Geophysical Prospecting **39**, 723–736, 1991

AN ALTERNATIVE STRATEGY FOR NON-LINEAR INVERSION OF SEISMIC WAVEFORMS[1]

M. S. SAMBRIDGE[2], A. TARANTOLA[3] and B. L. N. KENNETT[4]

ABSTRACT

SAMBRIDGE, M.S., TARANTOLA, A. and KENNETT, B.L.N. 1991. An alternative strategy for non-linear inversion of seismic waveforms. *Geophysical Prospecting* **39**, 723–736.

A common example of a large-scale non-linear inverse problem is the inversion of seismic waveforms. Techniques used to solve this type of problem usually involve finding the minimum of some misfit function between observations and theoretical predictions. As the size of the problem increases, techniques requiring the inversion of large matrices become very cumbersome. Considerable storage and computational effort are required to perform the inversion and to avoid stability problems. Consequently methods which do not require any large-scale matrix inversion have proved to be very popular. Currently, descent type algorithms are in widespread use. Usually at each iteration a descent direction is derived from the gradient of the misfit function and an improvement is made to an existing model based on this, and perhaps previous descent directions.

A common feature in nearly all geophysically relevant problems is the existence of separate parameter types in the inversion, i.e. unknowns of different dimension and character. However, this fundamental difference in parameter types is not reflected in the inversion algorithms used. Usually gradient methods either mix parameter types together and take little notice of the individual character or assume some knowledge of their relative importance within the inversion process.

We propose a new strategy for the non-linear inversion of multi-offset reflection data. The paper is entirely theoretical and its aim is to show how a technique which has been applied in reflection tomography and to the inversion of arrival times for 3D structure, may be used in the waveform case. Specifically we show how to extend the algorithm presented by Tarantola to incorporate the subspace scheme. The proposed strategy involves no large-scale matrix inversion but pays particular attention to different parameter types in the inversion.

[1] Received January 1990, revision accepted February 1991.

[2] Institute of Theoretical Geophysics, Departments of Earth Sciences and Applied Mathematics and Theoretical Physics, Downing Street, Cambridge CB2 3EO, U.K.

[3] Institut de Physique du Globe, 4 place Jussieu, 75252 Paris Cédex 05, France.

[4] Research School of Earth Sciences, A.N.U., P.O. Box 4, Canberra, ACT 2601, Australia.

We use the formulae of Tarantola to state the problem as one of optimization and derive the same descent vectors. The new technique splits the descent vector so that each part depends on a different parameter type, and proceeds to minimize the misfit function within the subspace defined by these individual descent vectors. In this way, optimal use is made of the descent vector components, i.e. one finds the combination which produces the greatest reduction in the misfit function based on a local linearization of the problem within the subspace. This is not the case with other gradient methods. By solving a linearized problem in the chosen subspace, at each iteration one need only invert a small well-conditioned matrix (the projection of the full Hessian on to the subspace). The method is a hybrid between gradient and matrix inversion methods. The proposed algorithm requires the same gradient vectors to be determined as in the algorithm of Tarantola, although its primary aim is to make better use of those calculations in minimizing the objective function.

INTRODUCTION

In the non-linear inversion of seismic waveform data, one attempts to obtain an earth model for which the predicted seismogram most closely resembles the observed seismogram. This process presents two problems. Firstly, given an earth model we must solve the forward problem of generating the predicted seismogram, and secondly, we must solve the inverse problem of obtaining the optimum earth model. This paper is concerned only with the inverse problem and not the forward one, which requires the numerical solution of the elastic wave equation. We describe an alternative strategy for the non-linear inversion of seismic waveforms which represents a natural extension to the technique used by Tarantola (1986) in the inversion of multi-offset seismic reflection data.

A convenient way of dealing with the inverse problem is to state it in terms of a large-scale optimization problem in a functional space. One usually defines some misfit function which describes quantitatively the discrepancy between observed and predicted seismograms, and then attempts to find the earth model for which this function is minimized. Numerical methods currently available to solve this type of problem are limited by computational resources. In non-linear inverse problems, involving only a few degrees of freedom, one may employ a systematic search within a predetermined range of models, e.g. in the earthquake location problem (Sambridge and Kennett 1986). However, as both the size and complexity of the problem increase, random (Monte Carlo) searching for an optimal earth model proves to be too expensive and therefore unfeasible. In such cases one usually employs an iterative method which is computationally less expensive. Of these, gradient methods form the only practical approach for very large scale problems. Consequently they are used extensively in waveform inversion.

In most waveform inversion studies one usually attempts to invert for more than one parameter type, e.g. P- and S-wave velocity (or impedance) and density (Tarantola 1986; Mora 1987), or possibly density, elastic coefficients and some source function or surface traction as in Tarantola (Lecture at Majorana International School of Applied Geophysics, Erice, Sicily, 1987). The inclusion of fundamentally different parameter types, which may be of different dimension or even

constrained by different data types, changes the nature of the inverse problem. Usually each parameter type affects the data by different amounts and so the level of constraint imposed on each may vary significantly. This aspect is well known. Most gradient methods deal with it either by iterating to convergence, in the hope that once the more dominant parameters have converged the data will effectively contain more constraint on the others, or by inverting for different parameter types sequentially, i.e. one at a time while keeping all other types fixed. In fact the use of a simple gradient technique may significantly bias the inversion in favour of one parameter type at the expense of the others.

Mora (1987) uses a simple preconditioned conjugate gradient method which attempts to invert for all parameter types simultaneously. Tarantola (1986) suggests a similar technique using a sequential type inversion. In his method a hierarchy is introduced between different parameter types by the choice of model parametrization. Each type is then optimized in turn using a gradient method. This assumes that the parameter types are effectively decoupled by the parametrization and so may be determined independently. Our objective is to demonstrate how a 'subspace' scheme (a technique previously applied to other non-linear inverse problems by Kennett, Sambridge and Williamson (1988), Williamson (1986, 1990) and Sambridge (1990)) may be applied to a non-linear waveform problem. The subspace scheme deals with the same multiparameter type problem but it neither optimizes each parameter type sequentially whilst constraining all others, nor does it combine all parameter types into a single gradient. Instead it allows all parameter types to be adjusted simultaneously and also removes the possibility of artificially biasing the inversion towards the more dominant parameter types, e.g. by excessive iterations for a single parameter type or by using some unrealistic weighting scheme. This is demonstrated by an example which shows how, in a simple problem involving two parameter classes, both the two previous algorithms lead to a non-optimal improvement in the model, whereas the new procedure results in the best combination adjustments to each parameter type and in this sense achieves the most unbiased step. The new scheme makes use of exactly the same gradients calculated in the original method and is therefore no more computationally expensive in this respect. Furthermore it has been found to improve convergence rates over that of simple gradient techniques in the inversion of seismic reflection data (Williamson 1986).

The proposed strategy is in fact only one particular application of a subspace method, a general class of methods which have recently been found to be very useful in large-scale non-linear optimization problems involving multiparameter classes (Kennett and Williamson 1987; Sambridge 1988, 1990). Since the alternative strategy applies only to the optimization problem, it has applications in many non-linear inverse problems. We will not describe in detail the theoretical background required for the non-linear inversion of seismic waveforms since it is not essential for an understanding of the inversion strategy proposed. In order to demonstrate the technique applied to waveform inversion, we consider the non-linear inversion of multi-offset seismic reflection data, using much of the formalism developed by Tarantola (1986, lecture at Majorana International School of Applied Geophysics, Erice, Sicily 1987). We show how the subspace method may be used to extend the algorithm of

Tarantola (1986) without significant reorganization. In addition to the parameter types considered by Tarantola (1986), i.e. P- and S-wave impedances and density variations, we also include source characteristics as inversion parameters. This is necessary in some cases of waveform inversion where the source function is not fully known. It also provides a useful demonstration of how the new strategy may be extended to include any number of parameter types.

We compare the proposed technique with the existing single and conjugate gradient methods currently in use in waveform inversion and suggest that it would be superior in terms of efficiently minimizing a non-quadratic misfit function. The method separates each parameter type, reducing bias, without introducing artificial constraints, i.e. holding parameters fixed while allowing others to vary. For this reason we suggest that it results in a more natural and less-biased solution to the non-linear inverse problem.

Non-linear Inversion

Most non-linear inverse problems may be stated in terms of an optimization problem. Usually one is faced with a general problem of the form: find the vector $\mathbf{m}$ which minimizes the functional $F(\mathbf{m})$ given by

$$F(\mathbf{m}) = \tfrac{1}{2}\{\| \mathbf{d}_{obs} - \mathbf{d}_{cal} \|^2 + \| \mathbf{m} - \mathbf{m}_0 \|^2\}, \tag{1}$$

where $\mathbf{d}_{obs}$ is a vector representing the observed data set, $\mathbf{m}$ is a vector representing an earth model, $\mathbf{m}_0$ is some reference *a priori* model, $\| . \|$ is an L_2-norm defined through a covariance operator C such that $\| \varphi \| = \langle C^{-1}\varphi, \varphi \rangle$, and $\langle .,. \rangle$ is a duality product (see Tarantola 1987). The precise definition of the duality product will depend on the nature of the data and the model parametrization involved (for the seismic reflection problem discussed below an explicit definition is given for each). Introducing C_D as the covariance operator describing data uncertainties, C_M as the covariance operator describing uncertainties in $\mathbf{m}_0$, we write

$$F(\mathbf{m}) = \tfrac{1}{2}\{\langle C_D^{-1}(\mathbf{d}_{obs} - \mathbf{d}_{cal}), (\mathbf{d}_{obs} - \mathbf{d}_{cal})\rangle + \langle C_M^{-1}(\mathbf{m} - \mathbf{m}_0), (\mathbf{m} - \mathbf{m}_0)\rangle\}.$$

In general the relationship between $\mathbf{m}$ and $\mathbf{d}$ is non-linear. We assume that, given any model $\mathbf{m}$, we can calculate the corresponding data $\mathbf{d}$ and represent it by

$$\mathbf{d}_{cal} = \mathbf{g}(\mathbf{m}),$$

where $\mathbf{g}$ is a non-linear operator describing the forward modelling.

The general formulation above may be applied to many inverse problems. In seismic experiments the vector $\mathbf{m}$ describes a model of the real earth. Usually this consists of a set of functions describing some physical properties of the earth. All vectors $\mathbf{m}$ belong to a functional space known as the 'model' space M. Similarly we call the space containing all data vectors $\mathbf{d}$, the data space D. In the inversion of multi-offset seismic reflection data, one is usually presented with a series of shots at $\mathbf{x}_s$ and a series of receivers at $\mathbf{x}_r$. For each shot the data may be represented by a set of seismograms, usually representing surface displacements $u^i(\mathbf{x}_r, t)$, $i = 1, \ldots 3$.

Tarantola (1986) uses three functions to describe an earth model, namely P-wave impedance $IP(\mathbf{x})$, S-wave impedance $IS(\mathbf{x})$ and density $\rho(x)$. In addition we include here the source function $\phi^i(\mathbf{x}, t)$, $i = 1, 3$, as parameters in the inversion. For algebraic simplicity we consider only a single source and use a vector $\boldsymbol{\phi}(\mathbf{x}, t)$ to represent the three components when convenient, but the treatment given here may easily be generalized to several sources. We may now write the model vector explicitly in terms of its components

$$\mathbf{m} = \{\phi(\mathbf{x}, t), IP(\mathbf{x}), IS(\mathbf{x}), \rho(\mathbf{x})\} \tag{2}$$

If we ignore the possibility of cross-variances between model parameter types, i.e.

$$C_M = \begin{pmatrix} C_{IP} & & & 0 \\ & C_{IS} & & \\ & & C_\rho & \\ 0 & & & C_\phi \end{pmatrix},$$

then (1) may be rewritten as

$$F(\mathbf{m}) = \tfrac{1}{2}\{\| \mathbf{u}_{obs} - \mathbf{u}_{cal} \|^2 + \| IP(\mathbf{x}) - IP_o(\mathbf{x}) \|^2 + \| IS(\mathbf{x}) - IS_o(\mathbf{x}) \|^2 + \| \rho(\mathbf{x}) - \rho_o(\mathbf{x}) \|^2 + \| \boldsymbol{\phi}(\mathbf{x}, t) - \boldsymbol{\phi}_o(\mathbf{x}, t) \|^2\}, \tag{3}$$

where the duality products in this case are written:

for the observations,

$$\| \mathbf{u}_{obs} - \mathbf{u}_{cal} \|^2 = \sum_r \int_0^T dt \int_0^T dt' [u^i(\mathbf{x}_r, t)_{obs} - u^i(\mathbf{x}_r, t)_{cal}] \times W^{ij}(t, t', \mathbf{x}_r)[u^j(\mathbf{x}_r, t')_{obs} - u^j(\mathbf{x}_r, t')_{cal}],$$

and for the model functions,

$$\| IP(\mathbf{x}) - IP_o(\mathbf{x}) \|^2 = \int_V dV(\mathbf{x}) \int_V dV(\mathbf{x}')[IP(\mathbf{x}) - IP_o(\mathbf{x})] \times W_p(\mathbf{x}, \mathbf{x}')[IP(\mathbf{x}') - IP_o(\mathbf{x}')],$$

with similar expressions for $IS(\mathbf{x})$, $\rho(\mathbf{x})$ and $\phi(\mathbf{x}, t)$ (the last of which involves integrals over space and time). The weighting functions $W^{ij}(t, t', \mathbf{x}_r)$ and $W_p(\mathbf{x}, \mathbf{x}')$ etc. are the integral kernels of the inverse of the covariance operators C_D and C_{IP} respectively. For instance if one has a known covariance function $C(\mathbf{x}, \mathbf{x}')$, the corresponding weighting function $W(\mathbf{x}, \mathbf{x}')$ is determined by the equation

$$\int_V C(\mathbf{x}, \mathbf{x}')W(\mathbf{x}, \mathbf{x}'') \, dV(\mathbf{x}') = \delta(\mathbf{x} - \mathbf{x}'').$$

(Tarantola (1987, ch. 7) gives several examples of common covariance functions and the corresponding weighting functions.) The functions $\boldsymbol{\phi}(\mathbf{x}, t)$, $IP(\mathbf{x})$, $IS(\mathbf{x})$, $\rho(\mathbf{x})$ are related to the observed data through the elastic wave equation. Given any model of

the medium we determine the corresponding surface displacements at the receivers by some numerical solution of the elastic wave equation (e.g. Vireux 1986). This constitutes the forward problem and will not be dealt with here. This paper is concerned only with the solution of the optimization of F given by (3). We attempt to find the set of functions $\{\phi(\mathbf{x}, t), IP(\mathbf{x}), IS(\mathbf{x}), \rho(\mathbf{x})\}$ such that the functional F is minimized.

Although we have represented the earth by a set of functions, ultimately the problem is discretized so that it is suitable for numerical computations. In practice we are faced with a very large scale problem ($\sim 10^5$–10^9 degrees of freedom). Since the functional $F(\mathbf{m})$ is in general non-quadratic, any techniques based on a local quadratic approximation of the objective functional F must be iterated. The second-order expansion of the misfit functional F about an *a priori* model $\mathbf{m}_0$ defines the gradient vector $\hat{\gamma}$ and the Hessian $\mathbf{H}$,

$$F(\mathbf{m} + \delta\mathbf{m}) = F(\mathbf{m}) + \langle \hat{\gamma}, \delta\mathbf{m} \rangle + \tfrac{1}{2}\langle \mathbf{H}\, \delta\mathbf{m}, \delta\mathbf{m} \rangle + \mathrm{O}(\| \delta\mathbf{m} \|^3). \tag{4}$$

The gradient $\hat{\gamma}$ is an element of the dual of the model space, where dual and model space elements are related through the model covariance operator C_M by

$$\gamma = C_M \hat{\gamma}, \tag{5}$$

where γ is the direction of steepest ascent in model space (we denote all elements of the dual by a hat ˆ). We are not concerned with the details of calculating γ for any particular problem. Tarantola (1986) shows how it may be determined for each parameter type in the reflection problem. We merely summarized the formulae here for completeness.

Essentially it requires the solution of two problems. First we take a model $\mathbf{m}$ and solve the forward problem, i.e. through some numerical solution of the elastic wave equation, to obtain the predicted surface displacements at the receiver positions $\mathbf{x}_r$, which we write as

$$\tilde{u}^i_{\mathrm{cal}}(\mathbf{x}_r, t), \qquad i = 1, \ldots 3.$$

This field is labelled the 'current' field. We then use the weighted residuals between observed and predicted displacements as sources for a second field

$$\delta\hat{u}^i(\mathbf{x}_r, t) = \frac{u^i_{\mathrm{cal}}(\mathbf{x}_r, t) - u^i_{\mathrm{obs}}(\mathbf{x}_r, t)}{\sigma^2(\mathbf{x}_r, t)}, \tag{6}$$

where we have assumed for ease of notation that errors are uncorrelated in the data set, which is represented by choosing the weighting functions

$$W^{ij}(t, t', \mathbf{x}_r) = \sigma^2(\mathbf{x}_r, t)\delta^{ij}\delta(t - t').$$

These sources are then back-propagated in time through the medium to produce the 'missing' field

$$\tilde{u}^i(\mathbf{x}, t), \qquad i = 1, \ldots 3$$

A time correlation between current and missing field yields the required gradient. (For a detailed description of this procedure see Tarantola (1986).) We obtain for

each parameter type:

$$\phi: \quad \delta\hat{\phi}^i(\mathbf{x}, t) = \tilde{u}_i(\mathbf{x}, t);$$

$$\text{P:} \quad \delta\widehat{IP}(\mathbf{x}) = -2\alpha(\mathbf{x}) \int_0^T dt \tilde{u}^{ii}(\mathbf{x}, t)\tilde{u}^{jj}(\mathbf{x}, t);$$

$$\text{S:} \quad \delta\widehat{IS}(\mathbf{x}) = -4\beta(\mathbf{x}) \int_0^T dt\{\tilde{u}^{km}(\mathbf{x}, t)\tilde{u}^{km}(\mathbf{x}, t) - \tilde{u}^{ii}(\mathbf{x}, t)\tilde{u}^{jj}(\mathbf{x}, t)\};$$

$$\rho: \quad \delta\hat{\rho}(\mathbf{x}) = \int_0^T dt\{\dot{\tilde{u}}^i(\mathbf{x}, t)\dot{\tilde{u}}^i(\mathbf{x}, t) + [\alpha^2(\mathbf{x}) - 2\beta^2(\mathbf{x})]\tilde{u}^{ii}(\mathbf{x}, t)u^{jj}(\mathbf{x}, t) + 2\beta^2(\mathbf{x})\tilde{u}^{km}(\mathbf{x}, t)\tilde{u}^{km}(\mathbf{x}, t), \tag{6A}$$

where $\dot{u} = \partial u/\partial t$, $u^{ij} = \partial u^i/\partial x^j$ and $\alpha(\mathbf{x})$ and $\beta(\mathbf{x})$ are the compressional P-wave and S-wave velocities respectively. From here on we shall assume that the parts of the gradient vector, i.e. $\delta\hat{\phi}(\mathbf{x})$, $\delta\widehat{IP}(\mathbf{x})$, $\delta\widehat{IS}(\mathbf{x})$, $\delta\hat{\rho}(\mathbf{x})$, may be determined without difficulty and for ease of notation we shall rewrite this vector below as $\hat{\gamma} = (\hat{\gamma}_\phi, \hat{\gamma}_{IP}, \hat{\gamma}_{IS}, \hat{\gamma}_\rho)$.

Inversion Algorithms Using Gradient Methods

Having determined the gradient $\hat{\gamma}$ of the misfit function, the steepest ascent direction is given by (5). A steepest descent algorithm is of the form

$$\mathbf{m}_{n+1} = \mathbf{m}_n - \lambda_n \gamma_n \tag{7}$$

or

$$\delta\mathbf{m}_n = -\lambda_n \gamma_n,$$

where λ_n is a scalar chosen so that $F(\mathbf{m})$ is minimized along the step direction (see below). At each iteration the model is updated by taking a step in the steepest descent direction. Iterations are halted when some convergence criterion is satisfied. This type of method is notoriously slow in converging. A more practical approach is to employ some kind of conjugate gradient method. Mora (1987) uses a preconditioned conjugate gradient technique to minimize $F(\mathbf{m})$. In this type of method the first iteration is exactly as above ($n = 1$). Thereafter the descent direction is modified to incorporate each of the previous directions, i.e. to γ_n one adds a vector proportional to $\delta\mathbf{m}_{n-1}$. In this way after n iterations γ_n has contributions from all previous n descent directions. Conjugate gradient methods have been found to speed convergence at practically no extra computational cost.

All these methods introduce only one new descent direction at each iteration given by (5). Therefore they essentially group together all components of γ, i.e. γ_ϕ, γ_P, γ_{IS}, γ_ρ, into a single direction. Introducing the components of $\mathbf{m}$ into (7) gives

$$\begin{aligned}
\delta\phi_n &= -\lambda_n\{C_\phi \delta\hat{\phi}_n + \phi_n - \phi_o\},\\
\delta IP_n &= -\lambda_n\{C_{IP}\delta\widehat{IP}_n + IP_n - IP_o\},\\
\delta IS_n &= -\lambda_n\{C_{IS}\delta\widehat{IP}_n + IS_n - IS_o\},\\
\delta\rho_n &= -\lambda_n\{C_\rho \delta\hat{\rho}_n + \rho_n - \rho_o\},
\end{aligned} \tag{8}$$

where the second terms on the r.h.s. of (8) are due to the quadratic model term in the definition of $F(\mathbf{m})$ in (3). The algorithm given by (8) may be thought of as a 1D subspace scheme, i.e. the optimization of $F(\mathbf{m})$ in the complete model space has been replaced by a 1D optimization of F down the direction of steepest descent. The single free parameter λ_n is a constant for each parameter type and so the adjustment $\delta\phi_n$, δIP_n, δIS_n, $\delta\rho_n$, to each model parameter type at the nth iteration is governed by the properties of the overall descent direction and not by its individual direction. This feature is quite common. Many gradient algorithms actually ignore the differences between fundamentally different parameter types and simply treat them equally.

Tarantola (1986) suggests an algorithm which does make explicit use of the individual gradients, but in a rather coarse manner. He defines a hierarchy between parameter types and then inverts for each one in turn using a preconditioned steepest descent algorithm similar to (7). Usually when we include source terms, as in our example, we must invert for these first, then for P-wave impedance, S-wave impedance and finally density. This produces an algorithm of the form:

1: invert for the source function $\phi(\mathbf{x})$ until convergence using

$$\delta\phi_n(\mathbf{x}, t) = -\lambda_n^{\phi}\{C_{\phi}\,\delta\hat{\phi}_n(\mathbf{x}, t) + \phi_n(\mathbf{x}, t) - \phi_0(\mathbf{x}, t)\};$$

2: invert for P-wave impedance $IP(\mathbf{x})$ using

$$\delta IP_j(\mathbf{x}) = -\lambda_j^{\mathrm{IP}}\{C_{\mathrm{IP}}\,\delta\widehat{IP}_j(\mathbf{x}) + IP_j(\mathbf{x}) - IP_0(\mathbf{x})\};$$

3: invert for S-wave impedance $IS(\mathbf{x})$ using

$$\delta IS_k(\mathbf{x}) = -\lambda_k^{\mathrm{IS}}\{C_{\mathrm{IS}}\,\delta\widehat{IS}_k(\mathbf{x}) + IS_k(\mathbf{x}) - IS_0(\mathbf{x})\};$$

4: invert for the density $\rho(\mathbf{x})$ using

$$\delta\rho_l(\mathbf{x}) = -\lambda_l^{\rho}\{C_{\rho}\,\delta\hat{\rho}_l(\mathbf{x}, t) + \rho_l(\mathbf{x}) - \rho_0(\mathbf{x})\}, \tag{9}$$

where the free parameters λ_n^{ϕ}, λ_n^{IP}, λ_n^{IS}, λ_n^{ρ} are found such that the functional F is minimized along each individual descent direction.

In this algorithm each parameter type is adjusted while keeping the others fixed either at their *a priori* values or their updated values. Although this does allow the update to each parameter type to be determined by its own gradient, it essentially assumes independence between parameter types, which is never really the case. The user must decide, usually without much guidance, when to stop iterations on one parameter type and move on to the next. Ideally when inverting for one parameter type we should not neglect all others even if the problem does contain some natural hierarchy, and certainly not if it contains none.

We claim that neither of the two approaches described above is optimal. Both require the calculation of the individual gradient components, $\hat{\gamma}_{\phi}$, $\hat{\gamma}_{\mathrm{IP}}$, $\hat{\gamma}_{\mathrm{IS}}$, $\hat{\gamma}_{\rho}$ but neither make the most efficient use of them. The subspace approach, on the other hand, makes optimal use of the gradient components, and is therefore more efficient than either of the other two methods. The differences are especially noticeable in cases where the contours of the objective function are strongly aligned along one of the axes. Returning to the full problem, we must find the optimal combination of

gradient vector components. This may be achieved quite easily using a quadratic approximation to the misfit function within the subspace defined by the descent directions γ_ϕ, γ_{IP}, γ_{IS}, γ_ρ. We proceed as follows. Given the misfit F at a point $\mathbf{m}$ in model space, we approximate its value at a point $(\mathbf{m} + \delta\mathbf{m})$ using the second-order expansion shown in (4). If we restrict the movement in model space to our chosen subspace, i.e.

$$\delta\mathbf{m} = \sum_{i=1}^{k} \alpha_i \mathbf{a}^{(i)}, \tag{10}$$

where k is the number of subspace directions and the vectors $\mathbf{a}^{(i)}$ are given by

$$\mathbf{a}^{(1)} = \begin{pmatrix} \gamma_\phi \\ 0 \\ 0 \\ 0 \end{pmatrix}, \qquad \mathbf{a}^{(2)} = \begin{pmatrix} 0 \\ \gamma_{IP} \\ 0 \\ 0 \end{pmatrix}, \qquad \mathbf{a}^{(3)} = \begin{pmatrix} 0 \\ 0 \\ \gamma_{IS} \\ 0 \end{pmatrix}, \qquad \mathbf{a}^{(4)} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \gamma_\rho \end{pmatrix}, \tag{11}$$

then we have, using (4),

$$F(\mathbf{m} + \delta\mathbf{m}) = F(\mathbf{m}) + \sum_{i=1}^{k} \alpha_i \langle \hat{\gamma}, \mathbf{a}^{(i)} \rangle + \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1}^{k} \alpha_i \alpha_j \langle \mathbf{H}\mathbf{a}^{(i)}, \mathbf{a}^{(j)} \rangle + \cdots \tag{12}$$

To find the optimal coefficients α_i we set $\partial F / \partial \alpha_i = 0$, for $i = 1, \ldots. k$, and obtain

$$\langle \hat{\gamma}, \mathbf{a}^{(i)} \rangle + \sum_{j=1}^{k} \alpha_j \langle \mathbf{H}\mathbf{a}^{(i)}, \mathbf{a}^{(j)} \rangle \approx 0, \qquad i = 1, \ldots k. \tag{13}$$

Since the duality product $\langle \mathbf{H}\mathbf{a}^{(i)}, \mathbf{a}^{(j)} \rangle$ is a scalar we may define the $k \times k$ matrix $\mathbf{H}$ such that

$$(\mathbf{H})_{ij} = \langle \mathbf{H}\mathbf{a}^{(i)}, \mathbf{a}^{(j)} \rangle \tag{14}$$

and the k-dimensional vector $\boldsymbol{\theta}$ where

$$\theta_i = \langle \hat{\gamma}, \mathbf{a}^{(i)} \rangle. \tag{15}$$

Rearranging (13) gives

$$\alpha_i = -(\mathbf{H})_{ij}^{-1} \theta_j, \qquad i = 1, \ldots, k, \tag{16}$$

or if $\boldsymbol{\alpha}$ is the vector of k dimensions with components α_i then

$$\boldsymbol{\alpha} = -\mathbf{H}^{-1}\boldsymbol{\theta}. \tag{17}$$

The update to the model is then given by (10). The $k \times k$ matrix $\mathbf{H}$ is determined using (14). Since k is the number of subspace directions which is usually quite small (4×4 in our case) then $\mathbf{H}$ is simple to invert. In general the coefficients α_i may be found without difficulty.

To demonstrate that the subspace approach makes optimal use of the gradient partition vectors and in this sense results in a less-biased step at each iteration, we consider a simple example problem where the model vector may be divided into two parameter types or classes (not two dimensions).

$$\mathbf{m} = \begin{bmatrix} \mathbf{p}(\mathbf{x}) \\ \mathbf{s}(\mathbf{x}) \end{bmatrix},$$

where $\mathbf{p}(\mathbf{x})$ and $\mathbf{s}(\mathbf{x})$ are vectors of some dimensions N_p and N_s respectively, and $\mathbf{m}$ has dimension $(N_p + N_s)$. The steepest descent direction is then given by

$$\gamma = -\begin{bmatrix} \gamma_p \\ \gamma_s \end{bmatrix},$$

where

$$\gamma_p = C_p \hat{\gamma}_p, \qquad \gamma_s = C_s \hat{\gamma}_s.$$

If we project the contours of the misfit function on to the subspace defined by the steepest descent vector components $-\gamma_p$ and $-\gamma_s$, then within this subspace, γ is in general non-optimal, i.e. does not pass through the minimum of the projected contours. This will still be the case, even in a linear problem where the contours are by definition elliptical. Figure 1 shows this more clearly, where the perpendicular axes represent the vectors $[-\gamma_p, 0]^T$ and $[0, -\gamma_s]^T$ and the steepest descent vector is given by their sum. A single iteration of a steepest descent algorithm seeks the minimum value of the objective function along the dotted line at 45° to the axis and arrives at the point P_{sd} while a sequential method arrives at P_{seq} (taking the P-parameter type first and then the S-parameter type). The subspace approach on the other hand performs a quadratic optimization of the objective function in this plane and therefore by definition will arive at P_{opt}.

By using the subspace technique we take a much better step at each iteration than those given by the single descent algorithm or the sequential type algorithm. Furthermore, the direction of the single descent term γ is dependent on the *a priori* choice of the covariance operators C_ϕ, C_{IP}, C_{IS}, C_ρ through (5). So the relative adjustments of each parameter type will also be influenced by this choice. Using the above approach we essentially perform a least-squares inversion within the subspace, i.e. we project the full Hessian on to the subspace and solve the least-squares problem. In this case the chosen descent direction is independent of the relative 'sizes' of the subspace vectors. In the analogous discrete inverse problem this effect is described well by Williamson (1986).) So it may be seen that the subspace technique to a certain extent removes the implicit biasing which occurs from grouping each model parameter type into a single gradient vector. Conjugate gradient methods, which employ the overall descent direction γ, will suffer from similar down-weighting problems as in the single gradient method. Additionally conjugate gradient methods work by making use of previous descent directions and so retain old information on the curvature of the misfit function. Subspace methods, on the other hand, always use current information and so, one presumes, will be more

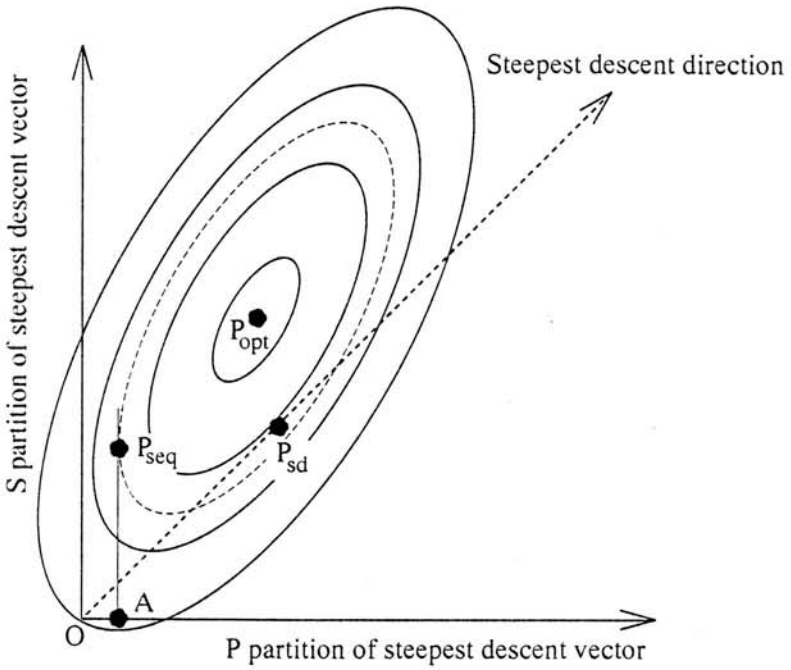

FIG. 1. Elliptical contours projected on to the 2D subspace formed by partition of the gradient vector. N.B. The overall steepest descent direction given by P_{sd} is non-optimal within the subspace.

useful in strongly non-linear problems where previous descent directions quickly become obsolete. Several numerical examples of subspace schemes applied to multi-parameter type optimization problems are available. Kennett *et al.* (1988) give examples of a non-linear traveltime reflection problem involving P-velocity and reflector depth parameters and a linear inversion of earthquake arrival times for P- and S-velocities and hypocentral location parameters. Williamson (1990) also gives examples of the reflection problem and Sambridge (1990) uses the subspace scheme in a non-linear inversion for 3D velocity fields and earthquake parameters. In each of these studies the subspace scheme is a useful tool in dealing with several parameter types simultaneously without the need of a sequential or single gradient method. Kennett *et al.* (1988) also compare a subspace algorithm with a standard steepest descent scheme and find it makes optimal use of the gradient vector partitions.

COMPUTATIONAL ASPECTS

In general, if we have a steepest descent algorithm of the form (7), then the 'optimal' step length is usually defined as the one which gives the greatest reduction in the

misfit S. If the problem is linear, i.e. $F(\mathbf{m})$ is quadratic in $\mathbf{m}$ then it is easy to show that

$$\alpha_n = \frac{\langle \hat{\gamma}_n, \gamma_n \rangle}{\langle \mathbf{H}_n \gamma_n, \gamma_n \rangle}, \tag{18}$$

where $\mathbf{H}$ is the Hessian of the misfit functional $F(\mathbf{m})$. Since the problem is non-linear, this value is only an approximation. If $F(\mathbf{m})$ is given by (3) then it may be shown that

$$\mathbf{H}_n \approx G_n^{\mathrm{T}} C_{\mathrm{D}}^{-1} G_n + C_{\mathrm{M}}^{-1}, \tag{19}$$

where G_n is the functional derivative of the non-linear operator $\mathbf{g}$ at the point $\mathbf{m}_n$ and G_n^{T} is the transpose operator (Morse and Feshbach 1953). In practice the explicit calculation of G_n is avoided. To show how this is done we rewrite the denominator in (18) as

$$\begin{aligned} \langle \mathbf{H}_n \gamma_n, \gamma_n \rangle &= \langle G_n^{\mathrm{T}} C_{\mathrm{D}}^{-1} G_n \gamma_n, \gamma_n \rangle + \langle C_{\mathrm{M}}^{-1} \gamma_n, \gamma_n \rangle \\ &= \langle C_{\mathrm{D}}^{-1} G_n \gamma_n, G_n \gamma_n \rangle + \langle C_{\mathrm{M}}^{-1} \gamma_n, \gamma_n \rangle. \end{aligned} \tag{20}$$

The second term may be calculated easily. For the first we require only the operation of the derivative operator on the descent direction. In the steepest descent algorithm we need to calculate the vector $G_n \gamma_n$. This may be done with explicit determination of G by solving the forward problem, i.e. we perturb the model by some small amount $\varepsilon \delta \mathbf{m}$ and determine the new predicted data set $\mathbf{g}(\mathbf{m} + \varepsilon \delta \mathbf{m})$, then use the expression

$$G \delta \mathbf{m} \approx \frac{\mathbf{g}(\mathbf{m} + \varepsilon \delta \mathbf{m}) - \mathbf{g}(\mathbf{m})}{\varepsilon}. \tag{21}$$

In the sequential inversion of Tarantola this process is repeated for each parameter type in order to calculate the step lengths λ_n^{ϕ}, λ_j^{IP}, etc. In the subspace scheme described above we perform exactly the same calculations. To demonstrate this we rewrite (14) using the Hessian approximation (20) and obtain

$$\begin{aligned} (\mathbf{H})_{ij} &= \langle G^{\mathrm{T}} C_{\mathrm{D}}^{-1} G \mathbf{a}^{(i)}, \mathbf{a}^{(j)} \rangle + \langle C_{\mathrm{M}}^{-1} \mathbf{a}^{(i)}, \mathbf{a}^{(j)} \rangle \\ &= \langle C_{\mathrm{D}}^{-1} G \mathbf{a}^{(i)}, G \mathbf{a}^{(j)} \rangle + \langle C_{\mathrm{M}}^{-1} \mathbf{a}^{(i)}, \mathbf{a}^{(j)} \rangle \\ &= \langle C_{\mathrm{D}}^{-1} \mathbf{b}^{(i)}, \mathbf{b}^{(j)} \rangle + \langle C_{\mathrm{M}}^{-1} \mathbf{a}^{(i)}, \mathbf{a}^{(j)} \rangle, \end{aligned} \tag{20}$$

where we have introduced the vectors $\mathbf{b}$ such that $\mathbf{b}^{(i)} = G \mathbf{a}^{(i)}$, each of which requires the solution of the forward problem. (In fact in calculating the vectors $\mathbf{b}^{(i)}$ we perturb only part of the model each time we solve the forward problem as each subspace direction depends on only one parameter type.) It is thought that in most problems the computation required to solve this series of partial forward problems would be comparable to that required by a single complete forward modelling. In

ase a partial forward solving is required to determine each of the step lengths $_i^P$, etc, in the Tarantola algorithm (9).

y substituting the vectors $\mathbf{a}^{(i)}$ from above into (20), the contribution to $\mathbf{H}$ from :cond term becomes

$$\begin{aligned}
\mathbf{H})_{ij} &= \langle C_\phi^{-1}\gamma_\phi, \gamma_\phi\rangle && \text{for } i,j = 1,\\
&= \langle C_{IP}^{-1}\gamma_{IP}, \gamma_{IP}\rangle && \text{for } i,j = 2,\\
&= \langle C_{IS}^{-1}\gamma_{IS}, \gamma_{IS}\rangle && \text{for } i,j = 3,\\
&= \langle C_\rho^{-1}\gamma_\rho, \gamma_\rho\rangle && \text{for } i,j = 4,
\end{aligned}$$

$$\mathbf{H})_{ij} = 0 \qquad \text{for } i \neq j.$$

by substituting (11) into (15) we see that these terms are also the components of ector $(\boldsymbol{\theta})_i$ $(i = 1, \ldots, 4)$. So having determined the vectors $\mathbf{b}^{(i)}$, $i = 1, \ldots, 4)$ (by al forward solving) we can evaluate all terms in $\mathbf{H}$ and $\boldsymbol{\theta}$ by evaluating the opriate duality products. Note for our earth model given by the set of functions $t)$, $IP(\mathbf{x})$, $IS(\mathbf{x})$, $\rho(\mathbf{x})\}$ the duality products become integrals over time (for the :e) and earth volume.

Summary of Proposed Inversion Algorithm

key features of one iteration of the proposed algorithm may be summarized as ws.

. From some *a priori* model $\{\phi(\mathbf{x}, t)$, $IP(\mathbf{x})$, $IS(\mathbf{x})$, $\rho(\mathbf{x})\}$ solve the forward lem to find the predicted surface displacements at the receiver positions $u^i_{cal}(\mathbf{x}_r$, en calculate the weighted residuals as given by (6).

. Back-propagate the weighted residuals to determine the missing field.

. Time-correlate the current and missing field to find the descent vectors with :ct to each parameter type $\gamma_\phi, \gamma_{IP}, \gamma_{IS}, \gamma_\rho$ as in (5) and (6A).

. Solve the forward problem after perturbing each parameter type and calculate ectors $\mathbf{b}^{(i)}$ $(i = 1, \ldots 4)$ given by (21).

. Determine the components of the 4×4 Hessian using the vectors $\mathbf{b}^{(i)}$ and Hence calculate the subspace coefficients α_i given by (17).

. Update all model parameters using (10).

Conclusion

paper demonstrates how the previously proposed subspace scheme may be ied to the non-linear inversion of multi-offset seismic reflection waveforms. It not make claims based on an actual application of this technique to a real data out instead restricts itself to a comparison with two previously used techniques demonstrates, using a simple example, how the subspace technique is superior

in this case. The application of the gradient subspace technique to the inversion of seismic reflection waveforms produces a simple algorithm which requires the same gradient vector calculations as the algorithm of Tarantola (1986) devised for the same problem (without sources as parameters). Dividing the steepest descent vector into its components and minimizing the misfit function within this 4D subspace produces optimal use of the individual descent directions (based on a local linearization of the problem within that subspace). This is the most important feature of the proposed scheme and it alone makes it preferable to single gradient methods. The algorithm uses only up-to-date local information at each iteration and does not retain previous descent directions. Finally it provides a natural way of incorporating the subspace strategy into the inversion of seismic reflection waveforms.

REFERENCES

KENNETT, B.L.N., SAMBRIDGE, M.S. and WILLIAMSON, P.R. 1988. Subspace methods for large scale inverse problems involving multiple parameter classes. *Geophysical Journal* **94**, 237–247.

KENNETT, B.L.N. and WILLIAMSON, P.R. 1987. Subspace methods for large scale nonlinear inversion. In: Mathematical Geophysics: a survey of recent developments in seismology and geodynamics. N. J. Vlaar, G. Nolet, M. J. R. Wortel and S. A. P. L. Cloetingh (eds), D. Reidel, Dordrecht, The Netherlands.

MORA, P. 1987. Elastic wavefield inversion. Ph.D. thesis, Stanford University, U.S.A.

MORSE, P.M. and FESHBACH, H. 1953. *Methods of Theoretical Physics.* McGraw-Hill Book Co.

SAMBRIDGE, M.S. 1988. Seismic inversion for earthquake location and 3D velocity structure. Ph.D. thesis, Australian National University, Canberra, Australia.

SAMBRIDGE, M.S. 1990. Non-linear arrival time inversion: constraining velocity anomalies by seeking smooth models in three dimensions. *Geophysical Journal International* **102**, 653–677.

SAMBRIDGE, M.S. and KENNETT, B.L.N. 1986. A novel method of hypocentre location. *Geophysical Journal of the Royal Astronomical Society* **87**, 679–697.

TARANTOLA, A. 1986. A strategy for the nonlinear elastic inversion of seismic reflection data. *Geophysics* **51**, 1893–1903.

TARANTOLA, A. 1987. *Inverse Problem Theory : Methods for Data Fitting and Model Parameter Estimation.* Elsevier Science Publishing Co.

VIREUX, J. 1986. P-SV wave propagation in heterogeneous media; velocity–stress finite-difference method. *Geophysics* **51**, 889–901.

WILLIAMSON, P.R. 1986. Tomographic inversion of travel time data in reflection seismology. Ph.D. thesis, University of Cambridge.

WILLIAMSON, P.R. 1990. Tomographic inversion in reflection seismology. *Geophysical Journal International* **100**, 255–274.