x_i) \quad i = 1, 2, \ldots, n$$
$$= \lim_{y_i \to g_0(x_i)} \frac{\psi[y_i - g_0(x_i)]}{[y_i - g_0(x_i)]} \quad \text{if} \quad y_i = g_0(x_i) \tag{9}$$

which transforms (7) to the well-known weighted least-squares normal equations

$$\mathbf{C}^{\mathrm{T}}\mathbf{W}_k\mathbf{C}\hat{\boldsymbol{\beta}}_k = \mathbf{C}^{\mathrm{T}}\mathbf{W}_k\mathbf{y} \qquad k = 0, 1, 2, \ldots \tag{10}$$

with $\mathbf{W}_k$ being the $n \times n$ diagonal matrix of 'weights' with diagonal elements $(w_{1k}, w_{2k}, \ldots, w_{nk})$ given by (9). This system of equations may be solved repeatedly using successive estimates of $\hat{\beta}_k$ to recalculate the weights until convergence is reached. The initial estimate for $\hat{\beta}_0$ may be provided by a LS fit to the data with all the $w_{ik} = 1$. When $\psi(\epsilon_i)$ no longer changes between successive iterations the parameter estimates have converged.

This procedure for solving the likelihood equations is known as W-estimation and suffers from the disadvantage that even when the equations have a unique solution the corresponding W-estimator can only be guaranteed to converge to this solution if $\psi(\epsilon)$ is non-decreasing. Heavy-tailed distributions will violate this condition and are said to have redescending ψ functions; their W-estimators will have infinitely many solutions. However, if the initial estimate for $\hat{\beta}$ is sufficiently near the true M-estimator solution, then the W-estimator will converge to the true solution. Thus in some cases it might be more appropriate to start with the one-norm solution as the initial model for $\hat{\beta}_0$, where this is closer to the true solution than the LS solution.

Huber (1972) has shown that the covariance matrix for $\hat{\beta}$ may be approximated by the asymptotic form

$$\operatorname{cov}(\hat{\beta}) = \frac{E([\psi - E(\psi)]^2)}{[E(\psi')]^2}(\mathbf{C}^{\mathrm{T}}\mathbf{C})^{-1}$$
$$\approx \frac{\frac{1}{n-p}\sum\left[\psi(\epsilon_i) - \frac{1}{n}\sum\psi(\epsilon_i)\right]^2}{\left[\frac{1}{n}\sum\psi'(\epsilon_i)\right]^2}(\mathbf{C}^{\mathrm{T}}\mathbf{C})^{-1}, \tag{11}$$

where E is the expectation opeator. This is similar to the expression obtained for LS covariance estimates, except that the variance term σ^2 multiplying the covariance matrix is replaced by the asymptotic form

$$\sigma_{as}^2 = \frac{E[\psi - E(\psi)]^2}{[E(\psi')]^2}. \tag{12}$$

This asymptotic variance provides a useful means of determining when the W-estimation has converged as it will then no longer change between successive iterations. A tolerance of 1 per cent was used in most cases, which reflected a similar tolerance between successive amplitude estimates. Initially it was thought that the difference between the sums of squares of the weighted residuals obtained from successive solutions to equation (10) could be used as an indication of whether the W-estimation had converged. However, in some cases where the p.d.f. was highly non-Gaussian this was not an adequate criterion—the weighted sums of the squares of the residuals converges more rapidly than the amplitude estimates. The asymptotic variance proved a more reliable indicator of when the estimation procedure had converged.

DENSITY FUNCTION ESTIMATION

The ML estimation procedure described above presumes throughout that the p.d.f. of the data errors is known. In practice this is often not the case and one has to make do

with a guess at the distribution of the errors associated with a given model (hence the widespread use of LS estimation). However, in the geomagnetic problem that we discuss here, as well as in many other geophysical problems, there are sufficient data available that we may use them to generate an estimate of the error p.d.f. Sometimes this may be done directly, when measurements of the noise distribution are available independently of the signal. For example, in controlled source electromagnetic sounding the noise distribution is determined by the telluric signal and electrode noise and could be monitored without the presence of the controlled source signal. In deep electrical soundings in South Africa, Van Zijl, Hugo & de Bellocq (1970) studied the statistical distribution of their measurements, and by use of a careful rejection procedure were able to work solely with noise that was approximately normally distributed. An alternative approach, which is advocated here, would be to model the actual noise distribution and use it for ML estimation as described in the previous section. In other cases, if we presume a knowledge of the form of the signal, we may estimate the distribution of the noise iteratively. Let us suppose initially that the error distribution is Gaussian i.i.d. Then a histogram of the residuals from a LS fit to the data should provide a good approximation to a normal p.d.f. as the number of data becomes sufficiently large. On the other hand if the Gaussian p.d.f. is a poor model for the errors, this should be manifest in the shape of the histogram. Then, however, we could use the histogram of residuals to provide an estimate of the true underlying p.d.f. for use in the ML procedure. This is a problem in adaptive estimation of the type discussed by Bickel (1982), where he derives sufficient conditions for adapative estimation of a Euclidean parameter in the presence of an infinite dimensional shape nuisance parameter, i.e. he shows that under the assumptions given here one may obtain parameter estimates that are as good as if one actually knew the true underlying noise distribution.

Over the last 30 years or so a variety of methods for probability density function estimation has been developed (see e.g. Tapia & Thompson 1978; Silverman 1986). The histogram of the LS residuals is in fact an estimate of the p.d.f., when suitably normalized. However, for the purposes of the ML estimation described in the previous section it lacks a few essential characteristics, such as differentiability and smoothness. It is desirable to find a flexible form for the representation of the p.d.f., that can model any asymmetry in the tails, as well as following closely the variation in the centre of the histogram. Smoothing splines (Reinsch 1967; Silverman 1985) provide a useful means of modelling arbitrary functional forms, and by a judicious choice of misfit it is possible to take into account that the histogram is only an estimate of the p.d.f. It would also be useful to be able to compute analytically the influence function $\psi(\epsilon)$ and its derivative for use in the ML estimation procedure. The B-spline representation for splines (de Boor 1978, chapter XIV) enables us to do this. For simplicity we may choose to fit $\rho(\epsilon) = -\ln f(\epsilon)$ rather than fitting the p.d.f. directly, i.e. we can find a representation for ρ of the form

$$\hat{\rho}(\epsilon) = \sum_{j=1}^{m} \alpha_j b_j(z_j, \epsilon) \tag{13}$$

that is the minimizer over twice continuously differentiable functions on $[x_1, x_n]$ of

$$\sum_{j=1}^{m} \left[\frac{q_j - \hat{\rho}(\epsilon_j)}{\sigma_j} \right]^2 + \lambda \int_{x_1}^{x_n} [\partial_\epsilon^2 \hat{\rho}(\epsilon)]^2 \, d\epsilon, \tag{14}$$

where b_j are B-spline basis functions at knots z_j, m is the number of histogram bins, q_j is the negative logarithm of the fraction of the measurements in the jth bin, and $1/\sigma_j$ is the weight applied to q_j in the fitting procedure. Fitting in the log domain has a number of advantages. Firstly, many of the histograms from non-Gaussian data show the p.d.f.s to be very narrow and possess high curvature near the mode of the p.d.f., making them difficult to fit with a smoothing spline in the linear domain without very careful knot placement. Secondly, the computations in the ML estimation are much simpler, and last, but not least, the necessity for a positivity constraint on f is eliminated. The data are weighted in inverse proportion to their amplitude, to compensate for fitting in log domain. This method is akin to the penalized likelihood approach to density estimation, first applied by Good & Gaskins (1971). They actually fit the p.d.f. directly, not the loss function, and suggest a different form for the roughness penalty. However, penalized likelihood estimation in the log domain is not without precedent. Silverman (1982) discusses it, and in a later paper (Silverman 1984) gives a heuristic argument relating this type of estimate to adaptive kernel estimates (Silverman 1986 provides a good review of kernel estimators for p.d.f.s). In density estimation it is not normal to bin the data first to form a histogram, as is done here; for the example discussed in the next section it is necessitated by the large number of data, which would otherwise make routine density estimation prohibitively expensive. This will be true in many situations where there are sufficient data available to make a reliable estimate of the p.d.f.

The estimated p.d.f.s are usually quite simple in form even though they do not correspond to known distribution functions. The computation of the smoothing spline model for $\rho(\epsilon)$ may be simplified by working with a depleted basis of B-splines, which no longer requires a knot to be positioned at every datum. These depleted basis representations, called penalized least-squares splines (PS), provide an excellent approximation to a smoothing spline, with a reduced number of parameters and at a lesser computational cost. Their use is described in Constable & Parker (1988). After fitting the PS, the p.d.f. f is normalized by performing a numerical integration using Gaussian quadrature. The expectation, mode and variance of the distribution may be computed by the same method. Using (12) we can also compute the asymptotic variance, σ_{as}^2, when the ML estimation procedure described in the previous section is used. These computations are facilitated by the fact that once ρ has been estimated by the PS, it is straightforward to differentiate it analytically and obtain $\hat{\psi}$ and $\hat{\psi}'$. Thus once we have a model for the p.d.f. of the errors we may estimate the improvement over LS in the size of the confidence intervals for β by comparing the actual variance of the distribution with the computed asymptotic variance.

Once the loss function $\hat{\rho}$ (and thus $\hat{\psi}$, $\hat{\psi}'$ and f) has been derived from the LS residuals, it is used in the ML

estimation described in the previous section. The density function estimation procedure is repeated until no significant change in the loss function is obtained. For the examples discussed here this was considered to have occurred when subsequent values of q_j (as given in equation (14)) were visually indistinguishable when they were plotted together or, equivalently, when the histograms of residuals from subsequent iterations were identical. This usually took no more than two or three iterations. The sum of the squares of the residuals from the ML fitting may be used to provide a quantitative measure of how much the residual distributions (and thus the estimated p.d.f.s) change between iterations.

APPLICATION TO LINE ESTIMATION FOR GEOMAGNETIC OBSERVATORY DATA

In this section an example of the application of the ML method to line amplitude estimation in geomagnetic observatory data is presented. Fig. 2 shows a sample section of some hourly average values of geomagnetic observatory data from Yellowknife, Canada. The effects of long-term secular variation in the internally generated magnetic field have been removed using the penalized least-squares spline algorithm described by Constable & Parker (1988). The problem is to determine the amplitudes and phases of a

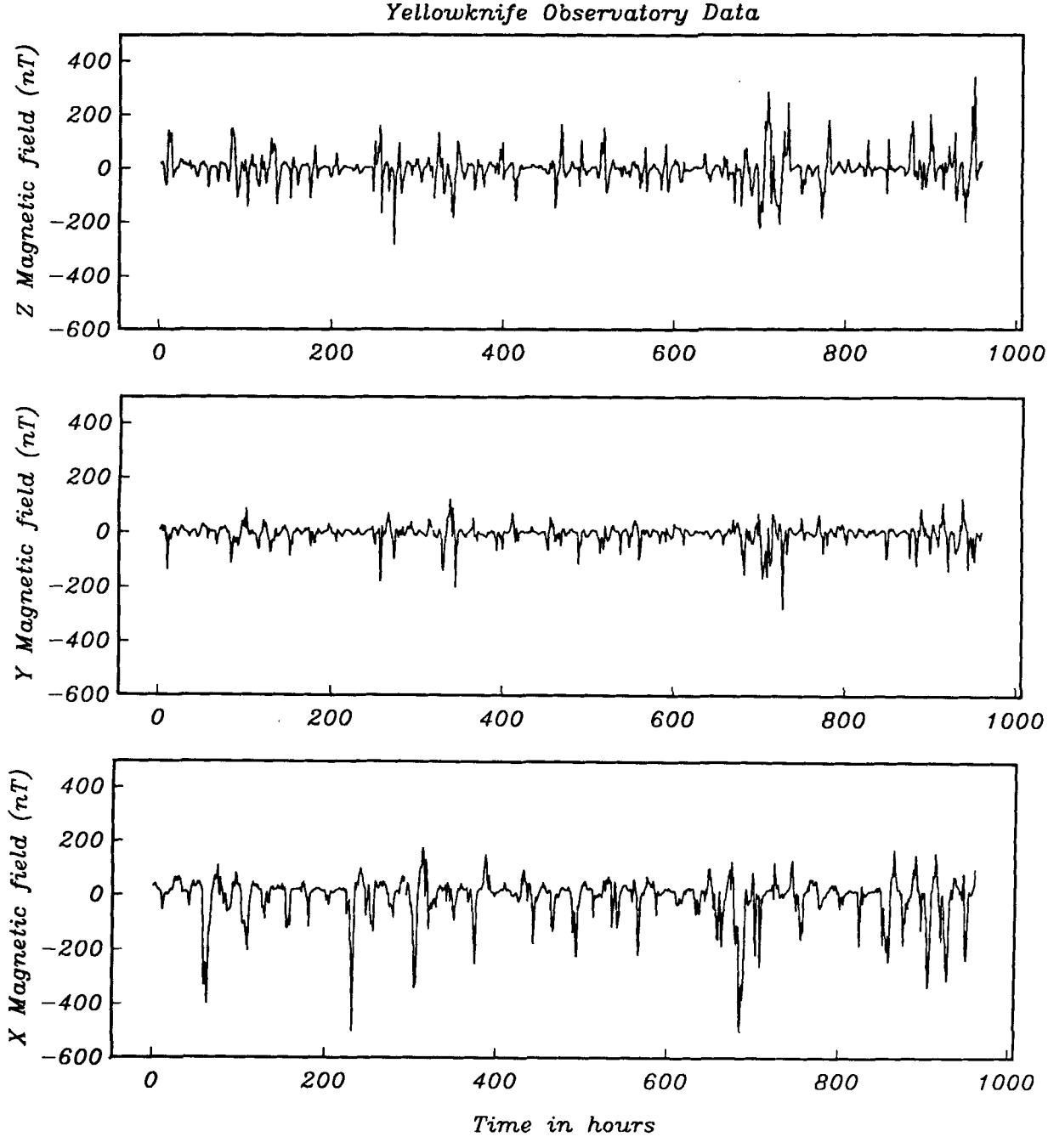

Figure 2. Sample of hourly average values of geomagnetic observatory data from Yellowknife, Canada.

number of presumed sinusoidal signals of known frequency generated by periodic variations in current flow in the magnetosphere, ionosphere and oceans. Some of these occur at a period of 24 h and its harmonics (known as S_1, S_2, S_3, etc.); others correspond to seasonal variations in the solar cycle, to tidal components in oceanic flow and other phenomena. Our ability to detect these signals will depend on their size, how close together their periods are and the length of the data series. Here we will look only at the period of one solar day and its first six harmonics in a data series that spans one year (this enables us to deal with a data set and parameter estimation problem of convenient size to illustrate the method discussed here). This is of interest, not just at this observatory, but also at other sites in North America, and the rest of the world, as it has application in conductivity studies of the Earth. Traditionally LS estimation has been used in this problem, either directly (e.g. Larsen 1968; Banks 1969; Malin & Schlapp 1980; Sellek & Malin 1982; Chave & Filloux 1984), or implicitly by using an FFT to compute the power spectrum (e.g. Currie 1966). Results from both direct LS and power spectral methods may be unduly influenced by outlying data if the noise distribution is non-Gaussian. It is evident from the figure that, in addition to a number of possibly periodic processes present in the record, there are geomagnetic storms that have a strong influence on the characteristics of

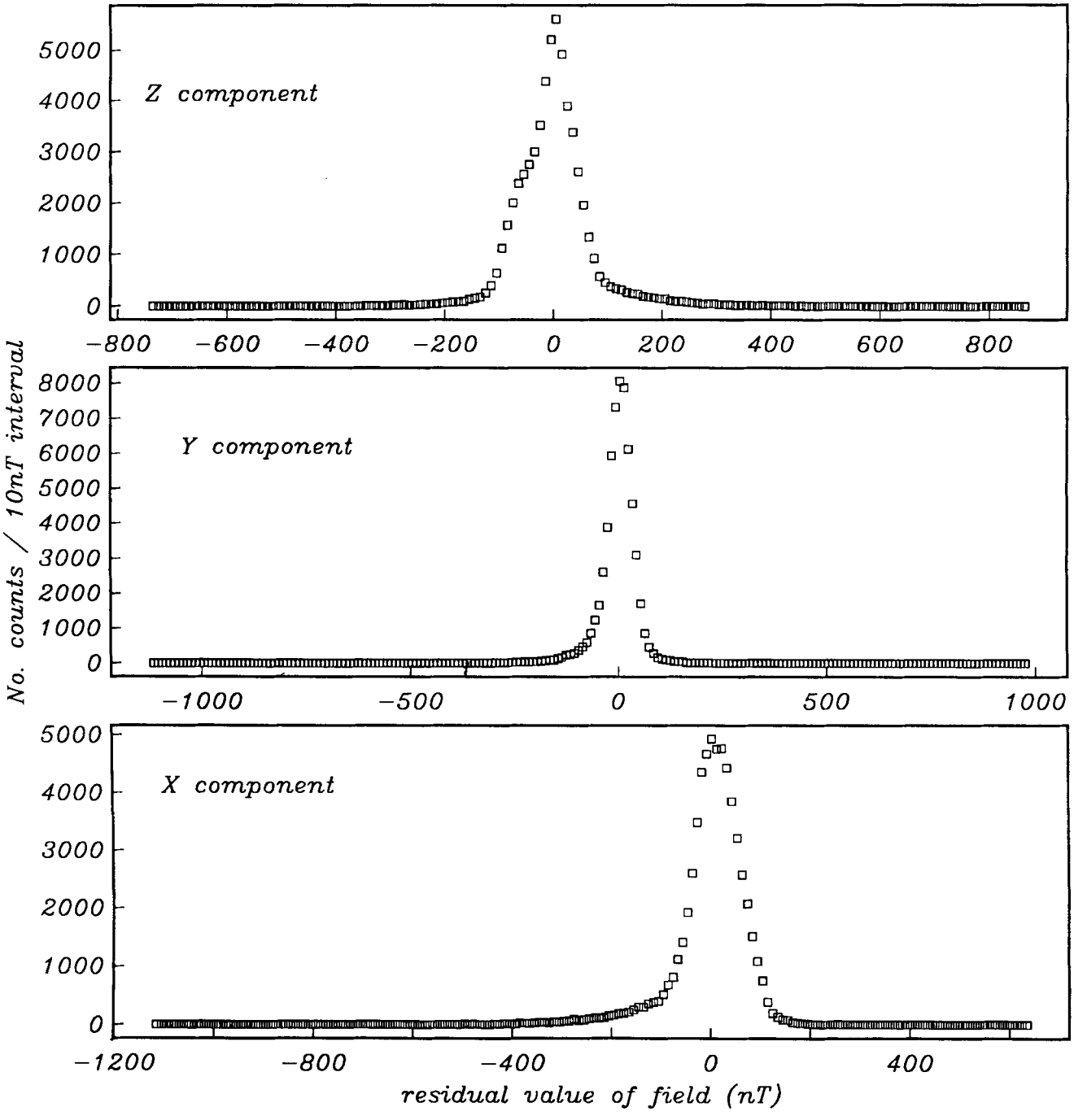

Figure 3. Histograms of residuals from LS fit of lines to Yellowknife data set.

the record. We regard these non-periodic variations as noise for the purposes of our model. They are not actually errors in the data, but simply generated by processes that we have not taken account of in our model. It is unlikely that a Gaussian model for the noise will be satisfactory in this case. Fig. 3 shows a histogram of the residuals from a LS fit to the data set of the lines of interest. These residuals are indeed non-Gaussian; the distribution is asymmetric and far too long-tailed. A statistical test such as the Kolmogorov–Smirnov test (Kendall & Stuart 1979, chapter 30) rejects the hypothesis that these data come from a normal distribution.

A further disturbing feature is apparent when we look at the time series of these residuals; they are clearly correlated. This correlation is due to the effect of geomagnetic storms, which arrive at irregular intervals and change the frequency content of the record. At first glance one might also think that the series is non-stationary, however, the records used here were sufficiently long that they may be regarded as stationary, but with patches of outliers generated by the geomagnetic storms. Fig. 4 shows the autocorrelation function for these residuals at lags up to 1900 hr. Clearly the assumption that the noise is uncorrelated is not a reasonable one. Much of the autocorrelation occurs at periods longer than those of interest, so we can attempt to remove it by

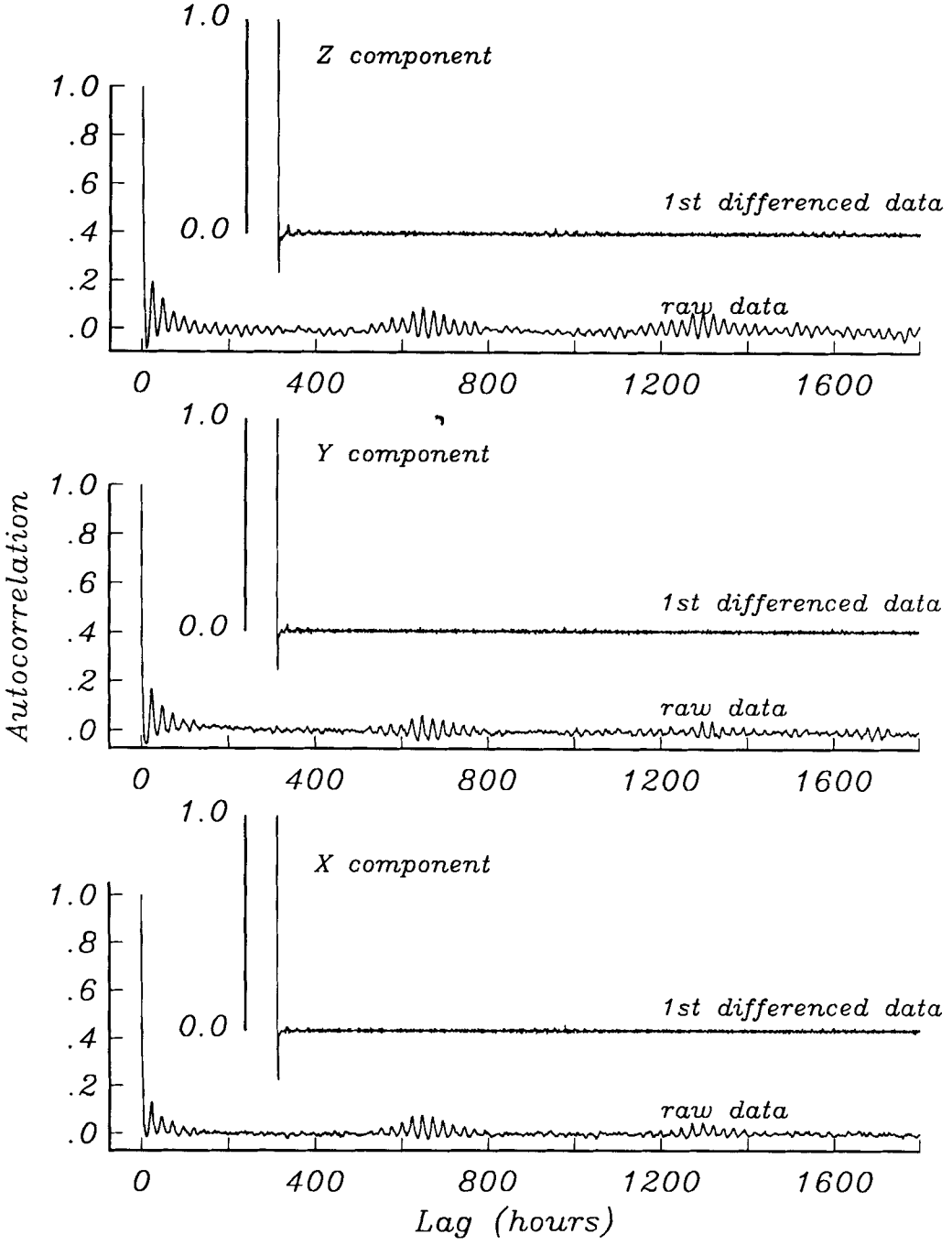

Figure 4. Autocorrelation functions for the residuals from a LS fit to the Yellowknife data. The lower curve in each part of the figure is for the unfiltered data, the upper part the autocorrelation for the residuals from the first differenced data.

prewhitening the data. Looking at an amplitude spectrum of the individual field components reveals that for sufficiently high frequencies it falls off approximately inversely with increasing frequency, with peaks superimposed at the periods of interest. For the noise to be uncorrelated, we require its spectrum to be flat over the frequency band of interest. This can be largely achieved for these data by differentiating the time series using a first differencing filter. The resulting autocorrelation functions for the least squares residuals are also shown in Fig. 4. The resulting amplitude estimates are easily compensated for this differencing procedure, since they change only by a factor of the period of the sinusoid in question. The residuals for the first differenced data are almost uncorrelated (and certainly a vast improvement over the original series); the autocorrelation has dropped to less than 0.05 by a lag of 4 hr, the shortest period of interest.

An example of the penalized spline fitting procedure described in the previous section is given in Fig. 5 using one year of hourly average values for the Z component of the

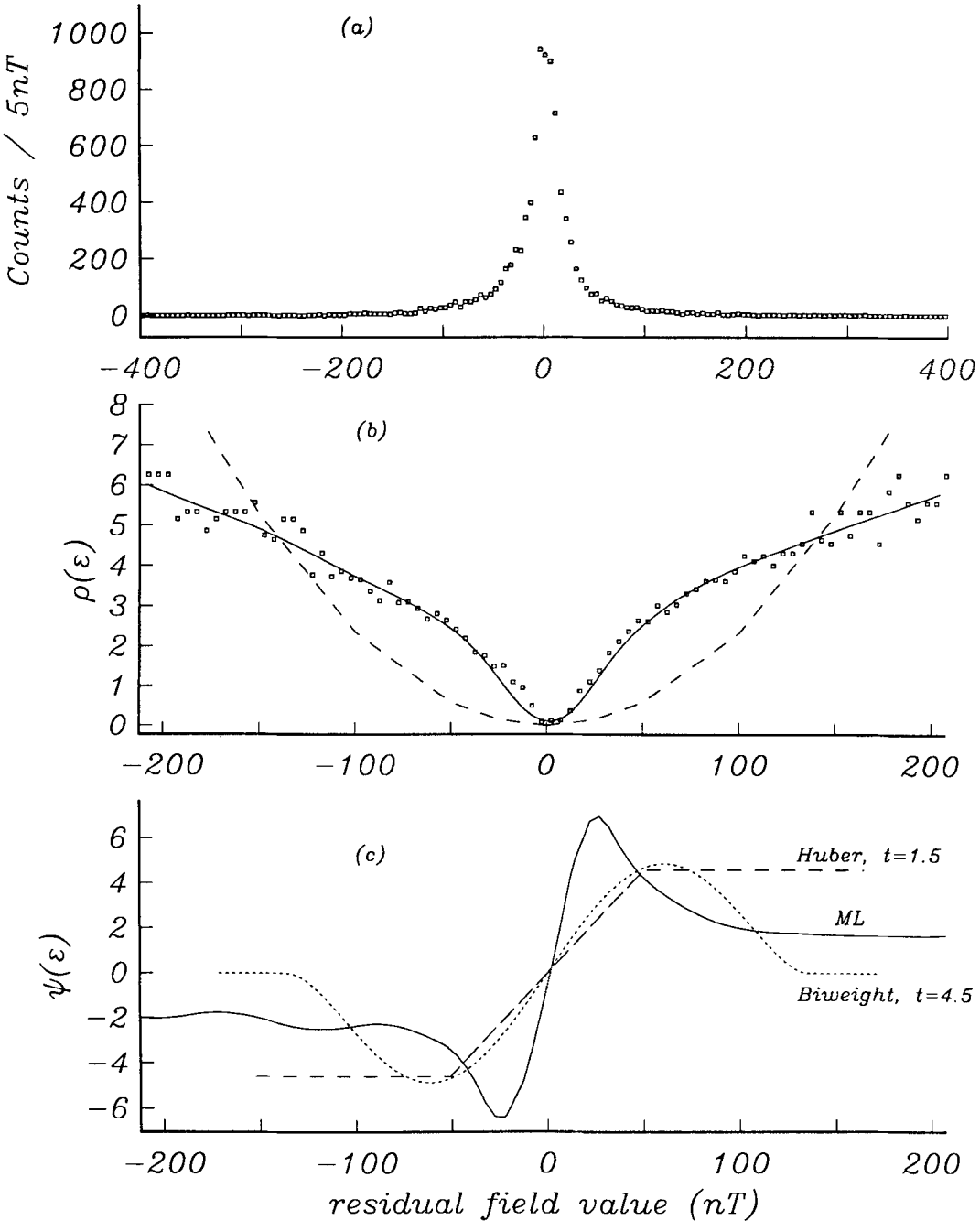

Figure 5. (a) Histogram of residuals from a LS fit to the first differencd Yellowknife Z component data. (b) Loss function for Yellowknife Z. Symbols represent the data, transformed from the histogram of (a). Solid line is the penalized cubic spline fit for the ML loss function. After initial ML estimation, subsequent iterations refine this estimate of the ML loss function. Dashed line is the loss function implied by a LS fitting procedure. (c) Solid line gives the ML influence function [the derivative of the loss function in (b)]. The dashed and dotted lines show the Huber and biweight influence functions respectively. The y-axis scale is arbitrary.

magnetic field at Yellowknife, i.e. 8760 data. Fig. 5(a) is a histogram of the residuals for the differenced data after least-squares fitting to the lines of interest. The distribution is very long-tailed, and slightly asymmetric: it falls off more steeply on the positive side than the negative. Fig. 5(b) shows $\rho(\epsilon)$ (squares) and the penalized spline $\hat{\rho}(\epsilon)$ (solid line) that has been fitted to it. In fitting the histogram data of Fig. 5(a) it is necessary to choose a cut-off point for the loss function estimation, beyond which the histogram is essentially zero. This point will depend on how many data are available, and how they are distributed; in this case the spline has been fit to the residuals between −212 and 238 nT, as the p.d.f. is essentially zero outside this region. In the ML estimation procedure the influence function is linearly extrapolated to zero outside this interval. The fitted function differs drastically from the sort of loss function used in least-squares estimation; *that* would be a parabola centred on the origin and is shown by the dashed line on the figure. This loss function's minimum is offset to the positive side of the origin, it initially rises more steeply than as the square of the residual, with the slope shallowing to be approximately linear in the size of the residual by about 80 nT (i.e. by about 2 standard deviations). The loss function is slightly asymmetric, reflecting the asymmetry in the distribution of the residuals. Fig. 5(c) shows $\hat{\psi}(\epsilon)$ and some of the influence functions of Fig. 1. Note the different shapes of the influence curves. Their absolute magnitudes will have no effect on the fit; what determines the result is the relative influence of data with different size residuals. All of the standard robust techniques will allow points at around two or three standard deviations from the mean to have a greater influence on the fit than ML estimation would. In the tails of the distribution the ML influence function is intermediate between the biweight and Huber functions. This is due to the approximately linear dependence of the loss function on residual size in this region (i.e. exponential tails of the p.d.f.).

A further advantage of this procedure is that it provides an estimate of location for the probability function for use in the ML estimation. This is particularly useful for this geomagnetic problem which has an asymmetric noise distribution. In this case the mean and mode of the distribution do not coincide and we use the mode as the estimate of location which will yield the ML solution. The p.d.f. model and influence function generated by this procedure are used in the ML estimation described in the previous section; then the residuals are recomputed and the process repeated until the residual distribution no longer changes. With the Yellowknife Z component data set the successive histograms of residuals were indistinguishable after two iterations.

COMPARISON WITH LS AND STANDARD ROBUST PROCEDURES

The ultimate justification for using the ML and p.d.f. estimation procedure should undoubtedly be that it provides parameter estimates superior in some way to those obtained by more conventional techniques such as ordinary LS or M-type robust estimation. It is of interest to compare the results obtained using the different methods and assess whether anything is gained by using the ML method.

Table 1. Line amplitude estimates for Yellowknife Z by LS and ML.

Method	Period hours	Amplitude ± 1σ nT	Bootstrap Average ± $1\sigma_B$ nT	σ_B/σ
LS	24.00000	33.01 ± 2.67	33.64 ± 2.57	0.96
ML		5.93 ± 0.96	5.89 ± 1.19	1.25
LS	12.00000	34.13 ± 1.34	33.92 ± 1.42	1.06
ML		19.03 ± 0.48	18.83 ± 0.73	1.52
LS	8.00000	3.16 ± 0.89	3.34 ± 0.74	0.83
ML		4.64 ± 0.32	4.62 ± 0.38	1.20
LS	6.00000	3.07 ± 0.67	3.12 ± 0.65	0.98
ML		2.11 ± 0.24	2.10 ± 0.32	1.35
LS	4.80000	2.60 ± 0.53	2.65 ± 0.60	1.12
ML		1.36 ± 0.19	1.35 ± 0.28	1.45
LS	4.00000	0.51 ± 0.45	0.68 ± 0.43	0.97
ML		0.44 ± 0.16	0.48 ± 0.23	1.45

Let us look again at the data from Yellowknife observatory. There are three components of the magnetic field measured at the site, and an estimate of the line amplitude and phase was made at the first six harmonics of the solar day. For the Z-component Table 1 lists those periods and the amplitudes of the lines estimated using both LS and ML (estimates were performed on the differenced data and then corrected back to give true line amplitudes). Fig. 6 shows the histograms of residuals for the Z component from a LS and a ML estimate of these lines over the interval from −200 to 200 nT. Note how the histogram changes shape with the method used. The ML histogram is narrower and peakier than that from LS, showing that more small residuals result from ML than from LS. In LS the outlying points tend to drag the fitted model towards them, thereby sacrificing the fit to the rest of the data. The mode of the ML residuals is at zero, whereas for LS the mean is at zero.

The asymptotic variance of the estimates decreases with successive iterations and the estimates of the line amplitudes change considerably as the noise distribution is approximated more closely. Fig. 7 shows the results of Table 1, the estimates of line amplitudes for the different methods. Some of these change drastically when the ML procedure is used, by far more than one would expect given the error bars, which are one standard error computed using (11). In some cases the line amplitudes do not change very much, but there is still a vast improvement in the confidence of the ML result compared with the others. Table 2 gives the asymptotic variances for the various methods as computed by (12). The ratios of these variances provide a measure of the relative efficiency of the different methods. The ML method provides a worthwhile improvement of a factor of about 8 over LS in the variance and thus in the confidence of the result. The standard robust techniques do not perform nearly as well.

Fitting lines at other observatories did not always yield such spectacular changes in the line amplitudes; 10 or 20 per cent changes were closer to the norm. The size of the change depends on how non-Gaussian the p.d.f.s are; if the p.d.f.s turn out to be Gaussian, then LS and ML estimation will yield the same result. However, in all cases the effect of using ML estimation was a substantial reduction in the asymptotic variance in the estimators, resulting in a great improvement in the confidence of the result. One question that should be addressed is to what extent the original assumptions about the data and model are likely to affect the results obtained. In particular the assumption that the noise distribution is independent and identically distributed

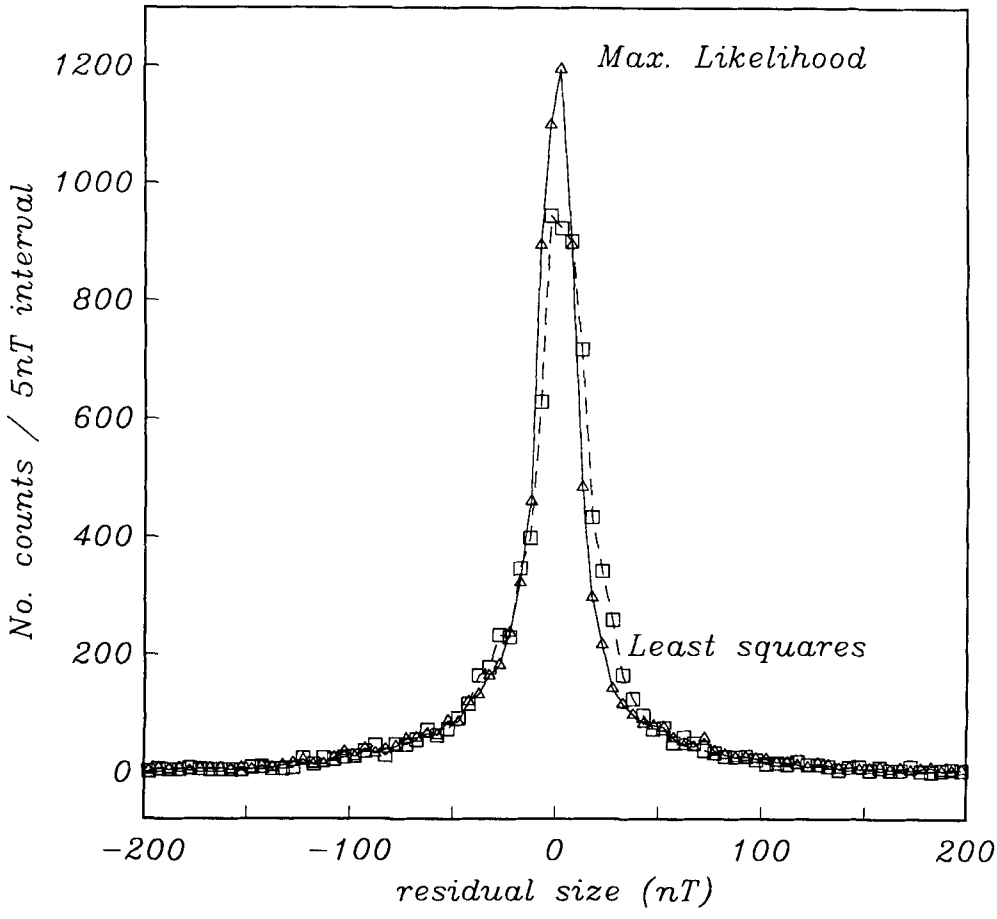

Figure 6. Comparison of the histograms of residuals for least-squares (dashed line and squares) and maximum likelihood (solid line and triangles) estimation.

for each datum is not exact, as can be seen from the autocorrelations for the differenced data. This may result in unduly optimistic values for the standard deviation in the parameters estimated (not only for the ML estimation, but also for the LS and robust procedures). One means of checking how much effect the assumption has, is by using a bootstrap technique to obtain alternative estimates for the standard deviations in the estimated parameters.

A review of bootstrap methods for obtaining standard errors, and confidence intervals is given by Efron & Tibshirani (1986). Let us suppose that the parameter estimates are distributed according to some unknown sampling distribution F and that the noise in the data y_i has a common distribution f as before. The bootstrap technique assumes that the empirical distribution of the data residuals is this distribution, f, and computes the standard deviation for the parameter estimates based on this assumption. The computation of the standard deviations is usually performed by a Monte Carlo type experiment, sampling from the empirical distribution $\hat{F}$ that approximates F. However, the method used here which computes the standard error in the parameters based on the estimate of the loss function ρ (and thus also f) could also be viewed as a sort of bootstrap method within a parametric framework. Its possible disadvantage is that the results may depend heavily on the initial assumptions about the data. This may be checked by performing a different type of bootstrap, which does involve a Monte Carlo type experiment. A random sample of size n is taken from the x_i values corresponding to the n first differenced data y_i, with replacement of data between sampling (i.e. repetition of data points occurs within the sample), and the desired parameters are computed for each sample of x_is and their associated y_is. If there are B bootstrap samples taken then the bootstrap estimate of the standard error in β will be given by

$$\hat{\sigma}_B = \left[\frac{\sum_{b=1}^{B} \{\hat{\beta}^*(b) - \hat{\beta}^*(.)\}^2}{B-1} \right]^{1/2}$$

$$\hat{\beta}^*(.) = \frac{\sum_{b=1}^{B} \hat{\beta}^*(b)}{B}.$$

As noted by Efron & Tibshirani (1986) this will not be sensitive to our initial assumptions in the ML estimation procedure, and the difference between the standard errors computed in this fashion and those computed using (12) will indicate how justifiable those initial assumptions were. In principle, if the number of bootstrap samples were allowed to approach infinity we could construct the true distribution function for the parameter estimates. For the large data set used in the geomagnetic example this would be prohibitively expensive; here we have simply tried to assess how reliable

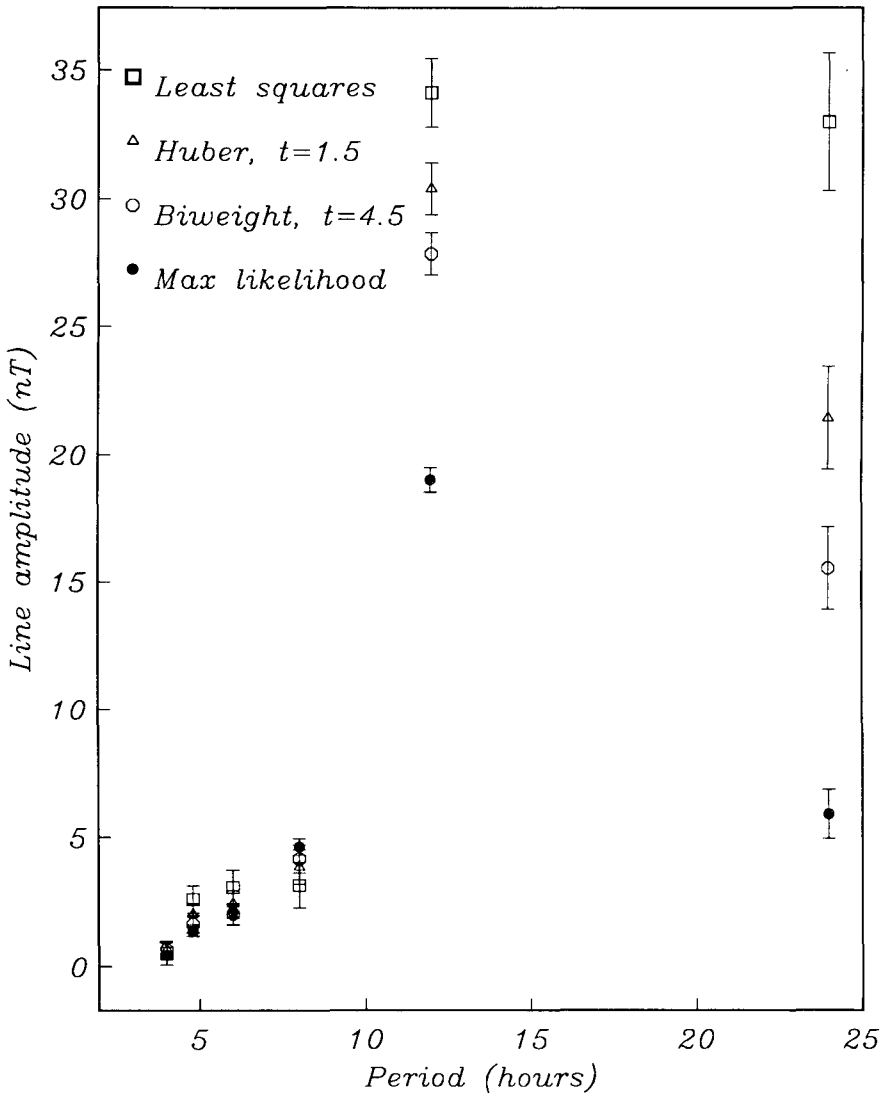

Figure 7. Line amplitudes obtained by the various methods discussed in the text. Error bars are one standard deviation as computed by (11).

the computed standard errors are, using 25 bootstrap samples from the original data. The results of this procedure are given in the last two columns of Table 1, where the average value of the parameter and σ_B are listed, as well as the ratio of the standard error computed using the bootstrap to the parametric estimate. It may be seen that for LS the bootstrap estimates are very similar to the parametric estimates. For ML the bootstrap estimates are consistently somewhat more variable than the parametric ones by a factor of one to one and a half. Nevertheless, the bootstrap ML errors are still substantially smaller than the standard errors obtained for LS, indicating that although the ML estimation is somewhat less robust to violations in our original assumptions about the data, it can still provide more efficient parameter estimates. These results must be regarded as tentative because of the small number of draws involved, however for smaller data sets, the method could provide a tractable means of defining the confidence limits for parameter estimates.

Table 2. Asymptotic variances computed for the various methods.

Method	Asymptotic Variance
LS	2132.3
Huber	1052.5
Biweight	833.9
ML	273.5

CONCLUSIONS

This paper has illustrated some of the differences between LS, ML and robust estimation procedures, by comparing the influence functions used in these procedures. From a theoretical point of view ML estimation is attractive, however, LS is still widely used because of its computational tractability, and because often little is known about the noise distribution. However, if the distribution of errors is known, it is in most cases straightforward to maximize the likelihood function and obtain the ML estimators for the parameters by solving iteratively the resulting non-linear

equivalent of the normal equations. This iterative procedure may be cast as a weighted LS problem, making the effort necessary to implement it minimal.

In the section on density function estimation a method is described, whereby the p.d.f. and loss function for a noise distribution can be obtained. When noise measurements can be obtained independently of the signal, the density function may be found directly (e.g. in controlled source electromagnetic sounding, where the noise arises from telluric currents), otherwise the noise distribution may be obtained in an iterative fashion, provided some assumptions are made about the form of the model. By using penalized least-squares splines to represent the ML loss function associated with any given noise distribution we are able to tailor the ML estimation procedure to any particular noise distribution.

The technique is illustrated with hourly values of geomagnetic observatory data, in which the amplitudes of a number of periodic processes were estimated. Geomagnetic storms and other phenomena in the record make the error distribution long-tailed, asymmetric and variable with location. By prewhitening the data it was possible to render the noise distribution largely uncorrelated, a necessary requirement for the ML estimation scheme presented here. A comparison of the ML method with standard and robust LS procedures shows that it can yield quite different results, both in the amplitudes for the fitted lines and in the decreased size of the confidence intervals. A bootstrap technique may be used to assess what effect any violation of our initial assumptions has on the reliability of the results. Although the LS estimates appeared more robust to the observed violations the ML results still compared very favourably in this example, as they always had smaller standard errors in the parameter estimates than for LS. For most geomagnetic data sets it is necessary to compute the loss function and p.d.f. and do the ML estimation once or twice before the distribution of residuals converges. How long each ML solution takes to converge using the W-estimation will depend on how non-Gaussian the p.d.f., f, is. In extreme cases eight or nine iterations in the W-estimation may be necessary; more typical distributions required two or three. The important distinction, however, is that unless the residuals are Gaussian, ML estimation is different from LS. The user must decide which estimate is preferable under the given circumstances.

ACKNOWLEDGEMENTS

Alan Chave suggested the original problem which motivated this work. Alan Chave, David Donoho, Bob Parker, John Rice and David Thompson took part in many useful discussions. Steven Constable and Philip Stark read the manuscript critically. An anonymous reviewer held me to higher standards than I would otherwise have achieved. I am grateful to them all for their assistance. Funding was provided by National Science Foundation grant EAR 84-126212.

REFERENCES

Banks, R. J., 1968. Geomagnetic variations and the electrical conductivity of the upper mantle, *Geophys. J. R. astr. Soc.*, **17,** 457–487.

Barnett, V. & Lewis, T., 1984. *Outliers in Statistical Data*, 2nd edn, Wiley, Chichester.

Bickel, P. J., 1982. On adaptive estimation, *Ann. Stat.*, **10,** 647–671.

Chave, A. D. & Filloux, J. H., 1984. Electromagnetic induction fields in the deep ocean off California: oceanic and ionospheric sources, *Geophys. J. R. astr. Soc.*, **77,** 143–171.

Constable, C. G. & Parker, R. L., 1988. Smoothing, splines and smoothing splines: their application in geomagnetism, *J. Comp. Phys.*, in press.

Currie, R. G., 1966. The geomagnetic spectrum – 40 days to 5.5 years, *J. geophys. Res.*, **71,** 4579–4598.

de Boor, C., 1978. *A Practical Guide to Splines*, Springer-Verlag, New York.

Efron, B. & Tibshirani, R., 1986. Bootstrap methods for standard errors, confidence intervals and other measures of statistical accuracy, *Stat. Science*, **1,** 54–77.

Good, I. J., & Gaskins, R. A., 1971. Non parametric roughness penalties for probability densities, *Biometrika*, **58,** 255–277.

Hampel, F. R., Rousseeuw, P. J., Ronchetti, E. M. & Stahel, W. A., 1986. *Robust Statistics, The Approach Based on Influence Functions*, Wiley, New York.

Huber, P. J., 1972. Robust Statistics: A review, *Ann. Math. Stat.*, **43,** 1041–1067.

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

Kendall, M. & Stuart, A., 1979. *The Advanced Theory of Statistics, Vol. 2, Inference and Relationship*, 4th edn, MacMillan, New York.

Larsen, J. C., 1968. Electric and magnetic fields induced by deep sea tides, *Geophys. J. R. astr. Soc.*, **16,** 47–70.

Malin, S. R. C. & Schlapp, D. M., 1980. Geomagnetic lunar analysis by least squares, *Geophys. J. R. astr. Soc.*, **60,** 409–418.

Montgomery, D. C. & Peck, E. A., 1981. *Introduction to Linear Regression Analysis*, Wiley, New York.

Priestley, M. B., 1981. *Spectral Analysis and Time Series*, Academic Press, London.

Reinsch, C., 1967. Smoothing by spline functions, *Numer. Math.*, **10,** 177–183.

Sellek, R. & Malin, S. R. C., 1982. Geomagnetic lunar analysis – the estimation of errors, *Geophys. J. R. astr. Soc.*, **70,** 793–796.

Silverman, B. W., 1982. On the estimation of a probability density function by the maximum penalised likelihood method, *Ann. Stat.*, **10,** 93–97.

Silverman, B. W., 1984. Spline smoothing: the equivalent variable kernel method, *Ann. Stat.*, **12,** 898–916.

Silverman, B. W., 1985. Some aspects of the spline smoothing approach to regression curve fitting, *J. R. Stat. Soc. B*, **47,** 1–52.

Silverman, B. W., 1986. *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, London.

Tapia, R. A. & Thompson, J. R., 1978. *Nonparametric Probability Density Estimation*, John Hopkins University Press, Baltimore.

Titterington, D. M., Smith, A. F. M. & Makov, U. E., 1985. *Statistical Analysis of Finite Mixture Distributions*, Wiley, New York.

Van Zijl, J. S. V., Hugo, P. L. V. & de Bellocq, J. H., 1970. Ultra deep Schlumberger sounding and crustal conductivity structure in South Africa, *Geophys. Prospect*, **18,** 615–634.