Inverse Problems 4 (1988) 1071–1091. Printed in the UK

Robust methods in inverse theory

John A Scales and Adam Gersztenkorn
Amoco Research Center, PO Box 3385, Tulsa, OK 74102, USA

Received 5 January 1988, in final form 22 March 1988

Abstract. Due to ill-conditioning of the linearised forward problem and the presence of noise in most data, inverse problems generally require some kind of 'regularisation' in order to generate physically plausible solutions. The most popular method of regularised inversion is damped least squares. Damping sometimes forces the solution to be smoother than it otherwise would be by raising all of the eigenvalues in an *ad hoc* fashion. In this paper an alternative is described, based upon the method of least-absolute deviation (LAD), which has a property known in the statistical literature as robustness. The history of LAD methods is even older than that of least squares, originating in the work of the 18th century polymath Roger Boscovich. But comparatively recent advances have made the use of such techniques feasible for the huge data sets encountered in areas such as inverse scattering, tomography, holography, and signal processing. An account of robust inversion methods—their history and recent computational developments—is given. The key computational technique turns out to be preconditioned conjugate gradient, an algorithm which had as its genesis 'the method of orthogonal vectors' published in 1948 by L Fox, H Huskey and J Wilkinson. Applications are illustrated from seismic tomography and inverse scattering, two of the most computationally intensive tasks in inverse theory.

1. Introduction

Forward problems and inverse problems can be thought of as mappings between linear spaces of model vectors and data vectors. Since one can only record finitely many data, and since one can only fit finitely many model parameters into a computer, it suffices in practice to consider model space and data space to be finite dimensional cartesian spaces. Since most physical observables are modelled by continuous functions, the latter assumption generally involves some sort of discretisation of the continuous observable. For example, the temperature on the surface of the Earth is a continuous function, but in order to run weather simulation programs one must divide the surface up into a finite number of cells, within each of which the temperature is constant. As to the mapping itself, it is almost always reduced to a linear operator since nonlinear problems are almost always solved by Newton's method. Given what has just been said about the discretisation of physical observables, this means that in most cases it suffices to consider the mappings to be represented as matrices.

Thus, the standard linearised inverse problem can be written $Ax = y + \varepsilon$, where $y + \varepsilon$ is a vector containing observed data, x is a vector of unknown model parameters, A is a matrix which represents the mapping from model space to data space, and ε is a vector whose components are independent random variables with approximately identical distributions. Although ε is not usually written explicitly, it is important to remember that it is there. For more details see, e.g., Huber (1981) and Nolet (1987).

0266-5611/88/041071+21 $02.50

The number of observations is seldom equal to the number of parameters, so a solution is defined to be a vector x which minimises $Ax-y$ in some sense. Since the time of Gauss and Legendre, the standard approach to this problem has been via the method of least squares, in which a vector x is computed which minimises the sum of the squares of $(Ax-y)_i$. This can be done in a linear fashion by solving the normal equations. The least-squares solution has an important statistical interpretation. Namely, if the data are normally distributed (i.e. have a gaussian probability density) and if the expectation value of the right-hand side is, say, μ, then the expectation value of the least-squares solution is just the inner product of the pseudo-inverse and μ (Stoer and Bulirsch 1980).

The distribution of errors in real data is seldom gaussian. Classical least-squares inversion tends to be rather sensitive to departures from normalcy; in the statistical lexicon, least-squares inversion is not robust (Huber 1981). This lack of robustness can result in solutions which are unphysical in the sense that they are disproportionately affected by spurious observations or exhibit large, high frequency oscillations, but which nevertheless fit the data (in a least-squares sense). It is the high frequency oscillations or roughness in the solution to which regularisation methods are usually directed.

Currently the most popular method of regularising the inverse problem for x is to apply damping to the least-squares inversion. Damping does not directly address the question of statistical robustness, but it does often result in smooth solutions by, in effect, raising all of the singular values of A by a fixed amount, the smallest singular values being substantially raised. The choice of the damping parameter is obviously problematical; the most efficient method may be trial and error and experience with like problems. More theoretically satisfying methods of regularisation are based on building smoothness into the minimisation *a priori*. This can be done in a variety of ways, for example by using error norms which penalise roughness (e.g. Sobolev space norms), using *a priori* information to constrain the solution, or by a maximum entropy approach. These methods result, roughly speaking, in the most featureless models consistent with the data. For an example of this approach see Constable *et al* (1987).

In this paper an inversion method will be discussed whose origins are even older than the method of least squares. Solutions will be sought which minimise not the sum of the squared errors, but rather the sum of their absolute values (least-absolute deviation or LAD). LAD optimisation is known to be resistant to certain (potentially large) errors in the data; a theoretical discussion of the robustness of LAD optimisation is given in Bloomfield and Steiger (1983). The basic idea, however, can be illustrated very simply. Suppose that one has n observations associated with a single parameter x. To compute the number which minimises the sum of the squared errors, differentiate the error $\Sigma_{i=1}^{n}|x-x_i|^2$ to get $x=(1/n)\Sigma_{i=1}^{n} x_i$. Thus the l_2 solution to the estimation problem (i.e. the number x which minimises the errors in the euclidean norm) is given by the mean of the data. On the other hand, by differentiating the sum of the absolute deviations and using $(\mathrm{d}/\mathrm{d}x)|x| = \mathrm{sgn}(x)$ one arrives at $\Sigma_{i=1}^{n}\mathrm{sgn}(x-x_i)=0$. This equation is satisfied if x is the median of the data. (The median of a set is defined as the middle of the set, once sorting with respect to size has been carried out. If the set has an even number of elements, the median is defined somewhat arbitrarily as the average of the two elements in the middle of the sorted set.) As a result, the least-squares solution will be adversely affected by outliers since the large deviations are averaged into the solution, whereas the LAD solution simply ignores these points. As an aside, note that for the same reason, median filtering is considerably more robust than mean filtering.

The terms LAD and l_1 minimisation or optimisation will be used interchangeably. More generally, consider the problem of l_p optimisation defined by

$$\min_x \sum_i \left| \sum_j A_{ij}x_j - y_i \right|^p \qquad p \geq 1$$

where A is a matrix of coefficients, y is the observation vector, x is a vector of model parameters, and the minimum is taken over all vectors in the domain of A. The lower bound on p is necessary since $(\Sigma|x_i|^p)^{1/p}$ is not a norm for p less than one.

2. The origin and development of robust inversion methods

The origin of LAD methods lies in the work of the Serbo-Croatian Jesuit Rudjer J Bošković (anglicised to Roger Boscovich) whom Željko Marković in the *Dictionary of Scientific Biography* describes as 'perhaps the last polymath to figure in an important way in the history of science'. Boscovich was from the beginning a mathematician, but his studies of natural philosophy were profound and, in some cases, astonishingly modern. Although Boscovich was one of the first continental proponents of Newton's work, he rejected Newton's theory of matter and instead proposed that 'continua' were made up of point particles (*puncta*) bound by a force field which was attractive at large distances (and asymptotically weak) but repulsive at short ones (being infinitely repulsive at zero distance). And further, that all phenomena arise from the spatial arrangement and relative displacement of the *puncta*. Thus he was able to reduce 'the three great fundamental principles of Newton (gravitation, cohesion and fermentation) to a single principle' (Jammer 1957). He also gave a profound explanation of the impulsive collision of bodies by showing that the velocities of the bodies must change continuously via a force which acted at a distance and that there was no actual contact†. Boscovich also made important contributions to pure mathematics, astronomy, geodesy, instrument design and engineering. And his *Theoria* contains many striking anticipations of modern field theory and special relativity. Accounts of the life and works of this singular figure are given in the book edited by Whyte (1961). We shall concentrate on Boscovich's work in the theory of robust estimation.

In a summary of his work with Christopher Maire (Maire and Boscovich 1755) on mapping and the determination of the figure of the Earth from geodetic measurements (no small feat in itself since the high-precision observations were made in difficult terrain), Boscovich proposed that among observed values of variables (x, y) connected by a linear relationship, the line that is most nearly in accord with the observations should satisfy two criteria (Eisenhart 1961):

the sum of the positive and negative corrections to the y values shall be equal, and

the sum of the absolute values of all of the corrections, positive and negative, shall be as small as possible.

Together these conditions require that the slope of the best-fitting line satisfy

$$\sum_{i=1}^{n} |(y_i - \bar{y}) - b(x_i - \bar{x})| = \text{minimum} \tag{1}$$

† '. . . *quae continuitatis lex cum (ut evinco) debeat omnino observari, illud infero, antequam ad contactum deveniant corpora, debere mutari eorum velocitates per vim quandam, quae sit par extinguendae velocitati, vel velocitatum differentiae, cuivis utcunque magnae.*' This work is summarised in Boscovich's *magnum opus*, the *Theoria philosophiae naturalis* (Boscovich 1763).

where $(\bar{x}, \bar{y})$ is the centroid of the observation points; in other words, equation (1) gives the line which minimises the absolute deviation among all lines passing through the mean of the data.

Boscovich restated these conditions in a prose appendix to the treatise on natural philosophy by Benedict Stay (1760) (written in Latin hexameters), and gave a clever geometrical method for finding the solution to equation (1). Laplace made extensive use of Boscovich's method and developed an analytical technique for determining the minimum absolute deviation line by using the median of the set $|x_i - \bar{x}|$. Laplace gave a rigorous account of these methods in his *Theorie Analytique des Probabilites*. He derived the probability density for the estimator which minimised the absolute values of the residuals and showed that this density approaches the normal density as the sample size increases. Laplace also gave necessary and sufficient conditions on the error distribution such that the median has smaller asymptotic variance than the mean (Stigler 1973).

Gauss was also aware of Boscovich's work and the potential value of minimum absolute deviation methods in his astronomical work. In his *Theoria Motus* (Gauss 1809) he describes an elegant algebraic method for computing the l_1 solution of an overdetermined linear system. Gauss's method is based on the observation that the l_1 solution will exactly satisfy one of the square subsystems, whose size equals the number of columns in the original matrix (assuming full column rank). According to Gauss, this result is easily shown; for a proof of this adumbration of the fundamental theorem of linear programming see Bloomfield and Steiger (1983).

The technical term 'robust' was not coined until 1953 by G E P Box. And as a result of a series of papers by Peter Huber in the 1960s, the professional statisticians began to focus their attention on the subject. But in spite of the growth of interest in robust statistics, until the mid-1970s most methods for computing l_1 solutions of overdetermined linear systems were based upon relatively slow linear programming extensions of Gauss's method (e.g. Barrodale and Roberts 1973). Schlossmacher (1973) published an algorithm for the l_1 problem which is based on the solution of a sequence of weighted least-squares problems, the weights being chosen iteratively. Suppose A is a rectangular matrix of full column rank. By differentiating the error function $(\partial_i \equiv \partial/\partial x_i)$

$$\sum_i \partial_k \left| \sum_j A_{ij} x_j - y_i \right|^p$$

and recalling that $\operatorname{sgn}(x) = x/|x|$, one has

$$\sum_i \partial_k \left| \sum_j A_{ij} x_j - y_i \right|^p \equiv \sum_i \partial_k |r_i|^p = \sum_i \operatorname{sgn}(r_i) p |r_i|^{p-1} A_{ik}$$

$$= \sum_i r_i p |r_i|^{p-2} A_{ik}$$

$$= [A^{\mathrm{T}} R (Ax - y)]_k$$

where the matrix R is defined to be diag $p|r_i|^{p-2}$, actually diag $|r_i|^{p-2}$ since the factor of p can be cancelled when one sets this last expression equal to zero. The result of this calculation is a nonlinear form of the normal equations

$$A^{\mathrm{T}} R A x = A^{\mathrm{T}} R y.$$

If $p=2$ the weights are unity and one is left with the usual normal equations, but if $p\neq 2$ the weighting is nonlinear.

If A has full column rank, a natural iteration scheme for solving the weighted normal equations is (at the ith iteration)

$$(A^{\mathrm{T}}R_{i-1}A)x_i = A^{\mathrm{T}}R_{i-1}y. \tag{2}$$

For the case $p=1$, this is precisely Schlossmacher's method. The iteration is started by choosing the weights to be unity and computing an ordinary least-squares solution. Since by Gauss's method it is known that there will always be some exactly zero residuals, iterative algorithms of this sort could become unstable. To avoid this, Schlossmacher simply eliminates the corresponding observation whenever a residual becomes very small. As he points out: This is justified since the effect of an almost zero residual on the total sum will be negligible.' As will be shown presently, there are a number of alternatives to simply deleting the observations, which nevertheless guarantee a stable algorithm. Equation (2) is referred to as IRLS for iteratively reweighted least squares.

Following close on the heels of Schlossmacher's paper was a lengthy paper by Beaton and Tukey (1974) considering robust estimation and smoothing from a very general point of view. Instead of the l_p problem, Beaton and Tukey consider minimisation in terms of a general 'loss function' ρ:

$$\min_x \sum_i \rho\left(\sum_i A_{ij}x_j - y_i\right).$$

A number of examples of loss functions are given in the survey paper by Holland and Welsch (1977).

We shall close this section with a few observations about the l_p minimisation problem. First, it follows straightforwardly from the Minkowski inequality that if x and y are solutions of the l_p minimisation problem, then the convex combination $tx+(1-t)y$, where $0<t<1$, is also a solution. In other words, either the problem is uniquely soluble or there are infinitely many solutions. And secondly, while it is clear that if A has full column rank then the least-squares problem is uniquely soluble since the inverse of $A^{\mathrm{T}}A$ exists, this is not the case for the LAD problem. In fact one can prove the following theorem (cf Rice (1964) for an L_p version based upon a slightly different argument).

Theorem. Assume that the matrix A has full column rank, $y\in R^m$ is given, $p>1$, and that there exists a vector $x\in R^m$ such that

$$\min_x \|Ax-y\| = \gamma$$

where $\|x\|$ denotes the l_p norm. Then x is unique.

To prove this one needs nothing more than the Minkowski inequality.

Lemma (Minkowski). Let p be strictly greater than 1. Then

$$\left(\sum_i |x_i+y_i|^p\right)^{1/p} \leq \left(\sum_i |x_i|_p\right)^{1/p} + \left(\sum_i |y_i|^p\right)^{1/p}.$$

Further, the equality holds if and only if $k_1 x = k_2 y$ for some positive constants k_1 and k_2.

Proof. Luenberger (1969 p 31).

The main theorem can now be shown easily.

Proof. Suppose, on the contrary, that the solution is not unique. Let z and w be supposed solutions; it will be shown that $z=w$. Thus

$$\min_x \|Ax-y\| = \|Az-y\| = \|Aw-y\| = \gamma.$$

Consider the convex combination $u=\alpha z+(1-\alpha)w$, $\alpha \in (0, 1)$. The error associated with u is

$$\begin{aligned}\|Au-y\| &= \|A[\alpha z+(1-\alpha)w]+\alpha y-\alpha y-y\| \\ &= \|\alpha(Az-y)+(1-\alpha)(Aw-y)\| \\ &\leqslant \alpha\|Az-y\|+(1-\alpha)\|Aw-y\| \\ &= \alpha\gamma+(1-\alpha)\gamma=\gamma.\end{aligned}$$

Thus, if there are two solutions, any convex combination is also a solution. But by definition γ is the smallest error. Therefore the error associated with u cannot be any less than this. As a result, the last inequality is actually a strict equality. By inference

$$\|\alpha(Az-y)+(1-\alpha)(Aw-y)\| = \|\alpha(Az-y)\|+\|(1-\alpha)(Aw-y)\|.$$

But the corollary to Minkowski says that equality holds if and only if the vectors are proportional

$$\alpha(Az-y) = \frac{k_2}{k_1}(1-\alpha)(Aw-y).$$

Taking norms on both sides of this last equation it follows that

$$\alpha\gamma = \frac{k_2}{k_1}(1-\alpha)\gamma$$

and therefore

$$\frac{k_2}{k_1} = \frac{\alpha}{1-\alpha}.$$

Inserting this back into the equation of proportionality yields

$$\alpha(Az-y) = \alpha(Aw-y)$$

which implies that

$$A(z-w) = 0.$$

And since A was assumed to have full column rank, z must equal w. Therefore the solution is unique. Note that this also takes care of the case in which $\gamma=0$, since

$$\|Ax-y\| = 0 \Rightarrow Ax-y=0,$$

which in turn means that the two supposed solutions satisfy

$$A(z-w) = 0.$$

This completes the proof. Needless to say, this result neglects the effect of finite precision arithmetic. It seems likely that the effects of rounding error will usually obliterate any difference between the computed solutions for $p=1$ and $p=1+\varepsilon$, provided ε is small. An example of the nonuniqueness of the LAD problem can be found in Scales and Treitel (1987).

A simple analytic example will illustrate some of the properties of the l_p inverse and the potential difficulties which arise as p approaches 1. Consider the linear system $Ax=y$ where $A=(1,\lambda)^T$, $y=(1,0)^T$, and where $\lambda \geq 0$. The l_p error criterion is

$$E_p(x)=|x-1|^p+\lambda^p|x|^p.$$

For $p=1$ the solution, by Gauss's method, is

$$x_{l_1}=\begin{cases}1 & \text{if } \lambda \leq 1\\ 0 & \text{if } \lambda \geq 1.\end{cases}$$

Since at $\lambda=1$ there are two solutions, $x=0$ and $x=1$, then the convex combination $x=t$, $t\in[0,1]$ is also a solution. For $p>1$ the critical points of E_p are readily calculated; restricting attention to the case $x\in(0,1)$, one finds that $\partial_x E_p(x)=0$ if and only if

$$\left(\frac{1-x}{x}\right)^{p-1}=\lambda^p.$$

From which it follows that

$$x_{l_p}=\frac{1}{1+\lambda^{p/(p-1)}}.$$

A plot of these functions (figure 1) shows, as expected, that the solutions are unique for $p>1$ and that they approach the step function solution given by Gauss's method as p approaches 1. Thus it appears that the l_1 solution may suffer from sensitivity to model parameters. This might be true in the solution to relatively small problems, amenable to exact solution via linear programming techniques. But for very large inversion problems, of the type encountered for example in tomography and inverse scattering, gradient methods are used which iteratively compute the solution for

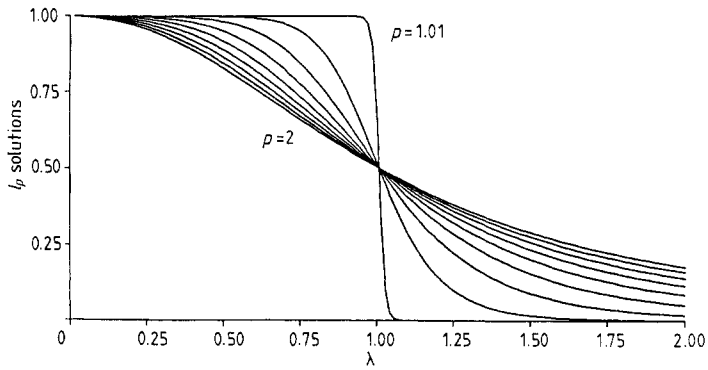

Figure 1. Analytic solutions for l_p optimisation $(1,\lambda)^T x=(1,0)^T$ as p approaches 1. The solution for $p=1$ is obtained by Gauss's method and is precisely the step function $1-H(x-1)$.

arbitrary $p \geq 1$. So there are several factors working to stabilise the problem. First, in the iterative methods (which rely on solving a sequence of weighted least-squares problems with weights depending on the residuals), one does not use an exact l_p error criterion in the objective function since this would involve dividing by zero whenever a basis vector was approached; instead, residuals smaller than a certain constant are set equal to that constant. Secondly, since the iterative methods are based on direct optimisation of the l_p objective functions and not Gauss's method, one can compute solutions for arbitrary $p \geq 1$. Thus one might use $2 > p > 1$ in order to achieve both robustness and unique solubility. And finally, for large inverse problems, with perhaps 10^5 to 10^6 equations, the effects of rounding error blur the exact analytic properties of the minimisation.

3. Iteratively reweighted least squares

Let us return now to the IRLS algorithm, equation (2). Notice that for $p = 2$ the weighting matrix R is the identity and one is left with the normal equations for least squares. But for $1 \leq p < 2$ the effect of R is to diminish the influence of large residuals: outliers in the data are effectively rejected for $p \approx 1$. There are three potential difficulties with the practical implementation of equation (2). First, some of the residuals will be zero (or, what amounts to the same thing on a computer, of the order of the machine precision or less); secondly, the matrix A may be rank deficient, so that the inverse of A^TA does not exist; and finally, equation (2) is not well suited to direct solution for large sparse problems since A^TA will usually be dense even when A is sparse.

Rather than eliminating observations associated with small residuals as Schlossmacher does, the absolute values of the residuals will be set equal to a small positive constant whenever these absolute values fall below this constant. In other words,

$$R_{ii} = \begin{cases} 1/|r_i|^{2-p} & \text{if } |r_i| > \varepsilon \\ 1/\varepsilon^{2-p} & \text{if } |r_i| \leq \varepsilon. \end{cases}$$

Generally one should also normalise the elementsof R to lie between $\varepsilon/r_{\max}$ and 1 in order to improve convergence (Byrd and Pyne 1979, Gersztenkorn *et al* 1986). Byrd and Pyne (1979) show that this choice of weighting matrix is sufficient to guarantee the convergence of IRLS. This is the least-squares weighting scheme proposed by Newcomb (1912), although not in an iterative context.

Now, by rank deficiency one simply means that the matrix A may have singular values which are of the order of the machine precision or less and therefore, for all practical purposes, the rank of A is less than the number of columns. One can avoid the problem of trying to solve the system explicitly with (A^TA) by using a solver which gives the pseudo-inverse solution at each step of the iteration; this could be a direct method such as SVD or an iterative method such as conjugate gradient or the row-action methods. An example of IRLS using direct methods is the paper by Coleman *et al* (1980). Here, iterative methods will be adopted to overcome the third difficulty mentioned above: namely, how to adapt IRLS for use on the very large sparse problems encountered in problems such as inverse scattering and tomography.

4. Implementation of IRLS via conjugate gradient

Conjugate gradient (CG) is an iterative method for solving linear systems or linear least squares. CG was invented by Hestenes and Stiefel (1952) but is in fact a special case of an algorithm termed by Fox *et al* (1948) 'the method of orthogonal vectors'. CG gives solutions to $Ax=b$, for A symmetric and positive definite, by minimising the quadratic form

$$f(x)=\tfrac{1}{2}(x, Ax)-b, x)+c$$

where $(.,.)$ denotes the Euclidean inner product. Notice that if z is a solution of $Az=b$ then for any other vector x

$$f(x)=f(z)+\tfrac{1}{2}((z-x), A(z-x))$$

and therefore, since A is assumed to be positive definite, z is the unique minimum of the quadratic form.

Conjugate gradient (Hestenes and Stiefel 1952). Let $x_0=0$, $\beta_1=0$, $r_0=p_1=b$. For $k=1, \ldots, n$; do until $r_k=0$

$$\beta_k=(r_{k-1}, r_{k-1})/(r_{k-2}, r_{k-2})$$
$$p_k=r_{k-1}+\beta_k p_{k-1}$$
$$\alpha_k=(r_{k-1}, r_{k-1})/(p_k, Ap_k)$$
$$x_k=x_{k-1}+\alpha_k p_k$$
$$r_k=r_{k-1}-\alpha_k Ap_k$$
$$x=x_k.$$

The solution vector x is computed by performing successive 1D minimisations along 'search vectors' p_k generated recursively from the residuals $r_k \equiv b-Ax_k$. The scale factors α_k are determined by the minimisation criterion, whereas the scale factors β_k are required to ensure that the search vectors p_k are 'A orthogonal' (or 'A conjugate') in the sense that $(p_i, Ap_j)=0$ provided that $i \neq j$. The A orthogonality in turn guarantees convergence (in exact arithmetic) to the unique solution in at most n steps, where n is the order of the matrix A. For more details on this and a host of related problems the reader is referred to the excellent monograph by Hestenes (1980).

To solve the linear least-squares problem one simply writes the normal equations in factored form

$$A^{\mathrm{T}}(Ax-y)=0.$$

One could apply the symmetric CG algorithm to this equation, using $A^{\mathrm{T}}y$ instead of y and calculating $A^{\mathrm{T}}(Ax)$ instead of Ax. Even though this avoids explicit calculation of the matrix product $A^{\mathrm{T}}A$, and therefore does not affect the sparsity of the problem, the algorithm suffers from the fact that the condition number of $A^{\mathrm{T}}A$ is the square of the condition number of A: stability and accuracy are adversely affected if A and A^{T} are applied in succession. This defect can be avoided if, whenever one has to calculate terms of the form

$$A^{\mathrm{T}}(Ax_i-y),$$

where x_i is the ith approximation to the solution, one first computes Ax_i-y and then applies A^{T} rather than subtracting $A^{\mathrm{T}}y$ from $A^{\mathrm{T}}Ax_i$. The difference between the two

approaches, due entirely to rounding error, may be dramatic in practice; for a careful analysis see Chandra (1978).

Now, to solve equation (2) simply write the l_p form of the normal equations in factored form

$$A^{\mathrm{T}}R(Ax-y)=0.$$

Using the least-squares form of CG it suffices to replace the routine for calculating A^{T} times a vector with one which does $A^{\mathrm{T}}R$ times a vector. Since R is diagonal this is easy; it does not affect the sparsity of the problems and results in only a modest increase in the number of floating point operations as compared to least-squares CG. Thus R becomes effectively a preconditioning matrix, and since it depends on the solution, the preconditioning is nonlinear.

The algebraic details of the IRLS with CG are given in Scales *et al* (1988). The basic idea, however, is quite simple. Begin by computing a 'small' number of unweighted, or least-squares, CG iterations. Limiting the number of iterations in CG is, practically speaking, equivalent to damping. The definition of 'small' depends on the assumed level of noise in the data. We use stopping criteria based on, among other things, the fractional decrease in the norm of the residual vector. Thus 'small' means small compared with the number of iterations required to make this fractional decrease of the order of the assumed level of the noise. Once the first unweighted iterations are complete, the residual vector (which is computed inductively in CG) is used to calculate the first weighting matrix. Another set of (now weighted) CG iterations is performed, at the end of which a new weighting matrix is computed, and so on. The weighted least-squares solutions need not be computed very accurately. The procedure terminates when the solution satisfies the global stopping criterion.

One *ad hoc* feature of the procedure is the selection of the Huber taper parameter ε. This scale factor does not work quite like a damping parameter since it only affects those data in which one has a high degree of confidence anyway, i.e. those associated with small residuals. It is simply a statement that beyond a certain point all observations are to be equally weighted. Some work has been done on computing scale factors *a priori* from the data; see for example Shanno and Rocke (1986). But at this point the cutoff is simply chosen to be min $(|r(i)|^{2-p})$ plus a fixed percentage (determined by the assumed level of noise) of

$$\max(|r(i)|^{2-p})-\min(|r(i)|^{2-p}).$$

This parameter is almost never changed in practice for a given type of problem. Experience indicates that the IRLS procedure is very sensitive to only one parameter, the number of unweighted CG iterations in the first step. Doing too many iterations in the first step results in the typical deleterious effects of undamped least squares. To be safe, the number of CG iterations is kept conservatively small, at the possible expense of a greater number of weighting steps. And to repeat, small here is defined relative to the number of iterations required to reduce $(\|r_n\|/\|r_0\|)$ to the assumed level of noise in the data.

Although in our work we have concentrated on the implementation of IRLS with conjugate gradient, it is important to realise that, in principle, any least-squares solver can be used. CG has been chosen primarily for its ease of use and proven track record in a variety of applications.

5. Applications to seismic inversion

To illustrate the potential advantages of l_1 inversion in seismic problems we shall consider two concrete examples. The first is the iterative Born inversion of 1D acoustic scattering data. Acoustic inverse scattering is the subject of a vast literature which we make no attempt to survey. Our goal is simply to illustrate how robust optimisation methods such as IRLS can be fitted easily into any inversion scheme which is based upon repeated solution of linear optimisation problems. Nevertheless it is worth pointing out, however briefly, some of the recent progress that has been made on this problem.

First, there are two important extensions of the classical Born inversion method in widespread use in seismic inversion. The first involves the use of heterogeneous background models from which perturbations are computed. In our approach we can use arbitrarily complicated background models—including multiple reflections (as in Gersztenkorn *et al* 1986). This is essentially equivalent to using a more realistic Green function in the underlying integral equation which describes the forward scattering problem. For a discussion of these methods see, for example, Clayton and Stolt (1981), Raz (1981), Gray (1984) and the references cited therein. The second important extension is to solve not for perturbations in the acoustic wavespeed or index of refraction, but rather for reflection coefficient. This approach is related to the use of the Bremmer series which expresses the total wavefield as a sum of waves that have been reflected $0, 1, \ldots, n$ times, and has the distinct advantage that the reflection coefficient is generally a small quantity in seismology. Further, the expression for the nth reflected wave can be obtained from the expression for the wave that has been reflected $n-1$ times using the impedance variation (Gray 1984). Raz (1981) shows that an inversion based upon a single backscatter event in a WKB background is equivalent to an implicit inversion based on a first-order Bremmer model. In any event, whether the inversion is made in the Born approximation, the Bremmer series, the Rytov approximation, or some other, the key step is a linearisation of the underlying equation relating medium properties with the scattered field. And thus robust methods such as IRLS are directly applicable. The fact that we shall consider a 1D example is of no particular significance. Many relevant problems in exploration seismology are approximately one dimensional, and the robust optimisation techniques that we propose will first easily into a multidimensional framework as we show in our second example.

The second example that we shall consider is the tomographic travel time inversion of a complicated 2D seismic reflection survey. There are at least three distinct approaches to the parametrisation and inversion of reflection seismic travel times. In the first approach (e.g. Bishop *et al* 1985) both the discretised velocity and reflector position values are included as parameters in a travel time inversion. In the second approach (e.g. Bording *et al* 1987) the travel times are used to invert for the velocity alone, while a waveform technique such as depth migration is used to update the position of the reflectors. Another approach, due to Stork (1988), is to include both velocity and reflector parameters in the travel time inversion in order to resolve velocity/reflector-depth ambiguity, but use migration to ultimately position the reflectors. The use of migration turns out to be important since it is relatively insensitive to some of the common tomographic artefacts, and tends to give more accurate reflector positions than those obtained by travel time inversion. Nevertheless, all of these formulations of the seismic tomography problem are nonlinear; as with all nonlinear

inverse problems convergence is difficult to characterise. It is clearly very important to have good starting models and to incorporate constraints and *a priori* information. Fortunately constraints are very easy to incorporate into IRLS. In practice, we find that good starting models are not difficult to obtain by conventional seismic processing techniques.

5.1. Inverse scattering

Consider the 1D wave equation with wavespeed $c = c(z)$ and an impulsive force term $\delta(z)f(t)$

$$\frac{\partial^2 U(z,t)}{\partial^2 z} - \frac{1}{c(z)^2}\frac{\partial^2 U(z,t)}{\partial t^2} = \delta(z)f(t).$$

Then, making the substitution $m(z) = (c_0/c(z))^2$, and Fourier transforming with respect to time, we are left with the following ODE:

$$\frac{\mathrm{d}^2 U(z,k,m)}{\mathrm{d}^2 z} + k^2 m(z) U(z,k,m) = \delta(z)f(k). \tag{3}$$

The boundary conditions for equation (3) are the usual free-surface condition at $z = 0$

$$\frac{\mathrm{d}U(0,k,m)}{\mathrm{d}z} = 0$$

and an outgoing wave at $z = H$

$$\frac{\mathrm{d}U(H,k,m)}{\mathrm{d}z} - \mathrm{i}kU(H,k,m) = 0.$$

Next, the ODE is recast as an integral equation by introducing a Green function G via

$$\frac{\mathrm{d}^2 G(z,\zeta,k,m_0)}{\mathrm{d}^2 z} + k^2 m_0(z) G(z,\zeta,k,m_0) = \delta(z-\zeta)$$

satisfying the same boundary conditions as the solution U. This definition of G requires that it be computed numerically. We used the synthetic seismogram program of Gutowski and Treitel (1987) to compute the impulse response directly.

With these definitions, one can show that the following integral equation (the Lippmann–Schwinger equation) holds:

$$U(z,k,m) = U(z,k,m_0) - k\int_0^H G(z,\zeta,k,m_0)\Delta m(\zeta) U(\zeta,k,m)\,\mathrm{d}\zeta. \tag{4}$$

For details of the derivation of equation (4) see Gersztenkorn (1984). The term Δm appearing in equation (4) represents a perturbation about some background model. When U appears with the argument m, the true model, it is interpreted as the observed scattering data, or strictly, a computed approximation thereto. Equation (4) is rigorously true for all Δm, but we shall now assume that $\|\Delta m\|$ is small, namely that we have a good initial approximation to the velocity–depth model, and that we can therefore replace $U(m)$ with $U(m_0)$ inside the integral: this is the Born approximation

$$U_L(z,k,m) = U(z,k,m_0) - k\int_0^H G(z,\zeta,k,m_0)\Delta m(\zeta) U(\zeta,k,m_0)\,\mathrm{d}\zeta. \tag{5}$$

In the time domain, taking advantage of the fact that the data are recorded at $z=0$, equation (5) leads to

$$U_L(0,t,m)=U(0,t,m_0)+\frac{1}{c_0^2}\int_0^H \partial_t^2 \int_0^t G(0,\zeta,t',m_0)\Delta m(\zeta)U(\zeta,t-t',m_0)\,\mathrm{d}t'. \tag{6}$$

The U appearing on the right hand of equation (6) is known, since m_0 is known. If one knew the true model $m(z)$, then $U(0,k,m)$ would correspond to the observed wavefield. And therefore, to the extent that $U_L \approx U$, equation (6) relates the observed data to the unknown model perturbations. The idea then is to invert equation (6) for Δm, given $U(0,t,m)$ at a finite number of times.

Equation (6) can be simplified by defining the integral kernel

$$K(\zeta,t)=\frac{1}{c_0^2}\partial_t^2\int_0^t G(0,\zeta,t',m_0)U(\zeta,t-t',m_0)\,\mathrm{d}t'.$$

Then equation (6) becomes

$$U_L(0,t,m)-U(0,t,m_0)=\int_0^H K(\zeta,t)\Delta m(\zeta)\,\mathrm{d}\zeta. \tag{7}$$

We shall use equation (7) to iteratively compute solutions U_L corresponding to approximate models, and compare the results with the recorded data U_D to check for convergence. The l_p problem is then to minimise

$$\min_m \sum_j |U_D(t_j)-U_L(0,t_j,m))|^p.$$

Defining $D(t_i)\equiv U_D(t_i)-U(0,t_i,m_0)$ and using equation (7) this becomes

$$\min_m \sum_i \left|D(t_i)-\sum_j K(z_j,t_i)\Delta m(z_i)\Delta z\right|^p \tag{8}$$

where the integral has been replaced by its Riemann sum. The weighted normal equations for equation (8) are

$$K^{\mathrm{T}}R(K\Delta m-D)=0 \tag{9}$$

where $K=K(z_j,t_i)$ is the kernel matrix, Δm is the solution vector of model perturbations, $D=U(t_i)-U(0,t_i,m_0)$ and R is the residual weighting matrix. Equation (9) is of a form suitable for the application of IRLS, which we now illustrate.

Figure 2(*a*) shows an acoustic trace generated for the velocity model shown in figure 2(*b*) as a heavy trace. Since the data are noise free, ordinary damped least-squares inversion works quite well; the least-squares solution is shown in the light trace in figure 2(*b*). Now, suppose we add isolated spikes to the trace, as in figure 2(*c*). The damped least-squares solution and the model are shown in figure 2(*d*). Note that, due to the wild fluctuations in the least-squares solution, figure 2(*d*) is of necessity drawn to a different scale than figures 2(*b*) or 2(*e*). On the other hand, if we apply IRLS (with minimisation norm equal to 1) we get the results shown in figure 2(*e*). Comparing this result with figure 2(*b*) one can see that IRLS has successfully reconstructed the velocity profile in spite of the signal contamination.

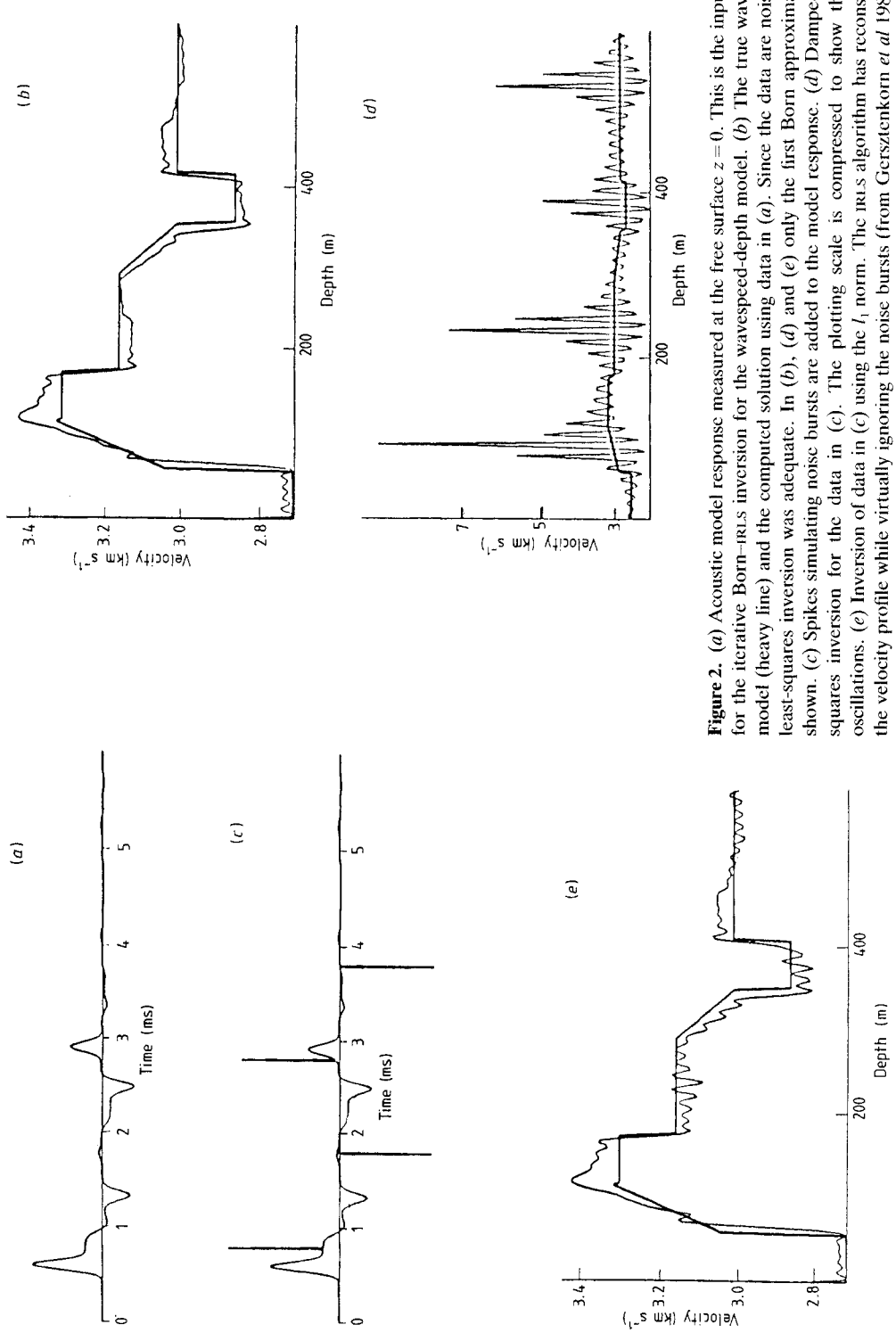

Figure 2. (*a*) Acoustic model response measured at the free surface $z = 0$. This is the input trace for the iterative Born–IRLS inversion for the wavespeed-depth model. (*b*) The true wavespeed model (heavy line) and the computed solution using data in (*a*). Since the data are noise free, least-squares inversion was adequate. In (*b*), (*d*) and (*e*) only the first Born approximation is shown. (*c*) Spikes simulating noise bursts are added to the model response. (*d*) Damped least-squares inversion for the data in (*c*). The plotting scale is compressed to show the wild oscillations. (*e*) Inversion of data in (*c*) using the l_1 norm. The IRLS algorithm has reconstructed the velocity profile while virtually ignoring the noise bursts (from Gersztenkorn *et al* 1986).

5.2. Tomography

Seismic travel tomography is another example of a nonlinear inverse problem. The data (travel times) are related to the model parameters (acoustic or elastic wavespeed), to the extent that ray theory is valid, via the integral

$$t(\text{ray}) = \int_{\text{ray}(s)} s(x, y, z)\, \mathrm{d}l \tag{10}$$

where x, y, and z are spatial coordinates, $\mathrm{d}l$ is the differential distance along the ray, and $s(x, y, z) = 1/v(x, y, z)$ is the slowness (reciprocal velocity) at the point (x, y, z). The travel time data are extracted from the raw seismograms; a reasonably large 2D problem might involve as many as 10^5 rays. Since the ray path depends on the unknown slowness, one must linearise this equation about some initial or reference slowness model. The linearised integral is then approximated as a sum; this requires discretising the slowness. We use cells or finite elements, within which the slowness is assumed to be constant. This results, for an ensemble of travel time observations, in a system of linear algebraic equations for the unknown slowness perturbation values

$$D\Delta s = \Delta t \tag{11}$$

where Δt is a vector whose components are the differences between the travel times computed for the model and the observed travel times, Δs is a vector whose components are the differences in slowness between the initial model and the updated model (referred to the cell-based description), and D is a matrix whose element D_{ij} is the distance the ith ray travels in the jth cell. As a consequence of Fermat's principle, perturbations in ray path are second order with respect to perturbations in slowness. In practice, the matrix D in equation (11) is large (perhaps 10^5 equations for a large 2D problem) and sparse (less than 1% nonzero).

We now show examples of how IRLS can be used to solve the linearised travel time equation (equation (11)). Here we shall consider an example of reflection tomography applied to clean, relatively noise-free, finite difference data. Figure 3(*a*) shows an acoustic velocity model for which finite difference synthetic seismograms were computed. Approximately 6600 travel times were extracted from the unstacked data. An initial model was used which consisted of flat, constant-velocity layers; the model was covered with about 3200 square cells (figure 3(*b*)). 6600 rays were traced through the initial model and the computed travel times were compared with those taken from the unstacked data; this gives the right-hand side of equation (11).

Due to the difficulty of interpreting reflection events from the deepest and most complex portions of the model, the bottom half of the model was unilluminated by rays in the inversion. The cell illumination is defined to be the sum of the ray length per cell divided by a characteristic cell length. A more sophisticated measure of the resolvability of a cell involves both the total ray length per cell and the angular spread of those rays, in the manner of the projection slice theorem. Ultimately one would like to be able to do a complete spectral decomposition of the matrix A. Direct methods such as the traditional SVD algorithm are out of the question for problems with 10^5 equations or more. Scales (1988) describes a technique whereby the conjugate gradient algorithm itself can be used to compute an SVD, taking full advantage of the sparsity of the matrix.

The inversion of these travel time data results in an updated approximation to the velocity model, figure 3(*c*). There is an 85% variance reduction in the data after one

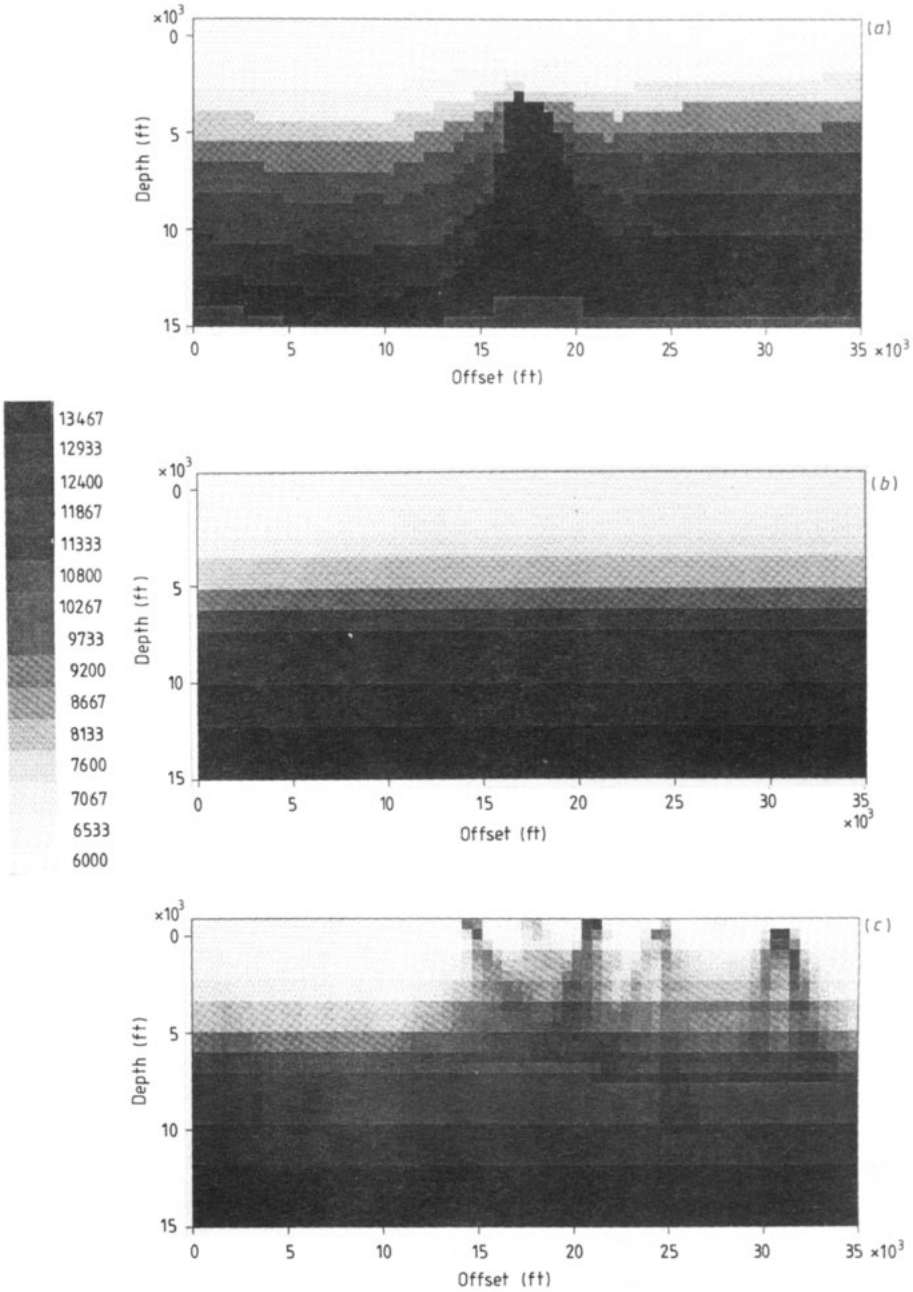

Figure 3. (*a*) Acoustic velocity-depth model for reflection tomography example. Finite difference synthetic seismograms were computed from which the travel times were extracted. (*b*) Initial velocity-depth model used in the travel time inversion. Rays were traced through this model. However, due to the difficulty in obtaining travel times for the deeper and more complex portions of the model, the ray illumination is non-uniform. The model was covered with about 3200 square cells within which (for purposes of inversion) it was assumed that the velocity was constant. (*c*) Updated velocity model produced by inverting travel time residuals in the l_1 norm using IRLS. (Note: the grey scale applies to (*a*), (*b*) and (*c*).)

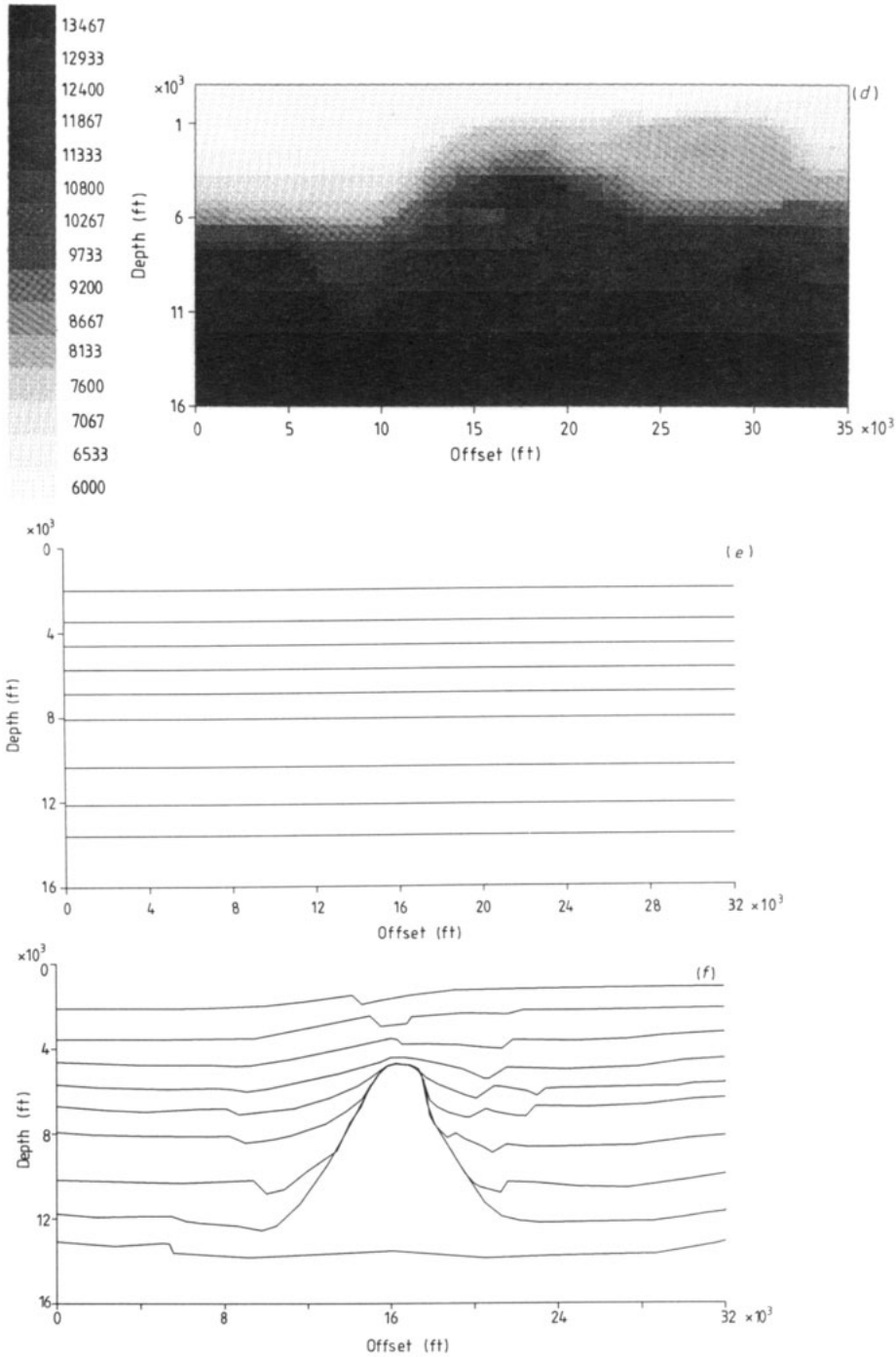

Figure 3. (*d*) The inversion procedure is an iterative, nonlinear one. In order to prepare the current velocity model for the next stage of the inversion procedure (more raytracing or migration) we often apply a 2D median filter. (*e*) Initial reflector position. Boundary value rays were traced from source to reflector to receiver corresponding to events interpreted in the raw seismic data. (*f*) Updated reflector position obtained by Kirchhoff migration using the tomographically determined velocity model. Figures 3(*g*)–(*i*) continue overleaf.

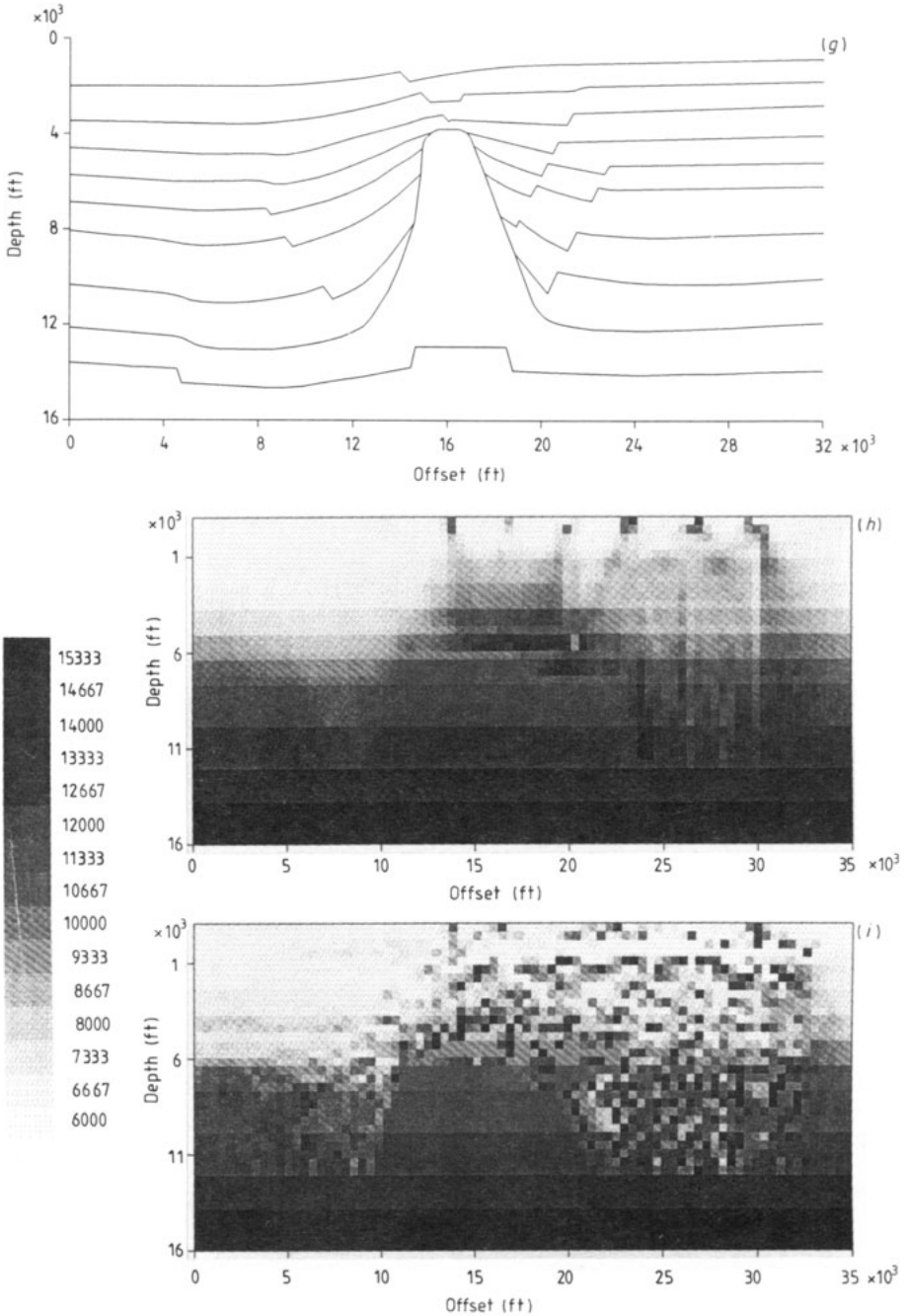

Figure 3. (*g*) Exact reflector position, used to produce finite difference synthetic seismograms which were taken to be the observed data. (*g*) and (*f*) overlay closely except near the dome feature in the lower centre of the model. (*h*) Updated velocity model produced by inverting travel time residuals in the $l_{1.3}$ norm using IRLS. The solution seems to retain the robust features for all norms near 1. (*i*) Updated velocity model produced by inverting travel time residuals in the l_2 norm using IRLS. The least-squares solution is quite rough as well as being more slowly convergent. (Note: the grey scale applies to (*h*) and (*i*).) (1 ft = 0.3 m.)

step. To smooth our velocity model for the next stage of the iterative inversion, we usually apply a fine-grain median filter. The median filtered computed tomogram is shown in figure 3(*d*). Median filters also have robust features: they reject outliers, whereas mean filters smear outliers into the solutions, and they preserve edges, which are ubiquitous in exploration seismology (Gersztenkorn and Scales 1988). In practice the smoothed tomogram shown in figure 3(*d*) is used as input to the next stage of an iterative inversion scheme.

Figures 3(*e*) (*f*) and (*g*) show the initial reflector positions, the reflector positions after one iteration (obtained via Kirchhoff migration) and the exact positions used in the modelling. The computed reflector positions overlay almost perfectly those of the exact solution, except in the lower-center of the model where no travel times could be obtained from the unstacked data and where the rays are strongly refracted by the high velocity of the salt dome. This situation should be greatly improved by using pre-stack migration techniques.

And finally, figures 3(*h*) and (*i*) show the solutions computed using the exact same iteration sequence as above but with optimisation norms respectively 1.3 and 2.0. When the norm is 2.0 (least squares) the weights are all unity. We can see that for norms near 1, the robust features of the algorithm are retained, but as the norm approaches 2 the solution begins to look rather rough.

We have also compared our l_1 IRLS results with results obtained with the popular damped least-squares code LSQR. First, IRLS (l_1 implemented as we have shown via preconditioned conjugate gradient) is marginally slower than LSQR (damped l_2). For the tomography problem we have shown here with 6600 equations in 3200 unknowns, the solution times for both methods were less than 15 CPU seconds on an IBM 3090 scalar computer. Secondly, for this problem, which has clean, easily interpreted seismograms, the data are relatively noise-free and there are no outliers. Thus, by choosing an appropriate damping parameter (by trial and error) we were able to get results with LSQR that were very similar to those achieved with IRLS. But for problems with outliers purposely added to the data, the l_1 method shone.

6. Conclusions

A brief account of the origins and development of robust inversion methods has been given. In addition, a strategy has been outlined for the efficient implementation of such methods on large-scale problems via conjugate gradient and iteratively re-weighted least squares. The use of these methods has been illustrated with examples from seismic reflection tomography and inverse scattering, two of the most computationally intensive tasks encountered in inverse theory.

Regularisation of inverse problems by LAD methods has obvious statistical advantages, and would seem to be less *ad hoc* than damped least squares. The implementation of LAD via IRLS is very attractive computationally as well. Implemented as shown, IRLS is nearly as fast as damped least squares, it is easy to program and use, and it is readily adapted to take advantage of the sparsity of problems encountered in tomography. Further, we have found that l_1 inversion via IRLS routinely works at least as well or better than the more familiar damped l_2 (least-squares) methods whenever noise and ill-conditioning are significant.

Much theoretical work remains to be done however. Current work includes investigating the use of statistical data analysis as part of the inversion procedure, e.g.

for the optimal choice of the minimisation norm and the residual taper, perhaps even allowing these parameters to vary during the inversion. Further, the stopping criteria that we use have been adopted *mutatis mutandis* from l_2 algorithms and could, perhaps, be sharpened. Nevertheless, the stabilising effects of IRLS are so useful, especially in the presence of short bursts of noise, that the method merits careful study.

Acknowledgments

We are grateful to our colleagues at Amoco with whom we have had many stimulating discussions: Drs Sven Treitel, Samuel H Gray, Laurence R Lines, and Mr Phillip Bording.

References

Barrodale I and Roberts F 1973 An improved algorithm for discrete l_1 linear approximation *SIAM J. Numer. Anal.* **10** 839–48

Beaton A and Tukey J 1974 The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data *Technometrics* **16** 147–92

Bishop T, Bube K, Cutler R, Langan R, Love P, Resnick J, Shuey R, Spindler D and Wyld H 1955 Tomographic determination of velocity and depth in laterally varying media *Geophys.* **50** 903–23

Bloomfield P and Steiger W 1983 *Least Absolute Deviation* (Boston: Birkhäuser) ch. 1

Bording R, Gersztenkorn A, Lines L, Scales J and Treitel S 1987 Applications of seismic travel time tomography *Geophys. J. R. Astron. Soc.* **90** 295–303

Boscovich R 1763 *Theoria philosophiae naturalis, redacta ad unicam legem virium in natura existentium* Venice

Byrd R and Pyne D 1979 Johns Hopkins University, Baltimore, MD *Technical Report* 313

Chandra R 1978 Conjugate gradient methods for partial differential equations *PhD thesis* Yale University

Clayton R and Stolt R 1981 A Born–WKBJ inversion method for acoustic reflection data *Geophys.* **46** 1559–67

Coleman D, Holland P, Kaden N, Klema V and Peters S 1980 A system of subroutines for iteratively reweighted least squares *ACM Trans. Math. Software* **6** 327–36

Constable S, Parker R and Constable C 1987 Occam's inversion: a practical algorithm for generating smooth models from electromagnetic sounding data *Geophys.* **52** 289–300

Eisenhart C 1961 Boscovich and the combination of observations *Roger Joseph Boscovich* ed. L L Whyte (London: Allen and Unwin) pp 200–12

Fox L, Huskey H and Wilkinson J 1948 Notes on the solution of algebraic linear simultaneous equations *Q. J. Mech. Appl. Math.* **1** 149–73

Gauss C F 1809 *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium* Hamburg (Engl. transl. by Charles Henry Davis, Little, Brown, and Company, Boston 1857)

Gerstenkorn A 1984 Iterative inversion for the one-dimensional acoustic wave equation *MS thesis* University of Tulsa

Gersztenkorn A, Bednar J B and Lines L 1986 Robust iterative inversion for the one-dimensional acoustic wave equation *Geophys.* **51** 357–68

Gersztenkorn A and Scales J 1988 Smoothing seismic tomograms with alpha-trimmed means *Geophys. J.* **92** 67–72

Gray S 1984 The relatiaonship between 'direct, discrete' and 'iterative, continuous' one-dimensional inverse methods *Geophys.* **49** 54–9

Gutowski P and Treitel S 1987 The generalized one-dimensional synthetic seismogram *Geophys.* **52** 589–605

Hestenes M 1980 *Conjugate Direction Methods in Optimization* (Berlin: Springer)

Hestenes M and Stiefel E 1952 Methods of conjugate gradients for solving linear systems *NBS J. Res.* **49** 409–36

Holland P and Welsch R 1977 Robust regression using iteratively reweighted least squares *Commun. Stat.* A **6** 813–27

Huber P 1981 *Robust Statistics* (New York: Wiley)
Jammer M 1957 *Concepts of Force* (Harvard)
Luenberger D 1969 *Optimization by Vector Space Methods* (New York: Wiley)
Maire C and Boscovich R 1755 *De Litteraria Expeditione per Pontificiam ditionem ad dimetiendas duas Meridiani gradus, et corrigendam mappam geographicam, jussu, et auspiciis Benedicti XIV Pont. Max. suscepta* Rome
Newcomb S 1912 Researches on the motion of the moon, II *Astron Papers* **9** 1–249 (US Nautical Almanac Office)
Nolet G 1987 Seismic wave propagation and seismic tomography *Seismic Tomography* ed. G Nolet (Dordrecht: Reidel) pp 1–23
Raz S 1981 A direct profile inversion: beyond the Born model *Radio Sci.* **16** 347–53
Rice J 1964 *The Approximation of Functions* (London: Addison Wesley)
Scales J 1988 On using conjugate gradient to calculate the eigenvalues and singular values of large, sparse matrices *Geophys. J.* (submitted)
Scales J, Gersztenkorn A and Treitel S 1988 Fast l_p solution of large, sparse linear systems: application to seismic travel time tomography *J. Comput. Phys.* **75** 314–33
Scales J and Treitel S 1987 On the relation between Gauss' method and IRLS for l_1 inversion *IEEE Trans. ASSP* **35** 581–2
Schlossmacher E 1973 An iterative technique for absolute deviations curve fitting *J. Am. Stat. Assoc.* **68** 857–65
Shanno D and Rocke D 1986 Numerical methods for robust regression: linear models *SIAM J. Sci. Stat. Comput.* **7** 86–97
Stay B 1760 *Philosophiae Recentioris, a Benedicto Stay in Romano Archigynasis Publico Eloquentare Professore, versibus traditae, Libri X, cum adnotationibus et Supplementis P. Rogerii Josephi Boscovich S. J., Tomus II* Rome
Stigler S 1973 Simon Newcomb, Percy Daniell, and the history of robust estimation 1885–1920 *J. Am. Stat. Assoc.* **68** 872–9
Stoer J and Bulirsch R 1980 *Introduction to Numerical Analysis* (New York: Springer)
Stork C 1988 Travel time tomographic velocity analysis of seismic surface reflection data *PhD thesis* California Institute of Technology
Whyte L L (ed.) 1961 *Roger Joseph Boscovich* (London: Allen and Unwin)