Geophysical Prospecting 32, 159–186, 1984.

TUTORIAL
A REVIEW OF LEAST-SQUARES INVERSION AND ITS APPLICATION TO GEOPHYSICAL PROBLEMS*

L.R. LINES and S. TREITEL**

ABSTRACT

LINES, L.R. and TREITEL, S. 1984, Tutorial: A Review of Least-Squares Inversion and its Application to Geophysical Problems, Geophysical Prospecting 32, 159–186.

Geophysical inversion involves the estimation of the parameters of a postulated earth model from a set of observations. Since the associated model responses can be nonlinear functions of the model parameters, nonlinear least-squares techniques prove to be useful for performing the inversion. A common type of inversion applies iterative damped linear least squares through use of the Marquardt-Levenberg method. Traditionally, this method has been implemented by solving the associated normal equations in conventional ways. However, Singular Value Decomposition (SVD) produces significant improvements in computational precision when applied to the same system of normal equations. Iterative least-squares modeling finds application in a wide variety of geophysical problems. Two examples illustrate the approach: (1) seismic wavelet deconvolution, and (2) the location of a buried wedge from surface gravity data. More generally, nonlinear least-squares inversion can be used to estimate earth models for any set of geophysical observations for which an appropriate mathematical description is available.

1. GEOPHYSICAL INVERSE PROBLEMS

Geophysical inversion may be viewed as an attempt to fit the response of an idealized subsurface earth model to a finite set of actual observations. We distinguish between a *model*, its *model parameters*, and the associated *model response*. A model consists of a set of relations representing a particular mathematical abstraction of an observed process. These equations in turn depend on a certain number, say p, of model parameters which we desire to estimate from the actual data. The model response consists of the synthetic data produced by a particular realization of

* Received December 1982, last material May 1983.
** Amoco Prod. Co., POB 59A, Tulsa, OK 74102, USA.

the model. The purpose of inversion is to extract the model parameter estimates from an attempted fit of the model response to the observed data. For example, the one-dimensional stratified acoustic medium obeying the acoustic wave equation is a *model*, whose *model parameters* are the layer impedances and layer thicknesses, and whose *model response* is the familiar normal incidence impulsive synthetic seismogram. At the risk of belaboring the obvious, we remark that if the model is physically irrelevant, the parameter estimates we may extract from it are meaningless. Because actual geophysical data are of finite extent and because they always contain components which we cannot explain in geologic terms, the inversion can *never* be unique—that is, more than one solution will satisfy the observations within a prescribed error.

In this treatment we make no pretense to an exhaustive survey of the art of geophysical inversion. Rather, we have chosen the least-squares approach because of its mathematical robustness when the recorded data are, in the words of Jackson (1972), "inaccurate, insufficient, and inconsistent." Numerous pertinent studies have appeared during the last two decades. These include the inversion of global earth data by use of free oscillations (Backus and Gilbert 1967, 1968, Wiggins 1972), the inversion of traveltime data (Crosson 1976, Neumann, Gjoystdal and Ursin 1981), the inversion of electromagnetic data (Inman 1975, Jupp and Vozoff 1975, Oristaglio and Worthington 1980), and the inversion of gravity data (Oldenburg 1974, Vigneresse 1977). A comprehensive summary of inverse methods for remote sensing has been given by Twomey (1977), while excellent treatments of the underlying mathematical concepts have been published by Jackson (1972, 1979) and Aki and Richards (1980).

Many geophysical model responses are nonlinear functions of the appropriate model parameters. The so-called "Marquardt-Levenberg" approach provides us with a powerful and versatile method to estimate such parameters in an iterative manner (Levenberg, 1944, Marquardt 1963). Our paper is an attempt to describe the salient features of least-squares inversion in relatively simple terms. However, we cannot avoid the use of matrix and vector algebra, so we must assume that the reader has the fortitude to bear with us.

2. Least-Squares Inversion

A model response can be either a linear or a nonlinear function of the model parameters. Thus for the linear convolutional model $y_t = a_t * b_t$, where $*$ denotes convolution, the data y_t are a linear function of the model parameters a_t and b_t, but if, say, $y_t = a_t^2$, then y_t is a nonlinear function of a_t. A comprehensive discussion of the relation between nonlinear and linear least-squares inversion methods can be found in the book by Draper and Smith (1981, chapter 10). A remark on terminology is in order: Draper and Smith are statisticians and use the word "regression" where we use "inversion". To complicate matters further, electrical engineers frequently refer to these methods as "systems identification" or "parameter identification". Other terms are also in use.

2.1. *Linear least-squares inversion*

Before proceeding to *nonlinear* inverse problems, we examine the *simpler* linear inverse problem. The extension of these ideas to the nonlinear case is introduced in section 2.3.

The basic strategy is to minimize the sum of squares of the errors between the model response and the observations. Let the n observations for a set of earth data be represented by the vector

$$\mathbf{y} = \text{col}\ (y_1, y_2, \ldots, y_n), \tag{1}$$

and let the model response be the vector

$$\mathbf{f} = \text{col}\ (f_1, f_2, \ldots, f_n). \tag{2}$$

The model is a function of p parameters which are elements of a vector $\boldsymbol{\theta}$,

$$\boldsymbol{\theta} = \text{col}\ (\theta_1, \theta_2, \ldots, \theta_p). \tag{3}$$

Let θ_j^0 be an initial estimate of the parameter θ_j, $(j = 1, \ldots, p)$, and let $\mathbf{f}^0$ be the initial model response. If the model response f is a *linear* function of the parameters, a perturbation of the model response about θ^0 can be represented by the first-order Taylor expansion:

$$\mathbf{f} = \mathbf{f}^0 + \sum_{j=1}^{p} \left.\frac{\partial \mathbf{f}}{\partial \theta_j}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^0} (\theta_j - \theta_j^0), \tag{4}$$

or, in matrix notation,

$$\mathbf{f} = \mathbf{f}^0 + \mathbf{Z}\boldsymbol{\delta}, \tag{5}$$

where $\mathbf{Z}$ is the $n \times p$ Jacobian matrix of partial derivatives with elements

$$Z_{ij} = \frac{\partial f_i}{\partial \theta_j}, \tag{6}$$

and $\boldsymbol{\delta} = \boldsymbol{\theta} - \boldsymbol{\theta}^0$ is the parameter change vector with elements δ_j representing the changes, or perturbations in the parameters θ_j, that is

$$\delta_j = \theta_j - \theta_j^0, \qquad (j = 1, \ldots, p).$$

Our choice of perturbations in $\boldsymbol{\theta}$ will be made so as to minimize the sum of squares of the errors between the model response and the data. Let $\mathbf{e}$ represent the error vector expressing the difference between the model response $\mathbf{f}$ and the observed data $\mathbf{y}$:

$$\mathbf{y} - \mathbf{f} = \mathbf{e}. \tag{7}$$

Combining (5) and (7) yields

$$\mathbf{y} - (\mathbf{f}^0 + \mathbf{Z}\boldsymbol{\delta}) = \mathbf{e}$$

or

$$\mathbf{y} - \mathbf{f}^0 = \mathbf{Z}\boldsymbol{\delta} + \mathbf{e}. \tag{8}$$

The vector $\mathbf{y} - \mathbf{f}^0$, which contains the differences between the initial model response and the observed data, is called the discrepancy vector $\mathbf{g}$, so that

$$\mathbf{g} = \mathbf{y} - \mathbf{f}^0$$

and

$$\mathbf{e} = \mathbf{g} - \mathbf{Z}\boldsymbol{\delta}. \tag{9}$$

Before we describe a least-squares method leading to an estimate of the parameter change vector $\boldsymbol{\delta}$ we note that geophysical inverse problems are generally not well posed, that is, the $n \times p$ Jacobian matrix $\mathbf{Z}$ is usually not square and of full rank. In fact, many of these problems are overdetermined, that is, the number of data points exceeds the number of model parameters, i.e., $n > p$; the matrix $\mathbf{Z}$ may or may not be of full rank. In the case of rank deficiency, techniques exist to alleviate the resulting problem. The Marquardt-Levenberg method (see section 2.2) can accomplish this very well. A lucid treatment of the problem has been given by Lanczos (1961, chapter III).

In the simplest least-squares or "Gauss-Newton" approach, we seek to minimize the cumulative squared error $S = \mathbf{e}^T\mathbf{e}$ with respect to the parameter change vector $\boldsymbol{\delta}$. From (9), we have

$$S = \mathbf{e}^T\mathbf{e} = (\mathbf{g} - \mathbf{Z}\boldsymbol{\delta})^T(\mathbf{g} - \mathbf{Z}\boldsymbol{\delta}). \tag{10}$$

Minimization of $\mathbf{S}$ with respect to $\boldsymbol{\delta}$ requires that

$$\frac{\partial S}{\partial \boldsymbol{\delta}} = 0. \tag{11}$$

(Differentiation with respect to a vector is described by Graybill (1969) and implies that $\partial S/\partial \delta_i = 0$ for all i.)

Substituting (10) into (11) gives

$$\frac{\partial}{\partial \boldsymbol{\delta}} (\boldsymbol{\delta}^T\mathbf{Z}^T\mathbf{Z}\boldsymbol{\delta} - \mathbf{g}^T\mathbf{Z}\boldsymbol{\delta} - \boldsymbol{\delta}^T\mathbf{Z}^T\mathbf{g} + \mathbf{g}^T\mathbf{g}) = 0. \tag{12}$$

Carrying out the differentiation with respect to $\boldsymbol{\delta}$, we obtain the so-called "normal equations"

$$\mathbf{Z}^T\mathbf{Z}\boldsymbol{\delta} = \mathbf{Z}^T\mathbf{g}, \tag{13}$$

whose solution for the parameter change vector $\boldsymbol{\delta}$ is

$$\boldsymbol{\delta} = (\mathbf{Z}^T\mathbf{Z})^{-1}\,\mathbf{Z}^T\mathbf{g}, \tag{14}$$

which is also known as the Gauss-Newton solution. This set of equations has had many important applications in inverse theory and digital filtering. The above results can also be derived from purely geometrical considerations—see, for example, Strang (1980, chapter 3). In fact, the term "normal" arises from the geometric property that the *least-squares* error vector $\mathbf{e}$ is perpendicular, or normal to the column vectors of the matrix $\mathbf{Z}$.

The actual value of the cumulative least-squares error $\hat{S}$ results by substituting the least-squares solution (14) into (10),

$$\hat{S} = (\mathbf{Z}\hat{\boldsymbol{\delta}} - \mathbf{g})^T(\mathbf{Z}\hat{\boldsymbol{\delta}} - \mathbf{g})$$

where $\hat{\boldsymbol{\delta}}$ is the least-squares solution (14). Substituting this expression for $\hat{\boldsymbol{\delta}}$ in the above equation and carrying out some matrix algebra yields

$$\hat{S} = \mathbf{g}^T(\mathbf{I}_n - \mathbf{Z}\mathbf{Z}_L^{-1})\,\mathbf{g},$$

where $\mathbf{Z}_L^{-1} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T$. The $p \times n$ matrix $\mathbf{Z}_L^{-1}$ is the least-squares, or Lanczos inverse of $\mathbf{Z}$ (see also section 3.1), and $\mathbf{I}_n$ is the $n \times n$ identity matrix. We observe that $\hat{S}$ becomes small whenever the product $\mathbf{Z}\mathbf{Z}_L^{-1}$ approaches $\mathbf{I}_n$. In the literature, the Lanczos inverse is also often referred to as the "natural" or "generalized inverse" of the $n \times p$ rectangular matrix $\mathbf{Z}$.

The least-squares solution (14) has resulted from the straightforward minimization of the cumulative squared error S. This so-called "unconstrained" least-squares solution may have some undesirable properties, which we now describe.

An obvious difficulty occurs when the inverse of $\mathbf{Z}^T\mathbf{Z}$ does not exist, that is, when the matrix $\mathbf{Z}^T\mathbf{Z}$ is singular. Even if $\mathbf{Z}^T\mathbf{Z}$ exists, we may well be faced with a diverging solution, or we may have to contend with slow convergence. This can happen whenever the initial estimate of the model, $\mathbf{f}^0$, is poor (Draper and Smith 1981). As soon as $(\mathbf{Z}^T\mathbf{Z})$ becomes nearly singular, the elements of the solution vector $\boldsymbol{\delta}$ tend to grow without bound. Smith and Shanno (1971) have described how a nearly singular $\mathbf{Z}^T\mathbf{Z}$ tends to catapult the updated parameter vector $\boldsymbol{\theta} = \boldsymbol{\theta}^0 + \boldsymbol{\delta}$ away from an acceptable solution.

2.2. *The Marquardt-Levenberg method*

In order to reduce the difficulties when the matrix $\mathbf{Z}^T\mathbf{Z}$ is nearly singular, we solve an alternate least-squares problem. We impose the constraining condition that the sum of the squares, or energy of the elements of the parameter change vector $\boldsymbol{\delta}$, be bounded by a finite quantity, say δ_0^2. This approach was introduced by Levenberg (1944) and later described in detail by Marquardt (1963). It is sometimes known as the method of damped least squares; others refer to it as "ridge regression" (Inman 1975). The effect of this constraint is to prevent unbounded oscillations in the solution; i.e., to smooth the parameter change vector $\boldsymbol{\delta}$.

The constrained least-squares solution arises by solving a Lagrange multiplier problem in which $\mathbf{e}^T\mathbf{e}$ is minimized subject to the constraint that $\boldsymbol{\delta}^T\boldsymbol{\delta} = \delta_0^2$. Thus, we choose $\boldsymbol{\delta}$ to minimize a cost function $S(\boldsymbol{\delta}, \beta)$,

$$S(\boldsymbol{\delta}, \beta) = \boldsymbol{e}^T\boldsymbol{e} + \beta(\boldsymbol{\delta}^T\boldsymbol{\delta} - \delta_0^2), \tag{15}$$

where β is a Lagrange multiplier.

Differentiation with respect to the vector $\boldsymbol{\delta}$ yields a modified form of the normal equations

$$(\mathbf{Z}^T\mathbf{Z} + \beta\mathbf{I})\boldsymbol{\delta} = \mathbf{Z}^T\mathbf{g}, \tag{16a}$$

so that

$$\boldsymbol{\delta} = (\mathbf{Z}^T\mathbf{Z} + \beta\mathbf{I})^{-1}\,\mathbf{Z}^T\mathbf{g}. \tag{16b}$$

Comparison of (14) with (16b) shows that the constraint has produced a method for avoiding singularities or near singularities in the matrix $\mathbf{Z}^T\mathbf{Z}$. By adding a constant, β, to the main diagonal of $\mathbf{Z}^T\mathbf{Z}$, we have effectively added a d.c. level to the eigenvalues of the matrix $\mathbf{Z}^T\mathbf{Z}$ so that none of the eigenvalues can vanish. Levenberg (1944) terms the Lagrange multiplier β a "damping factor", since it effectively damps out changes in the parameter vector $\boldsymbol{\theta}$ by limiting the energy in the parameter discrepancy vector $\boldsymbol{\delta}$.

The solution (16b) has several other interesting characteristics. It is hybrid, because it combines the so-called "method of steepest descent" with the method of least squares. The steepest descent solution is normal to a given cost function contour, for which $S(\boldsymbol{\theta}) = \mathbf{e}^T\mathbf{e} = \text{constant}$, or $dS(\boldsymbol{\theta}) = 0$ (Smith and Shanno 1971). This condition is satisfied by the steepest descent column vector with p components

$$\boldsymbol{\delta}_g = -\nabla S(\boldsymbol{\theta}),$$

where ∇ is the gradient operator $\partial/\partial\theta_j, j = 1, \ldots, p$. We note that $+\nabla S(\boldsymbol{\theta})$ is then the steepest ascent vector.

Since

$$S(\boldsymbol{\theta}) = \mathbf{e}^T\mathbf{e} = \sum_{i=1}^{n} e_i^2,$$

the jth element of the column vector $\nabla S(\boldsymbol{\theta})$ is

$$\frac{\partial S}{\partial \theta_j} = 2\sum_{i=1}^{n} \frac{\partial e_i}{\partial \boldsymbol{\theta}_j}\, e_i, \qquad j = 1, \ldots, p.$$

Because $\mathbf{e}_i = \mathbf{y}_i - \mathbf{f}_i$ (see (7)), we have

$$\frac{\partial S}{\partial \boldsymbol{\theta}_j} = -2\sum_{i=1}^{n} \frac{\partial \mathbf{f}_i}{\partial \boldsymbol{\theta}_j}\, e_i, \qquad j = 1, \ldots, p, \tag{17}$$

and hence the column vector $\boldsymbol{\delta}_g = -\nabla S(\boldsymbol{\theta})$ is

$$\boldsymbol{\delta}_g = 2\mathbf{Z}^T\boldsymbol{e}$$

or

$$\boldsymbol{\delta}_g = 2\mathbf{Z}^T(\mathbf{y} - \mathbf{f}),$$

where $\mathbf{Z}^T$ is the transpose of the $n \times p$ Jacobian matrix $\mathbf{Z}$. The iteration begins with an *initial model* whose response is $\mathbf{f} = \mathbf{f}^0$, so

$$\boldsymbol{\delta}_g = 2\mathbf{Z}^T(\mathbf{y} - \mathbf{f}^0).$$

But $\mathbf{y} - \mathbf{f}^0 = \mathbf{g}$ (see (8)), and the steepest descent vector is

$$\boldsymbol{\delta}_g = 2\mathbf{Z}^T\mathbf{g}. \tag{18}$$

Since $\boldsymbol{\delta}_g$ is in the direction of decreasing S, convergence tends to occur, but usually at a slow rate. Further, computational problems can occur when the step size $\boldsymbol{\delta}_j = \boldsymbol{\theta}_j - \boldsymbol{\theta}_j^0$ becomes too small.

Generally speaking, the steepest descent method is optimal when $S(\boldsymbol{\theta})$ is large, while the least-squares method becomes effective when $S(\boldsymbol{\theta})$ is small. The Marquardt-Levenberg approach thus exploits the useful properties of both methods. The technique determines the parameter change vector $\boldsymbol{\delta}$ at each step in the iteration. Marquardt (1963) proved that the vectors $\boldsymbol{\delta}$ and $\boldsymbol{\delta}_g$ must lie within at most 90° of each other, and that this is a necessary property for convergence. Comparison of (16b) with (18) shows that as β becomes large $\boldsymbol{\delta} \approx \beta^{-1}\mathbf{Z}^T\boldsymbol{g}$, i.e., $\boldsymbol{\delta}$ becomes proportional to the gradient vector $\boldsymbol{\delta}$. However, as β becomes larger and larger, the elements of the parameter change vector $\boldsymbol{\delta}$ become smaller and smaller. Marquardt also showed that the angle between $\boldsymbol{\delta}$ and $\boldsymbol{\delta}_g$ is a monotonically decreasing function of β.

These matters are sketched in fig. 1 for a model with two parameters θ_1 and θ_2. In such a simple case the curves of constant cumulative squared error $S(\theta_1, \theta_2)$ appear as contours in the $\theta_1 - \theta_2$ plane. Suppose our starting point is

$$\boldsymbol{\theta}^0 = (\theta_1^0, \theta_2^0),$$

and suppose the problem has only a single minimum at the point

$$\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\theta}_2),$$

which is then the desired iterative least-squares solution. Figure 1 illustrates the first step in our attempt to reach this minimum when we use either the steepest descent, the Marquardt-Levenberg, or the Gauss-Newton approaches. We observe that the steepest descent vector is perpendicular to the outermost error contour at the point $\boldsymbol{\theta}^0 = (\theta_1^0, \theta_2^0)$. The Marquardt-Levenberg vector is bracketed by the Gauss-Newton and the steepest descent vectors, and in a way offers a compromise between either extreme. Successive application of any one of these algorithms will eventually reach

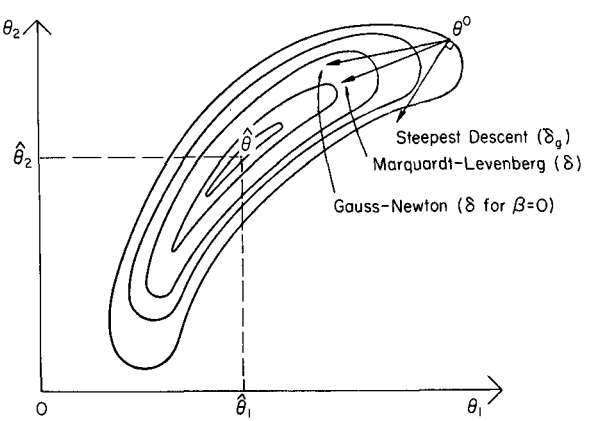

Fig. 1. The geometric relation between the Gauss-Newton, Marquardt-Levenberg, and Steepest Descent solutions for a two-parameter example (after Smith and Shanno 1971).

the minimum at $(\hat{\theta}_1, \hat{\theta}_2)$; in general, the Marquardt-Levenberg method will do so most efficiently (see also section 4.2).

A particular choice of β in (16b) allows either the linear least-squares method or the steepest descent method to dominate the parameter search. Setting $\beta = 0$ implies that the linear least-squares method predominates, while allowing β to increase moves the technique towards the method of steepest descent. Initially β is set to a large positive value, so that the good initial convergence properties of the steepest descent method can come into play. Then, as this happens, β is reduced by multiplying it by a constant factor <1 so that the linear least-squares method may take over in the region closer to a solution. If divergence occurs during a given iteration, β is divided once more by this factor until the error drops and convergence resumes (at least one hopes it does).

A further means to speed the convergence of (16b) involves scaling. Marquardt (1963) showed how to determine a diagonal matrix $\mathbf{D}$ such that $\mathbf{ZD}^{-1}$ places ones along the main diagonal of the scaled matrix $(\mathbf{ZD}^{-1})^T(\mathbf{ZD}^{-1})$. The element d_i of $\mathbf{D}$ is equal to the root mean sum of squares (r.m.s.) value of the elements in the ith column of the unscaled Jacobian matrix $\mathbf{Z}$. The resulting solution $\boldsymbol{\delta}^{(D)}$ must then be rescaled in the form

$$\boldsymbol{\delta} = \mathbf{D}^{-1}\boldsymbol{\delta}^{(D)}$$

Smith and Shanno (1971) give the details of this very useful scaling procedure.

One might conclude that the solution (16b) minimizes the cost function $S(\boldsymbol{\delta}, \beta)$ of (15) merely because $(\mathbf{Z}^T\mathbf{Z} + \beta\mathbf{I})$ is always positive definite (Marquardt 1963). However, Dennis (1977) shows that this is so only if the errors are linear functions of the model parameters. In particular, we note that $\partial S/\partial\boldsymbol{\delta} = 0$ is a necessary but not sufficient condition for the minimization of S. To establish sufficiency, we require knowledge of the second derivative or Hessian matrix $\mathbf{H}$, whose elements are

$$\mathbf{H}_{jk} = \frac{\partial^2 S}{\partial\theta_j\,\partial\theta_k}. \tag{19}$$

The matrix $\mathbf{H}$ must be positive definite in order that S be minimum (Fletcher 1980, pp. 10–11).

If we write

$$S = \sum_{i=1}^{n} e_i^2$$

from which

$$\frac{\partial S}{\partial\theta_j} = 2\sum_{i=1}^{n} e_i \frac{\partial e_i}{\partial\theta_j}, \qquad j = 1, \ldots, p,$$

then $\mathbf{H}_{jk}$ can be expressed in the form

$$\mathbf{H}_{jk} = \frac{\partial^2 S}{\partial\theta_j\,\partial\theta_k} = 2\sum_{i=1}^{n}\left(\frac{\partial e_i}{\partial\theta_j}\frac{\partial e_i}{\partial\theta_k} + e_i\frac{\partial^2 e_i}{\partial\theta_j\,\partial\theta_k}\right).$$

From (7), we obtain

$$\frac{\partial e_i}{\partial \theta_j} = -\frac{\partial f_i}{\partial \theta_j}$$

and then

$$\mathbf{H}_{jk} = 2\sum_{i=1}^{n}\left(\frac{\partial f_i}{\partial \theta_j}\frac{\partial f_i}{\partial \theta_k} + e_i\frac{\partial^2 e_i}{\partial \theta_j\,\partial \theta_k}\right) = 2\sum_{i=1}^{n}\left(\mathbf{Z}_{ij}\mathbf{Z}_{ik} + e_i\frac{\partial^2 e_i}{\partial \theta_j\,\partial \theta_k}\right),$$

which can be written in matrix form,

$$\mathbf{H} = 2\mathbf{Z}^T\mathbf{Z} + 2\mathbf{H}'. \tag{20}$$

If the errors are linear in the parameter changes, $2\mathbf{H}'$ vanishes and the Hessian matrix is simply $2\mathbf{Z}^T\mathbf{Z}$. In this case, the positive definiteness of $\mathbf{Z}^T\mathbf{Z}$ assures the least-squares minimization of S; if the errors are nonlinear in the parameter changes, the second term $\mathbf{H}'$ of (20) does not vanish, and minimization does not necessarily occur.

The method of constrained least squares is also relevant to the design of stabilized Wiener deconvolution filters. The addition of a positive constant to the main diagonal of the matrix $\mathbf{Z}^T\mathbf{Z}$ is known as "prewhitening" in digital filtering terminology, since the addition of white noise to a signal will add power to the main diagonal of its associated autocorrelation matrix. In the case of filter design, the constraint is placed on the power of the filtered noise. A discussion of constrained least squares in this context has been given by Treitel and Lines (1982).

One of the main advantages of the least-squares inversion method is its applicability to almost any problem for which a model can be constructed. It is generally easier to solve the "forward problem" that transforms a set of model parameters into a synthetic data set, than to proceed in the opposite direction and solve the inverse problem. Having found a method of finding the model response $\mathbf{f}$ from the parameters $\boldsymbol{\theta}$, we must compute the Jacobian matrix of partial derivatives, $\mathbf{Z}_{ij} = \partial f_i/\partial \theta_j$. These derivatives can be determined by formal differentiation if the model is simple enough. In other cases, the partial derivatives must be approximated by finite differences. This can be computationally expensive since we need to determine two model responses for every value of $\boldsymbol{\theta}_j$ even if we use the simplest forward difference formula which approximates $\partial f_i/\partial \theta_j$ by

$$\frac{f_i(\theta_j + \Delta\theta_j) - f_i(\theta_j)}{\Delta\theta_j}.$$

Clearly one should avoid using more parameters θ_j ($j = 1, \ldots, p$) than is absolutely necessary (the principle of parsimony!). Alternatively, one may have recourse to some special methods for finding partial derivatives, see e.g., Oristaglio and Worthington (1980). Apart from such difficulties, iterative least-squares modeling is very versatile and can be adapted to a wide range of geophysical inverse problems, as evidenced by the number of references previously given.

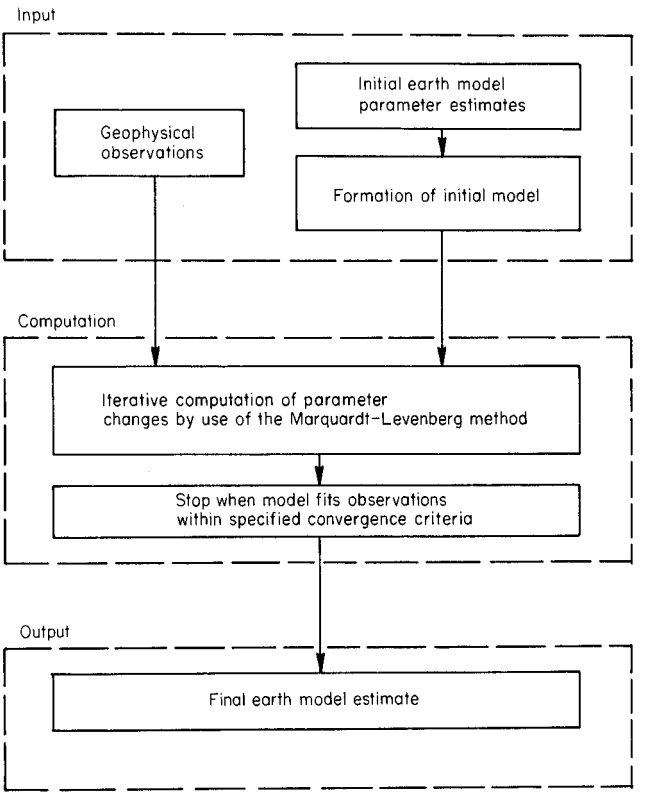

Fig. 2. Flow diagram of the nonlinear least-squares inversion method.

2.3. *Iterative least-squares and nonlinear problems*

In the preceding discussions we have outlined the Marquardt-Levenberg method for the constrained least-squares solution. Our nonlinear inversion approach uses constrained least squares iteratively to update the parameter vector for a given geophysical model. Although our model is often a nonlinear function of some or all of the parameters θ_i, we can often estimate the parameter vector $\boldsymbol{\theta}$ by a sequence of constrained *linear* least-squares estimates.

An outline of the iterative least-squares approach to geophysical inversion is described in the flow diagram of fig. 2. The parameter changes from the initial response estimates are determined by use of the fundamental relation (16b). We obtain an updated set of parameters, which are then used to compute a new model response estimate. At each stage, the sum of the squares of the error between the model response and the observation values is monitored. The iterative search for the parameter estimates terminates whenever either the squared error or a relative change in the squared error become less than a prespecified value. After these convergence criteria have been satisfied, we may say that the estimated geophysical parameters have produced a model which has matched the data within our specifi-

cations. *However, as will be seen in a later discussion on nonuniqueness, a good match between the model and the observations does not necessarily guarantee that the correct solution has been found.*

This so-called "iterative nonlinear least-squares inversion" approach is best illustrated by a set of examples (see section 4).

3. Singular Value Decomposition

3.1 Computational aspects

We have shown that the linear least-squares solution can be found by solving the normal equations (13)

$$\mathbf{Z}^T\mathbf{Z}\boldsymbol{\delta} = \mathbf{Z}^T\boldsymbol{g}.$$

This solution requires that $\mathbf{Z}^T\mathbf{Z}$ and $\mathbf{Z}^T\mathbf{g}$ be formed by matrix multiplication. $\mathbf{Z}^T\mathbf{Z}$ is a non-negative definite square symmetric matrix which can always be made positive definite by use of Marquardt's damping factor. We could therefore solve (13) by use of the Cholesky decomposition, so that no matrix inversions need be calculated explicitly (Lawson and Hanson 1974).

However, the formation of $\mathbf{Z}^T\mathbf{Z}$ and $\mathbf{Z}^T\mathbf{g}$ involves numerical inaccuracies. Such inaccuracies can be troublesome for large values of n and p. Golub and Reinsch (1970) recognized this problem and proposed that instead of dealing with (13), it is better to attempt a solution of the rectangular system

$$\mathbf{Z}\boldsymbol{\delta} = \mathbf{g}, \tag{21}$$

where we recall that the Jacobian matrix $\mathbf{Z}$ is generally $n \times p$. The formal solution of (21) is

$$\boldsymbol{\delta} = \mathbf{Z}^{-1}\mathbf{g},$$

but the conventional inverse $\mathbf{Z}^{-1}$ exists only if $\mathbf{Z}$ is square ($n = p$) and nonsingular. For most geophysical problems $n \gg p$, and in that case the inverse $\mathbf{Z}^{-1}$ must be identified with the so-called natural, or generalized inverse (see, e.g., Lanczos 1961, chapter III).

Golub and Reinsch (1970) have developed an efficient algorithm to solve (21), which makes use of the singular value decomposition (SVD). This procedure factors $\mathbf{Z}$ into a product of three matrices,

$$\mathbf{Z} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^T, \tag{22}$$

where $\mathbf{U}$ is an $n \times p$ matrix whose columns contain p of the total of n orthonormal "observation" eigenvectors $\mathbf{u}_i$, namely those which are *not* associated with the at least n-p null eigenvalues of $\mathbf{Z}\mathbf{Z}^T$. The u_i's satisfy

$$\mathbf{Z}\mathbf{Z}^T\mathbf{u}_i = \lambda_i^2\mathbf{u}_i, \qquad i = 1, 2, \ldots, p, p+1, \ldots, n \tag{23}$$

but with $\lambda_{p+1} = \lambda_{p+2} = \cdots = \lambda_n = 0$. Further, $\mathbf{V}$ is a $p \times p$ matrix whose columns contain the p orthonormal "parameter" eigenvectors $\mathbf{v}_i$ satisfying

$$\mathbf{Z}^T\mathbf{Z}\mathbf{v}_i = \lambda_i^2\mathbf{v}_i. \tag{24}$$

Finally, $\mathbf{\Lambda}$ is a $p \times p$ diagonal matrix containing the at most p non-null singular values $+\lambda_i$—or positive square roots of the eigenvalues λ_i^2—of $\mathbf{Z}^T\mathbf{Z}$. A more detailed description of the SVD method can be found in the book by Lawson and Hanson (1974), who also give beautifully documented FORTRAN subroutines for its numerical implementation. For the following SVD manipulations, we note that $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{V}\mathbf{V}^T = \mathbf{I}_p$, where $\mathbf{I}$ is the $p \times p$ identity matrix. Because $\mathbf{U}$ does *not* contain the eigenvectors $\mathbf{u}_i$ associated with the at least n-p null singular values of $\mathbf{U}\mathbf{U}^T$, we do *not* necessarily have that $\mathbf{U}\mathbf{U}^T = \mathbf{I}_n$, where $\mathbf{I}_n$ is the $n \times n$ identity matrix.

Wiggins, Larner and Wisecup (1976) have pointed out that the factorization of $\mathbf{Z}$ in terms of $\mathbf{U}$ and $\mathbf{V}$ is analogous to ordinary spectral decomposition in terms of sinusoids, which play the role of the observation and parameter eigenvectors $\mathbf{u}_i$ and $\mathbf{v}_i$.

Writing the equation $\mathbf{Z}\boldsymbol{\delta} = \mathbf{g}$ in terms of $\mathbf{U}$ and $\mathbf{V}$ gives

$$\mathbf{U}\mathbf{\Lambda}\mathbf{V}^T\boldsymbol{\delta} = \mathbf{g}.$$

Premultiplying both sides by $\mathbf{U}^T$, we obtain

$$\mathbf{\Lambda}\mathbf{V}^T\boldsymbol{\delta} = \mathbf{U}^T\mathbf{g}. \tag{25}$$

Following Wiggins et al. (1976), these expressions may be recast in terms of the individual non-null singular values λ_i and the eigenvectors $\mathbf{u}_i$ and $\mathbf{v}_i$. In other words, (25) can be written in the form of the p vector inner product relations

$$\lambda_i\mathbf{v}_i^T\boldsymbol{\delta} = \mathbf{u}_i^T\mathbf{g} \qquad \text{for} \quad i = 1, 2, \ldots, p. \tag{26}$$

The inner product (or vector dot product) $\mathbf{v}_i^T\boldsymbol{\delta}$ is the magnitude of the projection of the parameter change vector $\boldsymbol{\delta}$ onto the ith parameter eigenvector $\mathbf{v}_i$. Similarly, the dot product $\mathbf{u}_i^T\mathbf{g}$ is the magnitude of the projection of the discrepancy vector $\mathbf{g}$ onto the ith observation eigenvector $\mathbf{u}_i$. These projection magnitudes are *not* independent of each other for $i = 1, \ldots, p$; rather, they are coupled through the ith singular value λ_i, as (26) indicates. Whenever $\mathbf{u}_i^T\mathbf{g}$ is small, the implication is that the ith observation eigenvector $\mathbf{u}_i$ will contribute very little to the solution $\boldsymbol{\delta}$.

For the geophysical inverse problems treated here, we usually have $n > p$, that is, in our model responses the number of data points exceeds the number of parameters. However, experience indicates that some singular values λ_i may become small, but rarely do they vanish completely. Small singular values can affect the solution adversely. Among various methods existing to correct this condition, the use of Marquardt's damping factor is particularly useful, as we illustrate in Section 3.2 below.

For the cases $n > p$, an inverse for $\mathbf{Z}$ does not exist in the usual sense. However, the decomposition $\mathbf{Z} = \mathbf{U}\Lambda\mathbf{V}^T$ suggests use of the Lanczos (or least-squares) inverse

$$\mathbf{Z}_L^{-1} = \mathbf{V}\Lambda^{-1}\mathbf{U}^T \tag{27}$$

to solve $\mathbf{Z}\delta = \mathbf{g}$ (Jackson 1972), which we already encountered in section 2.1. We obtain

$$\delta = \mathbf{V}\Lambda^{-1}\mathbf{U}^T\mathbf{g} = \mathbf{Z}_L^{-1}\mathbf{g}. \tag{28}$$

This solution is in fact the least-squares solution (14), but written in terms of the singular value decomposition (22). We show this by substituting (22) into (14),

$$\begin{aligned}
\delta &= (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{g} \\
&= (\mathbf{V}\Lambda\mathbf{U}^T\mathbf{U}\Lambda\mathbf{V}^T)^{-1}\mathbf{V}\Lambda\mathbf{U}^T\mathbf{g} \\
&= (\mathbf{V}\Lambda^2\mathbf{V}^T)^{-1}\mathbf{V}\Delta\mathbf{U}^T\mathbf{g} \\
&= \mathbf{V}\Lambda^{-2}\mathbf{V}^T\mathbf{V}\Lambda\mathbf{U}^T\mathbf{g} \\
&= \mathbf{V}\Lambda^{-1}\mathbf{U}^T\mathbf{g} \\
&= \mathbf{Z}_L^{-1}\mathbf{g},
\end{aligned}$$

which is (28). Let us take a closer look at the SVD form of our solution $\delta = \mathbf{V}\Lambda^{-1}\mathbf{U}^T\mathbf{g}$, which we can expand as the weighted vector product sum

$$\delta = \frac{1}{\lambda_1}\mathbf{v}_1\mathbf{u}_1^T\mathbf{g} + \frac{1}{\lambda_2}\mathbf{v}_2\mathbf{u}_2^T\mathbf{g} + \cdots + \frac{1}{\lambda_p}\mathbf{v}_p\mathbf{u}_p^T\mathbf{g}.$$

Let $\alpha_i = \mathbf{u}_i^T\mathbf{g}$ $(i = 1, \ldots, p)$ be the magnitude of the projection of the discrepancy vector $\mathbf{g}$ onto the ith observation eigenvector $\mathbf{u}_i$, so that

$$\delta = \frac{\alpha_1}{\lambda_1}\mathbf{v}_1 + \frac{\alpha_2}{\lambda_2}\mathbf{v}_2 + \cdots + \frac{\alpha_p}{\lambda_p}\mathbf{v}_p.$$

We see that the solution vector δ is the weighted sum of the p parameter eigenvectors $\mathbf{v}_i$ with weights α_i/λ_i. In particular, if α_i/λ_i is small, the term $(\alpha_i/\lambda_i)\mathbf{v}_i$ has relatively little influence on the solution δ; this was already evident from (26). Further, should any λ_i be very small, α_i/λ_i will tend to be large (unless α_i is simultaneously very small), which means that the term $(\alpha_i/\lambda_i)\mathbf{v}_i$ will have a dominant influence on the solution δ. There is no cause for alarm unless λ_i is so small that we must question its numerical accuracy. In such cases we may opt to cast out the offending term altogether, or we may prefer the less drastic step of reducing its influence on δ, for example, by recourse to Marquardt's damping factor (see section 3.2).

The computational benefits of using the SVD solution (28) rather than the conventional solution (14) can best be appreciated with a small numerical example. Consider a simple well posed problem with two equations and two unknowns,

$$\begin{aligned}
4\delta_1 + 2\delta_2 &= 6.0, \\
2\delta_1 - \delta_2 &= 1.8.
\end{aligned}$$

The exact solution for this problem is the vector

$$\begin{bmatrix}\delta_1\\ \delta_2\end{bmatrix}=\begin{bmatrix}1.2\\ 0.6\end{bmatrix}.$$

We consider the effect of scaling on these solutions by multiplying the first equation by an arbitrary integral power of 10 and then computing the solution for

$$\mathbf{Z}\boldsymbol{\delta}=\mathbf{g},$$

where

$$\mathbf{Z}=\begin{bmatrix}4.0s & 2.0s\\ 2.0 & -1.0\end{bmatrix}$$

and

$$\mathbf{g}=\begin{bmatrix}6.0s\\ 1.8\end{bmatrix},$$

and where

$$s=10^m.$$

Table 1. *Numerical comparison between SVD and normal equation solutions.*

Scaling factors	Method (i) (Cholesky)	Method (ii) (Crout)	Method (iii) SVD
10^0	1.199 999 81	1.199 996 95	1.199 998 86
	0.600 002 62	0.599 998 45	0.599 994 87
10^1	1.200 029 37	1.199 920 65	1.199 997 90
	0.599 940 36	0.600 162 86	0.599 981 16
10^2	1.198 102 00	1.207 636 83	1.199 997 90
	0.603 794 99	0.584 688 12	0.599 998 65
10^3	3.510 852 80	0.772 354 245	1.199 999 81
	−4.021 701 81	1.455 289 84	0.600 000 14
10^{13}			1.199 999 81
			0.599 994 28

Three techniques were used to solve the normal equations (14) for $\boldsymbol{\delta}$, namely (i) Cholesky factorization, (ii) Gaussian elimination with Crout's method, and (iii), singular value decomposition (SVD) with equation (28). The results of our single precision computations are summarized in table 1.

Evidently, the SVD method is by far the most robust. In double precision, all methods produced accurate answers up to $s = 10^8$; then SVD proved superior. We conclude that SVD provides by far the greatest numerical robustness (see also Lawson and Hanson 1974; Stewart 1973).

3.2. *Singular value decomposition and Marquardt's factor*

Having discussed the advantages of Marquardt's method, we now describe a method of including it in the SVD formulation. This prescription for avoiding singularities has been treated by Lawson and Hanson (1974) and by Jupp and Vozoff (1975), among others.

We recall that the solution to the modified normal equations (16a) is

$$\boldsymbol{\delta} = (\mathbf{Z}^T\mathbf{Z} + \beta\mathbf{I})^{-1}\,\mathbf{Z}^T\mathbf{g}. \tag{29}$$

We write $(\mathbf{Z}^T\mathbf{Z})$ in terms of $\mathbf{U}$, $\boldsymbol{\Lambda}$, and $\mathbf{V}$:

$$\mathbf{Z}^T\mathbf{Z} = \mathbf{V}\boldsymbol{\Lambda}^2\mathbf{V}^T,$$

so that

$$(\mathbf{Z}^T\mathbf{Z})^{-1} = \mathbf{V}\boldsymbol{\Lambda}^{-2}\mathbf{V}^T. \tag{30}$$

The matrix $(\mathbf{Z}^T\mathbf{Z} + \beta\mathbf{I})$ becomes

$$\begin{aligned}(\mathbf{Z}^T\mathbf{Z} + \beta\mathbf{I}) &= \mathbf{V}\boldsymbol{\Lambda}^2\mathbf{V}^T + \beta\mathbf{I} \\ &= \mathbf{V}(\boldsymbol{\Lambda}^2 + \beta\mathbf{I})\mathbf{V}^T.\end{aligned}$$

Hence,

$$\begin{aligned}(\mathbf{Z}^T\mathbf{Z} + \beta\mathbf{I})^{-1} &= \mathbf{V}(\boldsymbol{\Lambda}^2 + \beta\mathbf{I})^{-1}\,\mathbf{V}^T \\ &= \mathbf{V}\ \mathrm{diag}\left(\frac{1}{\lambda_j^2 + \beta}\right)\mathbf{V}^T,\end{aligned} \tag{31}$$

where $(\boldsymbol{\Lambda}^2 + \beta\mathbf{I})^{-1}$ is a diagonal matrix of the form

$$(\boldsymbol{\Lambda}^2 + \beta\mathbf{I})^{-1} = \begin{bmatrix} \dfrac{1}{\lambda_1^2 + \beta} & 0 & \cdots & 0 \\ & & & \vdots \\ 0 & \dfrac{1}{\lambda_2^2 + \beta} & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & & \dfrac{1}{\lambda_p^2 + \beta} \end{bmatrix}.$$

We then substitute (31) and (22) into (16b) to obtain

$$\boldsymbol{\delta} = \mathbf{V}\ \mathrm{diag}\left(\frac{1}{\lambda_j^2 + \beta}\right)\mathbf{V}^T\mathbf{V}\boldsymbol{\Lambda}\mathbf{U}^T\mathbf{g},$$

$$\boldsymbol{\delta} = \mathbf{V}\ \mathrm{diag}\left(\frac{\lambda_j}{\lambda_j^2 + \beta}\right)\mathbf{U}^T\mathbf{g}. \tag{32}$$

Comparing (32) with (28), we note that SVD can be combined with Marquardt's method by replacing the element $1/\lambda_j$ in the $\mathbf{\Lambda}^{-1}$ matrix by the element

$$\frac{\lambda_j}{\lambda_j^2 + \beta},$$

where β is Marquardt's damping factor. It now becomes clear how β can obviate the problem of matrix singularities: even if $\lambda_j \rightarrow 0$, division by zero does not occur.

4. Solutions to Geophysical Inverse Problems

As remarked earlier, iterative least-squares methods have been used for many kinds of geophysical inverse problems. Let us describe some examples.

4.1. Deconvolution of normal incidence seismograms

The reflected seismic signal x_t for a vertically incident P (compressional) wave may be expressed as a convolution of a source-generated seismic wavelet w_t with the impulse response of a layered earth i_t (Robinson and Treitel 1980). A digitized version of this signal may be written

$$x_t = \sum_\tau w_\tau i_{t-\tau}. \tag{33}$$

The process of seismic deconvolution is an attempt to remove the source wavelet w_t from the trace x_t and subsequently estimate normal-incidence reflection coefficients (or acoustical impedance contrasts) for interfaces within the earth. Deconvolution is an example of a nonunique inverse problem, in which we seek to determine w_t and i_t from knowledge of x_t. This nonuniqueness can best be appreciated by considering the Fourier transform $X(\omega)$ of the trace

$$X(\omega) = W(\omega)I(\omega), \tag{34}$$

where $W(\omega)$ and $I(\omega)$ are the Fourier transforms of the wavelet and of the impulse response. Having knowledge of only $X(\omega)$ leaves us with one equation and two unknowns. Hence, deconvolution may be viewed more as a digital processing "art" in which we attempt to separate w_t and i_t by using the different statistical properties of the two time sequences. Some of these statistical methods have been discussed by Lines and Ulrych (1977). Nonlinear least-squares inversion gives us a method of perturbing classical solutions for w_t and i_t in order to provide a closer match between model responses and seismic trace data. Similar approaches have been used by Bilgeri and Carlini (1981), Lailly (1981), Cooke and Schneider (1983), and Bamberger, Chavent, Hemon and Lailly (1982).

To justify the application of a constrained linear least-squares method for seismic inversion, we recall the seismic trace model of equation (33). Because the trace x_t is the convolution of a source wavelet w_t with an impulse response i_t, the trace x_t is a *linear* function of the wavelet coefficients w_t. On the other hand, as

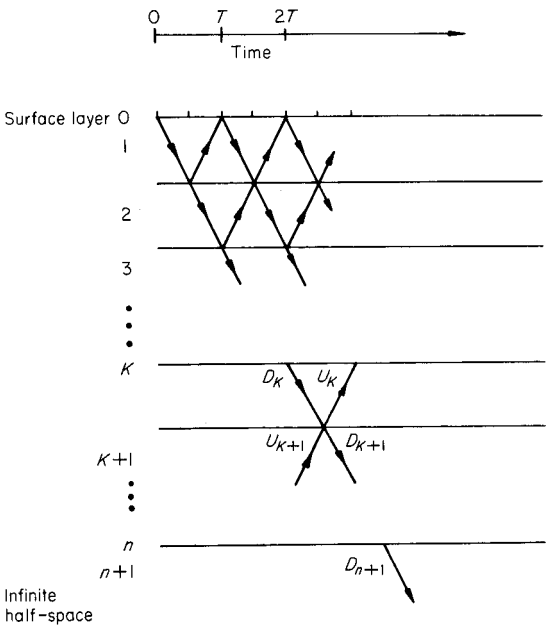

Fig. 3. Goupillaud layered earth model for modeling the seismic trace.

shown in detail below, the impulse response i_t is itself generally a *nonlinear* function of the reflection coefficients c_t. Nevertheless, if the reflection coefficients c_t are sufficiently small in magnitude, the impulse response i_t becomes a *quasilinear* function of the coefficients c_t, and for such cases the Taylor series expansions of the nonlinear model (4) becomes applicable.

Our layered earth model has been described by Goupillaud (1961) and is shown in fig. 3. It consists of n layers embedded between two half-spaces. The normal incidence response of this model has been treated in several papers, including Kunetz and d'Erceville (1962) and Treitel and Robinson (1966). Although all waves travel vertically in the present one-dimensional model, the wave paths in fig. 3 are drawn at nonvertical incidence to simulate the change of the wave's vertical position with time. Each layer is constructed to have unit two-way traveltime, so that the unit delay operator z represents two-way traveltime in each layer.

The pressure reflection coefficient at the boundary between the kth and the $k + 1$th layers is defined as the ratio of the pressure of the reflected wave to the pressure of the incident wave. If the incident wave is in the kth layer, this reflection coefficient is

$$c_k = \frac{\rho_{k+1} V_{k+1} - \rho_k V_k}{\rho_{k+1} V_{k+1} + \rho_k V_k}, \tag{35}$$

where ρ_k and V_k are the density and the velocity of compressional waves in the kth layer (Robinson and Treitel 1980, pp. 325–328). The transmission coefficient t_k for

pressure is

$$t_k = \frac{2\rho_{k+1}V_{k+1}}{\rho_{k+1}V_{k+1} + \rho_k V_k}. \tag{36}$$

By use of these reflection and transmission coefficients, the upgoing and downgoing waves in each layer can be related to the surface excitation. If $U_k(z)$ and $D_k(z)$ are the z-transforms of the upgoing and downgoing waves at the top of the kth layer, the relationship between the waves in each layer is

$$\begin{bmatrix} D_0(z) \\ U_0(z) \end{bmatrix} = \prod_{i=0}^{k} \mathbf{M}_i \begin{bmatrix} D_{k+1}(z) \\ U_{k+1}(z) \end{bmatrix}, \tag{37}$$

where

$$\mathbf{M}_i = \frac{1}{t_i} \begin{bmatrix} z^{-1/2} & c_i z^{-1/2} \\ c_i z^{1/2} & z^{1/2} \end{bmatrix} \tag{38}$$

is the communication (or layer) matrix.

For the n-layered model shown in fig. 3, it turns out that $U_{n+1}(z) = 0$ and $D_0(z) = 1 - c_0 U_0(z)$ (Robinson 1967, chapter 3). Here $U_0(z)$ is the z-transform of the impulse response of the layered earth. $U_0(z)$ contains all primary and multiple reflections from the layered medium as they are recorded just above the surface ($k = 0$). Since $U_0(z)$ contains products of reflection coefficients, it is a nonlinear function of these coefficients. However, under certain conditions this impulse response can be approximated as a linear function of the reflection coefficients. Consider the case of an isolated layer, say the kth. Then U_k is

$$U_k(z) = \frac{c_k + c_{k+1}z}{1 + c_k c_{k+1} z}. \tag{39}$$

The wave recorded just above the kth interface is a nonlinear function of the reflection coefficients but is approximately linear if

$$c_k c_{k+1} \ll 1.$$

In such cases,

$$U_k(z) \approx c_k + c_{k+1}z. \tag{40}$$

Thus $U_k(z)$ is the impulse response for a pressure geophone located just above the kth interface. Physically speaking, (40) says that the impulse response is dominated by primary reflections whenever the absolute value of the reflection coefficients is much less than unity. Since reflection coefficients are rarely larger than 0.3, the linearity assumption is often sound.

We illustrate our least-squares inversion approach to deconvolution with a simple model.

Figure 4a shows an example in which a seven-point wavelet is convolved with the impulse response due to the reflection coefficient sequence (fig. 4a):

$$C(z) = 0.200z^{10} - 0.300z^{15} + 0.100z^{16}.$$

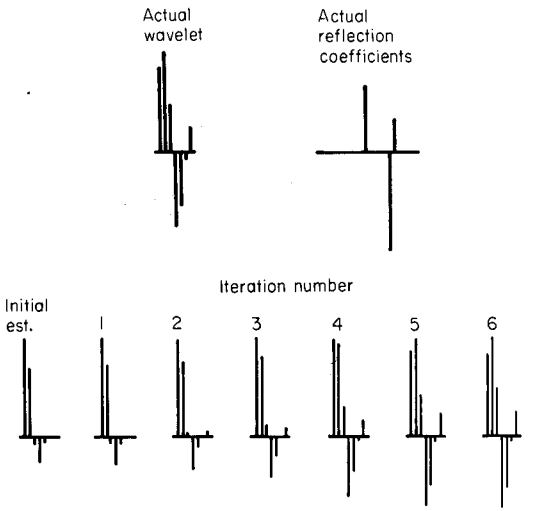

Fig. 4a. Example showing the evolution of wavelet estimates as a function of iteration number.

The initial wavelet guess, shown in fig. 4, was obtained with Kolmogorov's minimum phase estimation procedure (Lines and Ulrych 1977). For this first test, the reflection coefficient guesses were set to their true values. The w_t estimates converge within a specified error criterion after six iterations. Table 2 compares the convergence rates for SVD solutions versus solutions for the normal equations obtained by Cholesky factorization. Both methods provide excellent solutions to the problem, with SVD giving a slightly faster decrease in the normalized mean square error. Figure 4a shows the evolution of the wavelet estimates as a function of the number of iterations. The accurate reflection coefficient estimates remained essentially unchanged in these iterations. Figure 4b shows the similarity between the final model response and the actual trace.

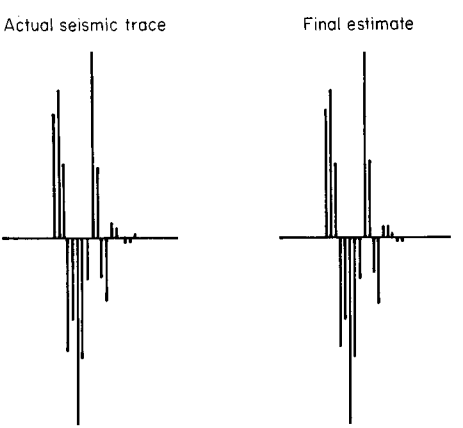

Fig. 4b. Comparison of seismic trace and final estimate.

Table 2. *Convergence rates for wavelet estimation problem.*

Iteration number	Marquardt damping factor, β	Error for SVD solution	Error for normal equations solutions
0		0.379 178 0	0.379 178 0
1	20.00	0.338 700 1	0.342 340 9
2	8.00	0.258 802 6	0.268 396 4
3	3.20	0.141 117 8	0.155 119 1
4	1.28	0.048 190 8	0.050 296 1
5	0.512	0.004 807 9	0.006 464 8
6	0.2048	0.000 158 5	0.000 247 2

It turns out that a suboptimal choice of β was used in the example of fig. 4a. In this case the reflection coefficient estimates are accurate, and the problem essentially reduces to simple *linear* regression, so $\beta = 0$ is an appropriate choice. Thus a single iteration is needed for convergence to occur. With an inaccurate initial choice of reflection coefficients or in the presence of additive noise, more than one iteration would have been required. (An initial choice of $C(z) = 0.10z^{10} + 0.10z^{15} + 0.10z^{16}$ requires two iterations.) Our experience has shown that β usually controls the number of iterations for convergence, not the final solution.

The addition of random noise to the seismic trace can alter the proper choice of β. When random noise is added to the trace in fig. 4b to produce a signal-to-noise ratio of 1.0, the initial choice $\beta = 0$ does a poorer job of reducing the error than the initial choice $\beta = 0.3$ with a reduction factor of 0.5. We recall that the β-reduction factor is some number of magnitude less than unity by which β for iteration step $i - 1$ is multiplied to produce β for iteration step i. The convergence rates for the first 10 iterations of these two procedures are shown in table 3.

Table 3. *Mean squared errors for different choices of the Marquardt damping factor β.*

Iteration Number	$\beta = 0$	Initial $\beta = 0.3$ β reduction factor = 0.5
0	1.124 639	1.124 639
1	0.404 531 8	0.422 032 2
2	0.353 364 2	0.336 310 7
3	0.327 818 3	0.309 483 2
4	0.330 218 7	0.302 397 1
5	0.345 368 7	0.296 603 3
6	0.670 667 9	0.291 081 0
7	0.417 752 8	0.234 934 4
8	0.633 654 1	0.287 833 6
9	0.428 456 5	0.289 834 1
10	0.388 203 0	0.282 991 9

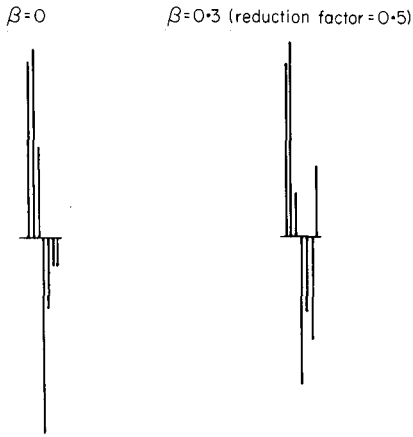

Fig. 4c. Wavelet estimates for noisy data example.

The mean squared error S for the nonzero value of β continues to decrease gradually to a plateau at $S = 0.262\,358\,4$. (The "plateau" is defined by $\Delta S/S \leq 0.001$.) The use of $\beta = 0$, on the other hand, causes the value of S to continue to fluctuate wildly between 0.301 987 1 and 1.458 158 for 40 iterations. Wavelet estimates for both procedures are shown in fig. 4c. The figure shows the wavelet for the

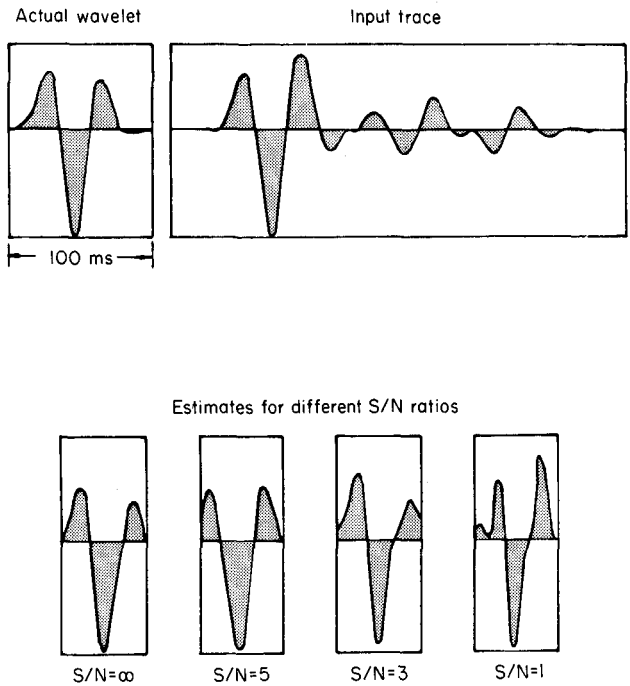

Fig. 4d. Wavelet estimation for different levels of additive noise. Wavelet estimates are shown for signal-to-noise ratios of ∞, 5, 3, and 1.

convergent case ($\beta \neq 0$) and the wavelet for the nonconvergent case ($\beta = 0$) after 40 iterations. This instability problem, which arises for the case of no damping in noisy environments, is analogous to the problem which occurs in seismic deconvolution when no prewhitening is used (Treitel and Lines 1982). Crosson (1976) discusses the arbitrary choice of the damping factor for inverse problems. This choice involves the usual tradeoff between resolution and stability.

Although there appear to be no a priori "foolproof" methods for choosing the damping factor in nonlinear problems, it is encouraging to see that our inverse problem solutions are reasonably robust with respect to the choice of β. This choice determines the rate of convergence, but in general does not adversely affect the final answer.

A physically more realistic wavelet estimation problem is pictured in fig. 4d. The seismic wavelet was convolved with the reflectivity function from a sonic log in order to produce the shown seismic trace. Random noise of different variances was added to this trace to produce synthetic traces with different signal-to-noise ratios. Nonlinear least-squares inversion was then used to produce wavelet estimates for signal-to-noise ratios of infinity (no noise), 5, 3, and 1. Estimates generally degrade in quality as noise increases, but do so gradually. Unlike some other inverse techniques, nonlinear least-squares inversion produces robust parameter estimates in noisy environments. This characteristic should not be underrated when dealing with geophysical data.

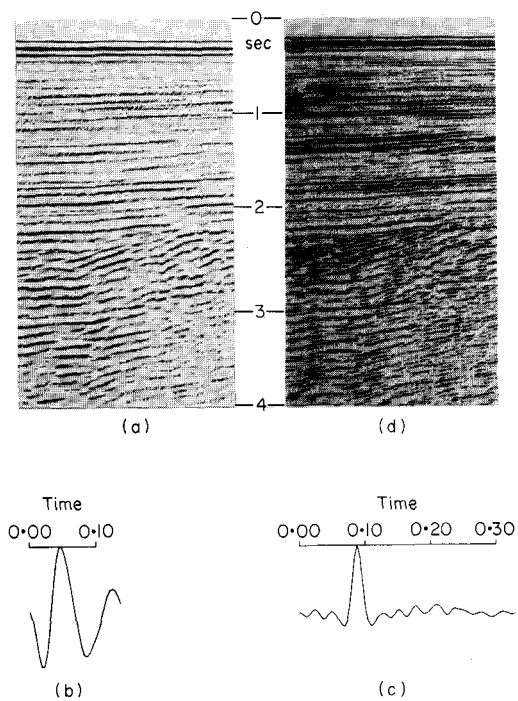

Fig. 5. (a) Input data, (b) wavelet estimate, (c) Wiener filter output, (d) wavelet deconvolution.

For a real data application of wavelet estimation by least-squares inversion we use the seismic section from Treitel, Gutowski and Wagner (1982). Figure 5a shows the input data set which has had plane wave processing applied. This processing decomposes a point-source seismic recording into a set of plane wave seismograms at various angles of incidence.

Figure 5b displays the wavelet estimate obtained by nonlinear inversion, and fig. 5c shows the "resolving kernel" or output obtained by convolving the wavelet with a Wiener spiking filter. Figure 5d shows the section resulting from the application of the Wiener filter to the seismic section in fig. 5a. It is apparent that the deconvolution has enhanced a wedge feature at about 2.2 s in the middle of the section and has vastly improved resolution on the section.

4.2. Inversion of gravity data for a buried wedge

Another geophysical example which illustrates the use of nonlinear least-squares inversion involves the determination of the position and density contrast of a buried wedge or slab of rock (see fig. 6a). The change in the vertical component of gravitational acceleration (far field) due to this slab is (see e.g. Dobrin 1976):

$$g_z = 2\gamma\rho t\left(\frac{\pi}{2} - \tan^{-1}\frac{x}{Z}\right), \tag{41}$$

where ρ is the density contrast of the anomalous slab of rock, t is the slab's thickness, x is the horizontal displacement from the edge of the slab, Z is the depth to the middle of the slab, and γ is the universal gravitational constant.

A possible inverse problem would be the following. Given a gravity profile,

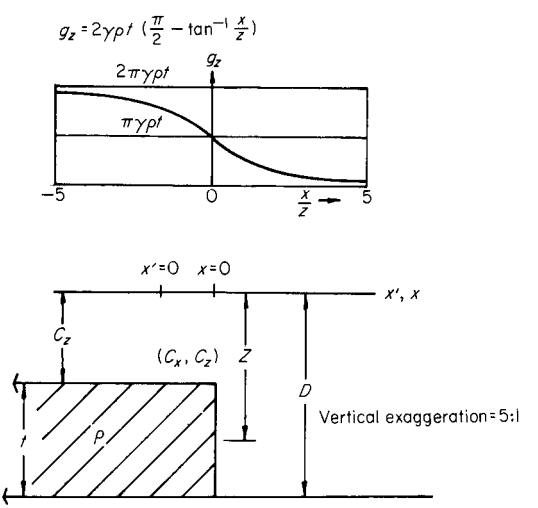

Fig. 6a. Gravity response of a buried wedge. (C_x, C_z) = location of wedge corner, $x = x' - C_x$ = horizontal distance of station from wedge station, D = depth to basement, $t = D - C_z$ = thickness of slab, $Z = (D + C_z)/2$, ρ = density contrast.

determine the position and density contrast for the slab. A direct means of determining the horizontal location of the wedge edge requires that we simply find the value of x for which $\partial^2 g_z/\partial x^2 = 0$. The determination of the rock density and of the wedge's vertical position are both nonunique.

As in the case of seismic deconvolution, we recognize that the nonuniqueness involved in estimating the slab's density contrast and thickness arises because we only know g_z, which by (41) means that we may at best solve for the product of slab density contrast and thickness, ρt. This represents an ambiguity commonly found in gravity interpretation. The problem was recognized by Parker (1974), who placed bounds on both density contrast and depth.

Solutions for the wedge's location and density contrast can be obtained by nonlinear least-squares inversion. Consider a problem in which the slab rests on a basement at depth D. The gravity model response g_z is a function of three parameters: the wedge corner location (C_x, C_z) and the density contrast ρ. Other variables may be defined as

$$x = x' - C_x,$$

where x' is the location of the gravity station

$$t = D - C_z,$$

and

$$Z = \frac{D + C_z}{2}.$$

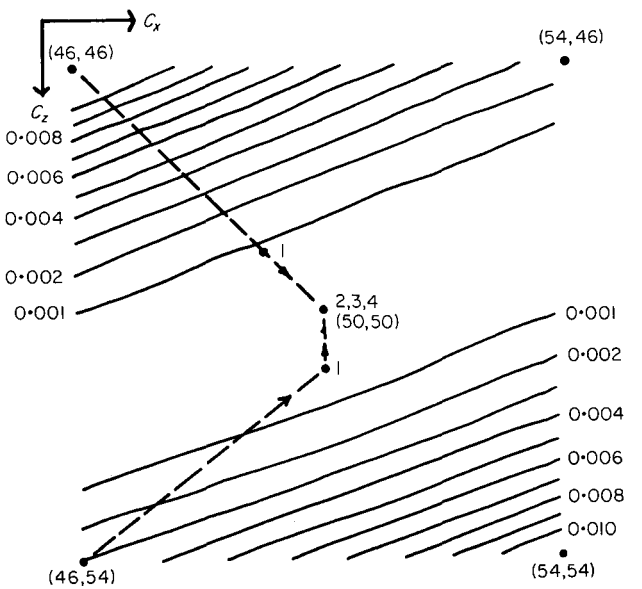

Fig. 6b. Gravity error contours for the buried wedge problem. Convergence paths are shown for examples 2 and 3 in table 4.

We can find analytic expressions for the Jacobian by computing $\partial g_z/\partial \rho$, $\partial g_z/\partial C_x$, and $\partial g_z/\partial C_z$.

The observations were created by placing our wedge corner at a depth of 1098 m (3600 ft) with a density contrast of 0.25 g/cm^3. A profile of 19 stations spaced 109.7 m (360 ft) apart was run over the top of the wedge. The true wedge corner position is at position (50, 50) on a 100 × 100 grid, with a grid spacing of 21.95 m (72 ft). The contour plot in fig. 6b illustrates the squared error S as a function of (C_x, C_z) for the correct density contrast. We see that there is a range of (C_x, C_z) values which produces low S values. This simply means that a shallow wedge with less total mass under the gravimeters gives almost the same response as a deeper wedge with more mass under the stations. Nevertheless, the parameters can converge to the correct answer if we examine some examples in terms of the initial values of ρ, C_x, and C_z. This is illustrated in table 4 which shows the convergence rates and solutions for various initial estimates of ρ and an initial guess of the corner location at $(C_x, C_z) = (46, 46)$. The "normalized mean square error" (MSE) is given by the quotient $\mathbf{e}^T\mathbf{e}$ divided by the sum of squares of the values $\mathbf{g}_z$.

Table 4. *Convergence for the wedge inversion problem*

Iteration No.	ρ	(C_x, C_z)	Normalized MSE
Example 1: Density contrast estimate low, wedge depth estimate shallow			
0 (initial model)	0.200 000	(46, 46)	$1.203\,279\ 10^{-2}$
1	0.248 564	(50, 51)	$6.712\,922\ 10^{-4}$
2	0.249 875	(50, 50)	$2.521\,883\ 10^{-7}$
3	0.250 007	(50, 50)	$8.723\,082\ 10^{-10}$
Example 2: Density contrast estimate high, wedge depth estimate deep			
0 (initial model)	0.3000 000	(46, 54)	$1.812\,316\ 10^{-2}$
1	0.251 544	(50, 51)	$2.031\,701\ 10^{-4}$
2	0.249 874	(50, 50)	$2.540\,213\ 10^{-7}$
3	0.250 008	(50, 50)	$9.821\,132\ 10^{-10}$
Example 3: Density contrast estimate high, wedge depth estimate shallow			
0	0.300 000	(46, 46)	$1.216\,016\ 10^{-1}$
1	0.248 564	(49, 49)	$4.747\,731\ 10^{-4}$
2	0.249 904	(50, 50)	$1.492\,059\ 10^{-7}$
3	0.250 008	(50, 50)	$9.348\,184\ 10^{-10}$
Example 4: Density contrast estimate low, wedge depth estimate deep			
0	0.200 000	(46, 54)	$5.960\,400\ 10^{-2}$
1	0.251 539	(52, 49)	$1.562\,625\ 10^{-4}$
2	0.250 492	(50, 50)	$3.863\,768\ 10^{-6}$
3	0.250 008	(50, 50)	$9.348\,184\ 10^{-10}$

All initial guesses have converged in a few iterations with the choice $\beta = 0$. The observed behavior suggests that our problem is quasilinear in the parameters C_x and C_z, but linear in ρ. This is evident for the case of C_x by inspection of the g_z versus x curve in fig. 6a. We note that the profile has been run from $(x/z) = -1$ to $(x/z) = 1$, and g_z is approximately linear in x over this range. Since $C_x = x' - x$, g_z is approximately linear in C_x also. A choice of nonzero β in this specific problem has generally caused convergence to be much slower than for $\beta = 0$.

Aside from the inevitable nonuniqueness question, nonlinear least-squares inversion provides a valid means to locate the wedge and to estimate its density contrast.

5. Conclusions

The essence of geophysical data inversion is to provide a suitable earth model from a given set of observations. A possible method of performing this task involves nonlinear least-squares inversion. The particular nonlinear approach we chose involves an iterative application of Marquardt's method. It was shown that the SVD solution to the least-squares problem has certain numerical advantages over conventional solutions to the normal equations. Moreover, the Marquardt damping factor can be included in the SVD approach by simple adjustment of the singular values. Two geophysical uses, namely seismic deconvolution and gravity inversion, were described to demonstrate the application of the method.

References

Aki, K. and Richards, P. 1980, Quantitative Seismology—Theory and Methods, Vol. 2, W. H. Freeman Co., San Francisco.

Backus, G.E. and Gilbert, J.F. 1967, Numerical application of a formalism for geophysical Journal of the Royal Astronomical Society, 13, 247–276.

Backus, G.E. and Gilbert, J.F. 1968, The resolving power of gross earth data, Geophysical Journal of the Royal Astronomical Society 16, 169–205.

Bamberger, A., Chavent, G., Hemon, C. and Lailly, P. 1982, Inversion of normal incidence seismograms Geophysics 47, 757–770.

Bilgeri, D. and Carlini, A. 1981, Non-linear estimation of reflection coefficients from seismic data Geophysical Prospecting 29, 672–686.

Cooke, D.A. and Schneider, W.A. 1983, Generalized inversion of reflection seismic data, Geophysics 48, 665–676.

Crosson, R.W. 1976, Crustal structure modeling of earthquake data, 1. Simultaneous least squares estimation of hypocenter and velocity parameters, Journal of Geophysical Research 81, 3036–3046.

Dennis, J. 1977, Nonlinear least squares and equations, in The State of the Art of Numerical Analysis, ed. D. Jacobs, Academic Press, London.

Dobrin, M. 1976, Introduction to Geophysical Prospecting, 3rd edn., McGraw-Hill, New York.

Draper, N.R. and Smith, H. 1981, Applied Regression Analysis, 2nd edn., Wiley, New York.

Fletcher, R. 1980, Practical Methods of Optimization, Vol. 1, Wiley Interscience, New York.

GJOYSTDAL, H. and URSIN, B. 1981, Inversion of reflection times in three dimensions, Geophysics 46, 972–983.

GOLUB, G.H. and REINSCH, C. 1970, Singular Value Decomposition and Least Squares Solutions: Handbook for Automatic Computation, II, Linear Algebra, eds. J. Wilkinson and C. Reinsch, Springer-Verlag, Berlin, Heidelberg, New York.

GOUPILLAUD, P.L. 1961, An approach to inverse filtering of near surface layer effects from seismic records, Geophysics 26, 654–760.

GRAYBILL, F.A. 1969, Introduction to Matrices with Applications in Statistics, Wadsworth Publishing Co. Inc., Belmont.

INMAN, J.R. 1975, Resistivity inversion with ridge regression, Geophysics 40, 798–817.

JACKSON, D.D. 1972, Interpretation of inaccurate, insufficient and inconsistent data, Geophysical Journal of the Royal Astronomical Society 28, 97–109.

JACKSON, D.D. 1979, The use of a priori data to resolve non-uniqueness in linear inversion, Geophysical Journal of the Royal Astronomical Society 57, 137–157.

JUPP, D.L.B. and VOZOFF, K. 1975, Stable iterative methods for the inversion of geophysical data, Geophysical Journal of the Royal Astronomical Society 42, 957–976.

KUNETZ, G. and D'ERCEVILLE, I. 1962, Sur certaines propriétés d'une onde plane de compression dans un milieu stratifié, Annales de Geophysique, 19, 351–359.

LANCZOS, C. 1961, Linear Differential Operators, Van Nostrand, Princeton. 665–679.

LAILLY, P. 1981, The inverse problem in 1-D reflection seismics, in The Solution of the Inverse Problem in Geophysical Interpretation, ed. R. Cassinis, Plenum Press, New York.

LAWSON, C.L. and HANSON, R.J. 1974, Solving Least Squares Problems, Prentice-Hall, Englewood Cliffs, New Jersey.

LEVENBERG, K. 1944, A method for the solution of certain nonlinear Problems in least squares, Quarterly of Applied Mathematics 2, 164–168.

LINES, L.R. and ULRYCH, T.J. 1977, The old and the new in seismic deconvolution and wavelet estimation, Geophysical Prospecting 25, 512–540.

MARQUARDT, D.W. 1963, An algorithm for least squares estimation of non-linear parameters Journal of the Society of Industrial and Applied Mathematics 11, 431–441.

NEUMANN, G. 1981, Determination of lateral inhomogeneities in reflection seismics by inversion of traveltime residuals, Geophysical Prospecting 29, 161–177.

OLDENBURG, D.W. 1974, The inversion and interpretation of gravity anomalies, Geophysics 39, 526–536.

ORISTAGLIO, M.L. and WORTHINGTON, M.H. 1980, Inversion of surface and borehole electromagnetic data for two dimensional electrical conductivity models, Geophysical Prospecting 28, 633–657.

PARKER, R.L. 1974, Best bounds on density and depth from gravity data, Geophysics 39, 644–649.

ROBINSON, E.A. 1967, Multichannel Time Series Analysis with Digital Computer Programs, Holden Day, San Francisco.

ROBINSON, E.A. and TREITEL, S. 1980, Geophysical Signal Analysis, Prentice-Hall, Englewood Cliffs, New Jersey.

SMITH, F.B. and SHANNO, D.F. 1971, An improved Marquardt procedure for nonlinear regressions, Technometrics 13, 63–75.

STEWART, G.W. 1973, Introduction to Matrix Computations, Academic Press, New York, London.

STRANG, G. 1980, Linear Algebra and its Applications, 2nd edn., Academic Press, New York, London.

TREITEL, S., GUTOWSKI, P.R. and WAGNER, D.E. 1982, Plane wave decomposition of seismograms, Geophysics 47, 1375–1401.

TREITEL, S. and LINES, L.R. 1982, Linear inverse theory and deconvolution, Geophysics 47, 1153–1159.

TREITEL, S. and ROBINSON, E.A. 1966, The design of high resolution digital rilters, IEEE Transactions on Geoscience Electronics GE-4, 25–38.

TWOMEY, S. 1977, Introduction to the Mathematics of Inversion in Remote Sensing and Indirect Measurements, Elsevier, Amsterdam.

VIGNERESSE, J.L. 1977, Linear inverse problem in gravity profile interpretations, Journal of Geophysics 43, 193–213.

WIGGINS, R.A. 1972, The generalized linear inverse problem: Implication of surface waves and free oscillations for earth structure, Reviews of Geophysics and Space Physics 10, 251–285.

WIGGINS, R.A., LARNER, K.L. and WISECUP, R.P. 1976, Residual static analysis as a general linear inverse problem, Geophysics 41, 922–938.