GEOPHYSICS, VOL. 38, NO. 5 (OCTOBER 1973), P. 826–844, 14 FIGS., 1 TABLE

ROBUST MODELING WITH ERRATIC DATA†

JON F. CLAERBOUT* AND FRANCIS MUIR‡

An attractive alternative to least-squares data modeling techniques is the use of absolute value error criteria. Unlike the least-squares techniques the inclusion of some infinite blunders along with the data will hardly affect the solution to an otherwise well-posed problem. An example of this great stability is seen when an average is determined by using the median rather than the arithmetic mean. Algorithms for absolute error minimization are often approximately as costly as least-squares algorithms; however, unlike least-squares, they naturally lend themselves to inequality or bounding constraints on models.

INTRODUCTION

The median and the mean are two kinds of statistical average. In a normal situation they behave in about the same way. At the present time, physical scientists almost always use the mean and, hence, tend to be unaware of the dramatic ability of the median to cast off the effect of blunders in the data. As an example, consider an expensive, all-day experiment which yields only one number for a result. On the first day, the result is 2.17, on the second day it is 2.14, and on the third and final day it is 1638.03. The mean of these results is 547.78 but the median (middle value) is 2.17. If you suspect a blunder on the third day you will obviously prefer the median. Statisticians call this the "robust" property of the median.

The objective of this paper is to show how many kinds of geophysical data fitting can be made to be robust. In particular, all the calculations we now do in solving overdetermined linear simultaneous equations by means of summed squared error minimization can be made robust, instead, by minimizing summed absolute values of errors. A computer algorithm to do this will be discussed. Computer time is comparable to that of least-squares methods. The algorithm solves a slightly broader class of problems than minimizing the summed absolute errors. Positive errors may be penalized with a different weight factor than negative errors. We call such an arrangement an asymmetric norm. A special case of an asymmetric norm is an inequality constraint. Inequalities or bounds may be applied to model parameters as well as measurement errors.

Perhaps we reveal a theoretician's bias when we speak of erratic *data*. An experimentalist could with equal validity claim that the data are fine, but the phenomenon they represent is far more complex than the theoretician either wants or is able to model. For example, when earthquakes are located by an untended computer which is fed from 100 telephone lines to remote seismometers, then the seismologist may be unable to make a noise model for all the various peculiarities of telecommunication difficulties and breakdowns. With robust modeling methods, we can often avoid the task of making a good noise model. The earthquake may be properly located even if it knocks down some of the telephone lines.

FIRST PRINCIPLES

First we will see why means and medians relate to squares and absolute values. Let x_i be an arbitrary number. We define m_2 by the value of

† Manuscript received by the Editor August 30, 1972; revised manuscript received January 5, 1973.

* Stanford University, Stanford, Calif. 94305.

‡ Chevron Oil Field Research Co., La Habra, Calif. 90631.

m which minimizes the sum of squared differences (called the L_2 norm) between m and x. We have

$$m_2 := m \left| \sum_{i=1}^{N} (m - x_i)^2 \quad \text{is min.} \right. \tag{1}$$

It is a straightforward task to find the minimum by setting the partial derivative of the sum with respect to m equal to zero. We obtain

$$0 = \sum_{i=1}^{N} 2(m_2 - x_i),$$

or

$$m_2 = \frac{1}{N} \sum_{i=1}^{N} x_i. \tag{2}$$

Obviously, m_2 is given by the usual definition of mean. Next let us define m_1 by minimizing the summed absolute values (called the L_1 norm). We have

$$m_1 := m \left| \sum_{i=1}^{N} | m - x_i | \quad \text{is min.} \right. \tag{3}$$

To find the minimum, we may again set the partial derivative with respect to m equal to zero;

$$0 = \sum_{i=1}^{N} \operatorname{sgn} (m_1 - x_i). \tag{4}$$

Here the sgn function is $+1$ when the argument is positive, -1 when the argument is negative, and for the moment undefined when the argument is zero. Equation (4) says that m_1 should be chosen so that m_1 exceeds x_i for $N/2$ terms; m_1 is less than x_i for $N/2$ terms; and if there is an x_i left in the middle, m_1 equals that x_i. This defines m_1 as a median. [For an even number N, the definition (3) requires only that m_1 lie anywhere between the middle two of the x_i.]

The number of additions required to compute the arithmetic mean of N numbers is $N-1$, where N is the number of points. The number of comparisons required to completely order a list of N numbers seems to be about $2N \ln N$ (Singleton, 1969), but complete ordering is not required for finding the median. Hoare (1962) provided an algorithm for finding the median which seems to require only about $(2+2 \ln 2)N$ comparisons.

Two other commonly known averages, which are of little use with most geophysical data, are the *mode* m_0, given by

$$m_0 := m \left| \sum_{i} (m - x_i)^0 \quad \text{is min,} \right. \tag{5}$$

where $0^0 = 0$ and $\alpha^0 = 1$, $\alpha \neq 0$, and

the *mid-range* m_∞, defined by the Chebyshev norm L_∞ as

$$m_\infty := m \left| \lim_{p \to \infty} \left(\sum_{i} (m - x_i)^p \right)^{1/p} \quad \text{is min.} \right. \tag{6}$$

The midpoint m_∞ bisects the distance between the extreme data points, thus minimizing the maximum error. In multiparameter problems, the L_∞ norm gives rise to "equal-ripple" approximations. Because L_1 and L_∞ stand on opposite sides of L_2, the philosophy behind L_∞ is somewhat the opposite of the philosophy behind L_1.

Now that we have seen the connection between means and squares and between medians and absolute values, it is natural to try to solve overdetermined simultaneous equations by minimizing absolute errors rather than squared errors. First we consider weighted medians. They are analogous to weighted sums.

Usually we take 2.17 as the median of the numbers (2.14, 2.17, 1638.03) because we implicitly apply weights (1, 1, 1). If we applied weights (3, 1, 1) it would be like having the numbers 2.14, 2.14, 2.14, 2.17, 1638.03 and the median would then be 2.14. A weighted median may be defined by the minimization

$$m_1 := m \left| \sum_{i} | w_i | | m - x_i | \quad \text{is min.} \right. \tag{7}$$

Obviously if the weight factors are all unity, this expression reduces to the earlier definition, whereas a weight factor equal to 3, for example, is just like including the same term three times with a weight of one. Figure 1 illustrates the definition (7) for a simple case. From Figure 1, it is apparent that a minimum is always found at a corner so the median can be taken equal to one of the x_i even if the weights are not integers. If the weights are all unity and there are an even number of numbers, then the error norm will be flat between the two middle numbers. Then any value in between satisfies our definition of median by minimizing the sum.

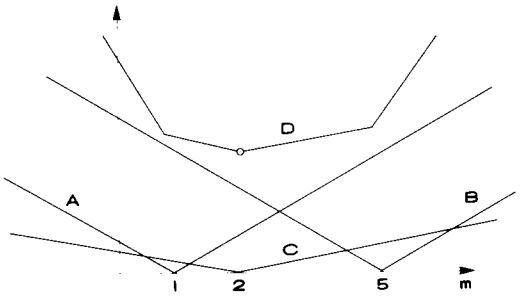

FIG. 1. A sum of weighted absolute value norms. The function labeled A is $.5|m-1|$, B is $.5|m-5|$, C is $.1|m-2|$, and D is the sum of A, B, and C. The sum D is minimized at $m=2$, a point which exactly solves $C=0=.1|m-2|$.

Let us rearrange (7) by bringing $|w_i|$ into the other absolute value function. We have

$$m_1 := m \left| \sum_i \left| |w_i| m - |w_i| x_i \right| \text{ is min} \right. \tag{8}$$
$$:= m \left| \sum_i |w_i m - w_i x_i| \text{ is min.} \right.$$

We will relabel the conventions of statistics to the usual conventions of simultaneous equations and linear programming. Let

$$a_i = w_i, \qquad d_i = w_i x_i, \tag{9a, b, c}$$

and

$$x = m.$$

With these new definitions, equation (8) becomes

$$x := x \left| \sum_i |a_i x - d_i| \text{ is min.} \right. \tag{9}$$

In other words, to solve the rank, one over-determined equations

$$[\mathbf{a}]x \cong [\mathbf{d}], \tag{10}$$

for x by minimizing the L_1 norm. This is, in effect, a weighted median problem. If (10) were solved by minimizing the L_2 norm (least squares), x would be the weighted average $x=(\mathbf{a}\cdot\mathbf{b})/(\mathbf{a}\cdot\mathbf{a})$.

Next we observe that the absolute value function is symmetric, i.e., $|e|=|-e|$. This property will not be required, so we will define the more general asymmetric norm with arbitrary upslope g_k^+ and downslope g_k^- shown in Figure 2. Note that a different penalty function may be applied for each error e_k in $e_k=d_k-\sum_j A_{kj}x_j$. Obviously, a sum of asymmetric penalty functions will be a piecewise linear function like the sum of absolute value functions in Figure 1. The *lower quartile* is like the median except that one-fourth of the data values lie below the quartile and three-fourths lie above. To find the lower quartile with an asymmetric norm, we would set $g_k^- = -3$ and $g_k^+ = +1$ for all k. Percentiles and other quantiles may be defined in a similar fashion.

An important property of the penalty functions we are dealing with is that they are convex downward. This means that sums of these functions are also convex. This property is sufficient to ensure that there is one unique minimum value for the error. The only possible nonuniqueness of solution arises if the bottom is flat.

The next step up the ladder of complexity is to consider two unknowns. The obvious generalization of (10) is

$$\begin{bmatrix} a_1 & c_1 \\ a_2 & c_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ a_k & c_k \\ \cdot & \cdot \\ \cdot & \cdot \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \cong \begin{bmatrix} d_1 \\ d_2 \\ \cdot \\ \cdot \\ d_k \\ \cdot \\ \cdot \end{bmatrix}. \tag{11}$$

We will assume the reader is familiar with the solution to (11) by the least-squares method. Solution by minimizing the sum of the absolute values of the errors begins in a similar way. We begin by defining the total error:

$$E = \sum_{k=1}^{N} |d_k - a_k x - c_k y|. \tag{12}$$

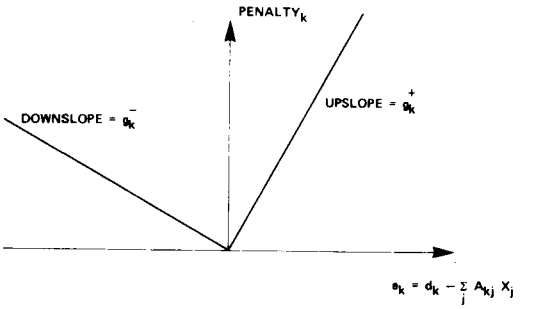

FIG. 2. The error norm is the sum of asymmetric penalty functions depicted.

Then we set the x derivative of the error equal to zero and the y derivative of the error equal to zero:

$$0=\frac{\partial E}{\partial x}=\sum_{k=1}^{N}-a_k \operatorname{sgn}(d_k-a_kx-c_ky), \quad (13a)$$

and

$$0=\frac{\partial E}{\partial y}=\sum_{k=1}^{N}-c_k \operatorname{sgn}(d_k-a_kx-c_ky). \quad (13b)$$

Now we run into a problem. If the sgn function always takes the value $+1$ or -1, then (13a) implies that the a_k may be divided into two piles of equal weight. Clearly many, indeed most, collections of numbers cannot be so balanced (e.g., if all the a_i except one are integers). The difficulty will be avoided if at least one of the equations of (11) is solved exactly so that sgn takes an indeterminate value for that term. Any algebraic confusion may be quickly dispelled by recollection of Figure 1 and the result that even with one unknown the minimum generally occurs at a corner where the first derivative is discontinuous. The same situation must again apply to (13b). The usual situation is that for N equations and M unknowns, precisely M of the N equations will be exactly satisfied in order to enable the error gradient to vanish at the minimum. Indeed, in the words of Gauss' *Theoria Motus Corporum Coelestium* which appeared in 1809 (Plackett, 1972):

> Laplace made use of another principle for the solution of linear equations, the number of which is greater than the number of unknown quantities, which had been previously proposed by Boscovich, namely that the differences themselves, but all of them taken positively, should make up as small a sum as possible. It can be easily shown, that a system of values of unknown quantities, derived from this principle alone, must necessarily (except the special cases in which the problem remains, to some extent, indeterminate) exactly satisfy as many equations out of the number proposed, as there are unknown quantities, so that the remaining equations out of the number proposed, as there are unknown quantities, so that the remaining equations come into consideration only so far as they help to *determine the choice.*

Today common usage in the field of linear programming is to refer to any nonsingular subset of M out of the N equations as a set of *basis equations*. The particular set of M equations which is solved when the error is minimized is called an *optimum basis*. Figure 3 shows some upper bounding fits of sums of sinusoids to a step. When M terms are used in the expansion, then the curve precisely fits the step (at least) at M points. Successive values of M are shown, odd values on the top row of graphs and even values on the bottom row. In each graph, precisely M points fit exactly (except the case $M=1$, which is degenerate and 20 points fit exactly). Surprisingly, where the model curve appears as a tangent to the data step there are always two adjacent points fitting exactly except at the ends of the interval.

EXAMPLES AND APPLICATIONS

Even the simplest solution in absolute value minimizations, namely, the median, can be expected to have many practical applications in exploration. We have many applications in which we sum seismic traces. In all of these we might consider whether the median would be better. Presently, before summing we must have editing programs to eliminate the frequently massive effects of air waves, ground roll, noise bursts, dead traces, ice breaks, etc. The editing is complicated enough and failures are not uncommon. Errors by the editors could be effectively eliminated if we used the median rather than the mean for trace averaging. In fact, with the median editors might not even be required.

The median has the property of stretch invariance. By this we mean that the pointer to the median points to the same place even if all numbers are scaled or are all exponentiated. If they are positive the numbers may all be squared, inverted, or have their logs taken. Essentially any monotonic transformation preserves order and leaves the median pointer unchanged. This property can be useful in experiments where the theory is unclear on how to average. For example, if an average sound velocity of a heterogeneous mixture is determined with the median, then it is unimportant whether the problem is parameterized with velocity, velocity squared, or inverse velocity (time average). Of course, a good theory for sound would be preferable.

The median may open some new possibilities in geophysical data analysis. In the past the forming of ratios of noisy data points was a commonly forbidden operation because the sum of such ratios is infinite if any one of the divisors is zero. In reality the difficulty lies not with the ratios but in the assumption that averaging must be done with sums. If the ratios are averaged with

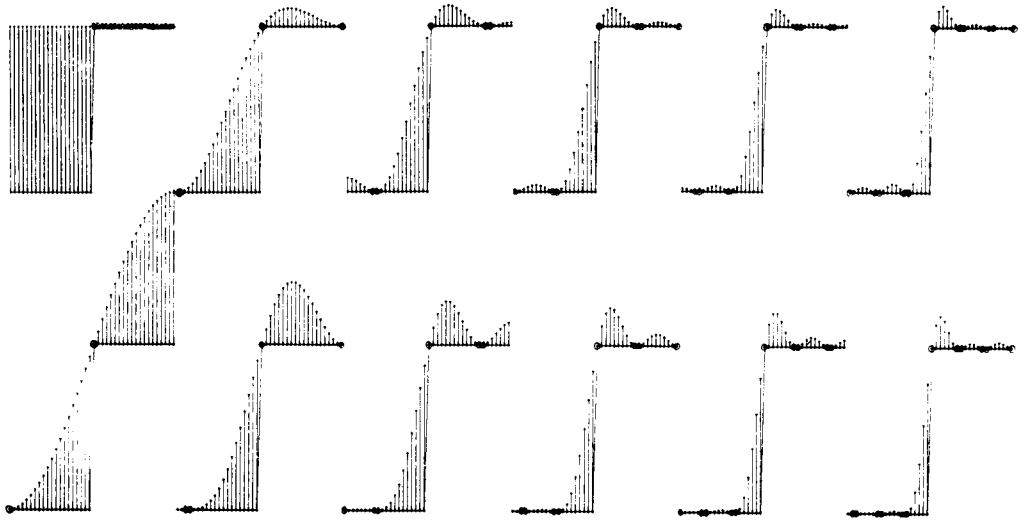

FIG. 3. Upper bounding fits to a step. We minimize

$$\sum_{i=1}^{N} |\, e(i) \,|$$

where $N=40$, $t=i-N/2-.5$, $e(i)$ negative for all i, and

$$e(i) = \text{step}\ (i - 20.5) - \sum_{\substack{k=0 \\ \text{even}}}^{M-1} a_k \cos \pi t/N - \sum_{\substack{k=1 \\ \text{odd}}}^{M-1} a_k \sin \pi t/N.$$

the median there is no problem. Notice also that the median of $\text{num}_i/\text{den}_i$ is the same as the inverse of the median of $\text{den}_i/\text{num}_i$.

Running means are often used for smoothing. In a running median one replaces each data value by the median value of it and its neighbors. A running median can be excellent for removing spikes in a time series. One problem with the running median which is not shared by the running mean is illustrated in Figure 4. In data which have a systematic variation, comparable to or greater than the random variation, the systematic variation should be removed before doing the running median; then it may be restored.

A problem not shared by the running mean arises when a desired smoothing window is wide enough that it includes trend or systematic variation of larger amplitude than the fuzzy noise which is to be smoothed. Then the effective width of the smoothing window is reduced. If a wide smoothing window is desired then the trend should first be removed.

Running medians may also be applied to complex data, but there are several options on how to proceed and should be tailored to the applications. Let $M_1(x_t)$ denote a running median on x_t, and $M_2(x_t)$ denote a running mean. If complex data are actually observed (as N_t+iW_t where N_t and W_t are the north component and the west component of earth tilt), then it may be suitable to apply the running medians directly to the real and imaginary parts, i.e., $M_1(N_t) + iM_1(W_t)$. On the other hand, in an application where a_k+ib_k represents a complex impedance as a function of frequency ω_k then it may be more suitable to take the logarithm first obtaining log amplitude and phase, i.e., $\log\,(a_k + ib_k) = \log\,|r_k| + i\phi_k$. Now the running median could be done separately on $\log r_k$ and the angles ϕ_k.

Notice that the same result is achieved if you smooth $\log r$, r, r^2, $1/r$, or $1/r^2$ with a running median, although drastically different results may occur if you smooth with a mean. For example, in magnetotellurics one observes a long time series of electric field E_t and another of perpendicular magnetic field H_t. These are Fourier transformed to complex numbers E_k and H_k, and the desired result is the ratio of the two. The problem is that because of randomness and noise some smoothing is necessary. Least squares leads one to the two different averages $M_2(E_k\overline{H}_k/M_2(E_k\overline{E}_k)$ and

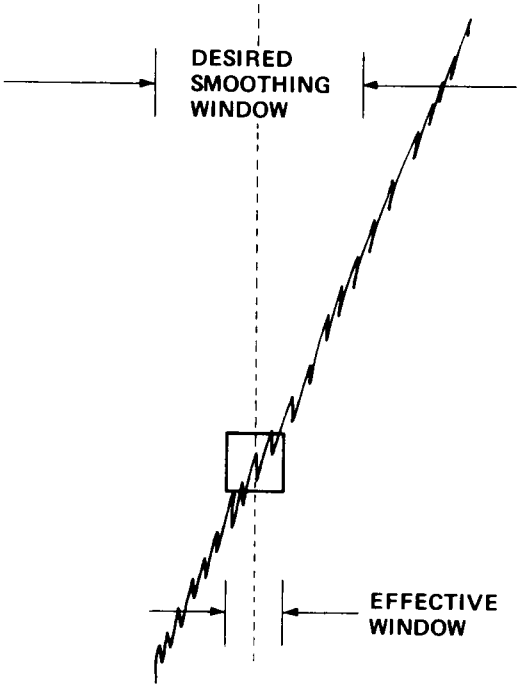

FIG. 4. Running median on fuzzy treading data.

$M_2(E_k\overline{H}_k)/M_2(H_k\overline{H}_k)$. Without noise these two would be inverses of one another and would relate to the electrical conductivity being measured. In the presence of noise the averages become biased differently and it is not clear how either relates to earth conductivity.

Alternately, one could consider smoothing log H_k/E_k with a running median and the exact inverse would be attained as from smoothing log E_k/H_k. A problem here is that when the data are sufficiently noisy, the 2π ambiguity in the phase may cause difficulty. A solution to this is to form, by finite differences, an approximation to $d/d\omega$ log $E/H=E_\omega/E-H_\omega/H$ and smooth it with a running median. As before, if there are any a priori or clearly observed trends, these should be removed before smoothing.

We have noted a curious fact about running medians. As indicated in Figures 5 and 6, the running median of a sin $(x)/x$ function has no side lobes at all if the window length is chosen equal to twice the zero crossing separation. This is not a special property of sin $(x)/x$, but a property stemming from the uniform spacing of zeros.

The asymmetric-linear norm can be specialized to the symmetric absolute value norm L_1 or it can be specialized to inequalities. In the first case, we have the usual least-squares applications, and in the second case we have the usual linear programming applications. An example of linear programming in GEOPHYSICS is in the editing of map data (Dougherty and Smith, 1966). There will undoubtedly be many important mixed applications (e.g., the location of earthquakes by minimizing summed absolute residuals with the inequality constraint that the earthquake must occur at a positive depth).

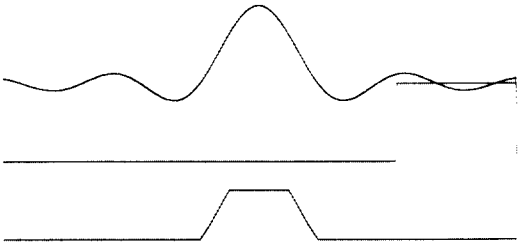

FIG. 5. Running median of sin $(x)/x$. Top is the sin $(x)/x$ function. Middle is the window of uniform weights used in a running median on the sin $(x)/x$ to give the running median at the bottom.

We should not overlook the possibility that there may be geophysical problems which deserve an asymmetric norm because they really do not fit into a least-squares, linear programming, or mixed framework. For example, consider the problem of determining the time of first arrival of a seismic wave on a seismogram as illustrated in Figure 7. Because of the presence of noise, the determination of the first arrival has a higher probability of being late than early. Thus, an asymmetric norm would be natural.

Another example is in the construction of depth to magnetic basement maps as indicated in Figure 8. The radii (diameters) of curvature of magnetic fields seen on the earth's surface are

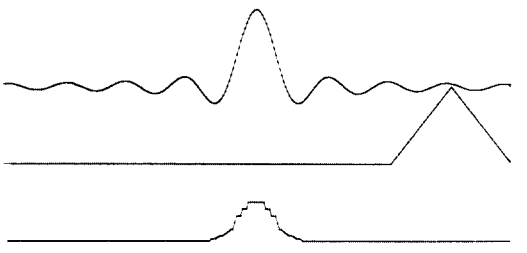

FIG. 6. Running median of sin $(x)/x$ with triangle weights.

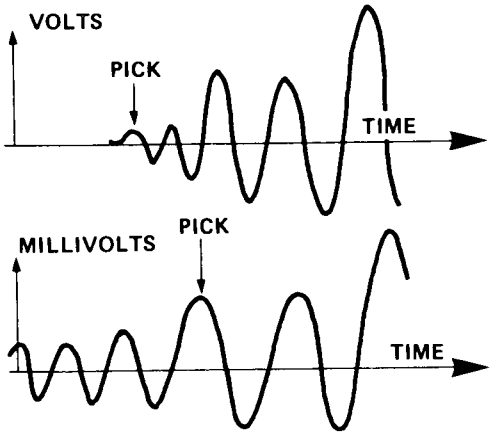

FIG. 7. In the presence of high ambient noise, a pick of first arrival time has a greater probability of being late than being early. This is a natural application for an asymmetric norm.

assumed to overestimate the depth to the top of magnetic monopole (dipole) sources. On first sight, this seems like a linear programming problem; however, the presence of nonbasement magnetic sources means that the field radius can be less than the depth to basement. Hence, depth should be taken as some low quantile rather than a lower bound of radii. This, incidentally, indicates how linear programming problems can be made more robust, namely, replace inequalities by a highly asymmetric norm.

Figure 9 illustrates the fitting of a sum of sinusoids to a step with norms of various asymmetries. Notice that even with infinite skewness the sinusoids still fit the step quite reasonably.

An illustrative example is the fitting of a straight line to scattered points. If there are only three points, we can quickly obtain a graphical solution. Let the points be denoted by (x_i, y_i), $i=1, 3$. Then we have three equations of the form $y_i \approx mx_i + b$ for the unknown slope m and unknown intercept b. If the absolute error is minimized, we know that there will exist an optimum basis, which means that two of the three equations will be exactly satisfied at the error minimum. In other words, the best line passes through two of the three points. Graphically we may connect all possible pairs of points by straight lines. Then we pick the line with the least error as illustrated in Figure 10.

From this example we see that when a traveler reaches a fork in the road, the L_1 norm tells him to take either one way or the other, but the L_2 norm instructs him to head off into the bushes. Likewise, a hunter when seeing two birds in the sky might not choose to shoot at the midpoint between them, especially if they are far apart. This is not to say that the L_1 norm is better than the L_2 norm; the L_1 norm is very different and can be much better than least-squares in some applications. The ubiquity of the square norm is explained by widespread acceptance of two questionable assumptions: 1) that the square norm is the only tractable norm, and 2) that most sensible (convex) norms can be expected to give about the same practical results. The latter assumption may be true when errors are small, in other words, when the signal-to-noise ratio is high.

Inconsistent data values are treated quite differently by the L_1 and L_2 norms. A common problem is the shifting of seismograms or cross-

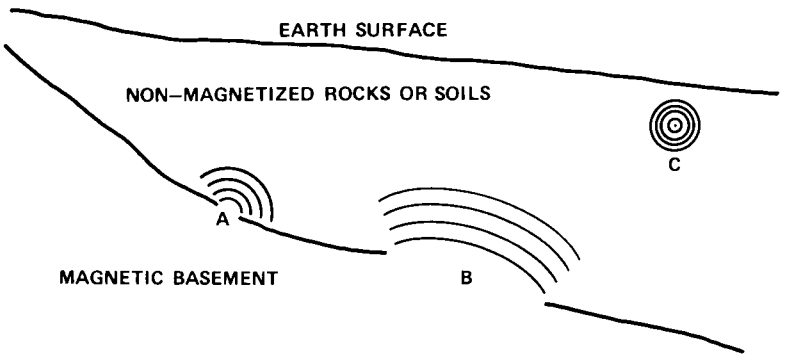

FIG. 8. The magnetic depth to basement may be found as a lower bound on the radius (diameter) of curvature of surface magnetic fields generated from basement monopole (dipole) sources. The source at A gives the correct depth. A distributed or deeper source at B causes an overestimate. A noise point at C indicates that the depth should be fit to a low quantile rather than a lower bound.

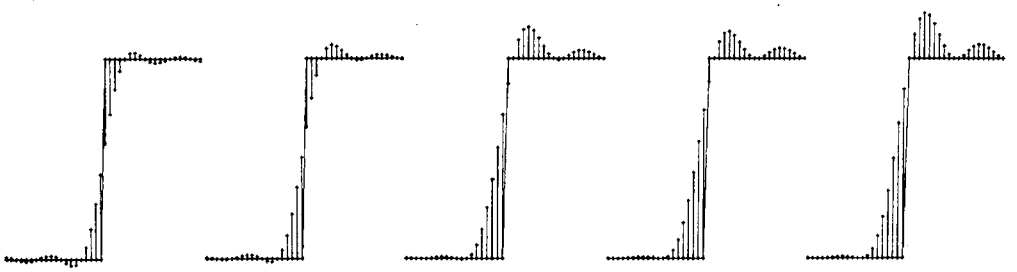

FIG. 9. Fitting of a few sines and cosines to a step by norms of increasing *skewness*. On the left the summed absolute error is minimized. On each step to the right the *skewness* is doubled. On the rightmost plot there are no model points below the step so the result is the same as if there were infinite *skewness*. In other words, the rightmost frame minimizes the area between model and data subject to the constraint that the model is always above the data.

correlation functions into best alignment. As illustrated in Figure 11, it is quite common to discover that the maximum of a crosscorrelation function cannot be unambiguously picked. Assume that the traces or crosscorrelations are to be aligned by placing a best fitting line through the maxima. Which of two ambiguous maxima should be included in the overdetermined set of equations? One answer is to include both (or all) of the ambiguous maxima. The L_1 norm will choose a line which picks either one or the other, or neither, when neither is consistent with the rest of the data. Least squares, on the other hand, when faced with two inconsistent data values (two linear equations which are the same except for the inhomogeneous part) effectively regards the two data values as one of double weight placed at the midpoint between the two. For the example at hand this is inappropriate because the midpoint is quite clearly *not* a maximum.

In many problems the square norm is the natural norm. This is often the case when measurements and errors have physical dimensions like voltage, current, force, or velocity. On the other hand, it is often unnatural to square things which are already positive like energy, power, mass density, compressability, probability, geometrical area, temperature, entropy, merchandise, etc. When such quantities occur as measurements, the asymmetric linear norm may well be the natural norm.

Now let us consider a numerical example in digital filter theory. Let the sampled waveform $(0, 0, \cdots, 0, 1, -1/2, 0 \cdots)$ be input into a filter with the two-point memory function (f_0, f_1). Then the output (omitting preceding and trailing zeros) is $(f_0, f_1 - f_0/2, -f_1/2)$. Suppose the filter is designed (numbers chosen for f_0 and f_1) so the output is a good approximation to $(1, 0, 0)$. Then the filter (f_0, f_1) is called a zero delay inverse filter to the filter $(1, -1/2)$ and equations for choosing f_0 and f_1 are·

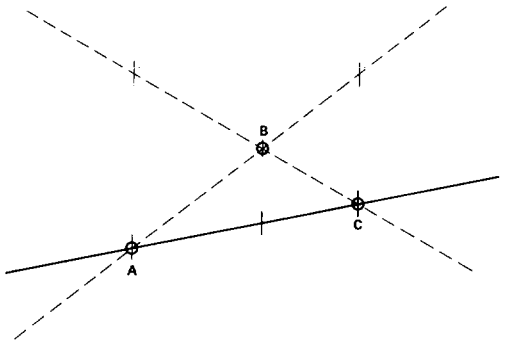

FIG. 10. The best fitting straight line passing through points A, B, and C under absolute error minimization (error measured along the vertical in this case) passes through points A and C.

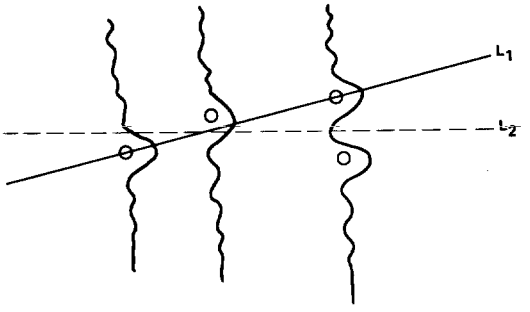

FIG. 11. Treatment of inconsistent data points. The maxima on each trace are picked. On the right-hand trace the maximum is ambiguous so two maxima are picked. Then L_1 and L_2 best fitting lines are made to pass through the picked points. The least-squares line L_2 tries to fit the midpoint between the ambiguous maxima. The least absolute value error line L_1 tends to pick the one of the ambiguous points which is most consistent with the rest of the data.

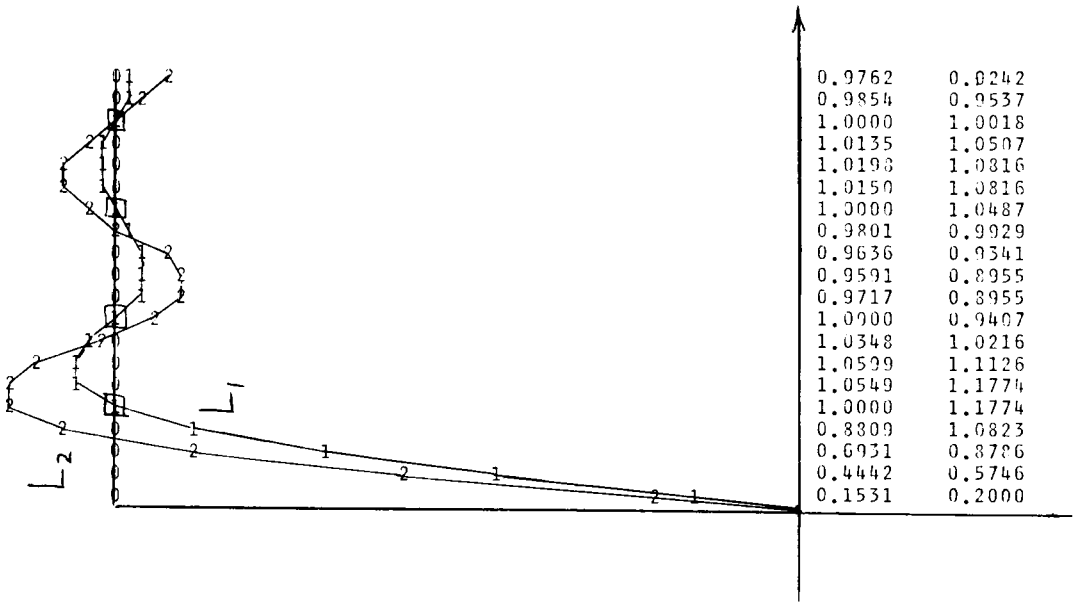

FIG. 12. Best fitting L_1 and L_2 sums of 4 sinusoids to a step. The L_2 fit is best near the discontinuity where squared error is high. The L_1 fit minimizes the area in error and is better than the L_2 fit away from the discontinuity. The L_1 approximation fits exactly at the points in the squares.

$$\begin{bmatrix} 1 & 0 \\ -.5 & 1 \\ 0 & -.5 \end{bmatrix} \begin{bmatrix} f_0 \\ f_1 \end{bmatrix} \approx \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}.$$

The least-squares solution is readily found to be $(f_0, f_1) = (20, 8)/21$, and the error is found to be $(1, -2, -4)/21$. The sum squared error is $1/21$ and the sum of the absolute values of the error is $1/3$. To obtain the least absolute values solution we solve each pair of equations and then find the one with minimum error. This is tabulated as

pair	(f_0, f_1)	*error terms*
1 and 2	(1, .5)	(0, 0, +.25)
1 and 3	(1, 0)	(0, +.5, 0)
2 and 3	(0, 0)	(1., 0, 0)

Thus we see that the best answer is (1, .5) which, surprisingly, turns out to be the truncation of the exact transform answer

$$1/(1 - .5z) = 1 + .5z + .25z^2 + \cdots.$$

Now let us look at the two-term zero delayed inverse to the nonminimum phase filter $(1, -2)$. This is defined by

$$\begin{bmatrix} 1 & 0 \\ -2 & 1 \\ 0 & -2 \end{bmatrix} \begin{bmatrix} f_0 \\ f_1 \end{bmatrix} \approx \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}.$$

The least-squares solution is $(f_0, f_1) = (5, 2)/21$ with an error sequence $(16, -8, 4)/21$, whereas the least absolute values solution is $(f_0, f_1) = (0, 0)$ with an output (0, 0, 0). The zero answer for the L_1 filter could be interpreted as a failure to attempt to solve the problem. This is not unreasonable in view of the theoretical impossibility of finding a convergent realizable inverse to a nonminimum phase filter.

Besides model fitting, the method of least squares has considerable theoretical importance. For example, it may be used to expand a step function into a linear combination of sinusoidal functions. The L_1 norm can be used in the same way. The results are somewhat different. Figure 12 illustrates that there is less propagation of error away from the discontinuity with the L_1 norm.

To the extent that a sum approximates an integral, the total error for L_1 in an example like Figure 12 is represented by the area between the data curve (the step) and the model curve.

Geometric area is invariant under rotation and translation. Thus, to the extent that summation approximates integration, the L_1 solution is invariant to physical translations and rotations of data and models. This property is not shared by L_2. L_2 has the property of invariance under rotations in the N-dimensional observation vector space. In Figure 12 we have a two-dimensional physical space and an $N=20$-dimensional observation vector space.

Presently the design of deconvolution filters is based on least squares. They could be designed with an absolute value criterion. If this were done the filter design would be far less affected by patches of highly unpredictable signal. When some event is unpredictable, and gives a big prediction error, least-squares tries harder to predict this event than the more predictable portion of the data. This is undesirable. The absolute value error criterion ensures an equal effort on predictable as on unpredictable portions of the trace. Thus, unpredictable events of interest, like primaries, can be expected to stand out larger after absolute value deconvolution. Since the L_1 method does not use an autocorrelation, as does least squares, we note that the stationarity assumption, always a questionable one with reflection data, is not implicit.

Before going any further, let us see how a model parameter may be kept positive. Consider for example the problem $\mathbf{Ax}=\mathbf{d}$ and $x_3 \geqq 0$ where the number of unknown M is 3. It may be set up as

$$\left[\begin{array}{c} \mathbf{A} \\ \hline 0 \quad 0 \quad 1 \end{array}\right] \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \approx \left[\begin{array}{c} \mathbf{d} \\ \hline 0 \end{array}\right].$$

Asymmetric weights will be set up for the error in the last "equation." The downward gradient for the error on the inequality when $x_3<0$ is $g^-_{(n)}$ It cannot be set equal to minus infinity in a computer program, but any arbitrarily large number, say -10^{30}, will do. Actually all that is required is that

$$-g^-(N) > \sum_{i=1}^{N-1} g^+(i) \left| a_{3i} \right| ,$$

because this will make the inequality downslope greater than all possible combined upslopes in $\mathbf{A}$. In our program the upslope error gradient for the inequality $x_3 \geqq 0$, $g^+(N)$ cannot be set precisely equal to zero, although any small value as 10^{-30} will do. The reason is that if $g^+(N)$ is set equal to zero, the last equation, when the inequality is slack, is confused with other equations which are candidates for entry into the basis. The last equation should be in the basis if $x_3=0$, but not if $x_3>0$.

With some special preparation equations which are underdetermined may also be solved. To $\mathbf{Ax}=\mathbf{d}$ we append some equations which ensure that the columns of $\mathbf{A}$ are linearly independent. For the 3×3 case,

$$\left[\begin{array}{ccc} & \mathbf{A} & \\ \hline a_1 & 0 & 0 \\ 0 & a_2 & 0 \\ 0 & 0 & a_3 \end{array}\right] \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \cong \left[\begin{array}{c} \mathbf{d} \\ \hline 0 \\ 0 \\ 0 \end{array}\right].$$

The last 3 equations will tend to drive $\mathbf{x}$ to zero. We will choose a_1, a_2, and a_3 very small so that the tendency toward zero is very weak and only strong enough to overcome the fact that $\mathbf{A}$ is not of sufficient rank to uniquely determine $\mathbf{x}$ all by itself. For example, if the first two columns of $\mathbf{A}$ are identical, we have equality in

$$\begin{bmatrix} \mathbf{A} \end{bmatrix} \begin{bmatrix} +\infty \\ -\infty \\ 0 \end{bmatrix} = \begin{bmatrix} \mathbf{0} \end{bmatrix}.$$

More generally, we are discussing matrices $\mathbf{A}$ for which $\mathbf{Ay}=\mathbf{0}$ has a nonzero solution for $\mathbf{y}$. In such cases algorithms which have not prepared for underdetermined sets tend to have a very large amount of $\mathbf{y}$ mixed in with their solution $\mathbf{x}$ to $\mathbf{Ax} \cong \mathbf{d}$. This is because although columns of $\mathbf{A}$ may be linearly dependent, they generally appear to be independent (though just barely) when viewed with the finite precision of a computer. Hence, the values a_1, a_2, and a_3 cannot be chosen to be arbitrarily small because they must overcome calculation precision errors in manipulation of A. We should choose

$$a_j = 10^{-5} \sum_i | A_{ij} | .$$

Perhaps we should solve all sets of equations as though they were underdetermined. Underdetermination can be observed by seeing some $a_j x_j = 0$ equations in the final basis or by the occurrence of some zero components in the solution **x**-vector. These zero components suggest that certain physical parameters cannot be determined with the given precision of **A**. Indeed, instead of using 10^{-5} in defining the a_j, there may be some other numbers which are indicative of the known precision of A_{ij}. In earthquake epicenter location, the A_{ij} derive from travel-time tables which, in turn, come from either long experience or theoretical velocity models. In either case, a precision can be assigned and it isn't as great as single-precision arithmetic.

Step-wise regression is a method of iteratively eliminating those model parameters which contribute least to fitting the data. A similar operation can be done simply by increasing the a_j in unison until the desired number of parameters x_j become zero (they become zero because as the a_j are increased, more of the $a_j x_j = 0$ equations come into the basis). We have used this to determine which coefficients in a prediction filter are the really important ones.

Many people contend that all geophysical problems are underdetermined because the earth is described by a continuum of unknowns, whereas we have only a finite number of measurements. In this point of view, the observations represent constraint equations, and some type of smoothness criterion is required to obtain a unique solution. Needless to say, the selected smoothness criterion has a profound effect on the results. The choice of the norm also has a strong effect on the results and the interpreted resolution of the results. The simplest smoothness criterion is that the solution **x** vector have minimum length. We set up the overdetermined simultaneous equations

$$\begin{bmatrix} \mathbf{A} \\ \hline \epsilon\mathbf{I} \end{bmatrix} [\mathbf{x}] \cong \begin{bmatrix} \mathbf{d} \\ \hline 0 \end{bmatrix} .$$

Here $\epsilon\mathbf{I}$ is a small constant times the identity matrix. Obviously, the bottom block of equations is trying to say $\mathbf{x} = \mathbf{0}$. If ϵ is taken small enough, all of the equations of **A**, if they are consistent, will be in the final basis. From the point of view of a computer program, ϵ is irrelevant; the constraint equations can simply be kept in the basis. Upon solving the set of overdetermined equations, generally there are exactly M equations in the final basis. Among these must be all K of the constraint equations. Among the equations $\epsilon\mathbf{I}\mathbf{x} \approx 0$ there must be $M-K$ in the basis and K not in the basis. This means that many, at least $M-K$, of the components of **x** vanish; at most K do not vanish. The situation is depicted in Figure 13a for $K=3$.

In deconvolving any observed seismic trace, it is rather disappointing to discover that there is a nonzero spike at every point in time regardless of the data sampling rate. One might hope to find spikes only where real geologic discontinuities take place. Perhaps the L_1 norm can be utilized to give an output trace like Figure 13a.

Another smoothness criterion is that the solutions x_i tend to a constant (as much as the constraint equations will allow). This is implemented by replacing the matrix **I** in $\epsilon\mathbf{I}$ by a matrix with the first difference operator (1, −1) along its main diagonal. This means that for all smoothness equations in the basis $x_i = x_{i+1}$, but for those K not in the basis, $x_i \neq x_{i+1}$, as illustrated in Figure 13b. Likewise having the second difference operator (1, −2, 1) on the main diagonal will amount to finding piecewise linear functions which satisfy the constraint equations as illustrated in Figure 13c.

Of course we can use any difference equation we wish to define the class of functions which we are fitting to our data. Inequalities may be used to ensure that the discontinuities are of appropriate sign, for example, with (1, −2, 1), inequalities on the discontinuities would ensure a convex or concave solution. Notice that there seems to be no intrinsic limit to the resolution attainable. This result stands in stark contrast to the L_2 norm in which **x** is made up of a linear combination of the rows of **A**. In the view of Backus and Gilbert (1967) these rows are regarded as the windows through which one can see **x**. If these windows, or more precisely, all linear combinations of the rows, lack deltaness (the ability to concentrate at one spot), the resolution is said to be poor. This measure of

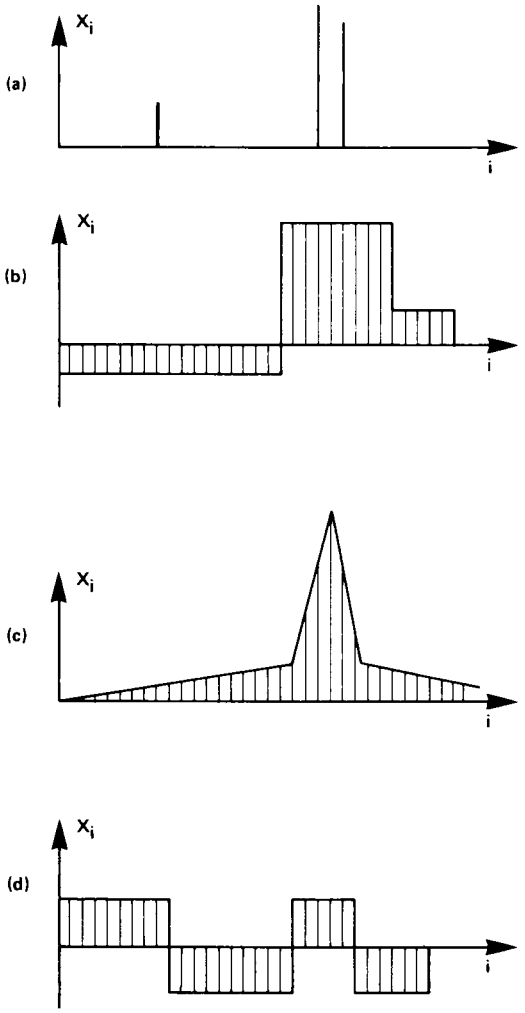

FIG. 13. Solutions to highly underdetermined asymmetric-linear norm problems where the smoothness criterion is taken to be minimization of the magnitude of (a) components of x, (b) first differences on x, (c) second differences on x, and (d) Chebyshev norm of x.

resolution is independent of the data. With asymmetric linear norms, the resolution is data dependent.

Let us consider an example in which the resolution becomes infinitely good if certain data values occur. Suppose the mass density as a function of radius inside a sphere is to be determined from the measured values of total mass, radius, and moment of inertia. If a data value of zero is found for the moment of inertia, then all the mass would be driven to the center of the sphere. In this example, the high resolution appears to be a result of the inequality constraints which computationally are a natural subset of asymmetric norms. The same resolution would result from least squares augmented by inequality constraints (quadratic programming); however, here again the resolution of the experiment becomes data dependent.

The Chebyshev norm L_∞ was not recommended for use on geophysical data; however, it might sometimes be appropriate for smoothing geophysical models. Recall that the Chebyshev norm of a vector (the infinite root of the sum of infinite powers of components) is the absolute value of the component of maximum magnitude. Therefore, we can easily solve L_∞ problems with asymmetric linear norm methods. This will be illustrated by the minimization of

$$E(x) = \|\mathbf{Ax} - \mathbf{d}\|_{L_1} + \|\mathbf{x}\|_{L_\infty}.$$

We begin by defining a new variable b (for biggest). We can arrange things so that b is the Chebyshev norm of $\mathbf{x}$ by setting up the inequalities

$$x_i + b \geq 0$$

and

$$x_i - b \leq 0$$

and then minimizing b. For the example where $\mathbf{x}$ has two components, the underdetermined set looks like

$$\begin{bmatrix} & \mathbf{A} & \mathbf{o} \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \\ 0 & 0 & \lambda \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ b \end{bmatrix} \cong \begin{bmatrix} \mathbf{d} \\ \\ \mathbf{o} \end{bmatrix}.$$

As with L_1 norm-smoothing criteria, if there are more smoothing equations than constraint equations, many of the smoothing equations will be in the final basis. This is illustrated in Figure 13d where most of the x_i are at the bounds b; typically only K would lie in between the bounds.

In time series analysis Widrow et al (1967) developed a simple method for least-squares filter adaptation to changing input data. Their method becomes even simpler with the absolute value norm. Instead of adjusting filters by a correction

which is proportional to the previous input times the present error, the correction is proportional to the previous input times the sign of the present error. Naturally this desensitizes the correction to large bursts of error.

Another application to time series analysis is a modification of the Burg algorithm (Ulrych, 1972) of spectral analysis by fitting prediction filters without end effects to finite data segments. The essence of the algorithm is computing a reflection coefficient c (in statistics called the partial autocorrelation coefficient) by the solution of the overdetermined set

$$\begin{bmatrix} \mathbf{e}^+ \\ \mathbf{e}^- \end{bmatrix} c \cong \begin{bmatrix} \mathbf{e}^- \\ \mathbf{e}^+ \end{bmatrix}$$

Here $\mathbf{e}^+$ and $\mathbf{e}^-$ are prediction error sequences in the forward and backward direction for filters of length M, and c is used with the Levinson recursion to generate a filter of length $M+1$. Burg's method is to solve the overdetermined set by means of least squares, obtaining

$$c = \frac{2(\mathbf{e}^+ \cdot \mathbf{e}^-)}{\mathbf{e}^+ \cdot \mathbf{e}^+ + \mathbf{e}^- \cdot \mathbf{e}^-} .$$

We propose to solve the overdetermined set using instead the absolute value norm. We have established that the absolute value norm also ensures $-1 \leqq c \leqq +1$, thereby preserving minimum phase. If the data consist of somewhat predictable noise along with occasional unpredictable bursts of signal, we expect L_1 to exhibit improved noise predictability. Also, we expect the estimated spectrum to be more related to the noise than to the signal plus noise.

The "flat bottom" nonuniqueness of skew norms is delightful for problems whose answers may not be uniquely determined from the data. Least squares would force such problems to have unique answers. A one-dimensional, everyday life example would be someone, a poor adder, who decides to verify his financial balance by working it out in four separate tries and then taking the median. If he obtains the same answer three or more times, then the median of the four numbers has a unique value. If he obtains four different answers then he knows he does not yet have enough information for a unique median and he can take appropriate action.

An important higher-dimensional example in seismology is the estimation of source and receiver time corrections. Here one has a set of observed traveltimes from the ith source to the jth receiver. After known systematic geometrical and velocity effects are removed, one has the time residual matrix t_{ij}. Then, near-source traveltimes s_i and near-receiver traveltimes r_j are estimated from the t_{ij} by minimizing the error e_{ij} in

$$e_{ij} = t_{ij} - s_i - r_j .$$

It is readily apparent that a trivial nonuniqueness arises in that an arbitrary constant added to all the s_i and subtracted from all the r_j will give the same residuals. What is not apparent (in fact its discovery amazed us), is that there is more nonuniqueness lurking in this problem. This will be illustrated numerically. Absolute error minimization could have reduced a 3-by-3 matrix of t_{ij} to the e_{ij} residual matrix

$$e_{ij} = \begin{bmatrix} 0 & -12 & 4 \\ 17 & 0 & 0 \\ 0 & 10 & 0 \end{bmatrix} .$$

As expected, there are 5 zeros representing the 5 independent unknowns of the 6 unknowns. Note that $\sum |e_{ij}| = 43$. Now modify source and receiver times by applying $+12$ to row 1 and -12 to column 1. We have

$$\begin{bmatrix} 0 & 0 & 16 \\ 5 & 0 & 0 \\ -12 & 10 & 0 \end{bmatrix} ,$$

still with $\sum |e_{ij}| = 43$. Now apply $+12$ to row 3 and -12 to column 3. We have

$$\begin{bmatrix} 0 & 0 & 4 \\ 5 & 0 & -12 \\ 0 & 22 & 0 \end{bmatrix} .$$

Furthermore, we can generate an infinite set of e_{ij} (and hence source and receiver corrections) all with the same $\sum |e_{ij}|$ by taking the given 3 and forming any convex combination (weighted combination where each weight is positive and the weights sum to one). From this example, one might conclude that it is better to use least squares to ensure a unique answer.

On the other hand, the existence of a sizeable nonuniqueness with absolute error minimization leaves the uncomfortable feeling that the unique-

ness of least squares is not real; perhaps it is only an artifact of the worst data point. The more prudent procedure would seem to be to examine the size and shape of the space of nonunique solutions. Then if it is unacceptably large, perhaps additional data can be found or perhaps further assumptions (like spatial smoothness or minimum $\sum|u_i|+|v_j|$) will make it unique.

Generally, if a system of equations has a unique L_2 solution, then the L_1 solution set is a bounded convex polyhedron whose vertices or extreme points correspond to particular subsets of equations in the system. The volume of this polyhedron gives some information on the resolvability of the solution which will be useful when the data consist of nearly correct values mixed with incorrect values (outliers). Then the polyhedron will be tiny if none of its vertices involve outliers.

Usually, however, the volume of the polyhedron will be far too optimistic an estimate of resolvability. For example, the standard error for an average of N Gaussian random numbers is expected to be much larger than the separation between the middle two numbers. Another example would be if the time residual matrix had turned out to be

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 7 & -11 \\ 0 & -3 & 8 \end{bmatrix}.$$

Then it would be unique, as the median of an odd number of points is unique. Just because it is unique you wouldn't say that the resolution of u_i and v_j is good. Estimates of the L_1 norm equivalent of standard error, the practical measure of uniqueness, are considered in the next section.

STATISTICAL ASPECTS

A property of the median is that for random variables with a symmetrical probability density function, the median m_1 coincides with the mean m_2. Also, the expected value of the sample median (the average value of many medians each estimated from a finite sample of data) equals the mean. Another important item is the expected variance of the sample median. In other words, for a sample of N points, the estimated median $\hat{m}_1$ is likely to differ from the true median m_1. As N tends to infinity, $\hat{m}_1$ gets closer to m_1. This is made specific and quantitative by defining the variance of the sample median $V(\hat{m}_1-m)$ by the following expectation $\mathcal{E}$:

$$V(\hat{m}_1 - m) = \mathcal{E}(m_1 - \hat{m}_1)^2.$$

This is a standard calculation in statistics. Asymptotically for Gaussian random variables and large N it becomes

$$V(\hat{m}_1 - m) = \frac{\pi}{2N}.$$

This is just slightly worse than the variance of the sample mean which is

$$V(\hat{m}_2 - m_2) = N^{-1}.$$

The best situation for the method of least squares is known to be when the errors have a Gaussian distribution. Then the median requires $\pi/2$ times as many data points to achieve the same standard deviation σ as the mean. In other words, for the same number of data points the standard deviation σ of the sample median will be about 25 percent greater than that of the mean. On the other hand, for non-Gaussian errors the standard deviation σ for the estimated median can be infinitely smaller than that of the mean.

Another interesting statistical property associated with L_1 is that the total error function contains everything in the probability function. By this we mean the following: Suppose the random numbers x_i are drawn from an amplitude density function $P(x)$. We previously defined the error norm function for the median as

$$E(\mu) = \sum_i |\mu - x_1|.$$

Now to avoid any confusion, let us redefine this as $\hat{E}(\mu)$, because it is estimated from a finite sample of data points

$$\hat{E}(\mu) = \sum_i |\mu - x_i|.$$

Now we will define $E(\mu)$ as an integral over the ensemble rather than a sum over the sample:

$$E(\mu) = \int_{-\infty}^{+\infty} |\mu - x| P(x)dx.$$

Differentiating this twice we obtain

$$\frac{\partial E}{\partial \mu} = \int P(x) \operatorname{sgn}(\mu - x)dx,$$

$$\frac{\partial^2 E}{\partial \mu^2} = \int P(x)\partial\delta(\mu - x)dx,$$

and

$$\frac{\partial^2 E}{\partial \mu^2} = \partial P(x),$$

which shows that differentiating the error norm enables one to calculate the probability function. It doesn't work this way with least squares. Then it is easy to show that the error norm may be deduced solely from the mean and variance of the probability density. Obviously, then, the probability density cannot be deduced from the error norm. A consequence of this is that surfaces of constant error, $E(\mu) = \text{const}$, are always ellipsoids with least squares but they will be somewhat more elaborate, perhaps highly skewed prismoids with asymmetric linear norms.

Integrating Poisson's equation $\partial^2 E/\partial\mu^2 = \partial P$ once, we see that the integral of the probability function is half the error gradient. This would seem to imply that a confidence region (integral of probability) could be defined as any region in which the error gradient has a magnitude less than a certain value. We might well suspect that in a multiparameter problem as $Ax \approx d$ the confidence region for x is where the error gradient vector has a magnitude less than a certain value.

In earlier examples, we expressed the idea that the solution is nonunique if the error function has a flat bottom. Then the region of uncertainty is where the error gradient vanishes. Now we are saying that the region of uncertainty is really the larger region in which the error gradient magnitude is sufficiently small. How small is sufficiently small? For the median of a sample of 100 random numbers, we expect about 50 percent probability that the true ensemble median lies between the 40th and 60th percentile. The reason is that the square root of 100 is 10 and 50 ± 10 is 40 or 60. Thus in the absence of a rigorous theory we believe that the 50 percent confidence region in a multiparameter problem is where the error gradient vector has a length less than the square root of the number of degrees of freedom, namely $(N/M)^{1/2}$.

Now let us determine the size, orientation and shape of the confidence region thus defined. In the problem of $\mathbf{Ax} \approx \mathbf{d}$, where one minimizes the sum of absolute values of $\mathbf{d} - \mathbf{Ax}$, let us suppose that we wish to discover the likely error in the first component x_1 of the unknown vector $\mathbf{x}$. To do this we solve two separate problems, one where we append $\mathbf{Ax} \approx \mathbf{d}$ with $a_1 x_1 \approx +\infty$, and one where we append it with $a_1 x_1 \approx -\infty$. The numerical value of a_1 should be typical of the elements of A_{i1}. The weight factors in the appended equations would be chosen $\mathbf{g}^+ = (N/M)^{1/2}$ and $\mathbf{g}^- = -(N/M)^{1/2}$. In other words, we would solve the $+$ problem and the $-$ problem:

$$\begin{bmatrix} \mathbf{A} \\ \hline a_1 \quad 00\cdots \end{bmatrix} [\mathbf{x}\pm] \cong \begin{bmatrix} \mathbf{d} \\ \hline \pm\infty \end{bmatrix},$$

where

$$\mathbf{g}^+ = \begin{bmatrix} 1 \\ 1 \\ \cdot \\ \cdot \\ \cdot \\ \sqrt{N/M} \end{bmatrix}, \quad \text{and} \quad \mathbf{g}^- = -\mathbf{g}^+.$$

We obtain vectors $\mathbf{x}^+$ and $\mathbf{x}^-$. From these vectors we see not only a range in the first component x_1 but also how the other components have reacted to the statistical force we have applied to x_1. Likewise we may apply $\sqrt{N/M}$ forces to the other components of the $\mathbf{x}$ vector to see how far they move.

ALGORITHM

An algorithm for L_1 norm minimization was given by Barrowdale and Young (1965). Their program seems to require computer time which is dependent on at least one term proportional to N squared, where N is the larger dimension of the coefficient matrix. This makes it more uneconomical than least squares in situations where N is large. We therefore developed the algorithm presented here.

We first considered an iterative weighted least-squares method with weights at the kth stage proportional to the inverse of the absolute values of the residuals at the $k-1$th stage. A problem with such an approach is that M of the weights must tend to infinity as a basis is

Table 1. Number of iterations required to fit N pseudorandom points in a plane to a straight line. The number of iterations seems to be increasing more slowly than $\log_2 N$. The computed time is in seconds on the IBM 360-67

N	ITERATIONS	LOG_2N	TIME
16	4	4	.06
32	4	5	.10
64	3	6	.16
128	5	7	.50
256	5	8	.94
512	7	9	2.83
1024	8	10	5.43
2048	8	11	11.13

attained. It was not clear how to do this in a rapid manner leading to the unique solution. Another method we considered and tried is a gradient descent method. The difficulty with this is that descent quickly brings one to an edge where gradient computation with the sgn function becomes ambiguous; the gradient then coincides with the edge. This led to the method to be described, which geometrically is to follow an edge (intersection of hyperplanes) to the point where the error is minimum (intersection with a new hyperplane). We then drop one of the old planes and make a new line out of the intersection of the remaining old hyperplanes and the new hyperplane. We follow this new line to a new minimum and repeat until motion ceases. The calculation takes a finite number of steps, but the number will not be known until the solution is found. The operation of forming a line and moving on it to a local minimum takes about $NM+M^2+4N$ operations and about $3M$ lines or iterations is typical for the cases we have studied. See Table 1 and Figure 14.

This $3NM^2$ for asymmetric linear norms may be compared to NM^2 for the squared norm. We will now construct an algebraic description of the geometric operations. The position $\mathbf{x}$ on a line

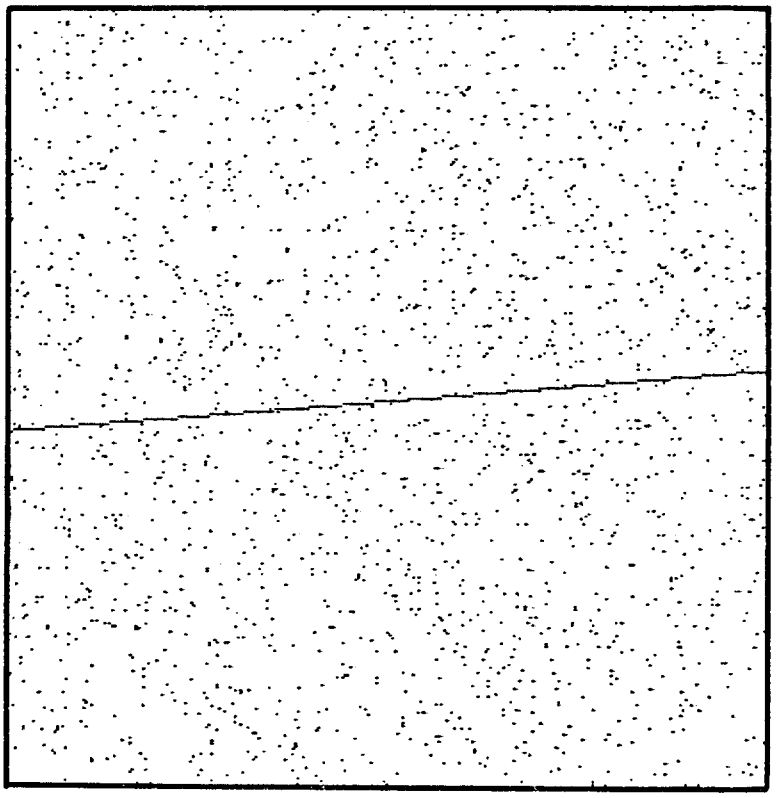

FIG. 14. The best fitting straight line to 2048 pseudorandom points. The absolute error minimized is measured vertically from each point to the line.

through $\mathbf{x}_0$ can be indicated by a scalar parameter t. The direction of the line can be specified by an M-component vector $\mathbf{g}$. Then any point $\mathbf{x}$ on the line may be represented as

$$\mathbf{x} = \mathbf{x}_0 + \mathbf{g}t. \tag{14}$$

Inserting (14) into the overdetermined set

$$\mathbf{A}\mathbf{x} \cong \mathbf{d}, \tag{15}$$

we obtain

$$\mathbf{A}(\mathbf{x}_0 + \mathbf{g}t) \cong \mathbf{d} \tag{16a}$$

and

$$(\mathbf{A}\mathbf{g})t \cong \mathbf{d} - \mathbf{A}\mathbf{x}_0. \tag{16b}$$

Defining $\mathbf{w}$ and $\mathbf{f}$ by

$$\mathbf{w} = \mathbf{A}\mathbf{g} \tag{17a}$$

and

$$\mathbf{e} = \mathbf{d} - \mathbf{A}\mathbf{x}_0, \tag{17b}$$

equation (16b) becomes

$$[\mathbf{w}]t \cong [\mathbf{e}]. \tag{17}$$

If we solve (17) by minimizing the summed absolute errors, we also obtain the minimum error along the line in (16a). But (17) is the weighted median problem discussed earlier.

Our method of solution to the median problem follows Hoare's (1962). A trial basis equation is picked, say $w_k t = e_k$. Then t is taken to be e_k/w_k. The equations are split into three groups, those with positive, negative, and zero residuals. If weights in the zero group can swing the balance of positive versus negative either way, t is the median. Otherwise we must pick a new trial basis equation from the stronger of the positive or negative group. The size of the group being inspected rapidly diminishes. When the right value for k is found, the kth equation in both (17) and (15) is satisfied. The kth equation is now considered to be a good candidate for the basis and we will show how to pick the vector $\mathbf{g}$ and continue to satisfy the kth equation (stay on the kth hyperplane) as we adjust t in the next iteration.

Now we need a set of basis equations. This is a set of M equations which is temporarily taken to be satisfied. Then as new equations are introduced into the basis by the weighted median solution, old equations are dropped out. The easiest strategy is merely to drop out the one which has been in longest. Let us denote our basis equations by

$$\mathbf{A}'\mathbf{x}_0 = \mathbf{d}'. \tag{18}$$

$\mathbf{A}'$ is a square matrix. The inverse of the matrix $\mathbf{A}'$ will be required and will be denoted by $\mathbf{B}$. Now suppose we decide to throw out the pth equation from the basis matrix $\mathbf{A}'$. Then for $\mathbf{g}$ we select the pth column of $\mathbf{B}$. To see why this works, note that since $\mathbf{A}'\mathbf{B}=\mathbf{I}$, the M-vector $\mathbf{A}'\mathbf{g}$ will now be the pth column from the identity matrix. Therefore, in the N-vector $\mathbf{w}=\mathbf{A}\mathbf{g}$, there is a component equal to $+1$, there are $M-1$ components equal to zero, and there are $N-M$ other unspecified elements. If the kth equation in (15) or (17) has been kept in the basis (18), then the kth equation in $\mathbf{A}\mathbf{g}t = \mathbf{d}-\mathbf{A}\mathbf{x}$ now reads

$$\text{zero } t = \text{zero}. \tag{19}$$

The left zero is an element from the identity matrix, and the right zero is from the statement that the kth equation is exactly satisfied. Clearly we can now adjust t as much as we like to attain a new local minimum and the kth equation will still be exactly satisfied. There is also one equation of the form

$$\text{one } t = \text{zero}. \tag{20}$$

It will be satisfied only if t is zero. Geometrically this means that if we must move to get to a minimum then this equation is not satisfied so we are jumping off from this hyperplane. This equation is the one leaving the basis. Of course if t turns out to be zero then it reenters the basis. The foregoing steps are iterated until such a time that for M successive iterations the equation thrown out of the basis by virtue of its age has immediately reappeared because $t=0$. This means that the basis can no longer be improved and we have arrived at the optimum basis and the final solution.

One of the peculiarities of the FIFO (first in first out) method just described of removing equations from the basis is that the error may stay constant over several iterations because the equation being removed from the basis immediately reenters. A gradient method was devised to pick the best equation to remove from the basis. Although the gradient method reduces the number of iterations required, unfortunately it approximately doubles the effort per iteration. The value of the gradient method thus becomes more ap-

parent in problems in which after a solution is obtained, the problem is perturbed and solved again. With the FIFO method, at least M iterations are required to be sure that none of the basis members must be replaced. This would make it as costly as least squares even if the perturbation did not introduce new basis equations.

With the gradient method, the fact that the basis remains the same is determined and the new exact solution may be found in one iteration. If only one basis equation needs to be changed this will often proceed and become verified in just two iterations. The gradient method for choosing a basis equation to be eliminated proceeds as follows.

First we will develop an expression for the error gradient in the vicinity of a point $\mathbf{x}_0$. Define $\mathbf{t}$ as a vector with all zero components except for the value t_k in the kth component corresponding to the equation leaving the basis. Moving from $\mathbf{x}_0$ in a direction and distance given by $\mathbf{t}$, we have $\mathbf{x}=\mathbf{x}_0+(\mathbf{A}')^{-1}\mathbf{t}$. Now we develop an expression for $\partial E/\partial t_k$ in order to see which equation to drop from the basis for fastest descent. We write out $E(t) = \sum |e_i| = \sum (\text{sgn } e_i)e_i$ in an expanded matrix form using the convention that double prime refers to those equations which are *not* in the basis:

$$E(t) = (\mathbf{h}, \text{sgn } \mathbf{e}'') \cdot \left\{ \begin{bmatrix} \mathbf{d}' \\ \mathbf{d}'' \end{bmatrix} - \begin{bmatrix} \mathbf{A}' \\ \mathbf{A}'' \end{bmatrix} [\{\mathbf{x}_0+\mathbf{A}'^{-1}\mathbf{t}\}] \right\}. \quad (21)$$

If $\mathbf{t}=0$ then $\mathbf{h}$ is ambiguous, since $\mathbf{e}'=\mathbf{0}$, and irrelevant, since it forms an inner product with the zero vector $\mathbf{d}'-\mathbf{A}'\mathbf{x}_0$. If we take $\mathbf{t}=(\pm t_1, 0, 0, \cdots)$ where $t_1>0$, there is error in the first component of $\mathbf{d}'-\mathbf{A}'\mathbf{x}_0$, so we may take $\mathbf{h}$ to be $\mathbf{h}=(\mp 1, 0, 0, \cdots)=\mp\delta(1)$. Thus, for t_k small and positive we have the gradient row vector (with a component for each k):

$$g_k^+ = \left.\frac{\partial E}{\partial t_k}\right|_{t_k>0} = [-\delta(l), \text{sgn } (\mathbf{e}'')] \begin{bmatrix} -\mathbf{I} \\ -\mathbf{A}''\mathbf{A}'^{-1} \end{bmatrix}, \quad (22)$$

or more simply

$$g_k^+ = 1 - [\text{sgn } (\mathbf{e}'')]\mathbf{A}''(\mathbf{A}')^{-1}, \quad (23)$$

and a like expression for negative t_k,

$$g_k^- = -1 - [\text{sgn } (\mathbf{e}'')]\mathbf{A}''(\mathbf{A}')^{-1}. \quad (24)$$

The optimum basis will be finally attained when for each k it is found that g_k^+ has the opposite sign as g_k^-. This is achieved when $-1 \leqq \text{sgn } (\mathbf{e}'')\mathbf{A}''(\mathbf{A}')^{-1} \leqq +1$. To reach the optimum as quickly as possible it seems reasonable to start off in the direction of greatest slope. First, exclude those directions with g_k^+ and g_k^- of opposite sign. Then, of each pair $|g_k^+|$ and $|g_k^-|$, exclude the larger because convexity implies descent in the direction of decreasing gradient. Select the direction of the largest remaining magnitude for fastest descent.

In programming this method we quickly discovered a degenerate case. Occasionally, when the M basis equations are satisfied there will be a few other "tag-along" equations which are satisfied too because they are linear combinations of basis equations. It is necessary to note that moving from $\mathbf{x}_0$ not only gives an error gradient from the dropped basis equation but may also introduce an addition to the gradient from the "tag-along" equations.

Degeneracy comes about when more than the expected M basis equations turn out to be satisfied. Of course this should never happen with "real" data but it happens very quickly, as in Figure 3, with integer test case data. Let us see how this creates a problem. Suppose we are iterating along with M basis equations. By casting out one equation at a time we have M directions in which to try to descend. If descent does not occur then we are at the bottom. Now if we discover that we are at an $\mathbf{x}_0$, where $M+1$ equations are exactly satisfied, then we must cast out all possible pairs of 2 from the $M+1$ equations to ensure descent if possible. In general, this can become quite involved and in practical cases may nearly always be unnecessary. Degeneracy is treated in all the standard linear programming texts.

In the hope of developing a fast general-purpose algorithm, we finally came to the following conclusions: Each application is likely to be best served by a different algorithm. A sure, but inefficient, method is to use standard linear programming packages as described in the next section.[1]

[1] One of the authors (JFC) expects to make available his program on receipt of a stamped, self-addressed envelope.

RELATION TO LINEAR PROGRAMMING

Any linear programming (LP) problem can be reposed as an asymmetric-norm problem. Likewise any asymmetric-norm problem can be posed in LP form. To see this, we split the unknowns $\mathbf{x}$ into two positive parts $\mathbf{x}=\mathbf{x}^+-\mathbf{x}^-$, likewise the errors $\mathbf{e}=\mathbf{e}^+-\mathbf{e}^-$. The overdetermined simultaneous equations are now written as equality constraints;

$$[\mathbf{A}\ -\mathbf{A}\ \mathbf{I}\ -\mathbf{I}]\begin{bmatrix}\mathbf{x}^+\\ \mathbf{x}^-\\ \mathbf{e}^+\\ \mathbf{e}^-\end{bmatrix}=[\mathbf{d}], \qquad (25a)$$

with the positivity constraints

$$\begin{bmatrix}\mathbf{x}^+\\ \mathbf{x}^-\\ \mathbf{e}^+\\ \mathbf{e}^-\end{bmatrix}\geq \mathbf{o}. \qquad (25b)$$

From the point of view of LP, equations (25a) and (25b) are constraint equations in the unknowns $\mathbf{x}^+$, $\mathbf{x}^-$, $\mathbf{e}^+$, and $\mathbf{e}^-$, and the objective function to be minimized is

$$\min = (\mathbf{o},\ \mathbf{o},\ \mathbf{g}^+,\ \mathbf{g}^-)\begin{bmatrix}\mathbf{x}^+\\ \mathbf{x}^-\\ \mathbf{e}^+\\ \mathbf{e}^-\end{bmatrix}. \qquad (25c)$$

The asymmetric-norm description may be unnatural when there are many positivity constraints; the LP form may be unnatural when variables need not be positive and hence must be doubled up. Barrowdale and Young's program (1965) is mainly an adaptation of LP so that the big sparse identity matrix and the $-A$ matrix need not be stored or manipulated. The cost of solving an LP problem with an N-by-M coefficient matrix is dominated by N^2M or M^2N depending on whether one works with the primal or with the dual.[2] Since least-squares costs are *dominatingly* NM^2 where $N\gg M$, one expects costs for LP to be about the same. Of course the authors cannot speak for all possible methods of LP; however, it appears that the lowest LP costs are actually more accurately of the order $NM^2+N^2+M^2$. The N^2 term is not shared by least squares and will dominate when $N>M^2$. The extreme case is finding a median. Here LP is much slower than Hoare's algorithim. In such cases we expect our asymmetric-norm algorithm to be much faster than an LP formulation of the asymmetric-norm problem.

[2] A discussion of duality is found in Wagner (1959).

CONCLUSION

In any application where averages are being formed or where the least-squares method is being used, there is a good chance that the job can be done better and perhaps more rapidly with robust methods and the asymmetric linear norm. When the computer cost does turn out to be greater, rarely will it exceed 2 or 4 times greater, which will often be justified by the better results. In studying numerous examples of data modeling, we have found no case in which the asymmetric norm method gave notably poorer results and many cases in which the results are very much better.

ACKNOWLEDGMENTS

We wish to express our thanks to the donors of the Petroleum Research Fund of the American Chemical Society, to the National Science Foundation (grant GA-30766), to Stanford University, to Princeton University, and to the Chevron Oil Field Research Company for their support in this work.

REFERENCES

Backus, G. E., and Gilbert, J. F., 1967, Numerical applications of a formalism for geophysical inverse problems: Geophys. J. Roy. Astr. Soc., v. 13, p. 247–276.

Barrowdale, Ian, and Young, Andrew, 1965, Algorithms for best L_1 and L_∞ linear approximations on a discrete set: Numerische Mathematik, v. 8, p. 295–306.

Dougherty, E. L., and Smith, S. T., 1966, The use of linear programming to filter digitilized map data: Geophysics, v. 31 no. 1, p. 253–259.

Hoare, C. A. R., 1962, Quicksort: Comput. J., v. 5, p. 10–15.

Plackett, R. L., 1972, Studies in the history of probability and statistics. Chap. XXIX, *in* The discovery of the method of least squares: Biometrica, v. 59, no. 2, p. 239–251.

Singleton, R. S., 1969, Algorithm 347 sort: Comm. ACM, v. 12.

Ulrych, T. J., 1972, Maximum entropy power spectrum of truncated sinusoids: J.G.R., vol. 77, no. 8, p. 1396–1400.

Wagner, H. L., 1959, Linear programming techniques for regression analysis: J. of the Amer. Statistical Assn., p. 206–212.

Widrow, B., Mantey, P. E., Griffiths, J. J., and Goode, B. B., 1967, Adaptive antenna systems: Proc. of IEEE, v. 55, p. 2143–2159.