Geophysical Journal (1988) **94**, 237–247

Subspace methods for large inverse problems with multiple parameter classes

B. L. N. Kennett*, M. S. Sambridge* and †P. R. Williamson

* *Research School of Earth Sciences, Australian National University, GPO Box 4, Canberra ACT 2601, Australia;*
† *Centre for Geophysical Exploration Research, Macquarie University, Sydney NSW 2109, Australia*

Accepted 1988 January 1. Received 1988 January 1; in original form 1987 October 9

SUMMARY

Most nonlinear inverse problems can be cast into the form of determining the minimum of a misfit functional of model parameters. This functional determines the misfit between observations and the corresponding theoretical predictions, subject to some regularization conditions on the form of the model. When there is only one type of parameter in the model, methods based on gradient techniques work well, especially when information on rate of change of gradients is included.

In the case of problems depending on multiparameter classes, simple gradient methods mix parameters of different character and physical dimensionality. This may lead to rather poor convergence and strong dependence on the scaling of the different parameter types. These difficulties can be overcome by replacing a gradient step by a local minimization in a subspace spanned by a limited number of vectors in model space. The basis vectors for the subspace should be chosen in the directions determined by the variation of the misfit functional with respect to each of the parameter types, with supplementation if required by additional vectors representing the rate of change of the gradient partitions. The construction of the perturbation requires the inversion of a matrix with the dimensions of the subspace which is easily accomplished.

Such a subspace scheme takes into account the different functional dependences on the various parameter types in a balanced way. The update to the current model does not depend on the scaling of the individual parameter classes. The subspace method is flexible and can be adapted to a wide range of choices of misfit criterion and modes of representation of the parameter classes. This style of iterative subspace procedure is well adapted to nonlinear problems with dependence on many parameters and can be successfully applied in a variety of problems, e.g. seismic reflection tomography, the simultaneous nonlinear determination of earthquake locations and velocity fields and in the inversion of full seismic waveforms.

Key words: Inverse problems, subspace methods, seismic tomography, waveform inversion

1 INTRODUCTION

In many geophysical problems the observable data depend on parameters of different types with varying character or physical dimensions. For example, the free oscillation frequencies for the Earth depend on the density and the *P*, *S* wavespeed distributions in an isotropic representation and even more parameters are introduced in an anisotropic model.

However, when an attempt is made to invert for a physical model comprising such a suite of parameter types, most inversion algorithms do not take the differences in the characters of the parameters into account. This is generally unimportant for small linear problems where generalized inverse methods are applicable. Poor conditioning of the matrices can usually be improved by numerical manipulation, e.g. column normalization. However, inappropriate relative scaling of different parameter types may adversely affect the inversion path in model space for problems which are large enough, or sufficiently non-linear, to require iterative schemes. This can retard convergence and may well result in a biased answer. A more severe problem arises with gradient methods in nonlinear inversion schemes.

Commonly, where data are not abundant, only those parameters which are expected to be most significant (or achievable) are determined in the inversion and the remainder are assumed known. Such a procedure has the disadvantage that a bias can be introduced into the inversion by imperfections in the representations of the fixed parameter types. An alternative strategy is to adopt a hierarchical approach to inversion and to determine parameter distributions sequentially in order of assumed importance. This procedure requires the different types of parameter to be essentially independent in their contributions to the observed data, and has the disadvantage of allowing build up of error in the successive inversions.

Unfortunately, there is frequently cross-dependence between two parameter types with consequent trade-offs in the character of the inversion depending on the precise sequence of operations. Such sequential inversions can be useful in the initial exploration of the character of a problem and where computational capacity is limited. The worst features can be avoided if the whole process is iterated, with only a partial inversion for each parameter type attempted at each step.

In this paper we present an approach which resolves the question of the weighting of the changes in different parameters in what seems to us to be a natural fashion. We formulate the inversion procedure as an optimisation problem requiring the minimum of a functional of the model parameters which assesses the concordance between observed and computed data values, and includes some regularization term, incorporating available *a priori* information, to prevent extravagant behaviour. The procedure is iterative and represents a hybrid between descent and matrix methods.

At each current model, we evaluate the gradient vector for the misfit functional and split it up into components, each involving a different parameter type. We then obtain an updated estimate of the model by minimizing the misfit functional within the subspace defined by the corresponding 'descent' vectors. Thus the weighting between parameters is determined solely by the misfit functional and corresponds in that sense to the best possible relative scaling. The computation required additional to that of a basic descents algorithm is small as it requires only the establishment and solution of a system of equations of the dimensionality of the subspace. At each step we need to solve a linearized problem involving the projection of the full Hessian onto the subspace. This approach is a particular application of a Subspace method in model space, such techniques have recently been found to be very effective in solving large-scale nonlinear inverse problems (Kennett & Williamson 1987).

We illustrate the method by showing how it can be adapted to a number of geophysical inversion problems involving a number of parameter types, specifically seismic reflection tomography, the simultaneous estimation of hypocentre parameters and the velocity distribution from earthquake travel times, and the problem of the inversion of full seismic waveforms. In seismic reflection tomography, both the shape of a reflector and the velocity field above it are to be estimated from the travel times of waves reflected from the interface. For the earthquake location and velocity estimation problem there are dimensional differences between the origin times and the spatial hypocentre coordinates and also differences in character between the P and S wavespeed distributions. These can be all be taken into account by a suitable choice of subspace. For full waveform inversion both the P and S wavespeed distributions as well as the density need to be found so that at least three parameter classes have to be found during the inversion.

2 INVERSION SCHEME

For simplicity and clarity we will confine our attention to discrete inverse problems, but the procedures we will describe can be readily adapted to the case of continuous parameter distributions (Sambridge *et al.* 1988).

We suppose that we are presented with a set of observations $\mathbf{d}_0\{d_{0r}, r = 1, \ldots, M\}$ and wish to use those observations to determine a discrete set of parameters $\mathbf{m}$ $\{m_k, k = 1, \ldots, N\}$. Since we are interested in the situation where the model is built up from a number of parameter types, we assume that we can partition the model into a number of subsets of parameters, with one for each type, so that we will set

$$\mathbf{m} = [\mathbf{m}_A, \mathbf{m}_B, \mathbf{m}_C, \ldots], \tag{2.1}$$

with a total of P parameter classes. The dimensionality of the subsets $\mathbf{m}_A$, $\mathbf{m}_B$ etc. will vary according to the nature of the problem. Within each subset, the parameters may describe the model directly or may be the coefficients in an expansion in terms of orthonormal functions (Nolet 1987a)

$$\mathbf{m}_A = \sum_{r=1}^{N_A} m_{Ar} h_r(\mathbf{x}),$$

where the basis functions h_r satisfy

$$\int_R \mathrm{d}^D\mathbf{x}\, h_i(\mathbf{x}) h_j(\mathbf{x}) = \delta_{ij},$$

over a region R of dimensionality D.

Corresponding to each of the observed data values d_{0r} we have to calculate the predicted value $g_r(\mathbf{m})$ which will be some functional of the model parameters. We will use a data misfit function $\Phi(\mathbf{d}_0, \mathbf{g}(\mathbf{m}))$ to assess the level of disagreement between observed and calculated data values, the choice of the function depends on the nature of the problem and the error statistics of the data. If it is reasonable to assume Gaussian statistics then we can adopt

$$\Phi(\mathbf{d}_0, \mathbf{d}) = (\mathbf{d}_0 - \mathbf{d})^T \mathbf{C}_d^{-1} (\mathbf{d}_0 - \mathbf{d}), \tag{2.2}$$

where the data uncertainties are introduced by the data covariance matrix $\mathbf{C}_d$. We will assume that any preconditioning to reduce the nonlinearity of the problem has been absorbed into the definitions of $\mathbf{d}$ and $\mathbf{g}(\mathbf{m})$ (see, e.g. Chapman & Orcutt 1985). In order to constrain the behaviour of the model parameters, we introduce a regularisation function $\Psi(\mathbf{m}, \mathbf{m}_s)$ in terms of some starting model $\mathbf{m}_s$; commonly this would be related to a norm on the model. For example, we may choose the quadratic form

$$\Psi(\mathbf{m}, \mathbf{m}_s) = (\mathbf{m} - \mathbf{m}_s)^T \mathbf{C}_m^{-1} (\mathbf{m} - \mathbf{m}_s), \tag{2.3}$$

where $\mathbf{C}_m$ is the model covariance matrix whose properties should be chosen to fit what is known about the situation. For the case with a number of parameter types

$$\mathbf{C}_m = \begin{pmatrix} \mathbf{C}_{AA} & \mathbf{C}_{AB} & \mathbf{0} \\ \mathbf{C}_{BA} & \mathbf{C}_{BB} & \\ \mathbf{0} & & \mathbf{C}_{CC} \\ & & \ddots \end{pmatrix}, \tag{2.4}$$

where we may need to introduce off-diagonal blocks to allow for trade-offs between different types of parameters. For example, in seismic model estimation we may expect correlation between the P- and S-wave velocities, and also with the density. The individual parameter covariance matrices $\mathbf{C}_{AA}$ etc. may also need off-diagonal contributions,

e.g. in tomographic work there are advantages in allowing some degree of nearest neighbour interaction between cells (Williamson 1986).

We now seek to minimise the discrepancy between observed and calculated data values whilst maintaining constraints on the character of the parameter distribution. We can do this by minimizing

$$F(\mathbf{m}) = \Phi(\mathbf{d}_0, \mathbf{g}(\mathbf{m})) + \Psi(\mathbf{m}, \mathbf{m}_s), \tag{2.5}$$

with respect to the model $\mathbf{m}$: explicitly we have with quadratic forms for Φ, Ψ

$$F(\mathbf{m}) = (\mathbf{g}(\mathbf{m}) - \mathbf{d}_0)^T \mathbf{C}_d^{-1}(\mathbf{g}(\mathbf{m}) - \mathbf{d}_0) + (\mathbf{m} - \mathbf{m}_s)^T \mathbf{C}_m^{-1}(\mathbf{m} - \mathbf{m}_s). \tag{2.6}$$

For perfect data we would aim for the global minimum of F, but with observed data one practical termination criterion is to stop once the data misfit term Φ is reduced below a preassigned threshold. Other possible termination criteria are discussed in Kennett (1988).

With many model parameters a direct search for the minimum is out of the question and so we aim to exploit the local behaviour of F to guide us to the desired minimum. If F is a smooth function of the model parameters we can make a locally quadratic approximation about some current model $\mathbf{m}_c$ by truncating the Taylors series for F

$$F^Q(\mathbf{m}_c + \delta\mathbf{m}) = F(\mathbf{m}_c) + \nabla_m F(\mathbf{m}_c) \cdot \delta\mathbf{m} + 1/2\, \delta\mathbf{m} \cdot \nabla_m \nabla_m F(\mathbf{m}_c) \cdot \delta\mathbf{m}, = F(\mathbf{m}_c) + \hat{\boldsymbol{\gamma}} \cdot \delta\mathbf{m} + 1/2\, \delta\mathbf{m}\hat{\mathbf{H}}\delta\mathbf{m}, \tag{2.7}$$

in terms of the gradient $\hat{\boldsymbol{\gamma}}$ and the Hessian matrix $\hat{\mathbf{H}}$. The gradient lies in a dual of the model space defined by our choice of the norm on the model space. If we adopt the quadratic norm $\Psi^{1/2}$, the equivalent vector in model space (the direction of steepest ascent $\boldsymbol{\gamma}$) is related by the action of the model covariance matrix (Tarantola 1987)

$$\boldsymbol{\gamma} = \mathbf{C}_m \hat{\boldsymbol{\gamma}}. \tag{2.8}$$

With our assumed form for F, the gradient

$$\hat{\boldsymbol{\gamma}} = \mathbf{G}^T \mathbf{C}_d^{-1}(\mathbf{g}(\mathbf{m}) - \mathbf{d}_0) + \mathbf{C}_m^{-1}(\mathbf{m} - \mathbf{m}_s), \tag{2.9}$$

where $G_{ij} = \partial g_i / \partial m_j$ and the Hessian matrix

$$\hat{\mathbf{H}} = \mathbf{G}^T \mathbf{C}_d^{-1} \mathbf{G} + \nabla_m \mathbf{G}^T \mathbf{C}_d^{-1}(\mathbf{g}(\mathbf{m}) - \mathbf{d}_0) + \mathbf{C}_m^{-1}. \tag{2.10}$$

The Frechet derivative G_{ij} can often be found in an analytical form. However, in many circumstances, the second derivative term $\nabla_m \mathbf{G} = \nabla_m \nabla_m \mathbf{g}$ is difficult to calculate, but since it appears with the data misfit its significance should diminish as minimization proceeds and it is often neglected at the outset.

2.1 Subspace methods

A class of very effective algorithms can be developed by restricting the local minimization of the quadratic approximation to the misfit functional F^Q to a relatively small n-dimensional subspace of model parameter space (Kennett & Williamson 1987).

We introduce n basis vectors $\{\mathbf{a}^{(j)}\}$ and a projection matrix $\mathbf{A}$ composed of the components of these vectors

$$A_{ij} = a_i^{(j)} \quad i = 1, \ldots, N, \quad j = 1, \ldots, n. \tag{2.11}$$

We now construct a perturbation to the current model in space spanned by the $\{\mathbf{a}^{(j)}\}$,

$$\delta\mathbf{m} = \sum_{j=1}^{n} \mu_j \mathbf{a}^{(j)}. \tag{2.12}$$

The coefficients μ are to be determined by minimizing F^Q for this class of perturbation for which

$$F^Q = F(\mathbf{m}_c) + \sum_{j=1}^{n} \mu_j \hat{\boldsymbol{\gamma}} \cdot \mathbf{a}^{(j)} + 1/2 \sum_{j=1}^{n} \sum_{k=1}^{n} \mu_j \mu_k \mathbf{a}^{(k)T} \hat{\mathbf{H}} \mathbf{a}^{(j)}, \tag{2.13}$$

and minimizing with respect to μ_j we require $\partial F^Q / \partial \mu_j = 0$, so that

$$\hat{\boldsymbol{\gamma}} \cdot \mathbf{a}^{(j)} + \sum_{k=1}^{n} \mu_k \mathbf{a}^{(k)T} \hat{\mathbf{H}} \mathbf{a}^{(j)} = 0. \tag{2.14}$$

We may now rewrite (2.14) in terms of the projection matrix $\mathbf{A}$ as

$$\mathbf{A}^T \hat{\boldsymbol{\gamma}} + \mathbf{A}^T \hat{\mathbf{H}} \mathbf{A} \boldsymbol{\mu} = \mathbf{0}.$$

The perturbation coefficients can thus be determined from the projection of the gradient and the Hessian matrix onto the subspace in the form

$$\boldsymbol{\mu} = -(\mathbf{A}^T \hat{\mathbf{H}} \mathbf{A})^{-1} \mathbf{A}^T \hat{\boldsymbol{\gamma}}. \tag{2.15}$$

The projected Hessian is a small $n \times n$ matrix, which is generally well conditioned with sensible choices for the basis vectors $\{\mathbf{a}^{(j)}\}$.

The model perturbation $\delta\mathbf{m}$ can be recovered by projecting back into the full model space, and for the choice of misfit functional F (2.6) can be represented as

$$\delta\mathbf{m} = -\mathbf{A}[\mathbf{A}^T(\hat{\mathbf{H}}_0 + \mathbf{C}_m^{-1})\mathbf{A}]^{-1} \mathbf{A}^T \hat{\boldsymbol{\gamma}}, \tag{2.16}$$

where $\hat{\mathbf{H}}_0$ is the Hessian of the data-fit term ($\nabla_m \nabla_m \Phi$). The structure of (2.16) is reminiscent of a projected Marquardt algorithm though $\mathbf{C}_m^{-1}$ need not be diagonal.

The basis vectors $\mathbf{a}^{(j)}$ will normally be related to the ascent vector $\boldsymbol{\gamma}$ and its rate of change and so (2.16) normally combines to some extent gradient and matrix techniques for minimizing F^Q. Once the local model update estimate $\delta\mathbf{m}$ is constructed from (2.16), a new current model is created and used to generate a further local quadratic approximation to the behaviour of F. The cycle of estimating $\delta\mathbf{m}$ and model construction is then iterated until a suitable termination criterion for the minimization of F is activated.

2.2 Subspace techniques for many parameter types

The subspace method we have just introduced is quite general and can be applied to a set of parameters of the same type, or to parameters of a number of different types, by appropriate choice of the basis vectors $\{\mathbf{a}^{(j)}\}$.

When we have different types of parameters, model space becomes a product space $\mathbf{M} = \mathbf{M}_A \times \mathbf{M}_B \times \mathbf{M}_C \times \ldots$, and we are faced with a scaling problem; as we change the relative sizes of the units for the different parameter classes the direction of gradient vector changes. A similar effect arises when working with dimensionless parameters under change of the choice of reference values. Further, to derive the ascent vectors in model space we need to invoke the action of the covariance matrix $\mathbf{C}_m$ on the gradients, and so

this matrix also has considerable significance. Often, our quantitative knowledge of a suitable choice for $\mathbf{C}_m$ is inadequate and a poor set of estimates for the entries may well slow convergence or introduce bias if an early termination is forced.

The subspace method provides a natural way in which to exploit the dependence of the objective function F on each parameter class. We partition the gradient vector $\hat{\gamma}$ into the contributions for each parameter type, so that

$$\hat{\gamma} = [\hat{\gamma}_A, \hat{\gamma}_B, \hat{\gamma}_c, \ldots], \tag{2.17}$$

where, e.g. $\hat{\gamma}_A = \partial F / \partial \mathbf{m}_A$, and it should be recalled that the partitions are not necessarily of equal dimensionality. In the interests of brevity we will write column vectors in a horizontal format as in (2.17).

Now, by the choice (2.12) of the form of possible model perturbations we have assumed that the basis vectors $\{\mathbf{a}^{(j)}\}$ lie in model space, but the gradient $\hat{\gamma}$ lies in the dual (gradient) space. If we concentrate on one class of model parameter at a time, we can construct the corresponding ascent vector in model space by the action of the model covariance matrix on just the gradient components corresponding to the particular parameter type. Thus we define for parameter type I

$$\gamma_I = \mathbf{C}_m[\ldots, \mathbf{0}, \hat{\gamma}_I, \mathbf{0}, \ldots]. \tag{2.18}$$

With the form of model covariance matrix introduced above in (2.4)

$$\begin{aligned} \gamma_A &= [\mathbf{C}_{AA}\hat{\gamma}_A, \mathbf{C}_{BA}\hat{\gamma}_A, 0, \ldots], \\ \gamma_B &= [\mathbf{C}_{AB}\hat{\gamma}_B, \mathbf{C}_{BB}\hat{\gamma}_B, \mathbf{0}, \ldots], \\ \gamma_C &= [\mathbf{0}, \mathbf{0}, \mathbf{C}_{CC}\hat{\gamma}_C, \mathbf{0}, \ldots], \\ &\ldots \end{aligned} \tag{2.19}$$

and the off-diagonal blocks in the model covariance allow for specific cross-coupling between parameter classes when this is a desirable feature of the problem. We now adopt the set of P ascent vectors as the directions of the set of basis vectors so that

$$\mathbf{a}^{(1)} = \|\gamma_A\|^{-1}\gamma_A, \qquad \mathbf{a}^{(2)} = \|\gamma_B\|^{-1}\gamma_B, \ldots \tag{2.20}$$

where we have normalized the basis vectors using the assumed quadratic norm (2.3) in model space. Thus

$$\|\gamma\|^2 = \gamma \mathbf{C}_m^{-1}\gamma = \hat{\gamma}\mathbf{C}_m\hat{\gamma},$$

and so

$$\|\gamma_A\| = (\hat{\gamma}_A^T \mathbf{C}_{AA}\gamma_A)^{1/2}.$$

Where cross-coupling exists between parameter sets—as in (2.19)—it is desirable to avoid linear dependence between the different $\mathbf{a}^{(j)}$. This can be achieved by orthogonalizing the basis vectors, and so we should modify the second vector to

$$\mathbf{b}^{(2)} = \mathbf{a}^{(2)} - \frac{(\mathbf{a}^{(1)}\mathbf{C}_m^{-1}\mathbf{a}^{(2)})}{(\mathbf{a}^{(1)}\mathbf{C}_m^{-1}\mathbf{a}^{(1)})}\mathbf{a}^{(1)}, \tag{2.21}$$

which will then need to be normalized. By this means we are able to build a P dimensional set of basis vectors, with each one corresponding to the variation of the data-misfit functional F with a particular parameter type.

The benefits of this adaptation of the subspace method are shown in Fig. 1. The plane defined by γ_a and γ_b in a problem with two parameter types (a, b) is shown, under the assumption of no off-diagonal blocks in $\mathbf{C}_m$. The contours of F^Q, the quadratic approximation to the objective functional at the current model $\mathbf{m}_c$ are superimposed. The dashed arrow denotes a typical descents direction so that the vector is perpendicular to a contour, the minimum value of F^Q along this path is reached at A_1. The solid arrow indicates the step prescribed by the subspace method using γ_a and γ_b as basis vectors which arrives at A_{opt}. The improvement over the descent step is readily apparent.

If the number of parameter classes P is greater than 4, the descent vectors derived from the gradients of the objective functional F will normally be a sufficient basis set. However, for a small number of parameter classes it is feasible to incorporate a further P^2 basis vectors representing the rate of change of the ascent vectors. We will illustrate the process by the example of two parameter classes. We partition the Hessian matrix into blocks by the dependence on parameter type

$$\hat{\mathbf{H}} = \begin{pmatrix} \hat{\mathbf{H}}_{AA} & \hat{\mathbf{H}}_{AB} \\ \hat{\mathbf{H}}_{BA} & \hat{\mathbf{H}}_{BB} \end{pmatrix}, \tag{2.22}$$

and then look at the rate of change of γ_A and γ_B with respect to both parameters. This generates four new vectors

$$\begin{pmatrix} \hat{\mathbf{H}}_{AA} & \hat{\mathbf{H}}_{AB} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\gamma_A, \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \hat{\mathbf{H}}_{BA} & \hat{\mathbf{H}}_{BB} \end{pmatrix}\gamma_A, \begin{pmatrix} \hat{\mathbf{H}}_{AA} & \hat{\mathbf{H}}_{AB} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\gamma_B, \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \hat{\mathbf{H}}_{BA} & \hat{\mathbf{H}}_{BB} \end{pmatrix}\gamma_B, \tag{2.23}$$

in the dual (gradient) space, which then have to be transfered back to model space by the action of the covariance matrix $\mathbf{C}_m$. They also have to be orthonormalized before addition to the basis set.

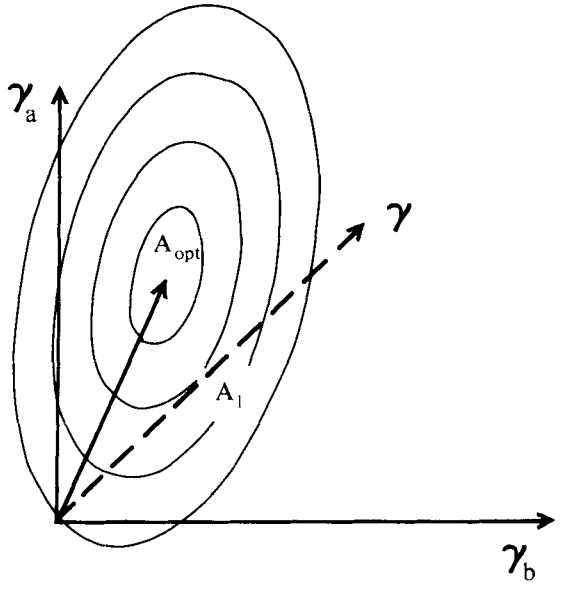


Figure 1. Elliptical contours of the local quadratic approximation to the objective functional F^Q projected onto the two dimensional subspace formed by partition of the gradient components. Note that the steepest descent direction is non-optimal within the subspace defined by its components: a step in the steepest descent direction arrives at A_1 whereas the 2-D subspace scheme arrives at A_{opt}.

Once we have set up the basis vectors $\{\mathbf{a}^{(j)}\}$, we have established the framework for using the subspace approach. With the local quadratic approximation for the mifist functional F, the perturbation to the current model should be estimated from (2.16)

$$\delta\mathbf{m} = -\mathbf{A}[\mathbf{A}^T(\hat{\mathbf{H}}_0 + \mathbf{C}_m^{-1})\mathbf{A}]^{-1}\mathbf{A}^T\hat{\boldsymbol{\gamma}}, \tag{2.16}$$

where A_{ij} is the projection matrix $\{a_i^{(j)}\}$. If the second derivative term in the Hessian ($\nabla_m\nabla_m\mathbf{g}$) can be neglected we need to evaluate terms like

$$\mathbf{K} = \mathbf{a}^{(i)T}[\mathbf{G}^T\mathbf{C}_d^{-1}\mathbf{G} + \mathbf{C}_m^{-1}]\mathbf{a}^{(j)}, \tag{2.24}$$

where the derivative $\mathbf{G}$ is evaluated at the current model. There is no need to construct the matrix $\mathbf{C}^T\mathbf{C}_d^{-1}\mathbf{G}$ since we can recast $\mathbf{K}$ into the form

$$\mathbf{K} = \mathbf{b}^{(i)T}\mathbf{C}_d^{-1}\mathbf{b}^{(j)} + \mathbf{a}^{(i)T}\mathbf{C}_m^{-1}\mathbf{a}^{(j)}, \tag{2.25}$$

where $\mathbf{b}^{(i)} = \mathbf{G}\mathbf{a}^{(i)}$, so only a single vector multiplication is required. When the basis vectors are just the P ascent vectors associated with the variations of the individual parameter classes, the vector $\mathbf{b}^{(I)}$ can be found directly from the action of a small change in the Ith parameter class on F.

If the model perturbation derived from (2.16) is so large as to move outside the likely range of the quadratic approximation for F, it is possible to regard (2.16) as defining a search direction and then only move partway towards the quadratic minimum by taking the update as $\mathbf{m}_c + \nu\delta\mathbf{m}$ with $\nu < 1$. However a preferable procedure is to modify the definition of F by adding a term $\varepsilon^2\|\mathbf{m} - \mathbf{m}_c\|^2$, where $\mathbf{m}_c$ is the current model. This does not affect the gradient but adds $\varepsilon^2\mathbf{I}$ to the Hessian and so decreases the step length in a way which will follow the true descents path as accurately as possible.

The subspace method essentially performs a least-squares inversion within the subspace, spanned by vectors which reflect the dependence on all the parameter classes. The model step generated is independent of the scaling of the particular parameter types (Williamson 1986). The weighting accorded to the different model types is determined solely by the behaviour of the objective functional. As a result we remove any bias that might be introduced by combining disparate parameter types in a single descent vector and achieve a effective balance between the information in the data (through Φ) and the *a priori* constraints imposed through Ψ.

Where information additional to the ascent directions for the individual parameter classes is desired it is preferable to generate this directly within the step, rather than to rely on information from previous iterations which may well not be relevant to the neighbourhood of the current model.

2.3 Comparison with gradient methods

From a current model $\mathbf{m}_c$, the simplest approach to updating the model in order to minimize the misfit functional F is to look for a model perturbation related to the descent vector in model space, so that

$$\delta\mathbf{m} = \mu\boldsymbol{\phi}(\mathbf{m}_c) \quad \text{with} \quad \boldsymbol{\phi} = \mathbf{S}_0\boldsymbol{\gamma}(\mathbf{m}_c). \tag{2.26}$$

The use of a matrix $\mathbf{S}_0$ differing from the unit matrix is termed 'pre-conditioning' by Tarantola (1987); he advocates an approximation to the initial curvature

$$\mathbf{S}_0 \approx [\mathbf{I} + \mathbf{C}_m\mathbf{G}^T(\mathbf{m}_s)\mathbf{C}_d^{-1}\mathbf{G}(\mathbf{m}_s)]^{-1}.$$

The parameter μ is to be chosen so that $F(\mathbf{m}_c + \delta\mathbf{m})$ is a minimum along the step direction. Within a local quadratic approximation, this is essentially a 1-D subspace approach and so

$$\mu = -\boldsymbol{\phi}^T\hat{\boldsymbol{\gamma}}(\boldsymbol{\gamma}^T\hat{\mathbf{H}}\boldsymbol{\phi}). \tag{2.27}$$

The disadvantage of such a scheme with many parameter classes is clearly demonstrated by recalling that $\boldsymbol{\gamma} = \sum_I \boldsymbol{\gamma}_I$, with summation over all the P parameter types. All the parameters are being treated the same way with differences in character ignored. The resulting direction and step will be affected by any rescaling of individual parameter classes.

The convergence of the steepest descents type of algorithm is comparatively slow and can be improved by using a conjugate gradient technique as employed by Mora (1987). In this case the search vector is built up from the gradient and the previous search directions. At the rth iteration

$$\boldsymbol{\phi}_r = \mathbf{S}_0\boldsymbol{\gamma}_r + \alpha_r\boldsymbol{\phi}_{r-1}$$
$$\text{where} \quad \alpha_r = (\boldsymbol{\gamma}_r - \boldsymbol{\gamma}_{r-1})^T\mathbf{C}_m^{-1}\mathbf{S}_0\boldsymbol{\gamma}_r/(\boldsymbol{\gamma}_{r-1}^T\mathbf{C}_m^{-1}\mathbf{S}_0\boldsymbol{\gamma}_{r-1}),$$

and so there are contributions from the previous r descent directions. This should give an improved choice of direction in which to look for a minimum. The actual minimization is however, once again, 1-D with comparable disadvantages to the steepest descent approach in terms of dependence on different parameter classes.

An alternative approach is to build up the total model perturbation as a sum of contributions corresponding to variations of one parameter at a time (cf. Tarantola 1986). We take

$$\delta\mathbf{m} = \sum_I \mu_I\boldsymbol{\gamma}_I, \tag{2.28}$$

with summation over parameter class. The weighting factors determined by minimization along the descent vector for each parameter class are

$$\mu_I = -\boldsymbol{\gamma}\mathbf{r}_I^T\hat{\boldsymbol{\gamma}}/(\boldsymbol{\gamma}_I^T\hat{\mathbf{H}}\boldsymbol{\gamma}_I). \tag{2.29}$$

Such an approach does begin to take account of the dependence of the misfit functional F on the different parameter classes and allows the update to each parameter type to be determined by its own gradient. However, this representation cannot take into account interactions between parameters and involves nearly as much computation as the subspace method.

The subspace method, on the other hand, makes full use of the information on the local dependence of F on the different parameter types and can allow for interdependence of parameter classes. The subspace method thus offers an effective and affordable means of handling inversion for multiple parameter types.

3 EXAMPLES OF THE USE OF SUBSPACE METHODS

We will describe three cases where the subspace method we have introduced in section 2.2 provides an effective

approach to the solution of an inverse problem. All of the examples are for seismological problems because these commonly involve parameters of different types; but the approach can certainly be used for a wide range of other geophysical inverse problems.

3.1 Simultaneous estimation of hypocentral parameters and velocity distributions from arrival times of earthquake phases

With a network of seismic stations, the times of arrival of various seismic phases can be estimated for an individual earthquake. But, before this information can be exploited to define the velocity structure in the neighbourhood of the network, the earthquake has to be located in space and time. Ideally, we should aim to locate earthquakes and determine the velocity structure simultaneously by making use of the data coverage from many events. The parameters required from the inversion clearly divide into two groups: the first consisting of the hypocentral parameters for all the events and the second consisting of those parameters needed to define the velocity distribution. This separation has been exploited in slightly different ways by Pavlis & Booker (1980) and Spencer & Gubbins (1980) in methods for linearized simultaneous inversion.

Within the context of the subspace approach, we could therefore envisage setting up two major parameter classes: $\mathbf{h}$ including all the hypocentral parameters and $\mathbf{v}$ for the velocity field. We can then partition the model as

$$\mathbf{m} = [\mathbf{h}, \mathbf{v}] \tag{3.1}$$

and exploit the 2-D subspace approach described in section 2.2. Such an approach can be made to work, but requires careful manipulation of the model covariance matrix to try to equalise the sensitivity of the misfit functional F to the parameters within each of the major groupings.

By treating the hypocentral parameters as a unit, we have implicitly mixed parameter dimensions by combining the spatial coordinates of the events with their origin times. The net result is that the shifts in the spatial coordinates are dominated by the adjustments to the origin times, since these have the largest gradient components. For a local velocity model for southeastern Australia, an upweighting of the spatial shifts by a factor of ten or more is needed to get flexibility in the inversion. A preferable solution is to recognize the distinction between the two types of hypocentral parameters and partition $\mathbf{h}$ into spatial and temporal parts,

$$\mathbf{h} = [\mathbf{h}_x, \mathbf{h}_t], \tag{3.2}$$

$\mathbf{h}_x$ then contains the spatial coordinates of all the events and $\mathbf{h}_t$ their origin times.

A similar problem arises for the velocity distribution parameters. The basic recorded times include both P- and S-wave phases and so the model representation must include both velocity fields. In addition, the position of a major interface, such as the Moho in regional work, can affect the travel times. It is therefore appropriate to split up the parameter set $\mathbf{v}$ into parts $\mathbf{v}_\alpha$, $\mathbf{v}_\beta$ associated with the P- and S-wave distributions and $\mathbf{v}_I$ arising from interfaces. Thus we should take

$$\mathbf{v} = [\mathbf{v}_\alpha, \mathbf{v}_\beta, \mathbf{v}_I]. \tag{3.3}$$

With the partitions of $\mathbf{h}$ and $\mathbf{v}$ as in (3.2), (3.3) the total model in the terminology of section 2 is

$$\mathbf{m} = [\mathbf{h}_x, \mathbf{h}_t; \mathbf{v}_\alpha, \mathbf{v}_\beta, \mathbf{v}_I]. \tag{3.4}$$

The corresponding subspace development would then be at least 5-D. Such a subspace development avoids the complications of having to worry about the relative sizes of the contributions from the different parameter types. It also means that at each iterative step, the dependance of the misfit functional F on all the different types of parameters have been treated in the same way in the estimate of the changes which have to be made to the current values. One may also wish to further subdivide $\mathbf{v}_\alpha$ and $\mathbf{v}_\beta$ to segregate different parts of the model, e.g. crustal and mantle velocities, and this would further increase the dimensionality of the subspace.

We will illustrate the merits of the subspace techniques by application to a study of the 3-D velocity structure of the southeastern corner of Australia, using earthquake and explosive sources (Fig. 2). Over 4200 travel times from 312 earthquakes were combined with 700 travel times from well-timed quarry blasts and refraction shots. The region was divided into cells of size one half a degree by one half a degree over the zone with adequate data coverage indicated by the heavy outline in Fig. 2. Separate P-wave distributions were taken for the crust and mantle, but the character of the S-wave readings precluded estimating S-wave mantle velocities, the depth of the crust/mantle interface was also allowed to vary. This leads to a total of 512 structural parameters of four different types and 1248 hypocentral parameters and the present analysis was based on a linearized treatment using a fixed set of ray paths based on a simple 1-D model for the region, as a preliminary study for a full nonlinear inversion with 3-D ray tracing.

In Fig. 3 we summarize the convergence behaviour of three different algorithms for this simultaneous hypocentre and structural inversion. In each case we attempt to minimize a data misfit functional of the form (2.2). The chain dotted curve (SD) shows the result of a steepest descent algorithm, i.e. a 1-D minimization along the steepest descent direction at each iteration. Such an approach makes no distinction between the different parameter types. The dashed curve (2D) is the result of using the 2-D subspace scheme described above, where the model parameters and split into hypocentral and structural sets. The basis vectors for the subspace are derived from the partitioned gradient vector as in (2.20). Convergence is both more rapid and smoother than for the simplest case. The solid curve (6D) shows the results of a 6-D subspace approach where the hypocentral parameters are divided into spatial and temporal components and the structural parameters comprise crustal and mantle P-wave speeds, crustal S-wave speeds and interface terms. The convergence per iteration is much more rapid than before with only a modest increase in computation time. Thus as the dimension of the subspace increases, the efficiency of the inversion algorithm per iteration is much improved. This is due to the

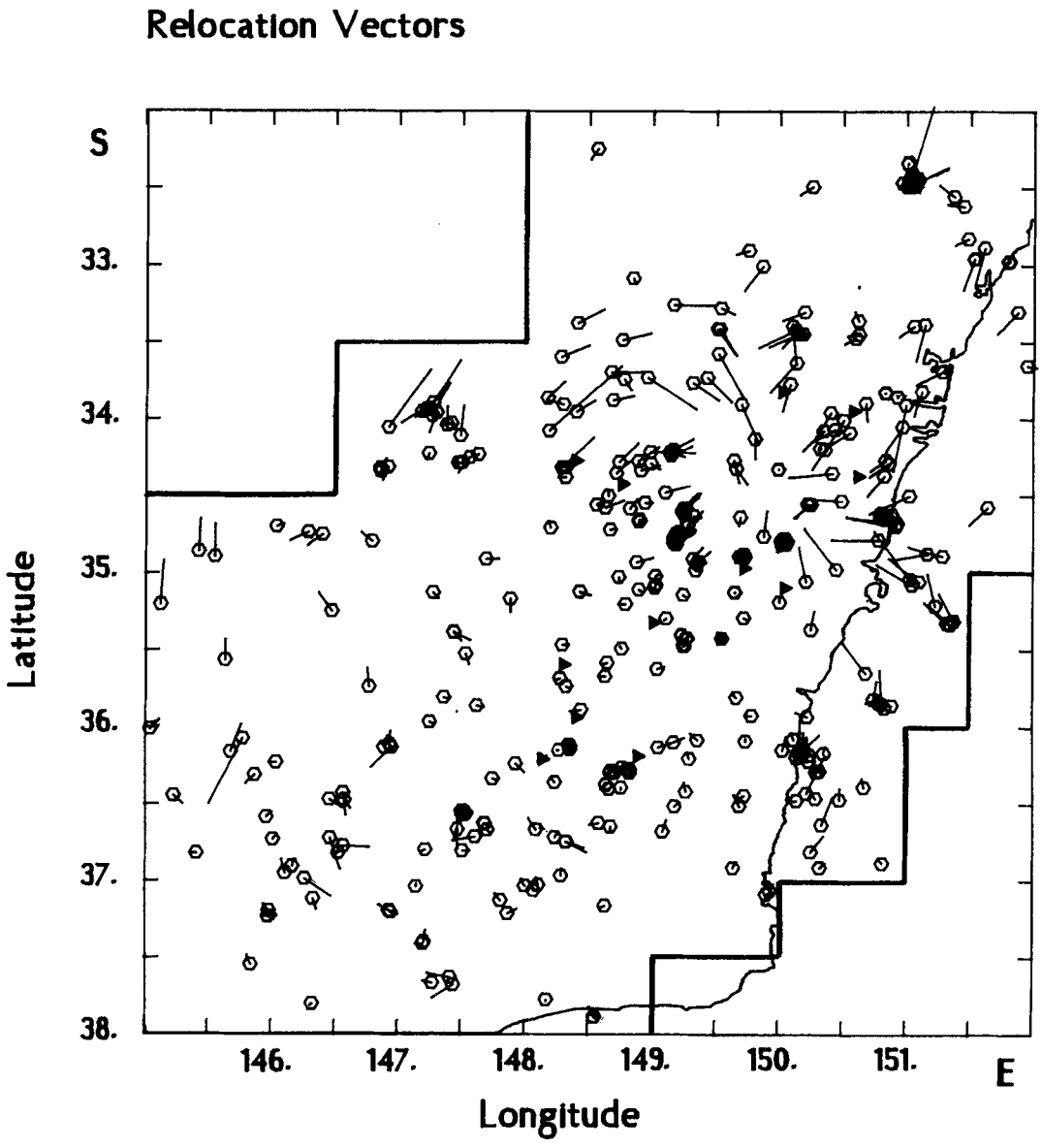
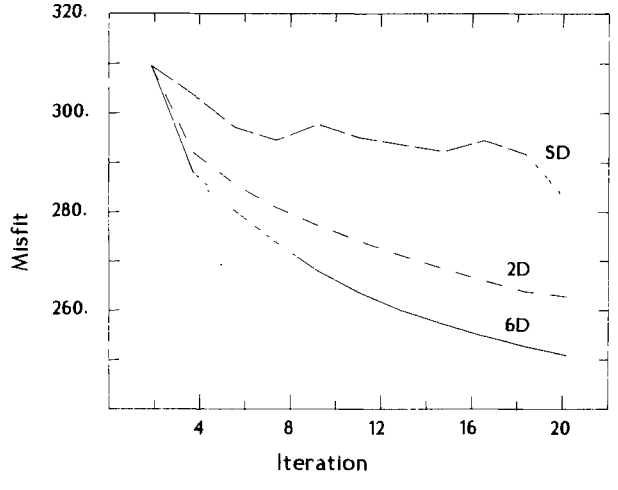


Figure 2. Distribution of earthquakes and recording stations used in the 3-D inversion for the southeast Australian region. The stations are shown by solid triangles and the earthquakes by open hexagons. The relocation vectors for the epicentres, determined from the 6-D subspace inversion, are indicated with a exaggerated scaling (× 15) in order to enhance the visibility of the smaller shifts.



higher dimensional schemes choosing a more optimal step than those of lower dimensionality.

Figure 4(a) shows the crustal *S*-wave velocity distribution recovered from an inversion with the two-dimensional subspace scheme. Outside the region containing the most significant anomaly to the NE, the inferred velocities are only slightly perturbed from their starting values (indicated

Figure 3. Comparison of convergence of different methods for solving the simultaneous hypocentre and velocity estimation problem. A simple descents scheme (SD) with no distinction between parameter types gives very slow convergence, whereas partitioning into a 2-D subspace scheme (2-D) with separation of the hypocentres and structural information results in much faster convergence. Further partitioning to separate each parameter class and so generate a 6-D subspace scheme (6-D) improves the rate of convergence even further.

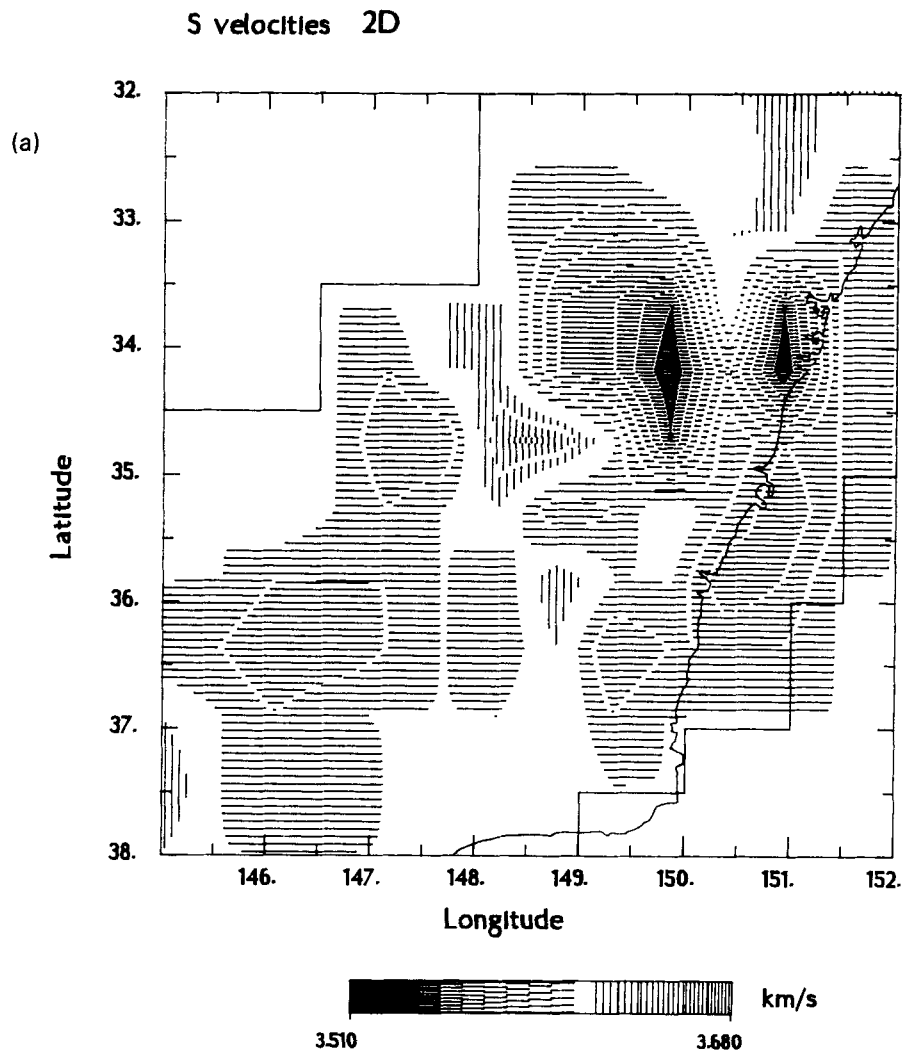


Figure 4. (a) Crustal S velocity pattern determined by inversion using a 2-D subspace scheme. Because the partial derivatives for S are downweighted relative to P there is a tendency for very little movement away from the reference value 3.61 km s^{-1} to occur. (b) Crustal S velocity pattern determined by inversion using a 6-D subspace scheme. Now with each parameter type allowed to vary independently the S-wave variations are no longer downweighted by the dominant P information and more complex velocity structure is inferred with a much better fit of the travel times to the original data.

by the white background). Indeed one can observe the cellular nature of the inversion in the contoured plot. The reason for this behaviour is that in the 2-D scheme we have lumped together all structural information into the same class and the contributions of the S-velocity parameters to the structural partition of the gradient vector are swamped by the much larger contributions from the crustal and mantle P velocities. The problem is avoided in the 6-D scheme, since the adjustments of the S-wave parameters are now independent of the relative sizes of the P- and S-wave gradient partitions. Fig. 4(b) shows the resulting crustal S-velocity map for the 6-D scheme, the size of the variations is much larger than before and more detail has been recovered in an inversion which gives a significantly better fit to the original data.

A similar downweighting effect was found to occur with the interface parameters and the epicentral coordinates of the earthquakes. In the 2-D subspace scheme, the origin times dominate the hypocentral shift to such an extent that no appreciable spatial movement was observed. Similarly the interface parameters were dominated by P-wave velocities and after inversion did not move far from their initial values. Only by adjusting the relative sizes of the entries in the model covariance matrix corresponding to different parameter types can the situation be improved within the 2-D subspace scheme. However, moving to the 6-D subspace scheme removes the problems and allows variation to occur for each of the different parameter types. The resulting epicentral shifts are illustrated in Fig. 2.

3.2 Seismic reflection tomography

Whereas the standard tomographic techniques used in seismic work are based on transmission problems (Nolet 1987a), in seismic reflection work it is necessary to use the times of arrival of reflected wave packets to infer the nature of the subsurface. In particular the nature of the near surface zone has a substantial effect on the character of the seismic records returned from depth. The reflection tomography problem is therefore to try to reconstruct the

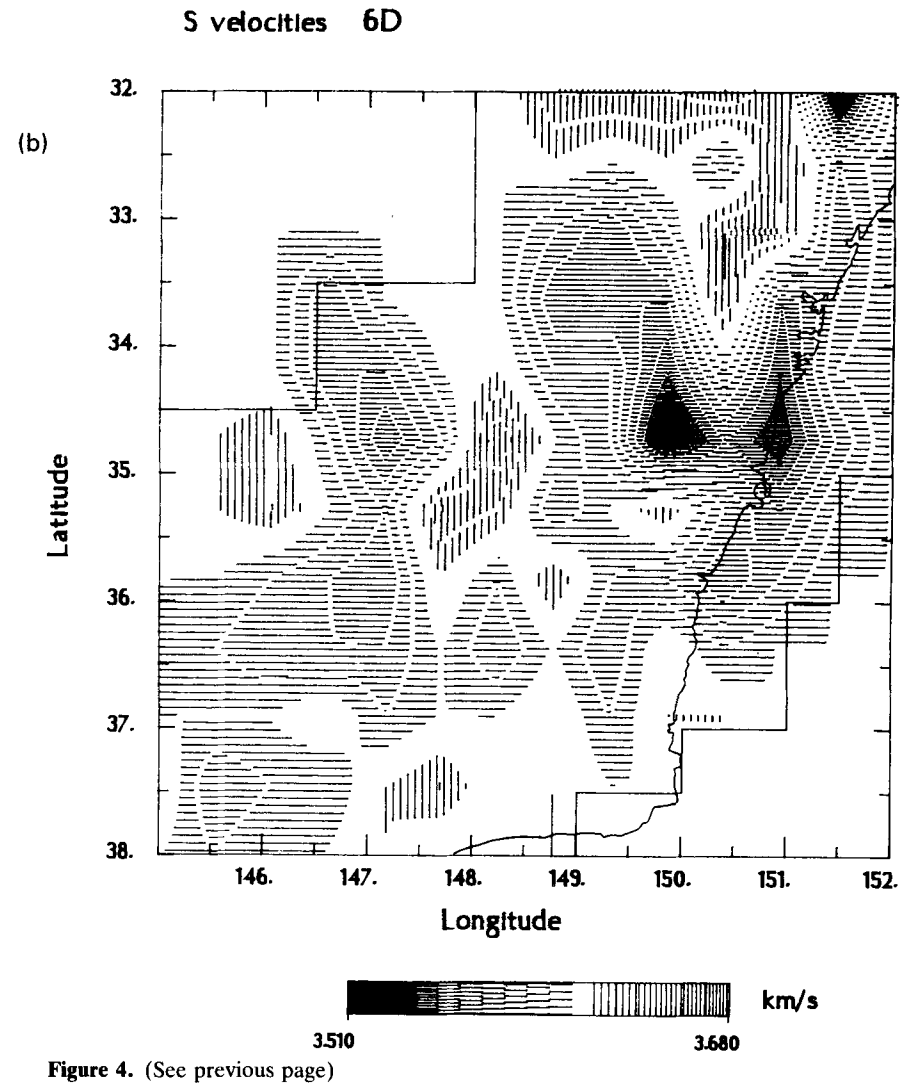


Figure 4. (See previous page)

shape of the first major reflector and the velocity field above the reflector. The seismic parameters required from the inversion therefore divide into two classes, (i) positional parameters describing the shape of the reflecting interface and (ii) velocity parameters for a cellular (or similar) partition of the velocity field.

It is very difficult in this highly nonlinear case to get satisfactory results using steepest descent methods (see Fig. 5). However, a subspace approach using the ascent directions for the two classes of parameters, supplemented by four vectors representing the rate of change of those vectors, can help to give good results with far superior convergence (Williamson 1986). This 6-D subspace approach requires more work per iteration than a simple descent scheme but the improvement in the data fit per step is much greater.

The nature of this nonlinear problem is such that it is easy for the minimization of the function F to be waylaid by the existence of local minima with rather different character to the true solution. Such apparent solutions may well be regarded as within acceptable levels of data when inverting for noisy data. These variant models are associated with a strong trade-off between the velocity close to the reflector and its position. Decreasing the velocity above the reflector or making it deeper will have similar effects on the reflected wave travel times, so that there can be a tradeoff between these two types of parameters which will depend on the starting model. This effect is illustrated in Fig. 5, where it has proved possible to achieve quite a good fit to the travel time data with a 6-D scheme but without a adequate recovery of the original model.

A partial cure can be provided for the local problem by modifying the scale of parametrization as the iteration proceeds (Williamson 1986). Initially a relatively coarse level of parameterization is employed and as the data fit improves a finer parameterization is introduced. Fig. 5 shows a successful application of this approach. However, even with this variable parametrization multiple inversions from different starting models may be needed to explore the character of acceptable solutions.

3.3 Inversion of seismic waveforms

A further class of problems in which there is a dependance on a number of different types of parameters arises in the inversion of seismic waveforms (Tarantola 1986; Nolet 1987b). For a complex isotropic region the complete waveforms recorded at a number of discrete receivers will

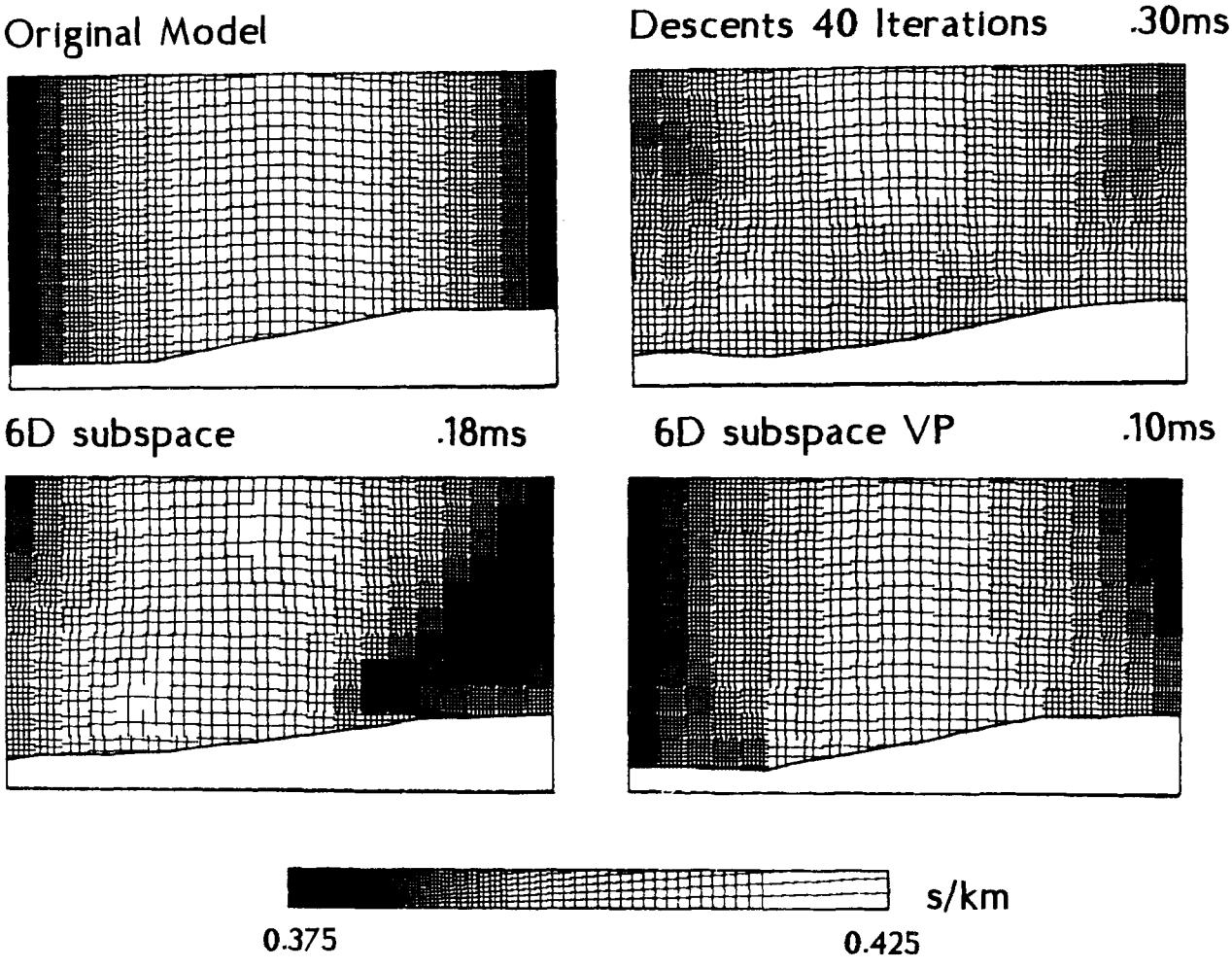


Figure 5. Comparisons of different inversions for the reflection tomography problem with unknown slowness field and reflector shape. The descents inversion uses rescaling of the model parameters by reference values and is only moderately successful. An application of the 6-D subspace approach with a fine parameterization gives a good fit to the data, but not a good recovery of the original model ('a local minimum'). The 6-D subspace method combined with variable parameterization, indicated by VP, gives very good recovery of the original model. The figures indicate the r.m.s. misfit in milliseconds between the original travel times and those computed for the postulated models.

depend on the *P*- and S-wave velocity and density distributions. In addition, the actual source character is often unknown so that a full inversion will require the estimation of four different aspects of the total model $\mathbf{m}$. We can write therefore

$$\mathbf{m} = [\mathbf{v}_\alpha, \mathbf{v}_\beta, \mathbf{v}_\rho, \mathbf{f}], \tag{3.5}$$

where the partitions $\mathbf{v}_\alpha$, $\mathbf{v}_\beta$, $\mathbf{v}_\rho$ represent the parameters describing the P, S and density fields and $\mathbf{f}$ the source parameters. This would then establish a 4-D subspace within which to set up the inversion procedure.

Tarantola (1986) has shown how the gradient terms can be evaluated by cross-correlating, at each point, the wavefield predicted for the current model with the back projections of the discrepancy between the observed and calculated waveforms at the receivers. For the case of reflection seismograms in a 2-D model, Tarantola also advocates working with *P*- and *S*-wave impedance rather than velocity, in order to try to improve the independence of the different sets of parameters. Such independence is essential for the sequential inversion scheme proposed by Tarantola (1986). For the waveform problem, a convenient representation of the field parameters is as continuous distributions and Sambridge, Tarantola & Kennett (1988) describe how the subspace method can be adapted to deal with this case.

4 DISCUSSION

The subspace method provides an algorithm for nonlinear inverse problems, with many parameter types, that can take into account the different functional dependencies in an equitable way. By associating at least one basis vector of the subspace with each parameter type, the solution for the update to the current model is produced in a way that does not depend on the scaling of the different parameter classes and thus the choice of covariance matrix relating model and dual space.

The mode of solution can be regarded as a cross between a gradient and a matrix approach. The model perturbation is built up from the local gradients of the nonlinear misfit functional with respect to each parameter type by a least-squares treatment in a subspace of small dimension. The dimensionality will typically be of the order of the

number of parameter types, although for two or three different classes, it may be worth bringing in additional information associated with the rates of change of the gradients. All the information needed is generated locally, and so is not dependent on bringing forward information from former models. As a result, the subspace approach is effective in many nonlinear problems. However, as for all nonlinear minimization routines, it cannot be guaranteed that the global minimum of the measure of data misfit can be found.

Throughout this paper we have illustrated the action of the subspace method with quadratic representations for the data misfit and regularization terms. Both of these forms are appropriate to the assumption of Gaussian statistics. When detailed information on the character of the error statistics are known the appropriate probability distributions should be employed in the construction of the data misfit or regularization terms, and the subspace method can be readily adapted to these new definitions.

REFERENCES

Chapman, C. H. & Orcutt, J. A. (1985). Least-squares fitting of marine seismic refraction data. *Geophys. J. R. astr. Soc.*, **82,** 339–374.

Kennett, B. L. N. (1988). Seismic velocity field estimation — strategies for a large-scale nonlinear inverse problem, *Exploration Geophys.*, **19,** 297–298.

Kennett, B. L. N. & Williamson, P. R. (1987). Subspace methods for large scale nonlinear inversion, in *Mathematical Geophysics: a survey of recent developments in seismology and geodynamics*, ed. Vlaar, N. J., Nolet, G., Wortel, M. J. R. & Cloetingh, S. A. P. L., D. Reidel, Dordrecht.

Mora, P. (1987) Elastic wavefield inversion, *PhD Thesis*, Stanford University.

Nolet, G. (1987a) Seismic wave propagation and seismic tomography, in *Seismic Tomography*, pp. 1–24, ed. Nolet, G., D. Reidel, Dordrecht.

Nolet, G. (1987b). Waveform tomography, in *Seismic Tomography*, pp. 301–322, ed. Nolet, G., D. Reidel, Dordrecht.

Pavlis, G. L. & Booker, J. R. (1980). The mixed discrete-continuous inverse problem: application to the simultaneous determination of earthquakes and velocity structure, *J. geophys. Res.*, **85,** 4801–4810.

Spencer, C. & Gubbins, G. (1980). Travel-time inversion for simultaneous earthquake location and velocity structure determination in laterally varying media, *Geophys. J. R. astr. Soc.*, **63,** 95–116.

Sambridge, M. S., Tarantola, A., & Kennett, B. L. N. (1988). An alternative strategy for the nonlinear inversion of seismic waveforms, *Geophys. Prospect.*, in press.

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, Amsterdam.

Williamson (1986) Tomographic inversion of travel time data in reflection seismology, *PhD Thesis,* University of Cambridge.