INVERSE PROBLEMS IN GEOPHYSICS

ROEL SNIEDER AND JEANNOT TRAMPERT
Dept. of Geophysics
Utrecht University
P.O. Box 80.021
3508 TA Utrecht
The Netherlands
email rsnieder@mines.edu

1. Introduction

An important aspect of the physical sciences is to make inferences about physical parameters from data. In general, the laws of physics provide the means for computing the data values given a model. This is called the "forward problem", see figure 1. In the inverse problem, the aim is to reconstruct the model from a set of measurements. In the ideal case, an exact theory exists that prescribes how the data should be transformed in order to reproduce the model. For some selected examples such a theory exists assuming that the required infinite and noise-free data sets would be available. A quantum mechanical potential in one spatial dimension can be reconstructed when the reflection coefficient is known for all energies [*Marchenko*, 1955; *Burridge,* 1980]. This technique can be generalized for the reconstruction of a quantum mechanical potential in three dimensions [*Newton,* 1989], but in that case a redundant data set is required for reasons that are not well understood. The mass-density in a one-dimensional string can be constructed from the measurements of all eigenfrequencies of that string [*Borg,* 1946], but due to the symmetry of this problem only the even part of the mass-density can be determined. If the seismic velocity in the earth depends only on depth, the velocity can be constructed exactly from the measurement of the arrival time as a function of distance

[0]Present address of R. Snieder: Dept. of Geophysics, Colorado School of Mines, Golden CO 80401, USA

[0]Reprinted from "Wavefield Inversion", Ed. *A. Wirgin*, Springer Verlag, New York, p. 119-190, 1999.

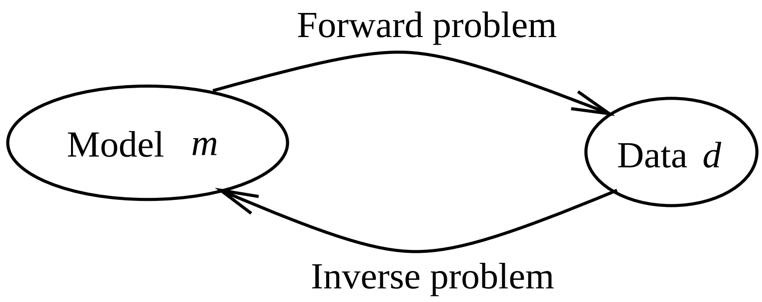

Figure 1. The traditional definition of the forward and inverse problems.

of seismic waves using an Abel transform [*Herglotz,* 1907; *Wiechert,* 1907]. Mathematically this problem is identical to the construction of a spherically symmetric quantum mechanical potential in three dimensions [*Keller et al.,* 1956]. However, the construction method of Herglotz-Wiechert only gives an unique result when the velocity increases monotonically with depth [*Gerver and Markushevitch,* 1966]. This situation is similar in quantum mechanics where a radially symmetric potential can only be constructed uniquely when the potential does not have local minima [*Sabatier*, 1973].

Despite the mathematical elegance of the exact nonlinear inversion schemes, they are of limited applicability. There are a number of reasons for this. First, the exact inversion techniques are usually only applicable for idealistic situations that may not hold in practice. For example, the Herglotz-Wiechert inversion presupposes that the velocity in the earth depends only on depth and that the velocity increases monotonically with depth. Seismic tomography has shown that both requirements are not met in the earth's mantle [*Nolet et al.,* 1994]. Second, the exact inversion techniques often are very unstable. The presence of this instability in the solution of the Marchenko equation has been shown explicitly by *Dorren et al.* [1994]. However, the third reason is the most fundamental. In many inverse problems the model that one aims to determine is a continuous function of the space variables. This means that the model has infinitely many degrees of freedom. However, in a realistic experiment the amount of data that can be used for the determination of the model is usually finite. A simple count of variables shows that the data cannot carry sufficient information to determine the model uniquely. In the context of linear inverse problems this point has been raised by *Backus and Gilbert* [1967, 1968] and more recently by *Parker* [1994]. This issue is equally relevant for nonlinear inverse problems.

The fact that in realistic experiments a finite amount of data is available to reconstruct a model with infinitely many degrees of freedom necessarily means that the inverse problem is not unique in the sense that there are many models that explain the data equally well. The model obtained from the inversion of the data is therefore not necessarily equal to the true model that one seeks. This implies that the view of inverse problems as shown in figure 1 is too simplistic. For realistic problems, inversion really consists of two steps. Let the true model be denoted by m and the data by d. From the data d one reconstructs an estimated model $\tilde{m}$, this is called the *estimation problem*, see figure 2. Apart from estimating a model $\tilde{m}$ that is consistent with the data, one also needs to investigate what relation the estimated model $\tilde{m}$ bears to the true model m. In the *appraisal problem* one determines what properties of the true model are recovered by the estimated model and what errors are attached to it. The essence of this discussion is that *inversion* = *estimation* + *appraisal*. It does not make much sense to make a physical interpretation of a model without acknowledging the fact of errors and limited resolution in the model [*Trampert*, 1998].

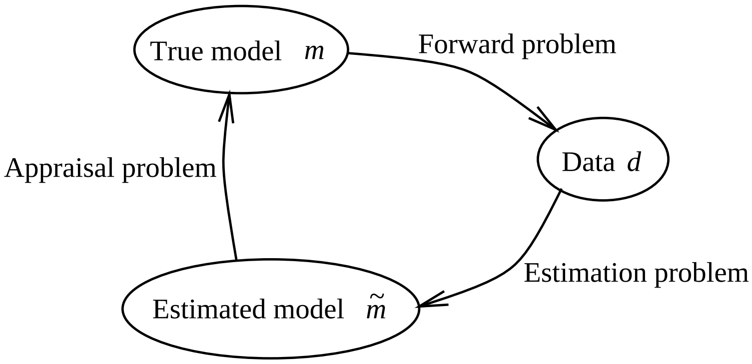

Figure 2. The inverse problem viewed as a combination of an estimation problem plus an appraisal problem.

In general there are two reasons why the estimated model differs from the true model. The first reason is the non-uniqueness of the inverse problem that causes several (usually infinitely many) models to fit the data. Technically, this model null-space exits due to inadequate sampling of the model space. The second reason is that real data (and physical theories more often than we would like) are always contaminated with errors and the estimated model is therefore affected by these errors as well. Therefore model appraisal has two aspects, non-uniqueness and error propagation.

Model estimation and model appraisal are fundamentally different for discrete models with a finite number of degrees of freedom and for continuous models with infinitely many degrees of freedom. Also, the problem of model appraisal is only well-solved for linear inverse problems. For this reason the inversion of discrete models and continuous models is treated separately, and the case of linear inversion and nonlinear inversion is also treated independently. In section 2 linear inversion for a finite number of model parameters is discussed. This is generalized in section 3 to deal with linear inverse problems for continuous models with infinitely many degrees of freedom. In reality many inverse problems are not really linear, but often these problems can be linearized by making a suitable approximation. In section 4 the single-scattering approximation is derived. This technique forms the basis of imaging tools used in reflection seismology. Rayleigh's principle, as treated in section 5, is the linearization that forms the basis for the inversion for the Earth's structure using normal-mode frequencies. The linearization technique of seismic travel time tomography is based on Fermat's principle, which is treated in section 6. Nonlinear inverse problems are significantly more difficult than linear inverse problems. It is shown in section 7 that non-linearity can be a source of ill-posedness. Presently, there is no satisfactory theory for the appraisal problem for nonlinear inverse problems. In section 8 three methods are presented that can be used for the nonlinear appraisal problem. However, neither of these methods is quite satisfactory, which indicates that nonlinear inverse problem theory is a field with important research challenges.

2. Solving finite linear systems of equations

As argued in the previous section, the inverse problem maps a finite number of data onto a model. In most practical applications in geophysics the model is a continuous function of the space coordinates and therefore has infinitely many degrees of freedom. For the moment we will ignore this and will assume that the model can be characterized by a finite number of parameters. We will return to the important case of models that are infinitely dimensional in section 3.

2.1. LINEAR MODEL ESTIMATION

For a finite-dimensional model, the model parameters can be ordered in a vector $\boldsymbol{m}$, and similarly the data can be ordered in a vector $\boldsymbol{d}$. The matrix $\boldsymbol{A}$ relates the data to the model through the product $\boldsymbol{Am}$. This matrix is often referred to as the theory operator. Indeed, it contains all the information on physics and mathematics we have chosen to model in the given problem. In practice, the data are contaminated with errors $\boldsymbol{e}$, so that the recorded data and the model are related by:

$$\boldsymbol{d} = \boldsymbol{Am} + \boldsymbol{e} \tag{1}$$

It should be noted there often there is an certain arbitrariness in the choice of the model parameters that are contained in the model vector $\boldsymbol{m}$. For example, if one wants to describe the density in the earth one could choose a model where the Earth's mantle and the core have a uniform density, in that case there are two model parameters. Alternatively, one could expand the density in the Earth in a large amount of eigenfunctions defined on the sphere such as spherical harmonics for lateral variations and polynomials for depth variations, in that case there are much more model parameters. These two different parameterizations of the same model correspond to different model parameters $\boldsymbol{m}$ and to a different matrix $\boldsymbol{A}$. This example illustrates that the model $\boldsymbol{m}$ is not necessarily the true model,[1] but that the choice of the model parameters usually contains a restriction on the class of models that can be constructed. Below we will refer to $\boldsymbol{m}$ as the true model regardless of the difficulties in its definition.

From the recorded data one makes an estimate of the model. Since this estimate in practice will be different from the true model the estimated model is denoted by $\tilde{\boldsymbol{m}}$. There are many ways for designing an inverse operator that maps the data on the estimated model [e.g. *Menke*, 1984; *Tarantola*, 1987; *Parker*, 1994]. Whatever estimator one may choose, the most general linear mapping from data to the estimated model can be written as:

$$\tilde{\boldsymbol{m}} = \boldsymbol{A}^{-g}\boldsymbol{d} \tag{2}$$

The operator $\boldsymbol{A}^{-g}$ is called the *generalized inverse* of the matrix $\boldsymbol{A}$. In general, the number of data is different from the number of model parameters. For this reason $\boldsymbol{A}$ is usually a non-square matrix, and hence its formal inverse does not exist. Later we will show how the generalized inverse $\boldsymbol{A}^{-g}$ may be chosen, but for the moment $\boldsymbol{A}^{-g}$ does not need to be specified. The relation between the estimated model $\tilde{\boldsymbol{m}}$ and the true model $\boldsymbol{m}$ follows by inserting (1) in expression (2):

$$\tilde{\boldsymbol{m}} = \boldsymbol{A}^{-g}\boldsymbol{A}\boldsymbol{m} + \boldsymbol{A}^{-g}\boldsymbol{e} \tag{3}$$

The matrix $\boldsymbol{A}^{-g}\boldsymbol{A}$ is called the *resolution kernel*, this operator is given by:

$$\boldsymbol{R} \equiv \boldsymbol{A}^{-g}\boldsymbol{A} \tag{4}$$

[1] We urge the reader to formulate a definition of the concept "true model." It is not so difficult to formulate a vague definition such as "the true model is the model that corresponds to reality and which is only known to the gods." However, we are not aware of any definition that is operational in the sense that it provides us with a set of actions that could potentially tell us what the true model really is.

Expression (3) can be interpreted by rewriting it in the following form:

$$\tilde{\boldsymbol{m}} = \boldsymbol{m} + \underbrace{\left(\boldsymbol{A}^{-g}\boldsymbol{A} - \boldsymbol{I}\right)\boldsymbol{m}}_{\text{Limited Resolution}} + \underbrace{\boldsymbol{A}^{-g}\boldsymbol{e}}_{\text{Error propagation}} \tag{5}$$

In the ideal case, the estimated model equals the true model vector: $\tilde{\boldsymbol{m}} = \boldsymbol{m}$ meaning that our chosen parameters, ordered in vector $\boldsymbol{m}$, may be estimated independently from each other. The last two terms in equation (5) account for *blurring* and *artifacts* in the estimated model. The term $\left(\boldsymbol{A}^{-g}\boldsymbol{A} - \boldsymbol{I}\right)\boldsymbol{m}$ describes the fact that components of the estimated model vector are linear combinations of different components of the true model vector. We only retrieve averages of our parameters and "blurring" occurs in the model estimation as we are not able to map out the finest details. In the ideal case this term vanishes; this happens when $\boldsymbol{A}^{-g}\boldsymbol{A}$ is equal to the identity matrix. With (4) this means that for perfectly resolved model parameters the resolution matrix is the identity matrix:

$$\text{Perfect resolution: } \boldsymbol{R} = \boldsymbol{I} \tag{6}$$

As noted earlier, usually there is a certain ambiguity in the definition of the model parameters that define the vector $\boldsymbol{m}$. The resolution operator tells us to what extend we can retrieve the model parameters independently from the estimation process. However, the resolution matrix does not tell us completely what the relation between the estimated model and the real underlying physical model is, because it does not take into account to what extent the choice of the model parameters has restricted the model that can be obtained from the estimation process.

The last term in (5) describes how the errors $\boldsymbol{e}$ are mapped onto the estimated model.[2] These errors are not known deterministically, otherwise they could be subtracted from the data. A statistical analysis is needed to describe the errors in the estimated model due to the errors in the data. When the data d_j are uncorrelated and have standard deviation σ_{d_j}, the standard deviation σ_{m_i} in the model estimate $\tilde{m}_i$, resulting from the propagation of data errors only, is given by:

$$\sigma_{m_i}^2 = \sum_j \left(A_{ij}^{-g}\sigma_{d_j}\right)^2 \tag{7}$$

Ideally, one would like to obtain both: a perfect resolution and no errors in the estimated model. Unfortunately this cannot be achieved in practice. The error propagation is, for instance, completely suppressed by using the

[2]As shown by *Scales and Snieder* [1998] the concept of errors in inverse problems is not as simple as it appears.

generalized inverse $\boldsymbol{A}^{-g} = 0$. This leads to the (absurd) estimated model $\tilde{\boldsymbol{m}} = 0$ which is indeed not affected by errors. However, this particular generalized inverse has a resolution matrix given by $\boldsymbol{R} = 0$, which is far from the ideal resolution matrix given in (6). Hence in practice, one has to find an acceptable trade-off between error-propagation and limitations in the resolution.

2.2. LEAST-SQUARES ESTIMATION

Let us for the moment consider the case where the number of independent data is larger than the number of unknowns. In that case, the system $\boldsymbol{d} = \boldsymbol{A}\boldsymbol{m}$ cannot always be satisfied for any given model $\boldsymbol{m}$ because of possible errors contained in the data vector making the equations inconsistent. As an example, let us consider the following problem. We have two masses with weight m_1 and m_2. The weighing of the first mass yields a weight of 1 (kilo). Then one measures the second mass to find a weight of 2. Next, one weighs the masses together to find a combined weight of 2. The system of equations that corresponds to these measurements is given by:

$$\begin{aligned} m_1 &= d_1 = 1 \\ m_2 &= d_2 = 2 \\ m_1 + m_2 &= d_3 = 2 \end{aligned} \tag{8}$$

The matrix $\boldsymbol{A}$ for this problem is given by:

$$\boldsymbol{A} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix} \tag{9}$$

It is clear that this system of equations cannot be satisfied. It cannot be true that the first mass has a weight $m_1 = 1$, and the second mass has a weight $m_2 = 2$ while the combined mass is equal to $m_1 + m_2 = 2$. Clearly errors have been made during the measurements, but there is no reason to discard one of the three equations in favor of the other two. This problem is illustrated graphically in figure 3. The three equations (8) correspond to the three solid lines in the (m_1, m_2)-plane. The fact that the three lines do not intersect in a single point signifies that the linear equations are inconsistent. The inverse problem of determining the two masses thus consists in reconciling these equations in a meaningful way.

A common way to estimate a model is to seek the model $\tilde{\boldsymbol{m}}$ that gives the best fit to the data in the sense that the difference, measured by the L_2-norm, between the data vector $\boldsymbol{d}$ and the recalculated data $\boldsymbol{A}\tilde{\boldsymbol{m}}$ is made as small as possible. This means that the least-squares solution is given by the model that minimizes the following cost function:

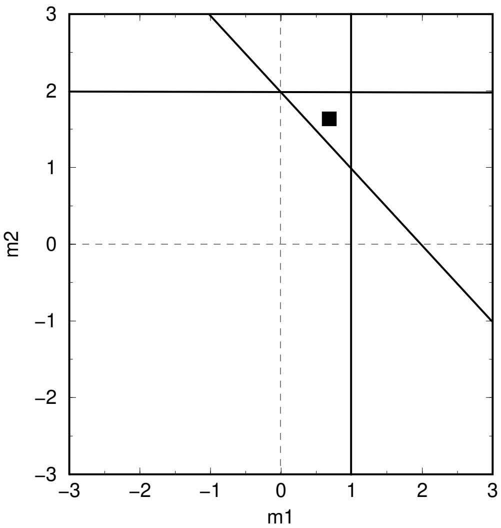

Figure 3. Geometrical interpretation of the linear equations (8).

$$S = \|\boldsymbol{d} - \boldsymbol{A}\boldsymbol{m}\|^2 \tag{10}$$

As shown in detail by *Strang* [1988] this quantity is minimized by the following model estimate:

$$\tilde{\boldsymbol{m}} = \left(\boldsymbol{A}^T\boldsymbol{A}\right)^{-1}\boldsymbol{A}^T\boldsymbol{d} \tag{11}$$

In the example of figure 3 the least-squares solution is the point in the (m_1, m_2)-plane that has the smallest distance to the three lines in that figure, this point is indicated by a black square. Using the matrix (9) one readily finds that the least-squares estimator of the problem (8) is given by:

$$\tilde{\boldsymbol{m}} = \left(\boldsymbol{A}^T\boldsymbol{A}\right)^{-1}\boldsymbol{A}^T\boldsymbol{d} = \frac{1}{3}\begin{pmatrix} 2 & -1 & 1 \\ -1 & 2 & 1 \end{pmatrix}\boldsymbol{d} \tag{12}$$

For the used data vector this means that the estimated model is given by:

$$\begin{aligned} \tilde{m}_1 &= 2/3 \\ \tilde{m}_2 &= 5/3 \end{aligned} \tag{13}$$

2.3. MINIMUM NORM ESTIMATION

In some problems the number of unknowns is less than the number of parameters. Consider for example the situation where there are two masses

m_1 and m_2 and one has measured only the combined weight of these masses:

$$m_1 + m_2 = d = 2 \tag{14}$$

The matrix that corresponds to this system of one equation is given by:

$$\boldsymbol{A} = \begin{pmatrix} 1 & 1 \end{pmatrix} \tag{15}$$

Graphically this problem is depicted in figure 4. Clearly any model vector lying on the solid line fits the equation (14) exactly. There are thus infinitely many solutions, provided the masses are positive, that exactly fit the data. A model estimate can be defined by choosing a model that fits the data exactly and that has the smallest L_2-norm, this model is indicated by in figure 4 by the black square.

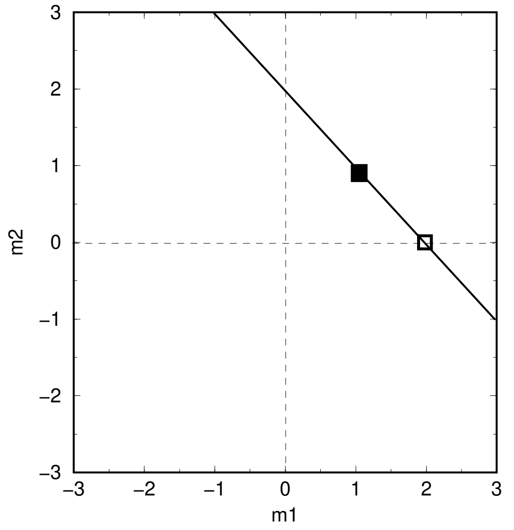

Figure 4. Geometrical interpretation of the linear equation (14) with two unknowns.

For a general underdetermined system of equations the minimum norm solution is defined as the model that fits the data exactly, $\boldsymbol{Am} = \boldsymbol{d}$, and that minimizes $\|\boldsymbol{m}\|^2$. Using Lagrange multipliers one can show that the minimum-norm solution is given by:

$$\tilde{\boldsymbol{m}} = \boldsymbol{A^T} \left(\boldsymbol{A} \boldsymbol{A}^T \right)^{-1} \boldsymbol{d} \ , \tag{16}$$

a detailed derivation is given by *Menke* [1984]. One readily finds that the minimum norm solution of “system” (14) is given by

$$m_1 = m_2 = 1 \ . \tag{17}$$

In the least-squares estimation, we assumed that we had enough information to evaluate all model parameters, even though contradictions occurred due to measurement errors. The problem is then purely over-determined and as a consequence $\boldsymbol{A}^T\boldsymbol{A}$ is regular. In the minimum norm solution, we assumed no contradictions in the available information, but we don't have enough equations to evaluate all model parameters. This is the case of a purely under-determined problem and here $\boldsymbol{A}\boldsymbol{A}^T$ is regular. The most common case, however, is that we have contradictory information on some model parameters, while others cannot be assessed due to a lack of information. Then neither $\boldsymbol{A}^T\boldsymbol{A}$ nor $\boldsymbol{A}\boldsymbol{A}^T$ can be inverted and the problem is ill-posed. Even if the inverse matrices formally exist, they are often ill-conditioned meaning that small changes in the data vector lead to large changes in the model estimation. This means that errors in the data will be magnified in the model estimation. Clearly a trick is needed to find a model that is not too sensitive on small changes in the data. To this effect, *Levenberg* [1944] introduced a damped least-squares solution. From a mathematical point of view, ill-posedness and ill-conditioning result from zero or close to zero singular values of $\boldsymbol{A}$.

Suppose one has a matrix $\boldsymbol{M}$ with eigenvalues λ_n and eigenvectors $\hat{\boldsymbol{v}}_n$:

$$\boldsymbol{M}\hat{\boldsymbol{v}}_n = \lambda_n \hat{\boldsymbol{v}}_n \tag{18}$$

One readily finds that the matrix $(\boldsymbol{M}+\gamma\boldsymbol{I})$ has eigenvalues $(\lambda_n + \gamma)$:

$$(\boldsymbol{M}+\gamma\boldsymbol{I})\,\hat{\boldsymbol{v}}_n = (\lambda_n + \gamma)\,\hat{\boldsymbol{v}}_n \tag{19}$$

This means that the eigenvalues of a matrix can be raised by adding the scaled identity matrix to the original matrix. This property can be used to define the *damped least-squares solution*:

$$\tilde{\boldsymbol{m}} = \left(\boldsymbol{A}^T\boldsymbol{A}+\gamma\boldsymbol{I}\right)^{-1}\boldsymbol{A}^T\boldsymbol{d} \tag{20}$$

Since the matrix $\boldsymbol{A}^T\boldsymbol{A}$ has positive eigenvalues[3] its eigenvalues are moved away from zero when the constant γ is positive. Alternatively, the solution (20) can be found by minimizing the following cost function:

[3]That the eigenvalues of $\boldsymbol{A}^T\boldsymbol{A}$ are positive follows from the following identity: $\left(\boldsymbol{x}\cdot\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x}\right) = \left(\boldsymbol{A}^{TT}\boldsymbol{x}\cdot\boldsymbol{A}\boldsymbol{x}\right) = (\boldsymbol{A}\boldsymbol{x}\cdot\boldsymbol{A}\boldsymbol{x}) = \|\boldsymbol{A}\boldsymbol{x}\|^2 \geq 0$. When $\boldsymbol{x}$ is the eigenvector $\hat{\boldsymbol{u}}^{(n)}$ of $\boldsymbol{A}^T\boldsymbol{A}$ with eigenvalue μ_n, this expression can be used to show that $\mu_n \left\|\hat{\boldsymbol{u}}^{(n)}\right\|^2 = \mu_n\left(\hat{\boldsymbol{u}}^{(\boldsymbol{n})}\cdot\hat{\boldsymbol{u}}^{(n)}\right) = \left(\hat{\boldsymbol{u}}^{(\boldsymbol{n})}\cdot\boldsymbol{A}^T\boldsymbol{A}\hat{\boldsymbol{u}}^{(n)}\right) \geq 0$, hence the eigenvalues μ_n are positive.

$$S = \|\boldsymbol{d} - \boldsymbol{A}\boldsymbol{m}\|^2 + \gamma \|\boldsymbol{m}\|^2 \tag{21}$$

This expression clearly shows what the effect of the damping is. Minimizing the first term of (21) amounts to finding the model that gives the best fit to the data. Minimizing the last term of (21) amounts to finding the model with the smallest norm. In general we cannot minimize both terms simultaneously, but in minimizing (21) we comprise in finding a model that both fits the data reasonably well and whose model size is not too large. The parameter γ controls the emphasis we put on these conflicting requirements and for this reason it is called the *trade-off parameter*.

For a number of applications the following matrix identity is extremely useful:[4]

$$\boxed{\left(\boldsymbol{A}^T\boldsymbol{B}^{-1}\boldsymbol{A} + \boldsymbol{D}^{-1}\right)^{-1}\boldsymbol{A}^T\boldsymbol{B}^{-1} = \boldsymbol{D}\boldsymbol{A}^T\left(\boldsymbol{B} + \boldsymbol{A}\boldsymbol{D}\boldsymbol{A}^T\right)^{-1}\ .} \tag{22}$$

In this expression $\boldsymbol{B}$ and $\boldsymbol{D}$ are regular square matrices, whereas $\boldsymbol{A}$ needs not to be square. This expression can be used to show that when damping or regularization is used, the least-squares solution and the minimum-norm solution (both supplied with a damping term) are identical. To see this use (22) with $\boldsymbol{B}^{-1} = \boldsymbol{I}$ and $\boldsymbol{D}^{-1} = \gamma\boldsymbol{I}$. It then follows that

$$\left(\boldsymbol{A}^T\boldsymbol{A} + \gamma\boldsymbol{I}\right)^{-1}\boldsymbol{A}^T\boldsymbol{d} = \frac{1}{\gamma}\boldsymbol{A}^T\left(\boldsymbol{I} + \boldsymbol{A}\frac{1}{\gamma}\boldsymbol{A}^T\right)^{-1}\boldsymbol{d} = \boldsymbol{A}^T\left(\boldsymbol{A}\boldsymbol{A}^T + \gamma\boldsymbol{I}\right)^{-1}\boldsymbol{d}\ . \tag{23}$$

The left hand side corresponds to the damped least-squares solution (20) while the right hand side is the damped version of the minimum-norm solution (16). This implies that when damping is applied the least-squares solution and the minimum-norm solution are identical.

2.5. THE CONSISTENCY PROBLEM FOR THE LEAST-SQUARES SOLUTION

The least-squares solution appears to provide an objective method for finding solutions of overdetermined problems. However, there is trouble ahead. To see this, let us consider the overdetermined system of equations (8).

[4]This identity follows from the identity $\boldsymbol{A}^T + \boldsymbol{A}^T\boldsymbol{B}^{-1}\boldsymbol{A}\boldsymbol{D}\boldsymbol{A}^T = \boldsymbol{A}^T\boldsymbol{B}^{-1}\boldsymbol{A}\boldsymbol{D}\boldsymbol{A}^T + \boldsymbol{A}^T$. Write the first term on the left hand side as $\boldsymbol{A}^T\boldsymbol{B}^{-1}\boldsymbol{B}$ and the last term on the right hand side as $\boldsymbol{D}^{-1}\boldsymbol{D}\boldsymbol{A}^T$. The resulting expression can then be written as $\boldsymbol{A}^T\boldsymbol{B}^{-1}\left(\boldsymbol{B} + \boldsymbol{A}\boldsymbol{D}\boldsymbol{A}^T\right) = \left(\boldsymbol{A}^T\boldsymbol{B}^{-1}\boldsymbol{A} + \boldsymbol{D}^{-1}\right)\boldsymbol{D}\boldsymbol{A}^T$. The expression (22) then follows by multiplying on the left with $\left(\boldsymbol{B} + \boldsymbol{A}\boldsymbol{D}\boldsymbol{A}^T\right)^{-1}$ and by multyplying on the right with $\left(\boldsymbol{A}^T\boldsymbol{B}^{-1}\boldsymbol{A} + \boldsymbol{D}^{-1}\right)^{-1}$.

Mathematically, this system of equations does not change when we multiply the last equation with a factor two. The following two systems of equations thus are completely equivalent:

$$\left.\begin{array}{rll} m_1 = & d_1 & = 1 \\ m_2 = & d_2 & = 2 \\ m_1 + m_2 = & d_3 & = 2 \end{array}\right\} \Leftrightarrow \left\{\begin{array}{rrrr} m_1 = & d_1 = & d_1^{'} = & 1 \\ m_2 = & d_2 = & d_2^{'} = & 2 \\ 2m_1 + 2m_2 = & 2d_3 = & d_3^{'} = & 4 \end{array}\right. \tag{24}$$

The matrices of the original system and the new equivalent system are given by

$$\boldsymbol{A} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix} \text{ and } \boldsymbol{A}^{'} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 2 & 2 \end{pmatrix} \tag{25}$$

In this section the unprimed quantities denote the original system of equations while the primed quantities refer to the transformed system of equations. One readily finds that the least-squares solution (11) of the original system and the transformed system are given by

$$\tilde{\boldsymbol{m}} = \frac{1}{3} \begin{pmatrix} 2 & -1 & 1 \\ -1 & 2 & 1 \end{pmatrix} \boldsymbol{d} \text{ and } \tilde{\boldsymbol{m}} = \frac{1}{9} \begin{pmatrix} 5 & -4 & 2 \\ -4 & 5 & 2 \end{pmatrix} \boldsymbol{d}^{'} , \tag{26}$$

Using the numerical values of the original data vector $\boldsymbol{d}$ and the transformed data vector $\boldsymbol{d}'$ this leads to the following estimates of the model:

$$\tilde{\boldsymbol{m}} = \begin{pmatrix} 2/3 \\ 5/3 \end{pmatrix} \text{ and } \tilde{\boldsymbol{m}} = \begin{pmatrix} 5/9 \\ 14/9 \end{pmatrix} \tag{27}$$

The problem is that these two estimators of the same model are *different*! This is surprising because the original system of equations and the transformed system of equations in (24) are mathematically equivalent. The reason that the two solutions are different is that the metric in the original data space and in the transformed data space has been changed by the transformation. This is a different way of saying that distances are measured in different ways in the least-squares criteria for solving the two systems of equations. Since the least-squares solution minimizes distances it makes sense that the least-squares solution changes when the metric (or measuring unit) of the data space is changed. This implies that the least-squares solution is not as objective as it appeared at first sight, because arbitrary transformations of the system of equations lead to different least-squares solutions!

For the least-squares solution the generalized inverse is given by $\boldsymbol{A}^{-g} = \left(\boldsymbol{A}^T\boldsymbol{A}\right)^{-1}\boldsymbol{A}^T$. One readily deduces that both for the original system and the transformed system the resolution kernel is equal to the identity matrix: $\boldsymbol{R} = \boldsymbol{A}^{-g}\boldsymbol{A} = \boldsymbol{I}$ and $\boldsymbol{R}^{'} = \boldsymbol{A}^{-g'}\boldsymbol{A}^{'} = \boldsymbol{I}$. Hence both systems have perfect resolution! The reader may be interested to pause and explain how this can be reconciled with the fact that the estimated models in (27) are different.

The reason for this discrepancy lies in the error propagation term $\boldsymbol{A}^{-g}\boldsymbol{e}$ in (5). We know that errors must be present in the data used in the systems defined in expression (24) because the equations are inconsistent. After scaling the equations, data and errors are reconciled in different ways in the two systems of equations, so that different model estimators are obtained. *It is thus the presence of inconsistencies in the system of equations caused by errors that creates a dependence of the least-squares solution to arbitrary scaling operations.*

Let us now consider the properties of the least-squares solution under transformations of the data vector and the model vector in a more general way. The initial system of equations is given by

$$\boldsymbol{Am} = \boldsymbol{d} \tag{28}$$

This expression is not quite correct because we ignored the errors $\boldsymbol{e}$ which will always be present. This is the reason why the above expression can not exactly be satisfied, and we will have to seek the least-squares solution to this system of equations. Let us consider a transformation of the model parameters through a transformation matrix $\boldsymbol{S}$:

$$\boldsymbol{m}^{'} = \boldsymbol{Sm} \ , \tag{29}$$

and a transformation of the data vector with a transformation matrix $\boldsymbol{Q}$:

$$\boldsymbol{d}^{'} = \boldsymbol{Qd} \ . \tag{30}$$

Assume that $\boldsymbol{S}$ has an inverse, the transformed system of equations then is given by

$$\boldsymbol{QAS}^{-1}\boldsymbol{m}^{'} = \boldsymbol{Qd} \ = \boldsymbol{d}'. \tag{31}$$

The original system of equations (28) has the least-squares solution

$$\boxed{\tilde{\boldsymbol{m}}^{(1)} = \left(\boldsymbol{A}^T\boldsymbol{A}\right)^{-1}\boldsymbol{A}^T\boldsymbol{d}} \tag{32}$$

The solution of the transformed system of equations (31) follows from the same expression, setting $\boldsymbol{A}^{'} = \boldsymbol{QAS}^{-1}$ and replacing $\boldsymbol{d}$ by $\boldsymbol{Qd}$. This, however, gives the solution to the transformed model vector $\boldsymbol{m}^{'}$. In order to compare this solution with the model estimate (32) we need to transform back to the original model space, using the relation $\boldsymbol{m} = \boldsymbol{S}^{-1}\boldsymbol{m}^{'}$. The least-squares solution $\tilde{\boldsymbol{m}}^{(2)}$ that follows from the transformed system is then given by:

$$\tilde{\boldsymbol{m}}^{(2)} = \boldsymbol{S}^{-1}\left(\boldsymbol{S}^{T-1}\boldsymbol{A}^T\boldsymbol{Q}^T\boldsymbol{Q}\boldsymbol{A}\boldsymbol{S}^{-1}\right)^{-1}\boldsymbol{S}^{T-1}\boldsymbol{A}^T\boldsymbol{Q}^T\boldsymbol{Q}\boldsymbol{d} \tag{33}$$

Assuming again that the appropriate inverses exist, this expression can be simplified by repeatedly applying the matrix identity $(\boldsymbol{N}\boldsymbol{M})^{-1} = \boldsymbol{M}^{-1}\boldsymbol{N}^{-1}$ to the term $\left(\boldsymbol{S}^{T-1}\boldsymbol{A}^T\boldsymbol{Q}^T\boldsymbol{Q}\boldsymbol{A}\boldsymbol{S}^{-1}\right)^{-1}$, giving $\left(\boldsymbol{S}^{T-1}\boldsymbol{A}^T\boldsymbol{Q}^T\boldsymbol{Q}\boldsymbol{A}\boldsymbol{S}^{-1}\right)^{-1} = \left(\boldsymbol{S}^{-1}\right)^{-1}\left(\boldsymbol{A}^T\boldsymbol{Q}^T\boldsymbol{Q}\boldsymbol{A}\right)^{-1}\left(\boldsymbol{S}^{T-1}\right)^{-1} = \boldsymbol{S}\left(\boldsymbol{A}^T\boldsymbol{Q}^T\boldsymbol{Q}\boldsymbol{A}\right)^{-1}\boldsymbol{S}^T$. The least-squares solution of the transformed system can then be written as:

$$\boxed{\tilde{\boldsymbol{m}}^{(2)} = \left(\boldsymbol{A}^T\boldsymbol{Q}^T\boldsymbol{Q}\boldsymbol{A}\right)^{-1}\boldsymbol{A}^T\boldsymbol{Q}^T\boldsymbol{Q}\boldsymbol{d}} \tag{34}$$

Comparing this with the least-squares solution $\tilde{\boldsymbol{m}}^{(1)}$ in expression (32) of the original system one finds that the least-squares solution is invariant for:

- Transformations of the model vector altogether.
- Transformations of the data vector if $\boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I}$.

The first property can be understood if we recall that the cost function in the least-squares problem does not minimize the model length but only the data misfit. The last property can be understood by comparing the quantities that are minimized in the original system and for the transformed system. For the original system one minimizes:

$$S = \|\boldsymbol{d} - \boldsymbol{A}\boldsymbol{m}\|^2 = \left((\boldsymbol{d} - \boldsymbol{A}\boldsymbol{m}) \cdot (\boldsymbol{d} - \boldsymbol{A}\boldsymbol{m})\right) \ , \tag{35}$$

while in the transformed system one minimizes

$$\begin{aligned} S' = \|\boldsymbol{Q}\boldsymbol{d} - \boldsymbol{Q}\boldsymbol{A}\boldsymbol{m}\|^2 &= \left(\boldsymbol{Q}\,(\boldsymbol{d} - \boldsymbol{A}\boldsymbol{m}) \cdot \boldsymbol{Q}\,(\boldsymbol{d} - \boldsymbol{A}\boldsymbol{m})\right) \\ &= \left(\boldsymbol{Q}^T\boldsymbol{Q}\,(\boldsymbol{d} - \boldsymbol{A}\boldsymbol{m}) \cdot (\boldsymbol{d} - \boldsymbol{A}\boldsymbol{m})\right) \end{aligned} \tag{36}$$

These two quantities are identical when the transformation $\boldsymbol{Q}$ is unitary, i.e. when $\boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I}$. This result stems from the property that unitary matrices do not affect the norm of a vector.

2.6. THE CONSISTENCY PROBLEM FOR THE MINIMUM-NORM SOLUTION

Consistency problems do not only arise for the least-squares solution, the minimum-norm solution suffers from the same problem. As an example let us return to the underdetermined "system" of equations (14). The minimum-norm solution of this problem is given by

$$\boxed{\tilde{m}_1 = 1 \qquad , \qquad \tilde{m}_2 = 1} \tag{37}$$

Now carry out a transformation from the model vector $\boldsymbol{m}$ to a new model vector $\boldsymbol{m}'$:

$$\begin{aligned} m_1' &= m_1 + m_2 \\ m_2' &= m_2 \end{aligned} \tag{38}$$

For this new model vector the "system" of equations is given by:

$$\tilde{m}_1' = d = 2 \tag{39}$$

Note that this transformed model vector brings out the fact that the system is undetermined much more clearly than the original system (14) because the new system imposes no constraint whatsoever on the model parameter m_2'. The minimum-norm solution of the transformed equation (39) is given by $\tilde{m}_1' = 2$, $\tilde{m}_2' = 0$. With the transformation (38) this solution in the transformed model space corresponds to the following solution in the original model space:

$$\boxed{\tilde{m}_1 = 2 \qquad , \qquad \tilde{m}_2 = 0} \tag{40}$$

This solution is shown by the open square in figure 4. Note that this solution differs from the minimum-norm solution (37) for the original system of equations. The reason for this discrepancy is similar to the consistency problem for the least-squares problem in section 2.5; the transformation (38) has changed the metric of model space, so that distances in the original model space and the transformed model space are measured in different ways. For this reason the minimum norm solution of the original problem and transformed problem are different.

We could carry out a similar general analysis for the transformation properties of the minimum norm solution under general transformations of the model vector and data vector as we carried out for the least-squares solution in section 2.5. However, in practice one applies regularization to the equations. As shown in equation (23) the damped least-squares solution and the damped minimum-norm solution are identical. For this reason the general transformation properties are treated in the next section for the damped least-squares solution.

2.7. THE NEED FOR A MORE GENERAL REGULARIZATION

The analysis of the transformation properties of the damped least-squares is completely analogous to the analysis of the undamped least-squares solution of section 2.5. Ignoring errors for the moment, the linear system of equations is given by (28): $\boldsymbol{Am} = \boldsymbol{d}$, and the transformation of the model vector and data vector is given by (29) and (30) respectively: $\boldsymbol{m}' = \boldsymbol{Sm}$ and $\boldsymbol{d}' = \boldsymbol{Qd}$. Assuming again that $\boldsymbol{S}^{-1}$ exists, the transformed system of equations is given by (31): $\boldsymbol{QAS}^{-1}\boldsymbol{m}' = \boldsymbol{Qd}$.

The damped least squares solution of the original system is given by:

$$\boxed{\tilde{\boldsymbol{m}}^{(1)} = \left(\boldsymbol{A}^T\boldsymbol{A}+\gamma\boldsymbol{I}\right)^{-1}\boldsymbol{A}^T\boldsymbol{d}} \tag{41}$$

Analogously to (34) the damped least-squares solution of the transformed equations is given by:

$$\tilde{\boldsymbol{m}}^{(2)} = \boldsymbol{S}^{-1}\left(\boldsymbol{S}^{T-1}\boldsymbol{A}^T\boldsymbol{Q}^T\boldsymbol{QAS}^{-1} + \gamma\boldsymbol{I}\right)^{-1}\boldsymbol{S}^{T-1}\boldsymbol{A}^T\boldsymbol{Q}^T\boldsymbol{Qd} \tag{42}$$

The damping parameter here is not necessarily equal to the damping parameter in the original damped least-squares solution, but for our purpose we do not need to make the distinction. Expression (42) can be simplified using the same steps as in the derivation of (34). Writing the term $\gamma\boldsymbol{I}$ as $\gamma\boldsymbol{I} =\gamma\boldsymbol{S}^{T-1}\boldsymbol{S}^T\boldsymbol{SS}^{-1}$, it follows that

$$\boxed{\tilde{\boldsymbol{m}}^{(2)} = \left(\boldsymbol{A}^T\boldsymbol{Q}^T\boldsymbol{QA}+\gamma\boldsymbol{S}^T\boldsymbol{S}\right)^{-1}\boldsymbol{A}^T\boldsymbol{Q}^T\boldsymbol{Qd}} \tag{43}$$

This expression points to a fundamental problem: the damping term in the model space $\boldsymbol{m}'$ is given by the identity matrix $\gamma\boldsymbol{I}$ (see (42)) and the damping term is $\gamma\boldsymbol{S}^T\boldsymbol{S}$ when expressed in terms of the original model vector $\boldsymbol{m}$ (see (43)). This implies that the damping $\gamma\boldsymbol{I}$ is not invariant for transformations of the model parameters. The terms $\boldsymbol{Q}^T\boldsymbol{Q}$ appear when a transformation of the data vector is carried out. This implies that the damped least-squares solution is in general not invariant under transformations of the data vector or model vector.

There is therefore a need for a more general regularization which allows to change model and data space in a consistent manner so that the solution is coordinate independent. Such a general regularization can be found from (43) by setting $\boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{W}_d$ and by defining $\boldsymbol{S}^T\boldsymbol{S} = \boldsymbol{W}_m$. The general least-squares solution is then given by:

$$\tilde{\boldsymbol{m}} = \left(\boldsymbol{A}^T\boldsymbol{W}_d\boldsymbol{A}+\gamma\boldsymbol{W}_m\right)^{-1}\boldsymbol{A}^T\boldsymbol{W}_d\boldsymbol{d}\ . \tag{44}$$

This solution minimizes the following cost function:

$$S = (\boldsymbol{d} - \boldsymbol{Am})^T\boldsymbol{W}_d(\boldsymbol{d} - \boldsymbol{Am}) + \gamma\boldsymbol{m}^T\boldsymbol{W}_m\boldsymbol{m} \tag{45}$$

This expression shows that in general the weight matrices $\boldsymbol{W}_d$ and $\boldsymbol{W}_m$ can be anything (as long as they are positive definite to ensure that S

has minima). Written in this way, γ may be seen as a trade-off parameter which compromises between two characteristics of the model: its size and its disagreement with the data. Both independent properties of the model cannot be arbitrary small together, hence there is a need for a balance. The choice of an optimum γ, however, is not an easy question. We have shown explicitly that when you start with a simple damped least-squares solution you can transform that problem into a more generally regularized least-squares solution in a different coordinate system and vice versa.

This implies that there is no reason to favor the damped least-squares solution over the more general least-squares solution (44). In fact, most inverse problems are ill-posed (partly underdetermined and partly overdetermined) and ill-conditioned (small errors in the data causes large variations in the model) which goes hand in hand with large null-spaces and hence non-unique solutions. Regularization is thus needed, but there is a large ambiguity in its choice [*Scales and Snieder*, 1997]. This reflects the fundamental difficulty that one faces in solving inverse problems: solving the system of equations is a minor problem, compared to choosing the regularization.

One approach is to use Bayesian statistics where one treats the inverse problem from a statistical point of view combining a-priori information about the data and the model with the data that are actually measured [*Tarantola and Valette,* 1982a; *Tarantola and Valette,* 1982b]. The weight matrices reflect true physical a-priori information (in a statistical sense) that one has of the data and the model, *independent* of the measured data. This includes for example the statistical noise characteristics of the instrument that produced the data, as well as information of the model and data that follow from other arguments. (For example, the mass-density in the Earth must be positive.) In such a Bayesian approach the weight matrices are given by

$$\boldsymbol{W}_d = \boldsymbol{C}_d^{-1} \qquad , \qquad \gamma \boldsymbol{W}_m = \boldsymbol{C}_m^{-1} \; , \tag{46}$$

where $\boldsymbol{C}_d^{-1}$ and $\boldsymbol{C}_m^{-1}$ are the a-priori covariance matrices for the data and model respectively:

$$\boldsymbol{C}_d = \left\langle (\boldsymbol{d} - \langle \boldsymbol{d} \rangle)\, (\boldsymbol{d} - \langle \boldsymbol{d} \rangle)^T \right\rangle \; , \tag{47}$$

$$\boldsymbol{C}_m = \left\langle (\boldsymbol{m} - \langle \boldsymbol{m} \rangle)\, (\boldsymbol{m} - \langle \boldsymbol{m} \rangle)^T \right\rangle \; . \tag{48}$$

In these expressions the brackets $\langle \cdots \rangle$ denote the expectation value. In this interpretation the estimator (44) corresponds to the most likely a-posteriori model when the error distribution is Gaussian. The statistical basis of Bayesian inversion leads to an objective solution if one respects the rule that the a-priori information has a true physical meaning. In practice however, one should realize that the choice of the a-priori distribution of the data and model is very often subjective as well. The reader can find further details in the column "To Bayes or not to Bayes" of *Scales and Snieder* [1997].

A different approach is to define the misfit function in such a way that it favours models with given properties (small, smooth, ...) [*Parker*, 1994]. Choosing a-priori information then amounts to defining an appropriate norm in which the data misfit and any given property of the model are measured. In our case, the weight matrices would then define a particular metric for the L_2-norm. As an example of choosing the weight matrices $\boldsymbol{W}_m$ the use of Occam's inversion is quite common [*Constable et al.*, 1987] where one seeks the smoothest model that is consistent with the data. Instead of putting a constraint on the model length, one seeks the square of its gradient to be as small as possible, i.e. the last term in (45) is a discretization of $\|\nabla m\|^2 = \int (\nabla m \cdot \nabla m)\, dV = -\int m\, \nabla^2\, m\, dV$,[5] and hence $\boldsymbol{W}_m$ corresponds to a discretized form of the Laplacian $-\nabla^2$.

2.8. THE TRANSFORMATION RULES FOR THE WEIGHT MATRICES

One of the fundamental requirements of an inverse solution should be that the results of the inversion are independent of arbitrary scalings applied to the model vector or data vector. Alas, this requirement is often ignored which can render comparisons of different models quite meaningless. For practical implications see *Trampert and Lévêque* [1990] and *Trampert et al.* [1992]. Here we derive how the weight matrices $\boldsymbol{W}_m$ and $\boldsymbol{W}_d$ should scale under such transformations for the least-squares solution to remain invariant.

Let us first consider the scaling (29) of the model vector: $\boldsymbol{m}' = \boldsymbol{S}\boldsymbol{m}$. Under this transformation the model term in the least-squares quantity (45) transforms as

$$\boldsymbol{m}^T \boldsymbol{W}_m \boldsymbol{m} = \boldsymbol{m}'^T \boldsymbol{S}^{T-1} \boldsymbol{W}_m \boldsymbol{S}^{-1} \boldsymbol{m}' = \boldsymbol{m}'^T \boldsymbol{W}'_m \boldsymbol{m}' \ , \tag{49}$$

with

$$\boldsymbol{W}'_m = \boldsymbol{S}^{T-1} \boldsymbol{W}_m \boldsymbol{S}^{-1} \ . \tag{50}$$

Under this transformation rule for the model weight matrix the least-squares criterion is unchanged, hence the least-squares solution is not changed when the model weight matrix $\boldsymbol{W}_m$ is transformed. It is of interest to note

[5]Note that we tacitly assumed in the last identity that there are no nonzero terms arising from the boundary conditions. A formal treatment based on Green's theorem allows for the incorporation of nonzero boundary terms.

that this rule implies that for Bayesian inversions, where the weight matrix is the inverse of the a-priori model covariance matrix ($\gamma \boldsymbol{W}_m = \boldsymbol{C}_m^{-1}$), the covariance matrix should transform as

$$\boldsymbol{C}_m^{'} = \boldsymbol{S}\boldsymbol{C}_m\boldsymbol{S}^T \tag{51}$$

One easily verifies from definition (48) that this is indeed the transformation rule for covariance operators.

Next let us consider how the transformation (30) for the data vector $\boldsymbol{d}^{'} = \boldsymbol{Q}\boldsymbol{d}$ affects the transformation of the data weight matrix $\boldsymbol{W}_d$. The matrix $\boldsymbol{A}$ scales under this transformation as $\boldsymbol{A}^{'} = \boldsymbol{Q}\boldsymbol{A}$. Under this transformation the data term in the least-squares quantity (45) transforms as

$$\begin{aligned} &(\boldsymbol{d} - \boldsymbol{A}\boldsymbol{m})^T \boldsymbol{W}_d (\boldsymbol{d} - \boldsymbol{A}\boldsymbol{m}) \\ &= \left(\boldsymbol{d}^{'} - \boldsymbol{A}^{'}\boldsymbol{m}\right)^T \boldsymbol{Q}^{T-1}\boldsymbol{W}_d\boldsymbol{Q}^{-1} \left(\boldsymbol{d}^{'} - \boldsymbol{A}^{'}\boldsymbol{m}\right) \\ &= \left(\boldsymbol{d}^{'} - \boldsymbol{A}^{'}\boldsymbol{m}\right)^T \boldsymbol{W}_d^{'} \left(\boldsymbol{d}^{'} - \boldsymbol{A}^{'}\boldsymbol{m}\right) \end{aligned} \tag{52}$$

with

$$\boldsymbol{W}_d^{'} = \boldsymbol{Q}^{T-1}\boldsymbol{W}_d\boldsymbol{Q}^{-1} \ . \tag{53}$$

For a Bayesian inversion the data weight matrix is the inverse of the data covariance matrix ($\boldsymbol{W}_d = \boldsymbol{C}_d^{-1}$), so that for a Bayesian inversion $\boldsymbol{C}_d$ should transform as

$$\boldsymbol{C}_d^{'} = \boldsymbol{Q}\boldsymbol{C}_d\boldsymbol{Q}^T \ . \tag{54}$$

Note again that this is the correct transformation rule for a covariance matrix defined in (47). This implies that the Bayesian viewpoint, where $\boldsymbol{W}_m$ and $\boldsymbol{W}_d$ are the inverses of the model and data covariance matrices, ensures that the solution is invariant under transformations of the model vector and/or the data vector.

Although we have derived in which way the weight matrices $\boldsymbol{W}_m$ and $\boldsymbol{W}_d$ should transform under transformations of model and data vectors, this does by no means imply that these matrices can be defined in a unambiguous way. An ill-posed and/or ill-conditioned inverse problem can only be solved if one is willing to control the solution by imposing a regularization term. In general, there is no unique recipe for choosing the weight matrices $\boldsymbol{W}_m$ and $\boldsymbol{W}_d$. It is the subjective input of the user that determines the choice of these matrices.

2.9. SOLVING THE SYSTEM OF LINEAR EQUATIONS

It should be noted that the least-squares solution always requires solving a set of linear algebraic equations. For instance, equation (44) may be written as

$$\left(\boldsymbol{A}^T\boldsymbol{W}_d\boldsymbol{A}+\gamma\boldsymbol{W}_m\right)\tilde{\boldsymbol{m}} = \boldsymbol{A}^T\boldsymbol{W}_d\boldsymbol{d}\ . \tag{55}$$

This represents a square system of linear equations, the so-called normal equations, of the form $\boldsymbol{Bx} = \boldsymbol{y}$. If we are merely interested in the estimation part of the problem, $\boldsymbol{B}$ doesn't need to be inverted. If we are also interested in the appraisal part of the problem (and we always should), it must be realized that $\boldsymbol{B}$ needs to be inverted at the cost of additional computer time. Many standard subroutine packages are available and *Press et al.* [1989] give a good and practical introduction to the subject. The reader should realize, however, that the system $\boldsymbol{Bx} = \boldsymbol{y}$ may become quite large for realistic geophysical problems which makes it worthwhile to consider a specialized routine best suited to the nature of $\boldsymbol{B}$ (symmetric, banded, sparse, ...). The dimension of the set of normal equations is also worth considering. Remember that the matrix $\boldsymbol{A}$ has the dimension $(N \times M)$, where N is the number of data and M the number of model parameters. System (55) has the dimension of the model space, but using (22) we may obtain a strictly equivalent system of the dimension of the data space. Choosing the smallest dimension to write the normal equation can save quite some computer time. Most techniques for solving the set of algebraic equations directly work with the matrix $\boldsymbol{B}$ as a whole requiring sufficient memory space to hold the matrix. But in global travel time tomography, for instance, these dimensions may become extremely large ($N > 10^6$ and $M > 10^5$) so that iterative methods need to employed which only work on parts of $\boldsymbol{B}$ at a time. Another problem which frequently occurs is that even though regularization is included in $\boldsymbol{B}$, it is singular or numerically very close to singular. A powerful technique, called Singular Value Decomposition (SVD), can diagnose precisely what the problem is and will give a useful numerical answer. SVD is the most effective tool in inverse theory to understand why a certain result has been obtained.

Iterative methods or SVD need not to work on square systems and may thus directly use the matrix $\boldsymbol{A}$. In this context it is useful to realize that the generalized least squares solution (44) is equivalent to the simple least squares solution of the system

$$\begin{pmatrix} \boldsymbol{W}_d^{1/2}\boldsymbol{A} \\ \cdots \\ \sqrt{\gamma}\ \boldsymbol{W}_m^{1/2} \end{pmatrix} \boldsymbol{m} = \begin{pmatrix} \boldsymbol{W}_d^{1/2}\boldsymbol{d} \\ \cdots \\ 0 \end{pmatrix}\ . \tag{56}$$

For a discussion on the meaning of a square root of a positive definite matrix the reader is referred to *Tarantola* [1987]. Keeping also in mind a certain freedom in choosing weighting matrices (see 2.7), the user might want to define directly $\boldsymbol{W}^{1/2}$ rather than $\boldsymbol{W}$. Expression (56) shows that regularization has the effect of adding extra rows to the system of linear equations, but the enlarged system is still of the form $\boldsymbol{Bx} = \boldsymbol{y}$, where the matrix $\boldsymbol{A}$ and the data vector $\boldsymbol{d}$ are augmented by extra rows that account for the regularization. $\boldsymbol{B}$ is not square anymore as in the case of normal equations. We will now illustrate in more detail the essence of singular value decomposition and iterative techniques as applied to the system $\boldsymbol{Bx} = \boldsymbol{y}$.

2.9.1. *Singular value decomposition*
Singular value decomposition was developed by *Lanczos* [1961], this technique is a generalization of the eigenvector decomposition of matrices to non-square matrices. Let us first consider a real symmetric $N \times N$-square matrix $\boldsymbol{B}$ with eigenvectors $\hat{\boldsymbol{v}}^{(n)}$ and eigenvalues λ_n. For such a matrix the eigenvectors form an orthonormal set, hence any vector $\boldsymbol{x}$ can be projected on these eigenvectors: $\boldsymbol{x} = \sum_{n=1}^{N} \hat{\boldsymbol{v}}^{(n)} \left(\hat{\boldsymbol{v}}^{(n)} \cdot \boldsymbol{x} \right)$. When $\boldsymbol{B}$ acts on this expression the result can be written as:

$$\boldsymbol{Bx} = \sum_{n=1}^{N} \lambda_n \hat{\boldsymbol{v}}^{(n)} \left(\hat{\boldsymbol{v}}^{(n)} \cdot \boldsymbol{x} \right) = \boldsymbol{y} \; . \tag{57}$$

Decomposing the vector $\boldsymbol{y}$ using the same eigenvectors $\hat{\boldsymbol{v}}^{(n)}$ gives $\boldsymbol{y} = \sum_{n=1}^{N} \hat{\boldsymbol{v}}^{(n)} \left(\hat{\boldsymbol{v}}^{(n)} \cdot \boldsymbol{y} \right)$, and inserting this into expression (57) yields the following expansion for the solution vector $\boldsymbol{x}$:

$$\boldsymbol{x} = \sum_{n=1}^{N} \frac{1}{\lambda_n} \hat{\boldsymbol{v}}^{(n)} \left(\hat{\boldsymbol{v}}^{(n)} \cdot \boldsymbol{y} \right) \; . \tag{58}$$

It can be seen that small eigenvectors can lead to instabilities in the solution $\boldsymbol{x}$. Singular value decomposition generalizes this expansion to non-square matrices. Details of this technique are given by *Lanczos* [1961] and by *Aki and Richards* [1980].

Now consider the following non-square system of equations:

$$\underbrace{\boldsymbol{B}}_{\substack{N \times M \\ matrix}} \quad \underbrace{\boldsymbol{x}}_{\substack{M \\ rows}} \quad = \quad \underbrace{\boldsymbol{y}}_{\substack{N \\ rows}} \tag{59}$$

Singular value decomposition is based on an expansion of $\boldsymbol{x}$ in an orthonormal set of eigenvectors $\hat{\boldsymbol{v}}^{(n)}$ and of $\boldsymbol{y}$ in an orthonormal set $\hat{\boldsymbol{u}}^{(n)}$. These vectors cannot be the eigenvectors of $\boldsymbol{B}$ because this matrix is not square, hence it does not have any eigenvectors. Instead, these vectors are related by the following relation:

$$\boldsymbol{B}\hat{\boldsymbol{v}}^{(n)} = \lambda_n \hat{\boldsymbol{u}}^{(n)} \qquad , \qquad \boldsymbol{B}^T \hat{\boldsymbol{u}}^{(n)} = \lambda_n \hat{\boldsymbol{v}}^{(n)} \tag{60}$$

It can easily be seen that the vectors $\hat{\boldsymbol{v}}^{(n)}$are the eigenvectors of $\boldsymbol{B}^T\boldsymbol{B}$ while the vectors $\hat{\boldsymbol{u}}^{(n)}$ are the eigenvectors of $\boldsymbol{B}\boldsymbol{B}^T$, hence these vectors can readily be determined. $\boldsymbol{B}^T\boldsymbol{B}$ and $\boldsymbol{B}\boldsymbol{B}^T$ share the same nonzero eigenvectors λ_n^2. The λ_n are called the singular values of $\boldsymbol{B}$. When $\boldsymbol{B}$ acts on the vector $\boldsymbol{x}$ the result can be written as

$$\boldsymbol{B}\boldsymbol{x} = \sum_{n=1}^{P} \lambda_n \hat{\boldsymbol{u}}^{(n)} \left(\hat{\boldsymbol{v}}^{(n)} \cdot \boldsymbol{x} \right) \ . \tag{61}$$

The upper limit of the summation is determined by the number of eigenvalues that are nonzero because the vanishing eigenvalues do not contribute to the sum. This number P can be significantly less than the dimension of the problem: $P \leq N$ and $P \leq M$.

It is convenient to arrange the vectors $\hat{\boldsymbol{u}}^{(n)}$ and $\hat{\boldsymbol{v}}^{(n)}$ as the columns of matrices $\boldsymbol{U}$ and $\boldsymbol{V}$, the eigenvectors from index P onwards correspond to zero eigenvalues and need to be included to make $\boldsymbol{U}$ and $\boldsymbol{V}$ complete:

$$\boldsymbol{U} = \left(\underbrace{\begin{matrix} \vdots & \vdots & & \vdots \\ \hat{\boldsymbol{u}}^{(1)} & \hat{\boldsymbol{u}}^{(2)} & \cdots & \hat{\boldsymbol{u}}^{(P)} \\ \vdots & \vdots & & \vdots \end{matrix}}_{\boldsymbol{U}_p} \ \underbrace{\begin{matrix} \vdots & & \vdots \\ \hat{\boldsymbol{u}}^{(P+1)} & \cdots & \hat{\boldsymbol{u}}^{(N)} \\ \vdots & & \vdots \end{matrix}}_{\boldsymbol{U}_0} \right) \ , \tag{62}$$

$$\boldsymbol{V} = \left(\underbrace{\begin{matrix} \vdots & \vdots & & \vdots \\ \hat{\boldsymbol{v}}^{(1)} & \hat{\boldsymbol{v}}^{(2)} & \cdots & \hat{\boldsymbol{v}}^{(P)} \\ \vdots & \vdots & & \vdots \end{matrix}}_{\boldsymbol{V}_p} \ \underbrace{\begin{matrix} \vdots & & \vdots \\ \hat{\boldsymbol{v}}^{(P+1)} & \cdots & \hat{\boldsymbol{v}}^{(M)} \\ \vdots & & \vdots \end{matrix}}_{\boldsymbol{V}_0} \right) \ , \tag{63}$$

The *orthogonality* of the eigenvectors implies that $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}$ and $\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}$. The *completeness* of the eigenvectors implies that $\boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{I}$ and $\boldsymbol{V}\boldsymbol{V}^T = \boldsymbol{I}$. Since the orthogonality of the eigenvectors also holds in the subspaces spanned by $\boldsymbol{U}_p$ and $\boldsymbol{V}_p$ we have $\boldsymbol{U}_p^T\boldsymbol{U}_p = \boldsymbol{I}$ and $\boldsymbol{V}_p^T\boldsymbol{V}_p = \boldsymbol{I}$. However, the vectors in these subspaces do in general not form a complete set, so that in general $\boldsymbol{U}\boldsymbol{U}_p^T \neq \boldsymbol{I}$ and $\boldsymbol{V}\boldsymbol{V}_p^T \neq \boldsymbol{I}$

The generalization of (61) to non-square systems can be written as

$$\boldsymbol{B} = \left(\begin{array}{cc} \boldsymbol{U}_p & \boldsymbol{U}_0 \end{array} \right) \left(\begin{array}{cc} \boldsymbol{\Sigma} & 0 \\ 0 & 0 \end{array} \right) \left(\begin{array}{c} \boldsymbol{V}_p^T \\ \boldsymbol{V}_0^T \end{array} \right) , \tag{64}$$

where the matrix $\boldsymbol{\Sigma}$ is given by:

$$\boldsymbol{\Sigma} = \left(\begin{array}{cccc} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & & \lambda_p \end{array} \right) \tag{65}$$

It follows from (61) that the eigenvectors $\hat{\boldsymbol{u}}^{(n)}$ that correspond to a vanishing eigenvalue do not contribute when $\boldsymbol{B}$ acts on a vector. These eigenvectors are ordered in the sub-matrix $\boldsymbol{U}_0$.This is equivalent to the statement that according to the representation (64) the matrix $\boldsymbol{B}$ can be constructed from $\boldsymbol{U}_p$ and $\boldsymbol{V}_p$ alone. $\boldsymbol{U}_0$ and $\boldsymbol{V}_0$ are dark spots of the space not illuminated by operator $\boldsymbol{B}$. Since $\boldsymbol{U}_0^T\boldsymbol{Bx} = \boldsymbol{0}$ the predicted data $\boldsymbol{Bx}$ are orthogonal to the subspace spanned by $\boldsymbol{U}_0$, see figure 5. This means that any components in the data vector that lie in $\boldsymbol{U}_0$ cannot be explained by *any* model. These components of the data vector necessarily correspond to errors in the data or errors in the operator $\boldsymbol{B}$ as a description of the physical problem. It is for this reason that $\boldsymbol{U}_0$ is called the *data-null-space*. In a least squares inversion one aims at minimizing the data misfit. Minimizing the data misfit then amounts to finding a model that produces a data vector in the subspace $\boldsymbol{U}_p$ that is closest to the true data. It follows from figure 5 that this is achieved by simply projecting the components of $\boldsymbol{U}_0$ contained in the data out of the problem. This is exactly what is done by limiting the sum over eigenvalues in (64) to the nonzero eigenvalues only. Of course, when $\boldsymbol{U}_0$ is empty, one can always find $\boldsymbol{x}$ which explains the data $\boldsymbol{y}$ exactly because $\boldsymbol{U}_p$ spans the complete data space.

In a similar way, the restriction of the summation over eigenvalues to the nonzero eigenvalues has the effect that the model estimate lies in the subspace spanned by $\boldsymbol{V}_p$, but that the estimated model has no component in $\boldsymbol{V}_0$. Any component of the model in $\boldsymbol{V}_0$ does not affect the data because $\boldsymbol{BV}_0 = \boldsymbol{0}$. This means that $\boldsymbol{V}_0$ defines the *model-null-space*. The data have no bearing on the components of the model vector that lie in $\boldsymbol{V}_0$. Setting the component of the model vector in the model-null-space equal to zero then implies that in the model estimation one only takes the nonzero eigenvalues into account. Expanding $\boldsymbol{x}$ in the vectors $\hat{\boldsymbol{v}}^{(n)}$ and $\boldsymbol{y}$ in the vectors $\hat{\boldsymbol{u}}^{(n)}$ and taking only the nonzero eigenvalues into account one can thus generalize

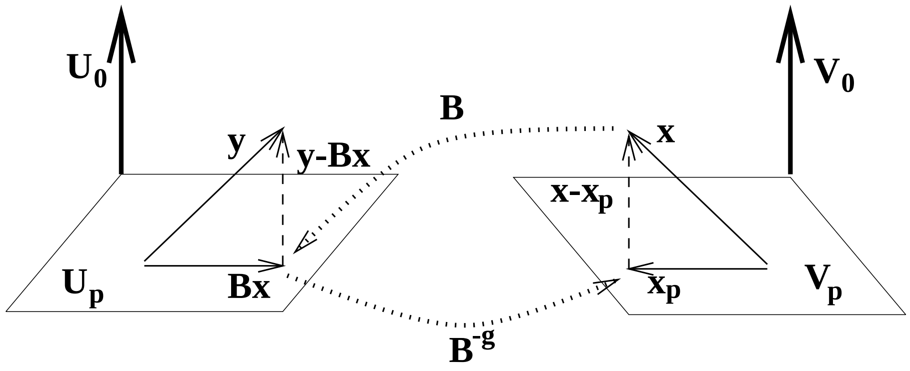

Figure 5. Geometrical interpretation of singular value decomposition. Note that starting from a model vector $\boldsymbol{x}$ and inverting the corresponding data $\boldsymbol{Bx}$ one obtains an estimated model vector $\boldsymbol{x}_p$ that in general differs from $\boldsymbol{x}$ because the resolution operator $\boldsymbol{R} = \boldsymbol{B}^{-g}\boldsymbol{B}$ in general is not equal to the identity matrix.

the solution (58) in the following way to non-square systems:

$$\boldsymbol{x} = \sum_{n=1}^{P} \frac{1}{\lambda_n} \hat{\boldsymbol{v}}^{(n)} \left(\hat{\boldsymbol{u}}^{(n)} \cdot \boldsymbol{y} \right) . \tag{66}$$

Using the matrices $\boldsymbol{U}_p$ and $\boldsymbol{V}_p$ this result can also be written as:

$$\boldsymbol{x} = \boldsymbol{V}_p \boldsymbol{\Sigma}^{-1} \boldsymbol{U}_p^T \boldsymbol{y} , \tag{67}$$

with $\boldsymbol{\Sigma}^{-1}$ given by

$$\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} 1/\lambda_1 & 0 & \cdots & 0 \\ 0 & 1/\lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/\lambda_p \end{pmatrix} \tag{68}$$

Similar to the forward problem, the inverse problem is not a function of $\boldsymbol{U}_0$ and $\boldsymbol{V}_0$. If both of these subspaces are zero, the operator $\boldsymbol{B}$ has an exact inverse. If $\boldsymbol{U}_0$ exists, one can show that the residual $\boldsymbol{y} - \boldsymbol{Bx}$ is perpendicular to $\boldsymbol{Bx}$ and hence the residual is minimum as for the least-squares solution. If $\boldsymbol{V}_0$ exists, solution (67) has no component in $\boldsymbol{V}_0$ and is therefore of minimum norm.

Clearly, small errors in $\boldsymbol{y}$ can lead to large errors in $\boldsymbol{x}$ when multiplied with $1/\lambda_n$ and the singular value is small. This process of error magnification can be controlled by limiting the summation in (66) to eigenvalues that differ significantly from zero. Alternatively, one can replace $1/\lambda_n$ by

$\lambda_n / \left(\lambda_n^2 + \gamma\right)$ with γ a positive constant. One can show that this is equivalent to the damped least-squares solution (20). See for example *Matsu'ura and Hirata* [1982] for a discussion on these different strategies. It should be noted that cutting off or damping small eigenvalues leads to different results. This makes it virtually impossible to quantitatively compare solutions of the same problem obtained with these fundamentally different strategies. One of the main reasons for the popularity of singular value decomposition is the control that one can exert over the error propagation in the solution. The drawback is that one needs to determine the eigenvectors of a matrix. For realistic large-scale problems ($P > 10^4$) this may require a prohibitive amount of CPU time. On the other hand, once the eigenvectors are calculated, resolution and error propagation are obtained at virtually no cost, since they merely involve the matrix multiplication $\boldsymbol{V}_p \boldsymbol{V}_p^T$.

2.9.2. *Iterative least-squares*

The least-squares solution of the system $\boldsymbol{Bx} = \boldsymbol{y}$ is, as shown in (11), given by $\boldsymbol{x} = \left(\boldsymbol{B}^T \boldsymbol{B}\right)^{-1} \boldsymbol{B}^T \boldsymbol{y}$. Note that the matrix $\boldsymbol{B}$ may contain some form of regularization as seen in expression (56). In practice, given the large-scale of many inverse problems, the matrix $\boldsymbol{B}^T \boldsymbol{B}$ may not fit in the computer memory. It is for this reason that one has developed iterative techniques which improve existing estimates of the solution.

Suppose one has in the n-th iteration of an iterative process a model estimate $\boldsymbol{x}_n$ and that one seeks an update $\delta \boldsymbol{x}_n$ such that the new model estimate $\boldsymbol{x}_{n+1} = \boldsymbol{x}_n + \delta \boldsymbol{x}_n$ is a better estimate of the model. Inserting this expression into the relation $\boldsymbol{Bx} = \boldsymbol{y}$ gives an expression for the model update:

$$\boldsymbol{B} \delta \boldsymbol{x}_n = (\boldsymbol{y} - \boldsymbol{B} \boldsymbol{x}_n) \tag{69}$$

Note that the right hand side of this expression is the residual of the model estimate $\boldsymbol{x}_n$, i.e. the difference $\boldsymbol{y} - \boldsymbol{B}\boldsymbol{x}_n$, is a measure to what extend the model estimate $\boldsymbol{x}_n$ does not explain the data. Expression (69) prescribes how the model should be updated in order to reduce the data residual. The least squares-solution of this expression is given by

$$\delta \boldsymbol{x}_n = \left(\boldsymbol{B}^T \boldsymbol{B}\right)^{-1} \boldsymbol{B}^T \left(\boldsymbol{y} - \boldsymbol{B} \boldsymbol{x}_n\right) \tag{70}$$

However, we have not gained anything yet because this expression is just as difficult to solve as the original equation and we still need to deal with the inverse of $\boldsymbol{B}^T \boldsymbol{B}$.

The advantage of solving the problem iteratively is that in such an approach one can replace the inverse $\left(\boldsymbol{B}^T \boldsymbol{B}\right)^{-1}$ by a suitably chosen estimate

$\boldsymbol{P}$ of this inverse, i.e. one computes a model update using the following expression:

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n + \boldsymbol{P}\boldsymbol{B}^T (\boldsymbol{y} - \boldsymbol{B}\boldsymbol{x}_n) \tag{71}$$

The operator $\boldsymbol{P}$ is called a preconditioning operator. If one sets $\boldsymbol{P} = \left(\boldsymbol{B}^T\boldsymbol{B}\right)^{-1}$ one retrieves the full solution in one step, but in that case one needs to compute $\left(\boldsymbol{B}^T\boldsymbol{B}\right)^{-1}$, which is what we wanted to avoid. Recognizing that $\boldsymbol{B}^T (\boldsymbol{y} - \boldsymbol{B}\boldsymbol{x}_n)$ is the direction of descent of the cost function at $\boldsymbol{x}_n$, one may, on the other end of the spectrum, choose $\boldsymbol{P} = c\boldsymbol{I}$, where c is a constant derived from a least-squares criterion to ensure the steepest possible descent [e.g. *Tarantola*, 1984]. In practice one has to find a balance between using an advanced preconditioning operator (that may be difficult to compute but that leads to a solution with a few iterations), and a simple preconditioning operator (that is easy to compute but that may require many iterations). The most commonly used algorithms in geophysics are SIRT (Simultaneous Iterative Reconstruction Technique) and LSQR (Least-Squares Conjugate Gradient). SIRT has the drawback of introducing implicit regularization into the solution [*van der Sluis and van der Vorst*, 1987] and if corrected for significantly decreases the convergence speed [*Trampert and Lévêque*, 1990]. A more successful balance is achieved by the LSQR algorithm [*Paige and Saunders,* 1982a, 1982b; *van der Sluis and van der Vorst*, 1987]. An iterative scheme that mimics the properties of SVD is given by *Nolet and Snieder* [1990].

Note that in the iterative least-squares algorithm (71) there is no need to invert a matrix, one only needs to multiply with the matrices $\boldsymbol{P}$ and $\boldsymbol{B}^T$, which can be done row by row. In many practical problems such as seismic tomography, the matrix $\boldsymbol{B}$ is very sparse which means that most of the matrix elements are zero, see section 6.1. For such a matrix the iterative least-squares algorithm is particularly efficient because the matrix multiplications only works on the nonzero matrix elements. The implementation of this idea in seismic tomography was developed by *Clayton and Comer*, [1983] and by *Nolet* [1985].

Despite the efficiency of the iterative least-squares problems one should be aware that the convergence of this process may be very slow. An example from *VanDecar and Snieder* [1994] is shown in figure 6. In this example one fits a function whose values are given by the dots in figure 6 by a model defined as the samples of that function at the x-values indicated by the tick marks along the horizontal axis. This problem obviously is ill-posed because the function is not defined between the data points. In order to define a solution, Occam's method is used where $\boldsymbol{W}_m$ in (56) is a discretized version of the gradient operator. The true regularized least-squares solution of this

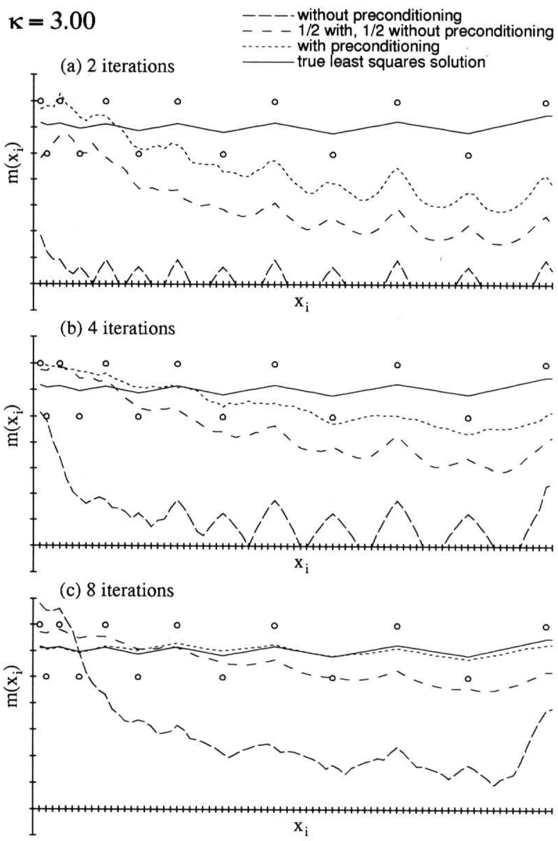

Figure 6. One-dimensional example of the convergence of a conjugate-gradient inversion at (a) iteration 2, (b) iteration 4 and (c) iteration 8. The solid line represents the true least-squares solution. The long-dashed lines are the interactive solutions without preconditioning while the dotted solutions are obtained using preconditioning.

problem is shown by the solid line. The iterative least-squares solution based on the conjugate-gradient algorithm after 2, 4 and 8 iterations is shown by the long-dashed line in the panels of figure 6.

It is clear that the convergence of the iterative least-squares algorithm is slow. The reason for this is easy to see. At every iteration, every point in the model interacts only with it's neighbouring points through the damping term $\boldsymbol{W}_m$ in the matrix. This means that a model element can only interact with a model element n positions further down the line after n iterations. In other words, the damping term imposes a global smoothness constraint on the solution, but in each iteration the model elements interact only locally.

For this reason an inordinate number of iterations are required. *VanDecar and Snieder* [1994] developed a preconditioning operator that leads to a very rapid convergence, this is indicated by the dotted lines in figure 6. Note that this iterative solution has superior convergence properties.

A general drawback of all iterative techniques is that most of its advantages are lost if one is interested in the appraisal part of the problem since this involves the knowledge of $(\boldsymbol{B}^T\boldsymbol{B})^{-1}$, which an iterative technique doesn't explicitly compute. A way around this is to solve the system M times, each time replacing the data by a column of $\boldsymbol{B}$ for the resolution matrix for instance [*Trampert and Lévêque*, 1990].

3. Linear inverse problems with continuous models

Up to this point, the model vector was finite dimensional. In many inverse problems the model is a continuous function of the space coordinates; it therefore has infinitely many degrees of freedom. For example, in inverse problems using gravity data one wants to determine the mass density $\rho(\boldsymbol{r})$ in the earth. This is in general a continuous function of the space coordinates. In realistic inverse problems one has a finite amount of data. A simple variable count shows that it is impossible to determine a continuous model with infinitely many degrees of freedom from a finite amount of data in a unique way.

In order to make the problem manageable we will restrict ourselves in this section to linear inverse problems. For such problems the data and the model are related by

$$d_i = \int G_i(x)m(x)dx + e_i \ . \tag{72}$$

The notation in this section is one-dimensional, but the theory can be used without modification in higher dimensions as well. The model is a continuous model but the data vector is discrete and in practice has a finite dimension. The kernel $G_i(x)$ plays the same role as the matrix $\boldsymbol{A}$ in (1). Note that the data are contaminated with errors e_i.

Since the forward problem is linear, the estimated model is obtained by making a linear combination of the data (this is the most general description of a linear estimator):

$$\tilde{m}(x) = \sum_i a_i(x)d_i \ . \tag{73}$$

The coefficients $a_i(x)$ completely specify the linear inverse operator. By inserting (72) in (73) one arrives again at a relation between the true model $m(x)$ and the estimated model $\tilde{m}(x)$:

$$\tilde{m}(x) = \underbrace{\int R(x,x')m(x')dx'}_{\text{Finite resolution}} + \underbrace{\sum_i a_i(x)e_i}_{\text{Error propagation}} \quad , \tag{74}$$

with the resolution kernel R given by

$$R(x,x') = \sum_i a_i(x)G_i(x') \ . \tag{75}$$

The first term in (74) accounts for averaging that takes places in the mapping from the true model to the estimated model. It specifies through the resolution kernel $R(x,x')$ what spatial resolution can be attained. In the ideal case the resolution or averaging kernel is a delta function: $R(x,x') = \delta(x - x')$. The resolution kernel, however, is a superposition of a finite amount of data kernels $G_i(x^{'})$. These data kernels are in general continuous functions, and since a delta function cannot be constructed from the superposition of a finite number of continuous functions, the resolution kernel will differ from a delta function. In Backus-Gilbert theory [*Backus and Gilbert,* 1967; 1968] one seeks the coefficients $a_i(x)$ in such a way that the resolution kernel resembles a delta function as well as possible given a certain criterion which measures this resemblance.

The second term in (74) accounts for error propagation. A proper treatment of this term needs to be based on statistics. It is shown by *Backus and Gilbert* [1970] that one cannot simultaneously optimize the resolution and suppress the error propagation and that one has to seek a trade-off between finite resolution and error propagation. The above mentioned work tries to explain the data exactly, even if they contain errors. *Gilbert* [1971] extended the theory to explain data within their error bars only.

As mentionned above, the Backus-Gilbert strategy finds the coefficients $a_i(x)$ in equation (73) by imposing a condition on the averaging kernel. *Taratola and Valette* [1982] solve the problem in a Baysian framework. They introduce prior information on the data (Gaussian error distribution) and prior assumptions (also Gaussian) on the unknown function $m(x)$ which also yields the coefficients $a_i(x)$. In this approach, the resolution or averaging kernel is a consequence of the *a priori* information, but generally different from a delta function. They further show that the Backus-Gilbert approach is contained in Tarantola-Valette solution.

In summary, in both strategies the infinite dimensional problem is transformed into a finite dimensional problem by seeking local averages of the true model.

Another approach is to evoke basis functions which amounts to changing the parameterization of the model. In general, any continuous model can be written as a superposition of a complete set of basis functions:

$$m(x) = \sum_{j=1}^{\infty} m_j B_j(x) \; . \tag{76}$$

In many global geophysical applications spherical harmonics are used to represent the seismic velocity or the density inside the earth, since they form a natural basis to describe a function on the sphere. In that case the $B_j(x)$ are the spherical harmonics and the sum over j stands for a sum over degree l and angular order m. The advantage of such an expansion is that one now deals with a discrete vector m_j of expansion coefficients rather than with a continuous function. However, the basis functions $B_j(x)$ only form a complete set when the sums is over infinitely many functions, and hence the problem has shifted from dealing with a continuous function to dealing with a vector with infinitely many components.

Inserting the expansion (76) into the forward problem (72) we may write

$$d_i = \sum_{j=1}^{\infty} A_{ij} m_j + e_i \; , \tag{77}$$

where the matrix elements A_{ij} are the projection of the data kernels onto the basis functions:

$$A_{ij} = \int G_i(x) B_j(x) dx \tag{78}$$

In practice, one cannot deal with infinitely many coefficients, and it is customary to ignore the fact that a model vector has infinitely many dimensions by taking only the first L basis functions into account. The resulting L-dimensional model vector will be denoted by $\boldsymbol{m}_L$, and a similar notation $\boldsymbol{A}_L$ is used for the first L rows of the matrix $\boldsymbol{A}$. The solution of the resulting finite dimensional model vector can be found using any technique shown in section 2. Recall that a general least-squares solution may be found by minimizing

$$S_L = \left(\boldsymbol{d} - \boldsymbol{A}_L \boldsymbol{m}_L\right)^T \boldsymbol{C}_d^{-1} \left(\boldsymbol{d} - \boldsymbol{A}_L \boldsymbol{m}_L\right) + \boldsymbol{m}_L^T \boldsymbol{C}_{mL}^{-1} \boldsymbol{m}_L \tag{79}$$

as a function of the truncated model vector $\boldsymbol{m}_L$. The weighting operators $\boldsymbol{C}_d$ and $\boldsymbol{C}_m$ may or may not be given the interpretation of covariance operators (see section 2.7). The resulting model estimate is then given by

$$\tilde{\boldsymbol{m}}_L = \left(\boldsymbol{A}_L^T\boldsymbol{C}_d^{-1}\boldsymbol{A}_L + \boldsymbol{C}_{mL}^{-1}\right)^{-1}\boldsymbol{A}_L^T\boldsymbol{C}_d^{-1}\boldsymbol{d} \tag{80}$$

3.2. SPECTRAL LEAKAGE, THE PROBLEM

Truncating the model vector after the first L elements may appear to be a convenient way to reduce the problem to a finite dimensional one. However, there are problems associated with this truncation. Analogously to (2) any linear estimator of the first L coefficients of the infinite vector $\boldsymbol{m}$ can be written as:

$$\tilde{\boldsymbol{m}}_L = \boldsymbol{A}_L^{-g}\boldsymbol{d}\ . \tag{81}$$

Let us now divide the sum over model elements in (77) as a sum over the first L elements that we are interested in and the remaining model elements:

$$\boldsymbol{d} = \boldsymbol{A}_L\boldsymbol{m}_L + \boldsymbol{A}_\infty\boldsymbol{m}_\infty + \boldsymbol{e}\ . \tag{82}$$

In this expression $\boldsymbol{m}_\infty$ denotes the infinitely dimensional vector with elements $(m_{L+1}, m_{L+2}, \ldots)$. Inserting this expression in (81) yields the relation between the estimated L model coefficients $\tilde{\boldsymbol{m}}_L$ and the true model $\boldsymbol{m}$ (for a discussion on the objectiveness of the concept of true model, see section 2.5):

$$\tilde{\boldsymbol{m}}_L = \boldsymbol{m}_L + \underbrace{\left(\boldsymbol{A}_L^{-g}\boldsymbol{A}_L - \boldsymbol{I}\right)\boldsymbol{m}_L}_{\text{Limited Resolution}} + \underbrace{\boldsymbol{A}_L^{-g}\boldsymbol{A}_\infty\boldsymbol{m}_\infty}_{\text{Spectral leakage}} + \underbrace{\boldsymbol{A}_L^{-g}\boldsymbol{e}}_{\text{Error propagation}} \tag{83}$$

The last three terms in this expression account for deviations in the estimated model from the true model. The second term and the last term are identical to the corresponding terms in expression (5) for finite-dimensional problems, accounting for finite resolution within the components of the vector $\boldsymbol{m}_L$ and error propagation respectively. The term $\boldsymbol{A}_L^{-g}\boldsymbol{A}_\infty\boldsymbol{m}_\infty$ has no counterpart in (5) and is responsible for a spurious mapping of the model coefficients $(m_{L+1}, m_{L+2}, \ldots)$ onto the estimated model coefficients $(\tilde{m}_1, \ldots, \tilde{m}_L)$. Since the basis function expansion in many problems is that of a spectral expansion, the mapping from the coefficients $\boldsymbol{m}_\infty$ onto the first L coefficients $\tilde{\boldsymbol{m}}_L$ will be referred to as *spectral leakage* [*Snieder et al.,* 1991].

An example of spectral leakage is shown in the figures 7-10. Suppose a function on a line, defined in the interval $0 \leq x < 1$, is expanded in normalized sines and cosines which have at most three wavelengths in the interval. This means that in this example $L = 7$. The function is sampled at given locations and the sampling is twice as dense on the subintervals $0 \leq x < 0.25$ and $0.5 \leq x < 0.75$ than on the remaining part of the line. The inverse problem consists in determining the expansion coefficients in the finite set of basis functions on the line given the sampled data points. Details of this example may be found in *Snieder* [1993].

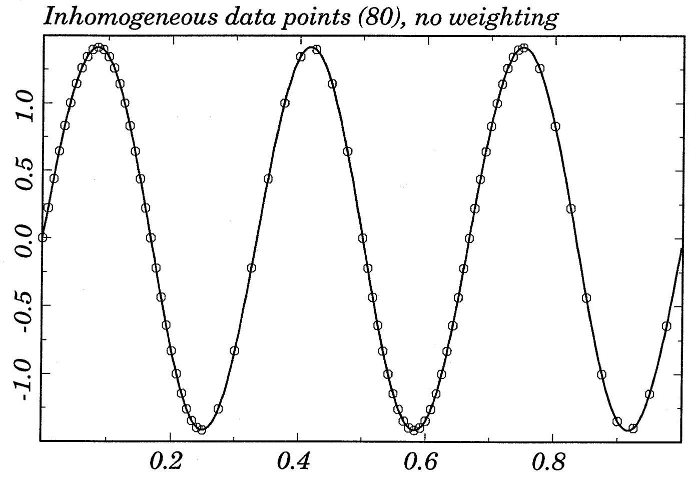

Figure 7. Unweighted least-squares fit of 80 data points (circles) sampling a sine wave with three wavelength along the interval. The estimated model is shown by the thick solid line, and is undistinguishable from the true projection on the basis functions.

In figure 7 the sampled function is a sine wave with exactly three wavelengths in the given interval. The sampling points are indicated by circles. The reconstructed function is given by the solid line and is indistinguishable from the true function. In figure 8 the input function is a sine wave with six wavelengths on the interval, this input function is indicated by the dashed line. Because of the orthogonality of the trigonometric functions, one would expect this function to have no projection onto the seven basis functions that are used in the inversion. Nevertheless, the reconstructed function shown by the solid line in figure 8 differs significantly from zero; it has about 50% of the magnitude of the input functions. As shown in *Snieder* [1993] the expansion coefficients have errors of about 100%! This difference between the estimated model and zero is entirely due to spectral leakage because the first L model components $\boldsymbol{m}_L$ are equal to zero and there are no data errors in this example so that only the term $\boldsymbol{A}_L^{-g}\boldsymbol{A}_\infty\boldsymbol{m}_\infty$ in (83) can give rise to errors in the model reconstruction.

An interesting observation is that in the interval $0 \leq x < 0.25$ where the sampling is dense the input model and the reconstructed model are in phase

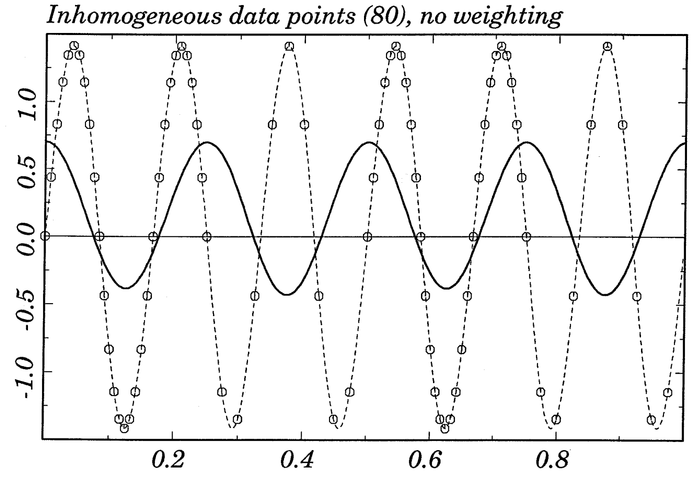

Figure 8. Unweighted least-squares fit of 80 data points (circles) sampling a sine wave with six periods (dashed line). The estimated model is shown by the thick solid line, the true projection on the basis functions is shown by the thin solid line.

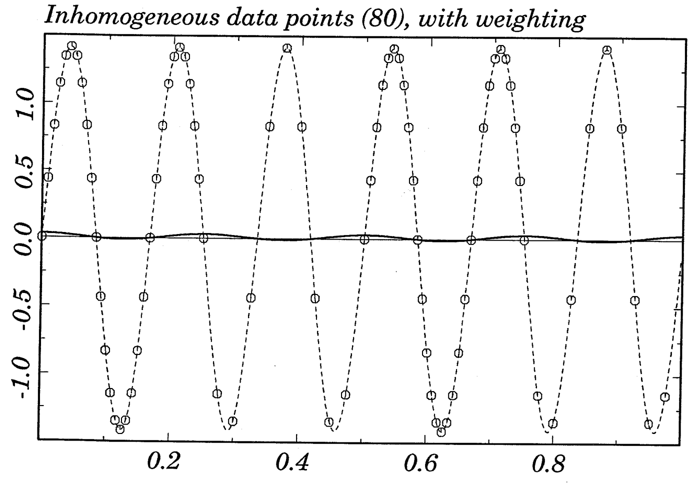

Figure 9. As the previous figure but for an inversion where each data point is with the inverse of the distance to neighbouring points.

whereas on the interval $0.25 \leq x < 0.5$ where the sampling is sparser the input model and reconstructed model are out off phase. This means that

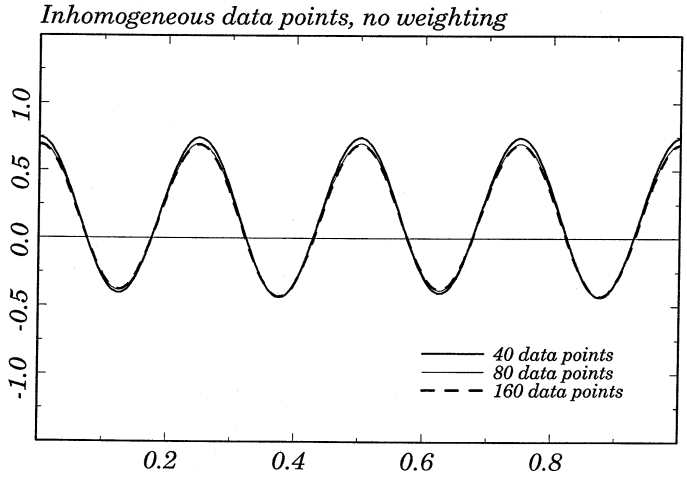

Figure 10. The estimated model when the true model is a sine wave with six wavelengths along the interval for inversions with different numbers of sampling points with the same sampling density along the line.

the least-squares criterion has selected a good fit in the densely sampled parts of the interval at the expense of a poorer fit in the sparsely sampled parts of the interval. This suggests a cure by down-weighting the contribution of the points in the densely sampled parts of the interval. Figure 9 shows the reconstructed function when the data points are weighted with weights that are inversely proportional to the sampling distance [*Snieder*, 1993], it can be seen that the reconstructed model indicated by the solid line is close to its true value (which is equal to zero). Weighting the data is the key to suppressing spectral leakage, we will return to this issue in section 3.3. The reason for the spectral leakage is that the orthogonality relation of the basis functions is weighted by the data kernels (i.e. $\sum_i \int G_i(x)B_j(x)dx \int G_i(y)B_k(y)dy$) and that the basis functions are not orthogonal for this inner product, i.e. this quantity is not equal to δ_{jk}.

Let us momentarily return to the unweighted inversion of figure 8. In that example 80 data points have been used. Note that the input function is not undersampled in any part of the interval, in other words there is no aliasing [*Claerbout*, 1976] occurring in this problem. In figure 10 the reconstructed function is shown when 40, 80 or 160 data points have been used in the inversion. In all examples the density of data points was the same as in figure 9. Spectral leakage is not a problem due to the number of data points. Adding more data points while keeping the sampling density constant does not help to suppress spectral leakage.

Spectral leakage is a fundamentally different issue than aliasing!

In some studies, the reliability of estimated models has been studied by dividing the data set randomly in two parts, keeping thus the sampling density constant, and comparing the models reconstructed from the two halved data sets [e.g. *Woodhouse and Dziewonski*, 1984]. The fact that spectral leakage does not reduce when more data points are used implies that this test does not detect artifacts due to the variability in the data coverage, and therefore may give an overly optimistic impression of the reliability of the estimated model.

3.3. SPECTRAL LEAKAGE, THE CURE

The method presented in this section for suppressing spectral leakage was developed by *Trampert and Snieder* [1996]. A key element in the analysis is to acknowledge that more than the first L basisfunctions are needed to describe the model. This is explicitly done by minimizing

$$\begin{aligned} S = (\boldsymbol{d} - \boldsymbol{A}_L \boldsymbol{m}_L - \boldsymbol{A}_\infty \boldsymbol{m}_\infty)^T \boldsymbol{C}_d^{-1} (\boldsymbol{d} - \boldsymbol{A}_L \boldsymbol{m}_L - \boldsymbol{A}_\infty \boldsymbol{m}_\infty) \\ + \boldsymbol{m}_L^T \boldsymbol{C}_{mL}^{-1} \boldsymbol{m}_L + \boldsymbol{m}_\infty^T \boldsymbol{C}_{m\infty}^{-1} \boldsymbol{m}_\infty \end{aligned} \tag{84}$$

The first term accounts for the minimization of the data misfit. Both, the first L basis functions contribute to the data misfit with $\boldsymbol{A}_L \boldsymbol{m}_L$, as well as the remaining basis functions through the term $\boldsymbol{A}_\infty \boldsymbol{m}_\infty$. The regularization terms $\boldsymbol{m}_L^T \boldsymbol{C}_{mL}^{-1} \boldsymbol{m}_L$ and $\boldsymbol{m}_\infty^T \boldsymbol{C}_{m\infty}^{-1} \boldsymbol{m}_\infty$ play a crucial role because they control to what extent the data misfit is distributed over the first L basisfunctions and over the remaining basis functions.

One has to minimize expression (84) both with respect to $\boldsymbol{m}_L$ and $\boldsymbol{m}_\infty$. Setting the derivatives $\partial S / \partial \boldsymbol{m}_L$ and $\partial S / \partial \boldsymbol{m}_\infty$ both equal to zero and eliminating $\boldsymbol{m}_\infty$ from these equations leads to the following estimate of the vector $\boldsymbol{m}_L$:

$$\tilde{\boldsymbol{m}}_L = \left(\boldsymbol{A}_L^T \boldsymbol{C}_{dL}^{-1} \boldsymbol{A}_L + \boldsymbol{C}_{mL}^{-1} \right)^{-1} \boldsymbol{A}_L^T \boldsymbol{C}_{dL}^{-1} \boldsymbol{d} \ , \tag{85}$$

where the new operator $\boldsymbol{C}_{dL}$ is given by

$$\boldsymbol{C}_{dL} = \boldsymbol{C}_d + \boldsymbol{A}_\infty \boldsymbol{C}_{m\infty} \boldsymbol{A}_\infty^T \ . \tag{86}$$

Expression (85) is very similar to the model estimate (80) obtained by simply ignoring all basis functions beyond degree L. The only difference is that the data weighting operator is given by (86) rather than by $\boldsymbol{C}_d$, but this difference is essential. The data weighting operator is positive definite. This implies that the new data weighting operator $\boldsymbol{C}_{dL}$ is always larger than the old operator $\boldsymbol{C}_d$. The new operator depends on the kernels $\boldsymbol{A}_\infty$ of the basisfunctions that are not inverted for, as well as the model weighting operator $\boldsymbol{C}_{m\infty}$ of these basis functions. From a Bayesian point of view these facts are all related by the observation that in this approach the data are partly explained by the infinite model vector $\boldsymbol{m}_\infty$. In the inversion for $\boldsymbol{m}_L$ only, the parts of the data which may be explained by $\boldsymbol{m}_\infty$ should be considered as noise, because they allow a variation in the data that is not cause by $\boldsymbol{m}_L$. For this reason this approach leads to an increase of the data variance $\boldsymbol{C}_{dL}$.

Although (86) looks simple, it is not trivial to evaluate because the product $\boldsymbol{A}_\infty \boldsymbol{C}_{m\infty} \boldsymbol{A}_\infty^T$ contains matrices of infinite dimensions. When the model weighting matrix $\boldsymbol{C}_{m\infty}$ is chosen to be proportional to the identity matrix ($\boldsymbol{C}_{m\infty} = \gamma \boldsymbol{I}$) there is a simple alternative because one only needs to evaluate the product $\boldsymbol{A}_\infty \boldsymbol{A}_\infty^T$. The ij-element of $\boldsymbol{A}_\infty \boldsymbol{A}_\infty^T$ can be written as $\left(\boldsymbol{A}_\infty \boldsymbol{A}_\infty^T\right)_{ij} = \sum_{k=L+1}^{\infty} A_{ik} A_{jk}$. Using expression (78) for the matrix elements and interchanging the summation and the integration one can show that

$$\left(\boldsymbol{A}_\infty \boldsymbol{A}_\infty^T\right)_{ij} = \int dx \int dy \; G_i(x) G_j(y) \sum_{k=L+1}^{\infty} B_k(x) B_k(y) \; . \tag{87}$$

The sum over k can be written as a sum over all k-values minus the sum over the first L values:

$$\sum_{k=L+1}^{\infty} B_k(x) B_k(y) = \sum_{k=1}^{\infty} B_k(x) B_k(y) - \sum_{k=1}^{L} B_k(x) B_k(y) \; . \tag{88}$$

The basis functions form a complete set. Because of the closure relation $\sum_{k=1}^{\infty} B_k(x) B_k(y) = \delta(x-y)$ [e.g. p. 157 of *Merzbacher*, 1970] this can be written as:

$$\sum_{k=L+1}^{\infty} B_k(x) B_k(y) = \delta(x-y) - \sum_{k=1}^{L} B_k(x) B_k(y) \; . \tag{89}$$

Inserting this in (87) leaves after carrying out the y-integration in the first term and using (78) for the second term:

$$\left(\boldsymbol{A}_\infty \boldsymbol{A}_\infty^T\right)_{ij} = \boldsymbol{\Gamma}_{ij} - \left(\boldsymbol{A}_L \boldsymbol{A}_L^T\right)_{ij} \; , \tag{90}$$

where the Gram matrix $\boldsymbol{\Gamma}$ of our inverse problem is given by

$$\Gamma_{ij} = \int G_i(x) G_j(x) dx \; . \tag{91}$$

The first term in (90) can be computed by direct integration, whilst the second term now only involves multiplication of matrices of *finite* dimension. It should, however, be pointed out that not all problems in geophysics possess a well defined Gram matrix.

3.4. SPECTRAL LEAKAGE AND GLOBAL TOMOGRAPHY

The burning question is of course if spectral leakage is a serious problem in geophysics. We believe that the answer to this question is yes and illustrate this with a surface wave tomography problem. A surface wave of a given frequency sees an average structure of the outer-most layer of the Earth. The penetration of the wave depends on its frequency. The mapping problem concerns only a two-dimensional quantity (phase velocity as function of latitude and longitude) and is thus easier to represent than a three-dimensional problem. To quantify the effect of leakage, we need to know the true answer to the problem.

We designed a synthetic experiment where we used a station-event distribution based on the global seismicity from 1989 together with the available global seismic stations. We have chosen arbitrarily 3000 paths giving a dense but uneven coverage (for details see *Trampert and Snieder*, [1996]). Next we took a Rayleigh wave phase velocity model [*Trampert and Woodhouse*, 1995] for a period of 80 seconds expressed in terms of spherical harmonics up to degree and order 40 and computed synthetic data for our chosen ray geometry. We added 10% random noise to the data and performed a classical least squares inversion up to degree and order 12. In figure 11 (bottom panel) we show the difference between the true degree 12 input model and the recovered model. This signal is far from zero and reaches in most places on the globe half the amplitude of the true model. If we applied the anti-leakage weighting defined in expression(86) the amplitudes of the artefacts are significantly reduced, see the top panel of figure 11. The anti-leakage operator does not completely prevent the bias because of the errors which we introduced in the synthetic data. To understand this, recall from equation (86) that the description of data errors and the anti-leakage operator add up linearly and thus any imperfect description of data errors will influence the anti-leakage. This suggests that our current tomographic models are likely to carry a bias from small-scale structures that are not accounted for by our current parameterizations.

4. The single scattering approximation and linearized waveform inversion

In the previous sections, the inverse problem was described when the forward problem was linear. Unfortunately, most wave propagation problems

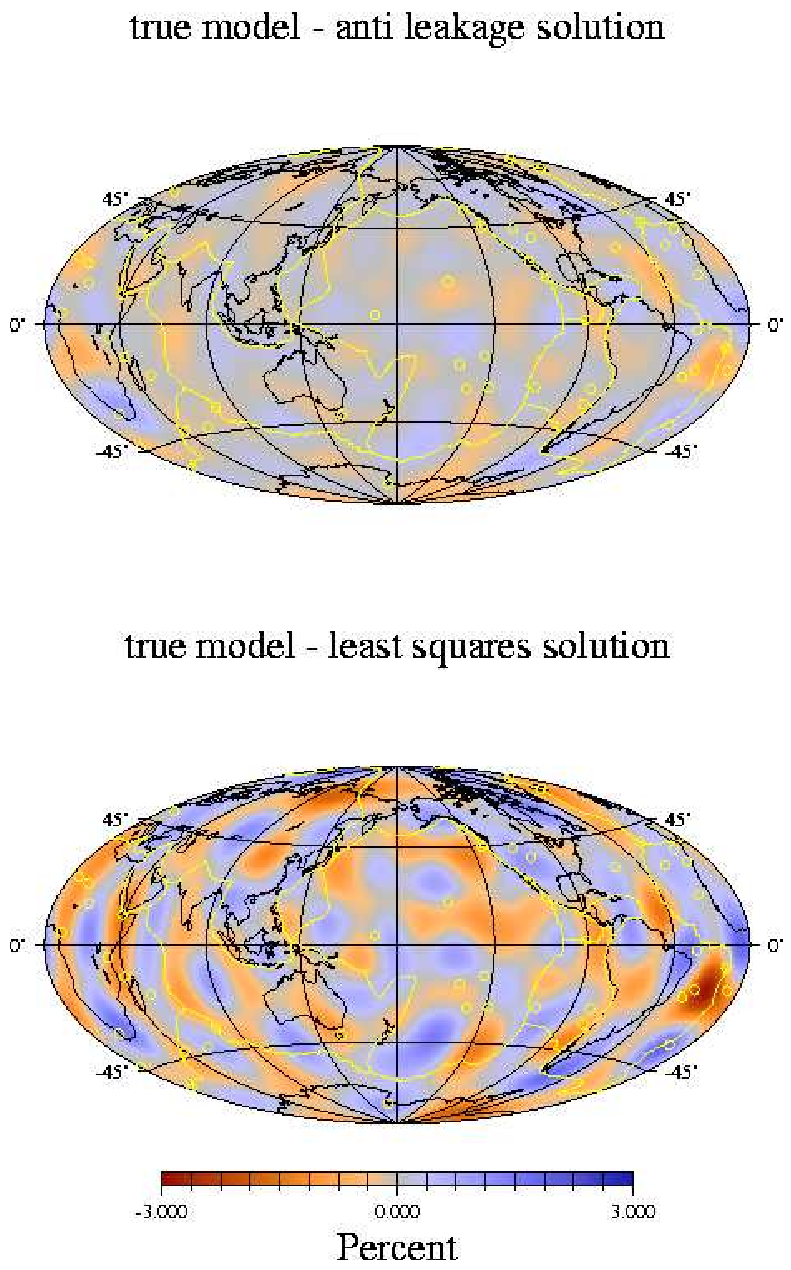

Figure 11. Example of phase velocity mapping with inhomogeneous sampling. The true model represents up to 6% perturbations with respect to a spherically average reference model. Plotted is the difference between true and estimated model: simple least-squares inversion (bottom) and anti-leakage inversion (top).

are nonlinear in the sense that the relation between the wave field and the medium is nonlinear. However, in many practical situations this relation can be linearized, notably with the single-scattering approximation (this section), with Fermat's theorem (section 6) and Rayleigh's principle (section 5).

4.1. THE BORN APPROXIMATION

Many wave propagation problems can symbolically be written in the form

$$Lu = F \tag{92}$$

In this expression u is the wave field, F accounts for the source that excites the waves and L is a differential operator that describes the wave propagation. For example, for acoustic waves in a medium with constant density (92) is given by

$$\left(\nabla^2 + \frac{\omega^2}{c^2(\boldsymbol{r})}\right) u(\boldsymbol{r}) = F(\boldsymbol{r}) \ , \tag{93}$$

so that $L = \nabla^2 + \omega^2/c^2(\boldsymbol{r})$. Often, the operator L can be decomposed into an operator L_0 that correspond to a reference medium for which we can solve the problem easily plus a small perturbation εL_1 that accounts for a perturbation of the medium:

$$L = L_0 + \varepsilon L_1 \ . \tag{94}$$

The small parameter ε has been introduced to facilitate a systematic perturbation approach. As an example, the operator L in (93) can be decomposed using

$$\frac{1}{c^2(\boldsymbol{r})} = \frac{1}{c_0^2}\left(1 + \varepsilon n(\boldsymbol{r})\right) \ , \tag{95}$$

where c_0 is a constant velocity that described the mean properties of the medium well while $n(\boldsymbol{r})$ accounts for the velocity perturbations. Using this decomposition one obtains $L_1 = n(\boldsymbol{r})\omega^2/c_0^2$.

The wavefield is perturbed by the perturbation of the medium, this perturbation can be written as a regular perturbation series:

$$u = u_0 + \varepsilon u_1 + \varepsilon^2 u_2 + \cdots \tag{96}$$

A systematic perturbation approach is obtained by inserting (94) and (96) in (92) and by collection the terms that are of equal power in the perturbation strength ε. The terms proportional to ε^0 and ε^n respectively lead to the following equations:

$$L_0 u_0 = F \ , \tag{97}$$

$$L_0 u_n = -L_1 u_{n-1} \qquad for \qquad n \geq 1 \; . \tag{98}$$

Equations (97) and (98) are of the form $L_0 u =$ *forcing term* and hence can be solved using the Green's function G of the unperturbed problem, i.e. by using the Green's function that satisfies

$$L_0 G(\boldsymbol{r}, \boldsymbol{r}') = \delta(\boldsymbol{r} - \boldsymbol{r}') \; . \tag{99}$$

For vector waves the Green's function should be replaced by a second order tensor, but the principles are the same. Using this Green's function the solution of (97) and (98) is given by

$$u_0 = GF \qquad \text{and} \qquad u_n = -GL_1 u_{n-1} \qquad (n \geq 1) \; . \tag{100}$$

The solution can now be constructed by solving the above equation recursively and by inserting the result in (96):

$$u = \underbrace{u_0}_{\substack{\text{Unperturbed}\\ \text{wave}}} - \underbrace{\varepsilon G L_1 u_0}_{\substack{\text{Single}\\ \text{scattered}\\ \text{wave}}} + \underbrace{\varepsilon^2 G L_1 G L_1 u_0}_{\substack{\text{Double}\\ \text{scattered}\\ \text{wave}}} + \cdots \tag{101}$$

In this formulation the total wavefield is written as a sum over the unperturbed wave, the waves that are scattered once by the inhomogeneity L_1, the waves that are scattered twice by the perturbation L_1 and all the higher order scattered waves. The series (101) is called the *Neumann series* in scattering theory.

The Born approximation u^{Born} consists in truncating this multiple scattering series after the single scattered wave:

$$u^{Born} = u_0 - \varepsilon G L_1 u_0 \; . \tag{102}$$

The great advantage of this approximation is that the scattered waves are given by $-\varepsilon G L_1 u_0$, hence the scattered waves now depend linearly on the perturbation of the medium that is contained in the operator L_1. This means that the scattered waves are related in this approximation linearly to the perturbation of the medium. This makes it possible to use the theory of linear inverse problems as shown in the section 2 for the solution of this problem. In doing so, one must keep in mind that the Born approximation ignores multiple scattering effects. When such effects are present in the data one should be extremely cautious in using the Born approximation.

4.2. INVERSION AND MIGRATION

As one of the simplest examples let us return to the acoustic wave equation (93) with the perturbed velocity given in (95). The unperturbed problem is then given by $(\nabla^2 + k^2)\, u = F$ where the constant wavenumber is given by $k = \omega / c_0$. The Green's function for this differential equation is given by

$$G(\boldsymbol{r}, \boldsymbol{r}') = -\frac{e^{ik|\boldsymbol{r} - \boldsymbol{r}'|}}{4\pi\, |\boldsymbol{r} - \boldsymbol{r}'|} \; . \tag{103}$$

When a point source with spectrum $F(\omega)$ is located in $\boldsymbol{r}_s$ the unperturbed wave is given by

$$u_0(\boldsymbol{r}) = -\frac{e^{ik|\boldsymbol{r} - \boldsymbol{r}_s|}}{4\pi\, |\boldsymbol{r} - \boldsymbol{r}_s|} F(\omega) \; . \tag{104}$$

In the Born approximation, the scattered waves are scattered only once, it follows from (102) that the single-scattered waves at a receiver position $\boldsymbol{r}_r$ are given by:

$$u_s(\boldsymbol{r}_r) = -\frac{1}{(4\pi c_0)^2} \int \frac{e^{ik|\boldsymbol{r}_r - \boldsymbol{r}|}}{|\boldsymbol{r}_r - \boldsymbol{r}|} n\,(\boldsymbol{r})\, \frac{e^{ik|\boldsymbol{r} - \boldsymbol{r}_s|}}{|\boldsymbol{r} - \boldsymbol{r}_s|} dV \, F(\omega) \; . \tag{105}$$

Suppose one measures these single-scattered waves, as is done in a seismic reflection experiment. The inverse problem then consists in finding the model perturbation $n(\boldsymbol{r})$ given the recorded data $u_s(\boldsymbol{r}_r)$. The theory of section 2 can be used for this when one discretizes the volume integral in (105). After discretization the scattering integral by dividing the volume in small cells, this integral can be written in the form

$$u_i = \sum_j A_{ij} n_j \; . \tag{106}$$

Since in a realistic imaging experiment one uses many source positions and records the wavefield at many receivers, u_i stands for the reflected wave for source-receiver pair and frequency component $\#i$. Because of the discretized volume integral, n_j denotes the perturbation of $1/c^2(\boldsymbol{r})$ in cell $\#j$. The matrix elements A_{ij} are given by

$$A_{ij} = -\frac{1}{(4\pi c_0)^2} \int_{cell\ j} \frac{e^{ik_i|\boldsymbol{r}_{ri} - \boldsymbol{r}|}}{|\boldsymbol{r}_{ri} - \boldsymbol{r}|} \frac{e^{ik_i|\boldsymbol{r} - \boldsymbol{r}_{si}|}}{|\boldsymbol{r} - \boldsymbol{r}_{si}|} dV \, F(\omega_i) \; . \tag{107}$$

In principle, one can solve the linear system (106) by brute force. However, in many realistic problems the size of the system of equations is so large that this is practically speaking impossible. This is notably the case in seismic exploration where the number data and the number of model parameters are exceedingly large. For problems of this scale, iterative solutions of the linear system of equations seems the only realistic way of obtaining a solution. In fact, it is in practice only possible to carry out the first step of such an iterative process. Let us find an estimated model by using the first step of the iterative solution (71) of section 2.9.2 using the preconditioning operator $\boldsymbol{P} = const. \cdot \boldsymbol{I}$. In this approach the estimated model is given by:

$$\tilde{\boldsymbol{n}}(\boldsymbol{r}) \sim \boldsymbol{A}^{\dagger}\boldsymbol{d} \sim \sum_{\substack{sources \\ receivers \\ frequencies}} \frac{e^{-i\omega(|\boldsymbol{r}_r - \boldsymbol{r}| + |\boldsymbol{r} - \boldsymbol{r}_s|)/c}}{|\boldsymbol{r}_r - \boldsymbol{r}|\,|\boldsymbol{r} - \boldsymbol{r}_s|} d_{rs}(\omega) \;, \qquad (108)$$

where $d_{rs}(\omega)$ denotes the scattered wave for source s, recorded at receiver r with frequency ω. For simplicity it is assumed here that the source signal is a delta-pulse in the time domain, so that $F(\omega) = 1$, and that all cells have equal volume: $V_j = const$. Note that the transpose $\boldsymbol{A}^T$ has been replaced by the Hermitian conjugate $\boldsymbol{A}^{\dagger}$.[6] The reason for this is that the analysis of section 2 was for real matrices. A similar analysis for complex matrices shows that the results are identical provided the transpose $\boldsymbol{A}^T$ is replaced by its complex conjugate. The summation in the matrix product effectively leads to a summation over all sources, receivers and frequency components because all these were labelled by the index i in (106).

It is instructive to consider the summation over frequencies in (108). At each frequency the data $d_{rs}(\omega)$ are multiplied with $\exp(-i\omega\tau)$ with $\tau = (|\boldsymbol{r}_r - \boldsymbol{r}| + |\boldsymbol{r} - \boldsymbol{r}_s|)/c$. This means that the frequency summation has the form of a Fourier transform so that up to a constant, the frequency summation gives the data in the time domain at time $t = (|\boldsymbol{r}_r - \boldsymbol{r}| + |\boldsymbol{r} - \boldsymbol{r}_s|)/c$:

$$\tilde{\boldsymbol{n}}(\boldsymbol{r}) \sim \sum_{\substack{sources \\ receivers}} \frac{1}{|\boldsymbol{r}_r - \boldsymbol{r}|\,|\boldsymbol{r} - \boldsymbol{r}_s|} d_{rs}\left(t = \frac{(|\boldsymbol{r}_r - \boldsymbol{r}| + |\boldsymbol{r} - \boldsymbol{r}_s|)}{c}\right) \qquad (109)$$

This expression implies that the image can be constructed by summing the data over all source-receiver pairs, and by considering for each target point the data at a time needed to travel from the source location $\boldsymbol{r}_s$ to the target point $\boldsymbol{r}$ to the receiver location $\boldsymbol{r}_r$.

This is the procedure followed in the imaging of seismic reflection data called "Kirchhoff migration" [*Claerbout, 1985*; *Yilmaz*, 1987]. Effectively one sums in such an approach over all the available data that have a travel time consistent with a scatterer at the target point $\boldsymbol{r}$. The only difference with the classical Kirchhoff migration is the presence of the geometrical

[6] The complex conjugate of the transpose is by definition the Hermitian conjugate: $A^{\dagger}_{ij} = A^{*}_{ji}$.

spreading terms $1/\left(|\boldsymbol{r}_r - \boldsymbol{r}|\,|\boldsymbol{r} - \boldsymbol{r}_s|\right)$ that are not included in Kirchhoff migration. In non-destructive testing this approach is known as Synthetic Aperture Focussing Technique (SAFT) [*Mayer et al.*, 1990]. The main conclusion is that Kirchhoff migration (up to some constants) corresponds to the first step of an iterative solution of the linearized scattering equation. The derivation of this section is a simplified version of the derivation given by *Tarantola* [1984] who incorporates other source signals than a delta pulse.

4.3. THE BORN APPROXIMATION FOR TRANSMISSION DATA

There is a widespread belief that the Born approximation can only be used for truly scattered waves. However, the Born approximation can also be used to account for effects of medium perturbations on transmitted waves, but with a domain of applicability to transmission problems that is smaller than for reflection or true scattering problems. To see this, assume that the velocity has a small constant perturbation δc and that a wave propagates over a distance L. The wave will then experience a phase shift $\exp i\varphi$, given by $\varphi = -\left(\omega/c^2\right) L\delta c$. In the Born approximation this perturbation is replaced by $\exp i\varphi \approx 1 + i\varphi$. This in only a good approximation when the phase shift φ is much less than a cycle. In practice, this limits the use of the Born approximation for the inversion of transmission data. Note that even when the velocity perturbation is small, the requirement that the phase shift is small breaks down for sufficiently large propagation distances L.

This does not imply that the Born approximation cannot be used for the inversion of transmitted waves. For surface waves propagating in the earth, the wavelength can be very long. (A Rayleigh wave at a period of $75s$ has a wavelength of about $300km$.) This means that these waves do not travel over very many wavelengths when they propagate for a few thousand kilometers. In addition, the heterogeneity in the Earth's mantle is not very large. This makes it possible to carry out inversions of transmitted surface wave data using the Born approximation.

The Born approximation for surface waves with the resulting scattering and mode-conversion coefficients is derived for a flat geometry by *Snieder* [1986a, 1986b], the generalization to a spherical geometry can be found in *Snieder and Nolet* [1987]. Although the Green's function and the scattering and mode-conversion coefficients are different for elastic surface waves than for acoustic waves, the Born approximation leads to a linearized relation between the perturbation δu of the waveform and the perturbation m of the Earth model:

$$\delta u_{rs}(\omega) = \iiint K_{rs}(\boldsymbol{r}, \omega) m(\boldsymbol{r}) dV \ , \tag{110}$$

where $K_{rs}(\boldsymbol{r}, \omega)$ contains the surface wave Green's function of the incoming and outgoing wave at the scatterer location $\boldsymbol{r}$ for source-receiver pair "rs" and frequency ω. The perturbation $m(\boldsymbol{r})$ stands for the perturbation in the elasticity tensor and/or the density. After discretization the volume integral can be written as a linear system of equations:

$$\delta u_i = \sum_j K_{ij} m_j \, , \tag{111}$$

where δu_i stands for the perturbation of the surface wave at source-receiver pair $\#i$ and frequency ω_i, while m_j stands for the perturbation of the Earth model in cell $\#j$. This linear system can be solved numerically, and the specific implementation for the inversion of surface wave data is shown by *Snieder* [1988a].

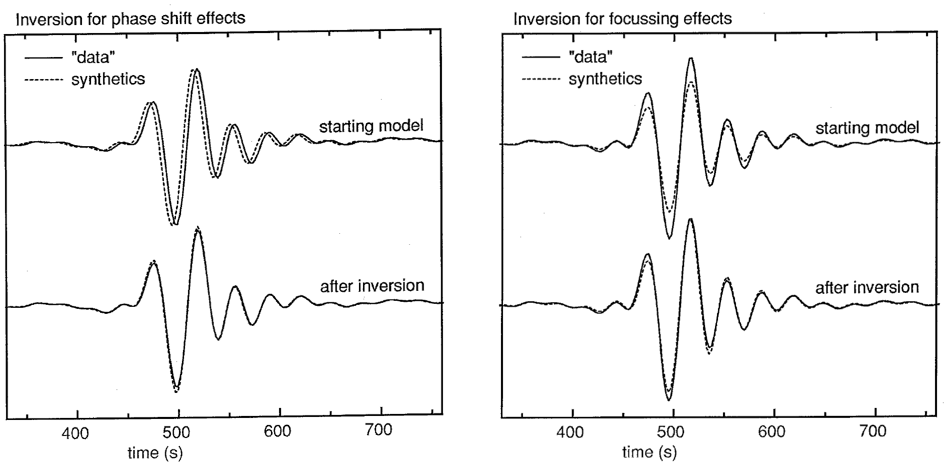

Figure 12. Phase shifted data (left panel) and data with an amplitude anomaly (right panel) before and after inversion. The synthetic data are shown with a solid line, the synthetics before and after inversion with a dashed line.

As an example how this algorithm operates consider the two upper seismograms shown in the two panels of figure 12. The seismogram of the left panel has been shifted in time with respect to the seismogram of a laterally homogeneous Earth model, whereas the amplitude of the seismogram of the right panel has been increased by 20%. Both seismograms are inverted simultaneously, and the corresponding model is shown in figure 13. (The map is shown only to display the scale of the problem but has otherwise no specific meaning.) The fit of the wave-forms after inversion is shown by the lower seismograms in figure 12, it can be seen that both the time shift and the amplitude change are accounted for well.

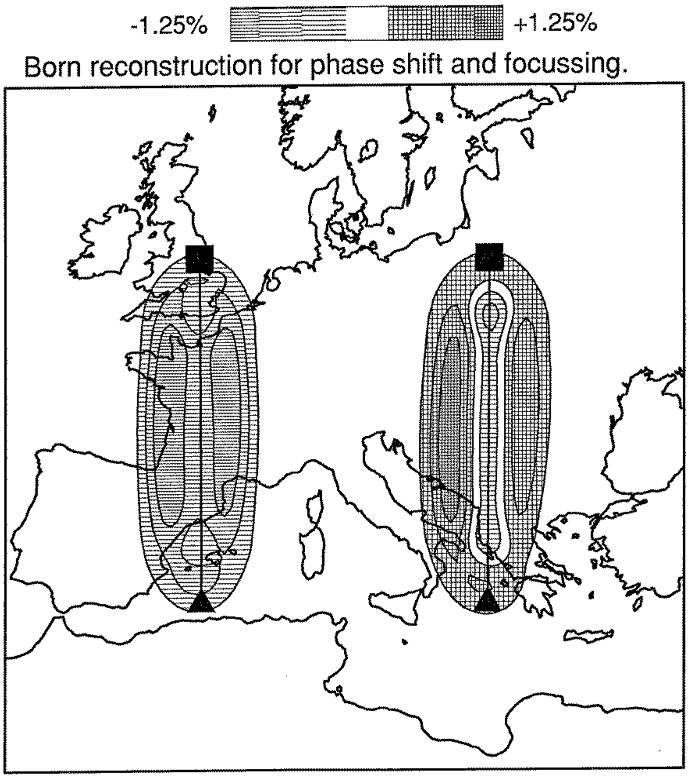

Figure 13. The velocity model obtained from the inversion of the seismograms in the previous figure. The source locations are marked with triangles, the receivers with squares. The source-receiver pair on the left is for the phase shifted seismogram, the pair on the right for the seismogram with an amplitude error. The map serves only to fix the scale.

The triangle and square on the left are the source and receiver respectively of the time-shifted seismogram while the same symbols on the right are for the seismogram with the amplitude change. The velocity anomaly for the time-shifted seismogram is a negative velocity perturbation straddling the line joining the source and receiver on the left in figure 13. This negative velocity change gives the required time-shift. Note that this velocity perturbation is not confined to the source-receiver line; its finite extent is due to the fact that rays have no physical meaning and that a wave is influenced by the velocity perturbation averaged over the first Fresnel zone [*Snieder and Lomax*, 1996].

The velocity structure on the right is for the seismogram with the perturbed amplitude. The perturbation is negative on the source-receiver line, but slightly further away from this line the velocity perturbation is positive. Since waves are deflected from areas of high velocity towards regions of low velocity, the wave energy is deflected towards the receiver. This means that the algorithm has realized a fit of the amplitude perturbation by creating a "surface wave lens" that produces just the right amount of focussing at the receiver to account for the amplitude perturbation.

Note that the model of figure 13 was constructed by numerically solving the linear system of equations (111). The input of the system of equations simple consisted of the real and imaginary components of the perturbation of the surface waves in the frequency domain. At no point in the inversion has the phase or amplitude of the waves been prescribed explicitly. However, the Born approximation contains all the relevant physics needed to translate the perturbation of the real and imaginary components of the waves into the physics of focussing and phase retardation.

4.4. SURFACE WAVE INVERSION OF THE STRUCTURE UNDER NORTH-AMERICA

The surface waveform inversion has been applied to infer the shear-velocity structure under Europe and the Mediterranean by *Snieder* [1988b]. In this section a model for the shear-velocity under North-America is shown that was made by *Alsina et al.* [1996]. The shear-velocity perturbation in three layers with depths between 25 and 300 km is shown in figure 14. Red colours indicate slow velocities that are indicative of high temperature while green colours indicate high velocities correspond to low temperature.[7]

It is instructive to consider the west-coast of North America. Under the Gulf of California the velocity is very low. In this region, the East-Pacific rise meets the continent. (The East-Pacific rise is a spreading center in the ocean bottom comparable to the Mid-Atlantic ridge in the Atlantic Ocean.) Since in a spreading center hot material wells up from the mantle, one expects the temperature to be low, which is consistent with the slow velocity in that region.

Further north at the state of California the velocity perturbation is close to zero. In this area the Pacific plate slides horizontally along the North-American plate. This so called "strike-slip motion" gives rise to earthquakes in California. However, since the plate simply slide along each other, the temperature field is not perturbed very much, which is reflected by the neutral velocity perturbation.

However, northward of Cape Mendocino under the states of Oregon and Washington there is a distinct positive velocity perturbation. In this region oceanic plates slide eastward under the continents. This process is called

[7]The identification of velocity anomalies with temperature perturbations should be treated with care. Apart from temperature perturbations the seismic velocity is also affected by variations is compostion, the presence of volatiles such as water and possibly also pressure.

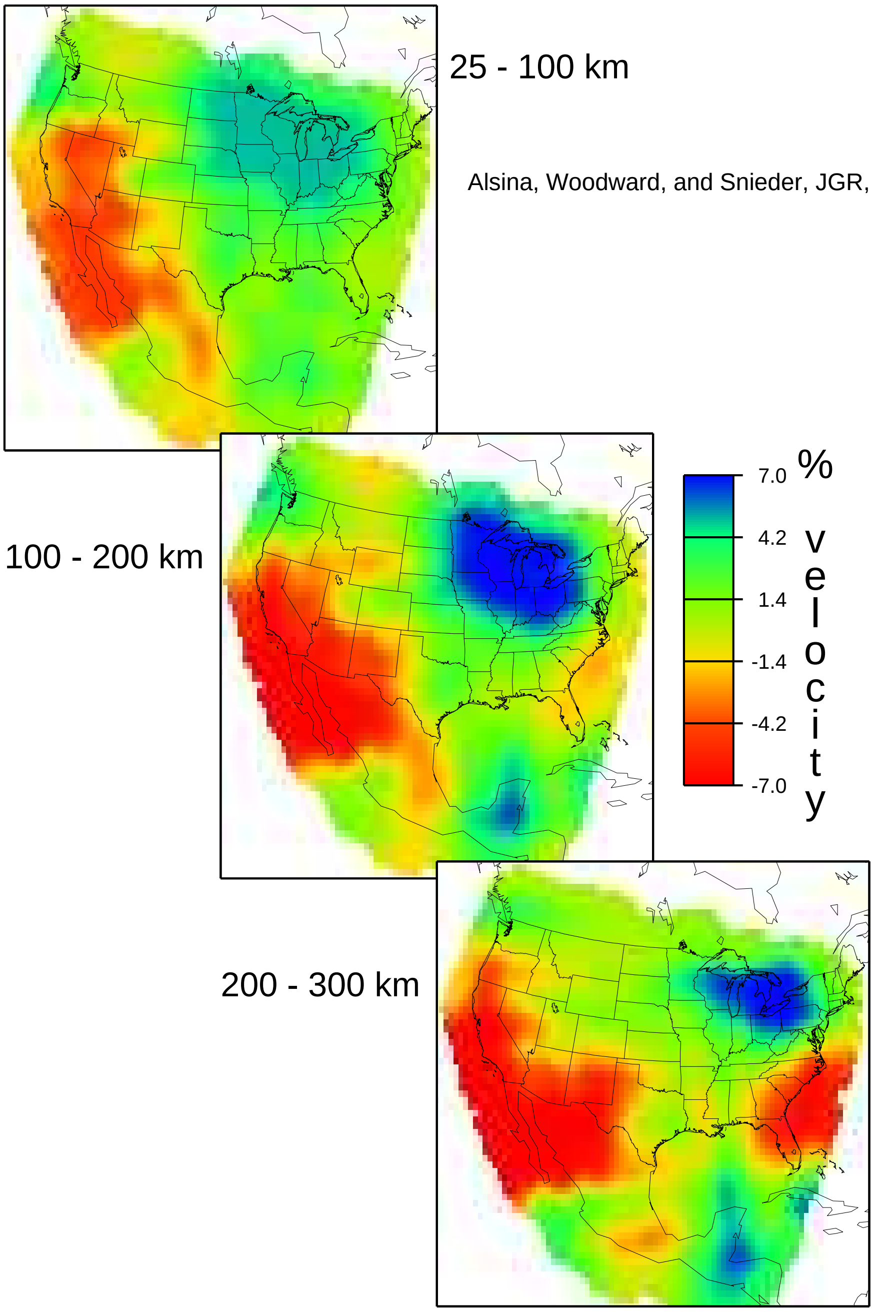

Figure 14. The relative shear velocity perturbation under North America in three layers at depths between 25 and 300 km as determined by [*Alsina et al.*, 1996].

"subduction." Since the subsiding plate is colder than the surrounding material this is reflected in the image as positive velocity anomalies.

The model shown in figure 14 and it's relation with the different tectonic regimes at the west coast of North-America show that with tomographic techniques as shown here it is possible to see plate-tectonics in action.

5. Rayleigh's principle and perturbed eigenfrequencies

Important information about the structure of the Earth's interior follows form observations of perturbations of the frequencies of the free oscillations of the Earth as well as from measurements of the perturbation of the phase or group velocity of surface waves. This information can most easily be retrieved from observations when the relation between the frequency shift (or surface wave phase velocity) and the perturbation of the Earth model can be linearized. This can be achieved by applying Rayleigh-Schrödinger perturbation theory.

5.1. RAYLEIGH-SCHRÖDINGER PERTURBATION THEORY

Consider an eigenvalue problem of the following form:

$$Hu_n = \lambda_n r u_n \ . \tag{112}$$

In this expression H is an operator, for wave-propagation problems it usually is a differential operator. The eigenfunctions of this operator are denoted by u_n and the corresponding eigenvalues by λ_n. The function r denotes a positive weight function. For the Earth the equation of motion is given by

$$\partial_j \left(c_{ijkl} \partial_k u_l \right) = -\rho \omega^2 u_i \ , \tag{113}$$

where c_{ijkl} is the elasticity tensor. A comparison with (112) shows that $H\left(\bullet\right)_i = \partial_j \left(c_{ijkl} \partial_k \left(\bullet\right)_l \right)$, that the weight function r is given by the density ρ and that λ corresponds to $-\omega^2$. In this case, H is a Hermitian operator:

$$H^\dagger = H \ . \tag{114}$$

For the special case of elastic wave propagation, this property stems from the symmetry properties of the elasticity tensor and the stress-free boundary conditions at the Earth's surface [*Nolet*, 1981; *Dahlen and Tromp*, 1998]. For a Hermitian operator the eigenvalues are real and the eigenfunctions are orthonormal with respect to the following inner product

$$\langle u_n | r u_m \rangle = \int u_n^* r u_m dV = \delta_{nm} \ . \tag{115}$$

Let the operator H and the weight function r be decomposed in a reference operator H_0 and weight function r_0 for which we know the eigenfunctions $u_n^{(0)}$ and eigenvalues $\lambda_n^{(0)}$ and perturbations εH_1 and εr_1:

$$H = H_0 + \varepsilon H_1 \qquad , \qquad r = r_0 + \varepsilon r_1 \ . \tag{116}$$

Under this perturbation the eigenfunctions and eigenvalues are perturbed as well:

$$u_n = u_n^{(0)} + \varepsilon u_n^{(1)} + \varepsilon^2 u_n^{(2)} + \cdots \tag{117}$$

$$\lambda_n = \lambda_n^{(0)} + \varepsilon \lambda_n^{(1)} + \varepsilon^2 \lambda_n^{(2)} + \cdots \tag{118}$$

The goal of this analysis is to find the first order perturbation $\lambda_n^{(1)}$ of the eigenvalues. This can be achieved by inserting the expressions (116) through (118) in (112) and by collecting the terms of order ε^1, this gives:

$$H_1 u_n^{(0)} + H_0 u_n^{(1)} = \lambda_n^{(1)} r_0 u_n^{(0)} + \lambda_n^{(0)} r_1 u_n^{(0)} + \lambda_n^{(0)} r_0 u_n^{(1)} \tag{119}$$

The problem with this expression is that we are only interested in $\lambda_n^{(1)}$, but that in order to retrieve this term from (119) we need the first order perturbation $u_n^{(1)}$ of the eigenfunctions as well. The perturbation of the eigenvalue can be extracted by taking the inner product of (119) with the unperturbed eigenfunction $u_n^{(0)}$:

$$\begin{aligned} &\left\langle u_n^{(0)} | H_1 u_n^{(0)} \right\rangle + \underbrace{\left\langle u_n^{(0)} | H_0 u_n^{(1)} \right\rangle}_{(\heartsuit)} \\ &= \lambda_n^{(1)} \left\langle u_n^{(0)} | r_0 u_n^{(0)} \right\rangle + \lambda_n^{(0)} \left\langle u_n^{(0)} | r_1 u_n^{(0)} \right\rangle + \underbrace{\lambda_n^{(0)} \left\langle u_n^{(0)} | r_0 u_n^{(1)} \right\rangle}_{(\spadesuit)} \end{aligned} \tag{120}$$

Note that the perturbed eigenfunctions $u_n^{(1)}$ only appear in the terms marked $(\heartsuit)$ and $(\spadesuit)$. Using the fact that H_0 is Hermitian and that the eigenvalues $\lambda_n^{(0)}$ are real one finds that these terms are equal:

$$\begin{aligned} (\heartsuit) &= \left\langle u_n^{(0)} | H_0 u_n^{(1)} \right\rangle = \left\langle H_0^\dagger u_n^{(0)} | u_n^{(1)} \right\rangle = \left\langle H_0 u_n^{(0)} | u_n^{(1)} \right\rangle \\ &= \left\langle \lambda_n^{(0)} u_n^{(0)} | u_n^{(1)} \right\rangle = \lambda_n^{(0)} \left\langle u_n^{(0)} | u_n^{(1)} \right\rangle = (\spadesuit) \end{aligned} \tag{121}$$

This means that the terms containing the perturbed eigenfunctions cancel. Solving the remaining equation for the perturbation of the eigenvalue gives:

$$\lambda_n^{(1)} = \frac{\left\langle u_n^{(0)} | \left(H_1 - \lambda_n^{(0)} r_1 \right) u_n^{(0)} \right\rangle}{\left\langle u_n^{(0)} | r_0 u_n^{(0)} \right\rangle} \ . \tag{122}$$

The perturbation of the eigenvalue thus follows by evaluating the inner product of the perturbed operators sandwiched between the unperturbed eigenfunctions. This is one form of Rayleigh's principle: the eigenvalues are stationary to first order for perturbations of the eigenfunctions. The crux is that according to 122 the first-order perturbation of the eigenvalues depends linearly on the perturbations of the operator H and weight r and hence linear inverse theory as exposed in the previous sections may be applied to infer the perturbations of H and r from the measured eigenvalues of the system.

5.2. THE PHASE VELOCITY PERTURBATION OF LOVE WAVES

Expression (122) can be used to evaluate the perturbation of the phase velocity of surface waves due to perturbations of the Earth model. This is shown here for the simplest example of Love waves in a plane geometry. The reader is referred to *Aki and Richards* [1980] for the extension to Rayleigh waves, while the analysis for a spherical geometry is treated by *Takeuchi and Saito* [1972]. As shown in *Aki and Richards* [1980] the Love wave eigenfunctions satisfy the following differential equation:

$$\partial_z(\mu\partial_z v) + \left(\rho\omega^2 - \mu k^2\right) v = 0 \ , \tag{123}$$

where μ is the shear modulus. The surface waves modes vanish at large depth ($z \rightarrow \infty$) and are stress-free at the surface. This gives the following boundary conditions:

$$\partial_z v(z=0) = 0 \qquad , \qquad v(z=\infty) = 0 \ . \tag{124}$$

In this analysis we assume that the frequency ω is a given constant. The operator H is given by $H = \partial_z \mu \partial_z + \rho\omega^2$, the weight function r is given by the shear modulus ($r(z) = \mu(z)$) while k^2 is the eigenvalue $\left(\lambda = k^2\right)$. An integration by parts using the boundary conditions (124) can be used to show that the operator H is Hermitian. The theory of section 5.1 can then be used to find the perturbation δk of the wavenumber due to a perturbation μ_1 in the shear modulus and a perturbation ρ_1 in the density. Using the first order relation $\delta k^2 = 2k\delta k$ one obtains from (122) that:

$$\delta k = \frac{\int \left(\rho_1\omega^2 - \mu_1 k^2\right) v^2 dz - \int \mu_1 \left(\partial_z v\right)^2 dz}{2k \int \mu_0 v^2 dz} \ . \tag{125}$$

Since the phase velocity is given by $c = \omega/k$ the phase velocity perturbation follows to leading order from this expression by using the relation $\delta c/c = -\delta k/k$. This means that the first order effect of the perturbation of the Earth model on the phase velocity of Love waves can readily be computed once the unperturbed eigenfunctions $v(z)$ are known. A similar result can be derived for Rayleigh waves.

In general, the perturbation in the phase velocity due to perturbation in the density, P-wave velocity α and S-wave velocity β can be written as:

$$\frac{\delta c}{c} = \int K_\beta(z)\frac{\delta\beta(z)}{\beta(z)}dz + \int K_\alpha(z)\frac{\delta\alpha(z)}{\alpha(z)}dz + \int K_\rho(z)\frac{\delta\rho(z)}{\rho(z)}dz \,. \quad (126)$$

For Love waves the kernel $K_\alpha(z)$ vanishes because Love waves are independent of the bulk modulus. Examples of the kernels K_β, K_α and K_ρ are shown in figure 15 for the fundamental mode Rayleigh waves of periods of $100s$ (left panel) and $30s$ (right panel) respectively. It can be seen that the sensitivity of Rayleigh waves to the P-velocity α is much less than the sensitivity for changes in the shear velocity β and that the sensitivity kernel K_α decays more rapidly with depth than the other kernels. Both phenomena are a consequence of the fact that the compressive component of the motion become evanescent at much shallower depth than the shear component.

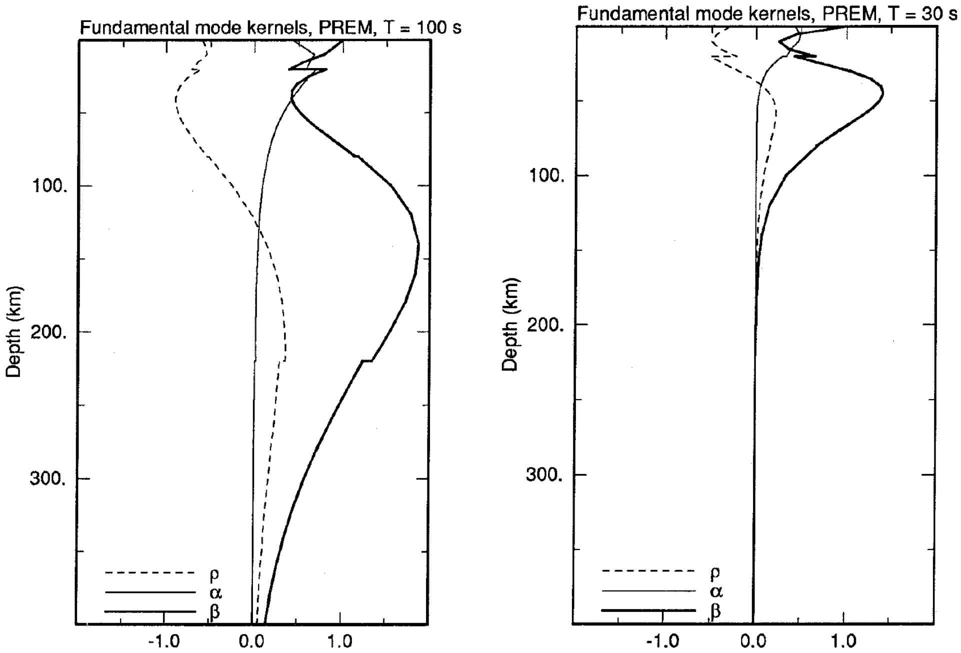

Figure 15. The sensitivity kernels K_α (thin solid line), K_β (thick solid line) and K_ρ (dashed line) for the fundamental mode Rayleigh wave for a period of $100s$ (left panel) and $30s$ (right panel). The kernels are computed for the PREM model [*Dziewonski and Anderson*, 1981].

For larger periods (left panel) the sensitivity kernels penetrate over a greater depth range than for shorter periods (right panel). This makes is it possible to obtain a depth resolution in surface wave inversions. By measuring the phase velocity of surface waves at different periods, one obtains according to (126) and figure 15 the inner product of the perturbation of the Earth model with different weight functions $K(z)$. The theory of linear inversion in section 2 can then be used to determine the perturbation of the Earth model as a function of depth. Of course, only a finite depth resolution can be obtained, but using the expressions (4) or (75) one can determine the depth resolution that can be obtained. The depth resolution can be increased by using higher modes as well as the fundamental mode surface waves and by using both Love waves and Rayleigh waves [*Nolet,* 1977; *Cara*, 1978; *van Heijst and Woodhouse*, 1997].

The theory of this section can be extended to include anisotropic perturbations in the Earth as well [*Tanimoto*, 1986; *Montagner and Nataf*, 1986]. The depth-dependence of the sensitivity kernels of the Rayleigh wave phase velocity show a dependence of the anisotropic P-wave velocity that is fundamentally different than that of the isotropic P-wave velocity [*Muyzert and Snieder*, 1999].

6. Fermat's theorem and seismic tomography

Travel-time tomography is a technique where one aims to reconstruct the velocity structure of a body given the measurement of travel times of waves that have propagated through that body. The travel time along a ray is given by

$$T = \int_{\boldsymbol{r}[u]} u(\boldsymbol{r}) ds \; . \tag{127}$$

In this expression, u is the slowness which is defined as the reciprocal of the velocity: $u = 1/c$. The slowness is used rather than the velocity because now the integrand is linear to the quantity we aim to retrieve. It is tempting to conclude from (127) that the relation between the travel time and the slowness is linear. However, this is wrong! The reason for this is that the integration in (127) is along the path on which the waves travel. The rays are curves of stationary travel time, and hence the ray location depends on the slowness as well. Travel time tomography is thus a nonlinear inverse problem: the unknown slowness is present both in the integrand and it determines the ray position $\boldsymbol{r}[u]$ in the travel time integral (127). Linear inversion techniques can only be used when the relation between T and u is linearized.

Traditionally this is achieved by invoking Fermat's theorem which states that the travel along a ray does not change to first order when this ray is perturbed [e.g. *Nolet,* 1987; *Ben-Menahem and Singh*, 1981]. The proof of Fermat's theorem is based on variational calculus and invokes the equations of kinematic ray tracing. However, the same result can be derived in a much simpler way when one starts from the eikonal equation.

6.1. FERMAT'S THEOREM, THE EIKONAL EQUATION AND SEISMIC TOMOGRAPHY

The derivation given in this section was formulated by *Aldridge* [1994] and is based on a perturbation analysis of the eikonal equation. Here, only the first order travel time perturbation will be derived, but *Snieder and Aldridge* [1995] have generalized the analysis to arbitrary order. Starting point is the eikonal equation which governs the propagation of wavefronts:

$$|\nabla T|^2 = u^2(\mathbf{r}) \; . \tag{128}$$

Consider a reference slowness $u_0(\mathbf{r})$ that is perturbed by a perturbation $\varepsilon u_1(\mathbf{r})$, where the parameter ε serves to facilitate a systematic perturbation approach:

$$u(\mathbf{r}) = u_0(\mathbf{r}) + \varepsilon u_1(\mathbf{r}) \; . \tag{129}$$

Under this perturbation the travel changes, and we assume that the travel time can be written as a regular perturbation series:[8]

$$T = T_0 + \varepsilon T_1 + \varepsilon^2 T_2 + \cdots \tag{130}$$

Inserting (129) and (130) in the eikonal equation (128) and collecting terms proportional to ε^0 and ε^1 gives:

$$|\nabla T_0|^2 = u_0^2(\mathbf{r}) \; , \tag{131}$$

$$(\nabla T_0 \cdot \nabla T_1) = u_0 u_1 \; . \tag{132}$$

The first equation is the eikonal equation for the reference travel time. Let the unit vector $\hat{\boldsymbol{t}}_0$ be directed along the gradient of T_0, then using (131) this gradient can be written as

$$\nabla T_0 = u_0 \hat{\boldsymbol{t}}_0 \; . \tag{133}$$

Taking the inner product of this expression with $\hat{\boldsymbol{t}}_0$ gives $\left(\hat{\boldsymbol{t}}_0 \cdot \nabla T_0\right) = u_0$, which can also be written as

$$\frac{dT_0}{ds_0} = u_0 \; . \tag{134}$$

[8] The assumption that the travel time perturbation is regular is not valid when caustics are present and the relation between the slownes perturbation and the travel time surface is not analytic.

In this expression $d/ds_0 = \hat{\boldsymbol{t}}_0 \cdot \nabla$ is the derivative along the unit vector $\hat{\boldsymbol{t}}_0$. Expression (134) can be integrated to give

$$T_0 = \int_{\boldsymbol{r}_0[u_0]} u_0(\boldsymbol{r}_0) ds_0 \ , \tag{135}$$

where $\boldsymbol{r}_0$ is the position of the ray in the reference slowness field.

Using (133), expression (132) for the travel time perturbation can be written as $\left(\hat{\boldsymbol{t}}_0 \cdot \nabla T_1\right) = u_1$, but since $\hat{\boldsymbol{t}}_0 \cdot \nabla = d/ds_0$ this can also be written as

$$\frac{dT_1}{ds_0} = u_1 \ . \tag{136}$$

This expression can be integrated to give

$$T_1 = \int_{\boldsymbol{r}_0[u_0]} u_1(\boldsymbol{r}_0) ds_0 \ . \tag{137}$$

The main point of this derivation is that the integration in (137) is along the *reference ray* $\boldsymbol{r}_0$ rather than along the true ray in the perturbed medium. Expression (137) constitutes a linearized relation between the travel time perturbation T_1 and the slowness perturbation u_1. When one divides the model in cells where one assumes the slowness perturbation is constant, then the discretized form of (137) can be written as

$$\delta T_i = \sum_j L_{ij} u_j \ . \tag{138}$$

In this expression, the subscript i labels the different travel times that are used in the inversion while j is the cell index. It follows from figure 16 that L_{ij} is the length of ray $\#i$ through cell $\#j$. Equation (138) forms a linear system of equations as discussed in section 2.

The matrix L_{ij} is in general very sparse for tomographic problems because every ray intersects only a small fraction of the cells, see figure 16. This is in particular the case in three dimensions where the relative number of intersected cells is much smaller than in two dimensions. This makes it particularly attractive to use iterative solutions for the linear system of equations as presented in section 2.9.2. It follows from expression (71) that in this iterative approach the solution is constructed by multiplication with the matrices $\boldsymbol{L}$ and $\boldsymbol{L}^T$, but that the inverse $\left(\boldsymbol{L}^T \boldsymbol{L}\right)^{-1}$ (which is not sparse) is not needed. In such an approach there not even a need to store the matrix $\boldsymbol{L}$, it is much efficient to store only the relatively few nonzero matrix elements and keep track of the indices of these elements. This approach

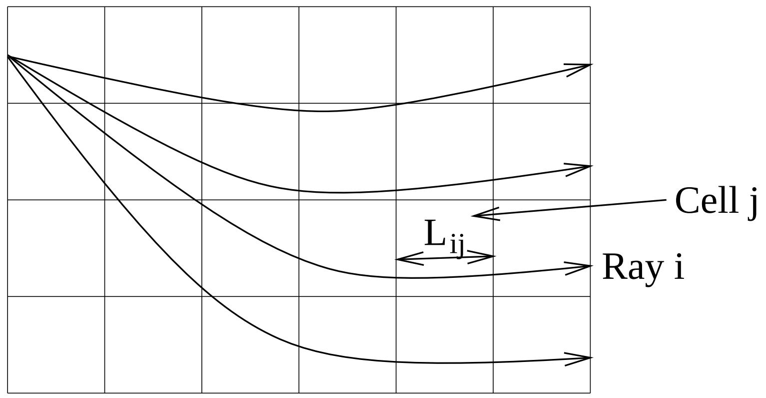

Figure 16. Diagram of a tomographic experiment.

has been developed by *Nolet* [1985] and has been used successfully for extremely large-scale tomographic inversion for the interior of the Earth's mantle [e.g. *Spakman et al.,* 1993; *van der Hilst et al.*, 1997].

It should be remarked that there is no strict need to use cells to discretize the slowness (and the same holds for the treatment of linearized waveform inversion in the sections 4.2 and 4.3). As an alternative the slowness perturbation can be expanded in a finite set of basis functions $B_j(\boldsymbol{r})$:

$$u_1(\boldsymbol{r}) = \sum_{j=1}^{L} m_j B_j(\boldsymbol{r}) \ . \qquad (139)$$

Inserting this in (137) one again arrives at a linear system of the form (138), but the matrix elements are now given by

$$L_{ij} = \int_{ref\ ray\ j} B_j(\boldsymbol{r}) ds_0 \ , \qquad (140)$$

where the integration now is along reference ray $\#j$. However, one should realize that when the basis functions have global character, such as spherical harmonics, the matrix $\boldsymbol{L}$ is in general not sparse with this model parameterization. This means one cannot fully exploit the computational efficiency of iterative solutions of linear systems.

6.2. SURFACE WAVE TOMOGRAPHY

For most people Fermat's principle goes hand in hand with travel time tomography, but surface wave tomography actually relies on it since the early twenties when *Gutenberg* [1924], using data collected by *Tams* [1921], explained the dispersion differences between surface waves propagating along continental and oceanic paths in terms of properties of the Earth's crust.

Surface wave tomography clearly means tomography based on surface waves. There are several ways of doing this. One may directly invert the waveform for structural parameters, provided the sensitivities are known. Generally, the structural parameters are non-linearly related to the waveforms and one way (among others) to linearize the problem has been outlined in section 4.3. A more classical approach, consists of the so-called two-step inversion where one first constructs models of phase or group velocity as a function of frequency. The information contained in these maps is then inverted for depth structure. This is possible because expression (126) gives a liner relation between the phase velocity as a function of frequency and the perturbation of the medium as a function a depth. The mapping from phase velocity to depth then reduces to a linear inverse problem to which the techniques of section 2 can be applied.

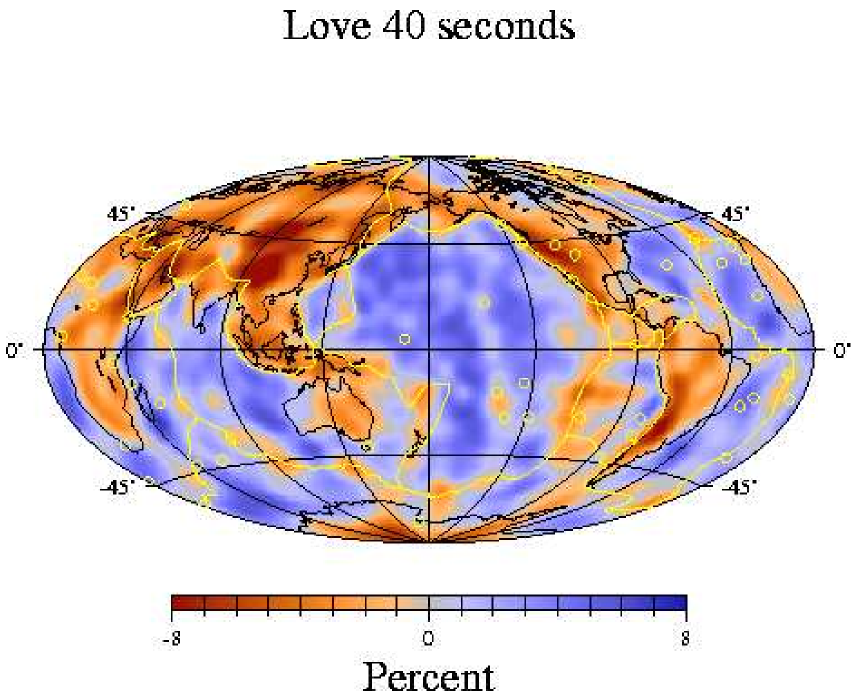

Figure 17. Love wave phase velocity perturbations at a period of 40 seconds. The variations are given in percent with respect to an average reference model. Yellow lines represent plate boundaries and yellow circles are hotspots.

It is this first step which commonly assumes Fermat's principle. If we assume that the Earth structure is sufficiently smooth compared to the wavelength of a propagating wave, locally any wave of frequency ω may be approximated by a plane wave. We are then in the realm of ray theory and are able to measure the mean phase (or group) velocity along the ray path. If furthermore Fermat's principle holds, the measurements correspond to path integrals along the minor and major arc segments of the great circle connecting station and event. It is undoubtedly the case that there are many examples of off-great-circle propagation and non-theoretical effects. The extend to which such effects corrupt the models constructed under the simple assumptions of Fermat's principle can only be determined by numerical modelling, and is still an open question.

It is commonly accepted that Fermat's principle is valid for surface wave periods longer than 150 seconds roughly, and most work until the mid-nineties concentrated on these longer periods. *Trampert and Woodhouse* [1995, 1996] extend these classical methods in two ways. Firstly, they applied it to much shorter periods in the range of 40 to 150 seconds. Secondly, they exploited the by now tremendous wealth of assembled digital data in a systematic fashion by developing fully automatic methods for extracting path-averaged phase velocities of surface waves from recorded waveforms. The most original feature of these studies is the phase velocity maps at short periods which gives new global information on the uppermost structure of the Earth. As an example, the model for the phase velocity of Love waves at a period of 40 seconds is shown in figure 17. The Love-wave model shows a remarkable correlation with surface topography and bathymetry, and hence with crustal thickness. This is too be expected because these waves sample the surface with an exponentially decaying sensitivity. The reference model used in the inversion has a crustal thickness of 24.4 km. If the true Earth has a thicker crust, mantle material of the reference model is replaced with slower crustal material resulting in a slow velocity perturbation. This is the case for the continents. In oceans, the opposite takes place. The sensitivity of 40 second Love waves decays very rapidly with depth and hence tectonic signatures from the uppermost mantle are less pronounced. Nevertheless, a slow velocity anomaly is observed along most oceanic ridges, where hot, and hence slow, material is put into place.

7. Nonlinearity and ill-posedness

Nonlinearity complicates the estimation problem considerably. In many practical problems, nonlinear inversion is treated as a nonlinear optimization problem where a suitably chosen measure of the data misfit is reduced as a function of the model parameters. There is a widespread belief in the inverse problem community that the dominant effect of nonlinearity is the creation of secondary minima in the misfit function that is minimized. This point of view is overly simplistic. Nonlinearity affects both the estimation problem and the appraisal problem. In sections 7.1 and 7.2 it is shown how non-linearity can be a source of ill-posedness of inverse problems. In section 8 the appraisal problem for nonlinear inverse problems is discussed.

7.1. EXAMPLE 1, NON-LINEARITY AND THE INVERSE PROBLEM FOR OF THE SCHRÖDINGER EQUATION

The inverse problem of the estimation of a quantum mechanical potential in one dimension from the measurement of the reflection coefficient of waves reflected by the potential is an interesting tool for studying nonlinear inversion because the inverse problem has a stunningly simple solution [*Marchenko,* 1955; *Burridge,* 1980]. This inverse problem is of direct relevance in the earth sciences; both the inverse problem of geomagnetic induction [*Weidelt,* 1972] as well as the seismic reflection problem [*Burridge,* 1980; *Newton*, 1981] can be reformulated similar to the problem treated in this section. For the Schrödinger equation the wavefield ψ satisfies the following differential equation:

$$\psi_{xx} + \left(k^2 - V(x)\right)\psi = 0 \ . \tag{141}$$

Let the reflection coefficient after a Fourier transform to the time domain be denoted by $R(t)$. The potential $V(x)$ at a fixed location x follows from the reflection coefficient by solving the Marchenko equation:

$$K(x,t) + R(x+t) + \int_{-t}^{x} K(x,\tau)R(\tau+t)d\tau = 0 \ , \tag{142}$$

for $K(x,t)$ and carrying out a differentiation:

$$V(x) = -2\frac{dK(x,x)}{dx} \ . \tag{143}$$

In figure 18 the results of a numerical solution of the Marchenko equation is shown. The potential is shown by a thick solid line. For this potential the reflection coefficient is computed and the potential is reconstructed by numerically solving the Marchenko equation (142) and carrying out the differentiation (143), details of the calculation can be found in [*Dorren et al.,* 1994]. The reconstructed potential of figure 18 (top panel) is on the scale of the figure indistinguishable from the true potential. In figure 18 (bottom panel) the same synthetic experiment is shown, the only difference is that the potential has been multiplied with a factor 2.5. If the problem would be linear, the reflection coefficients would be 2.5 times as strong as for the potential in the top panel and the reconstructed potential would also be 2.5 times as strong. This would lead to a near-perfect reconstruction of the potential. However, the reconstructed potential is given by the dashed line.

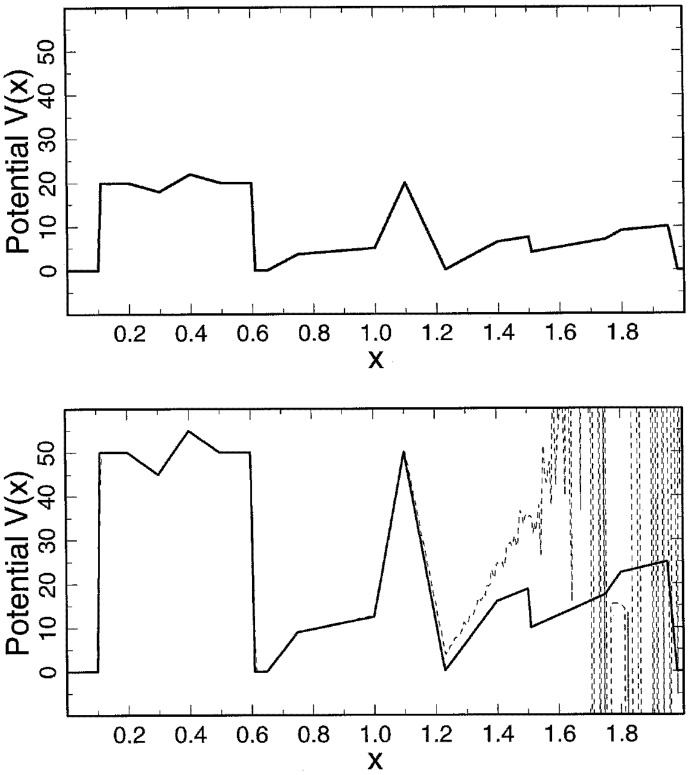

Figure 18. The original potential (solid lines) and the potential reconstructed with the Marchenko equation (dashed lines). The potential in the bottom panel is 2.5 times as strong as the potential in the top panel.

It can be seen that the left part of the potential (the side from which the waves are incident) the potential is reconstructed quite well, but that the part of the potential on the right is very poorly reconstructed. According to the reasoning above, the potential would have been reconstructed quite well if the problem had been linear. This implies that the instability in the reconstructed potential is due to the non-linearity of the problem.

The physical reason for this instability can relatively easy be understood. It follows from the Schrödinger equation (141) that the effective wavenumber is given by $\sqrt{k^2 - V(x)}$. When $k^2 < V(x)$ the wavenumber is complex which reflects the fact that the waves are evanescent when the potential energy is larger than the total energy. In that case the wavefield decays exponentially within the potential. For a fixed energy k^2 the wavefield is more evanescent for the potential in the bottom panel of figure 18 than for the potential in the top panel of that figure, simply because the potential energy is 2.5 times as high. This has the result that the wave-field penetrates further in the potential in the top panel than in the potential in the bottom panel of figure 18. Obviously, the potential in a certain region is not constrained by the recorded wavefield if the wavefield does not sample the potential in that region. The strong evanescence of the waves in the potential in the bottom panel of figure 18 imlpies that that parts of that potential are effectively not sampled by the waves. In that case, the numerical details of the algorithm, including the numerical round-off error determine the reconstructed potential in that region. The essential point is that the values of the model parameters affect the way in which the wavefield interrogates the model.

Physically, the instability in the reconstructed potential in figure 18 can thus be understood. However, what does this imply for the inverse problem? What happens physically if the potential on the left side is high, is that the wavefield is prevented from sampling the potential on the right part. This means that for some values of the model parameters (that describe how high the potential is on the left), other model parameters (that describe the potential on the right) are unconstrained by the data. In terms of a misfit function this implies that the misfit does not depend on the model parameters that describe the right side of the potential (when the left side of the potential is high). In other words, the misfit function does not have a minimum, but is has a broad plateau. Note that as an additional complexity this only occurs for certain values of other model parameters (that describe how high the potential is on the left).

7.2. EXAMPLE 2, NON-LINEARITY AND SEISMIC TOMOGRAPHY

A second example of the ill-posedness introduced by the non-linearity in inverse problems is seismic tomography. Consider a cross-borehole tomographic experiment where rays travel from a source in a well to a string of receivers on another well. The case where the velocity is homogeneous is shown in the left panel of figure 19. In that case the rays are straight lines that travel from the source on the left to receivers $R1$ through $R5$ on the right. If the velocity is not homogeneous, the rays are curved. This implies that the value of the model parameters determine the way in which the rays interrogate the model.

Suppose that a low-velocity anomaly is present, see the middle panel of figure 19. Since the rays of first arrivals are curves of minimal travel time, the rays curve around the slow-velocity anomaly. If the velocity anomaly is sufficiently slow, all the rays may completely curve around the slow-velocity. In that situation the anomaly would not be sampled by any rays and the velocity within the anomaly cannot be determined because it is not sampled by any rays. At best one could derive an upper bound for

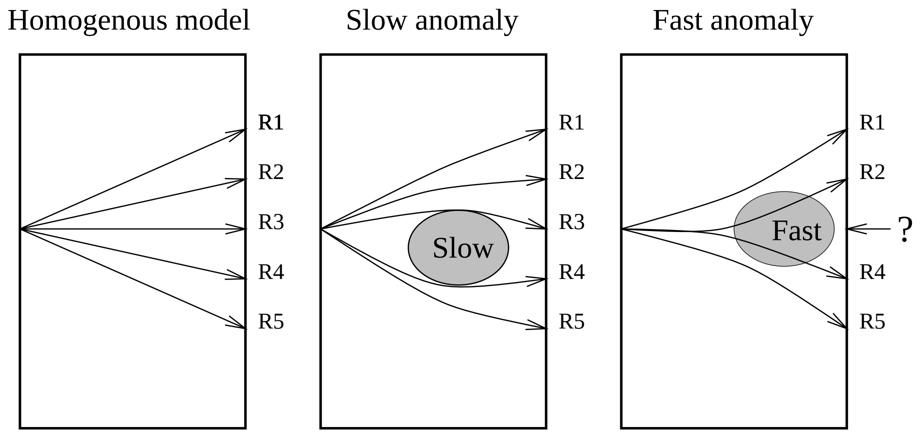

Figure 19. (a) Tomographic experiment where the velocity is homogeneous and the rays are straight. (b) Tomographic experiment where the rays curve around a low-velocity body. (c) Tomographic experiment where a high-velocity anomaly causes a shadow zone at the middle receiver.

the velocity within the anomaly. Just as in the previous section, the model parameters affect the way in which the probe (in this case the rays) samples the model. In terms of the misfit function this implies that for a certain range of model parameters, the misfit function does not depend on the model parameters at all, in other words: the misfit function in completely flat over an extended range of parameter values.

Let us now consider the opposite situation where a high-velocity anomaly is present, see the right panel of figure 19. Rays will be defocused by a high velocity body and a shadow-zone is formed behind the anomaly. This may mean that there is no ray that will hit the receiver $R3$, see the question mark in figure 19. This means that for some values of the model parameters it is impossible to compute the data because the travel time cannot be determined when there is no ray that hits the receiver. This means that for some values of the model parameters it is impossible to compute the corresponding data values given the theory that one uses. In this sense one can say that some values of the model parameters are "forbidden." It should be noted that this is not a complexity created in some exotic thought-experiment; the problem of the "missing rays" is a real problem in seismic travel time tomography [*Sambridge,* 1990], as well as the tomographic imaging of the temperature field in ovens [*Natterer et al.*, 1997].

The critical reader might remark at this point that the fact that for certain values of the velocity model there are no rays hitting the receiver is due to the fact that ray theory is an approximation to a true theory (wave theory), and that wave theory predicts that some energy will diffract into the shadow-zones. Although this is correct, one should also note that the wavefield in the shadow zones is very weak (that is why they are called shadow zones) and that in practice this diffracted wavefield is usually not detectable.

8. Model appraisal for nonlinear inverse problems

In the previous section the effect of non-linearity on model estimation has been discussed. In this section an attempt is made to describe the effect of non-linearity on the appraisal problem where one wants to describe how the estimated model is related to the true model (see figure 2). However, it should be stressed that there is presently no general theory to deal with the appraisal problem for a truly nonlinear inverse problem with infinitely many degrees of freedom. In practice, one often linearizes the problem around the estimated model and then uses linear theory to make inferences about the resolution and reliability of the estimated model. The lack of a general theory for the appraisal problem should be seen as a challenge for theorists! In this section three attempts are described to carry out model assessment for a nonlinear inverse problem. These attempts follow the lines of formal theory (section 8.1), a numerical approach (section 8.2) and a pragmatic approach (section 8.3).

8.1. NONLINEAR BACKUS-GILBERT THEORY

Linear Backus-Gilbert theory is based on the equations (72)-(74) for linear inverse problems for continuous models. The model estimate $\tilde{m}(x)$ at location x is constructed by making the linear superposition (73) of the data. The resolution kernel in equation (74) specifies the relation between the estimated model and the true model. In the ideal case the resolution kernel is a delta function. *Backus and Gilbert* [1967, 1968] have shown that the criterion that the resolution kernel should resemble a delta function as much as possible can be used to determine the coefficients $a_i(x)$ in equation (73) which prescribes how a datum d_i affects the estimated model at location x.

Backus-Gilbert theory has been generalized by *Snieder* [1991] for the special case that the forward problem can be written as a perturbation series:

$$d_i = \int G_i^{(1)}(x)m(x)dx + \iint G_i^{(2)}(x_1, x_2)m(x_1)m(x_2)dx_1dx_2 + \cdots \quad (144)$$

In a number of applications such a perturbation series arises naturally. Important examples are the Neumann series (101) in scattering theory where the scattering data are written as a sum of integrals that contain successively higher powers of the model perturbation, or ray perturbation theory where the travel time of rays is written as a sum of integrals with increasing powers of the slowness perturbation [e.g. *Snieder and Sambridge,* 1993; *Snieder and Aldridge,* 1995]. When the forward problem is nonlinear, the inverse problem is non-linear as well, this suggest that for a non-linear inverse problem the linear estimator (73) should be generalized to include terms that are nonlinear in the data as well:

$$\tilde{m}(x) = \sum_i a_i^{(1)}(x) d_i + \sum_{i,j} a_{ij}^{(2)}(x) d_i d_j + \cdots \tag{145}$$

The key of non-linear Backus-Gilbert theory is to insert the expansion of the forward problem (144) in the estimator (145). The result can then be written as:

$$\tilde{m}(x) = \int R^{(1)}(x; x_1) m(x_1) dx_1 + \iint R^{(2)}(x; x_1, x_2) m(x_1) m(x_2) dx_1 dx_2 + \cdots \; . \tag{146}$$

This expression generalizes the linear resolution kernel of equation (74) to nonlinear inverse problems. The kernel $R^{(1)}(x; x_1)$ describes to what extent the estimated model is a blurred version of the true model. The higher order kernels such as $R^{(2)}(x; x_1, x_2)$ can be interpreted as nonlinear resolution kernels that describe to what extent there is a spurious nonlinear mapping from the estimated model onto the true model in the inversion process.

In the ideal case, the estimated model is equal to the true model: $\tilde{m}(x) = m(x)$. This is the case when the linear resolution kernel $R^{(1)}(x; x_1)$ is a delta function $\delta(x - x_1)$ and when the nonlinear resolution kernels are equal to zero: $R^{(n)}(x; x_1, \cdots, x_n) = 0$ for $n \geq 2$. However, as in equation (75) the linear resolution kernel $R^{(1)}(x; x_1)$ can be written as a sum of a finite amount of data kernels $G^{(1)}(x_1)$. Since a delta function cannot be obtained by summing a finite amount of smooth functions, the linear resolution kernel can never truly be a delta function. This reflects the fact that with a finite amount of data the estimated model will be a blurred version of the true model. *Snieder* [1991] treats the inverse problem of the determination of the mass-density of a vibrating string from the eigenfrequencies of the string. He shows that if only a finite amount of eigenfrequencies are available, the nonlinear resolution kernels cannot be zero. This implies that the finiteness of the data set leads to a spurious nonlinear mapping from the true model to the estimated model. This can be related to the symmetries of the problem and to mode coupling "off the energy shell." The finite width of the linear resolution kernel and the fact that the nonlinear resolution kernels are nonzero imply not only that the estimated model is a blurred version of the true model, but also that the estimated model is biased. The reader is referred to *Snieder* [1991] for details. In that work it is also described how the coefficients $a^{(i)}$ in the estimator (145) can be determined.

Although nonlinear Backus-Gilbert theory is a new tool for dealing with the assessment problem for nonlinear inverse problems, one should realize that the theory can only be applied to weakly nonlinear problems where a (very) few orders are sufficient for an accurate description of both the forward and the inverse problem. In addition, the theory of *Snieder* [1991] is so complex that a reformulation is needed to make the theory applicable to the large-scale inverse problems that are being treated in practice.

It is important to realize that any regular (nonlinear) mapping from data d_i to an estimated model $\tilde{m}(x)$ can be written in the form of expression (145). The details of the employed algorithm then determine the coefficients $a^{(n)}_{i_1 \cdots i_n}$. The resolution analysis shown in equation (146) and the subsequent discussion therefore is applicable to the estimated model. The conclusions concerning the linear and nonlinear resolution kernels can thus be used for *any* regular mapping from the data to the estimated model.

8.2. GENERATION OF POPULATIONS OF MODELS THAT FIT THE DATA

Another approach to asses the reliability of estimated models is to generate not a single model that fit the data within a certain tolerance but to obtain an ensemble of models that fits the data within a certain tolerance [e.g. *Lomax and Snieder*, 1995]. An alternative approach is to compute the misfit for a very large class of models and to use the data fit, possibly in combination with Bayesian statistics to make inferences about the range of models that explain the data in a certain sense [*e.g. Mosegaard and Tarantola*, 1995; *Gouveia and Scales,* 1998, *Mosegaard,* 1998]. Obviously, this approach requires a numerical approach to create such ensembles, but with present day computers significance progress has been made.

An important concept in the generation of ensembles of models is the randomness in the search method that one employs. A descent method contains no element of randomness whatsoever, whereas a Monte Carlo search where one randomly samples model space is completely random. In between are algorithms which have both, a random component as well as a mechanism to prefer models that fit the data well. Examples of such algorithms are simulated annealing [*Krikpatrick et al.,* 1983*; Rothman,* 1985] or genetic algorithms [*Sambridge and Drijkoningen,* 1992; *Sen and Stoffa,* 1992; *Lomax and Snieder,* 1995]. A promising technique is the adaptive search [*Mosegaard and Tarantola,* 1995] where in the process of carrying out a random search, information about the misfit function is being built up, and this misfit function is then used to drive the random search in an intelligent way.

The main merit of algorithms that determine an ensemble of models together with information on how well each model explains the data is that this ensemble can be used to make inferences about the model. These inferences may or may not be statistical. The possibility to analyse such ensembles of models in a meaningful way has not yet been fully exploited.

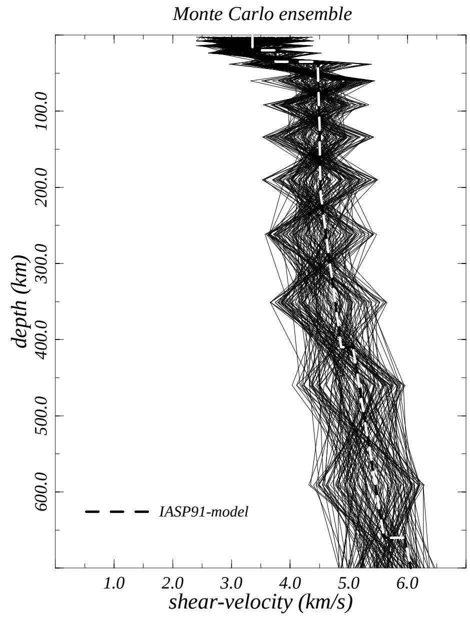

Figure 20. The true velocity model (broken curve) and an ensemble of models generated by a Monte Carlo search that fit surface wave group velocity data with a realistic tolerance (solid lines).

An example is shown from a study of *Douma et al.* [1996]. In their study synthetic group velocity data of the fundamental mode Rayleigh wave that propagates along the earth's surface were computed for periods between 10 and 300 s The true velocity model is shown as the dashed line in figure 20 as a function of depth. A Monte Carlo search was used to find models of the S-velocity that were consistent with the data within a realistic measurement error. The resulting population of models is shown in figure 20. In this study, the S-velocity was deliberately over-parameterized. As a result, the resulting models are highly oscillatory and contain strong trade-offs. The aim of the study of *Douma et al.* [1996] was to extract the robust features of the velocity model from the ensemble of models shown in figure 20. This was achieved by computing "Empirical Orthogonal Functions" (EOF's) from this ensemble. These functions give the patterns with the different degrees of variability within an ensemble.

The EOF's can be used to re-parameterize the model space in an intelligent way that reflects how well model perturbations are constrained by the data. Alternatively, the EOF's could be used to carry out a statistical analysis of the ensemble of models that explains the data. However, as noted by *Douma et al.* [1996] the EOF techniques is only useful for inverse problems that are weakly nonlinear.

8.3. USING DIFFERENT INVERSION METHODS

In the previous sections, a theoretical and a numerical method for model appraisal were presented. Apart from these more formal approaches, "common sense" is a powerful tool for carrying out model assessment. An important way to asses the reliability of a model is to determine the model in different ways. In the ideal case, different data sets are used by different research groups who use different theories to estimate the same properties. The agreement of disagreement between these models can be used as an indicator of the reliability of these models. It is admittedly difficult to quantify the sense of reliability that is thus obtained, but in the absence of an adequate theory to carry out model assessment for nonlinear inverse problems (and remember that we don't have such a theory) this may be the best approach.

An example of this approach is shown in figure 21. Nonlinear waveform inversion using the Partitioned Waveform Inversion of *Nolet* [1990] has been used by *van der Hilst and Kennett* [1997] to determine a three-dimensional model of the S-velocity under Australia. A cross section of the model at a depth of 140 km is shown by the colors in figure 21. In such an inversion, averages of the earth structure along long paths are used to determine the local variations in the S-velocity. Effects such as errors in the source mechanisms of the employed earthquakes, seismic anisotropy that is not accounted for in the inversion, variations in the path coverage and the path direction over the domain of inversion can lead to artifacts in the inversion. As a way of verifying the reliability of the solution, *Passier et al.* [1997] determined the local S-velocity using the waveform inversion method described by *Passier and Snieder* [1995]. In this method, a consistency requirement of the waveforms recorded in two nearby stations is used to estimate the horizontally averaged velocity between the stations as a function of depth. Just as the technique used by *van der Hilst and Kennett* [1997] this technique may lead to models with artifacts, but in general these artifacts will be different, see *Passier et al.* [1997] for a discussion.

The interstation S-velocity models are shown in the panels in figure 21. In these panels the depth range is from the surface down to a depth of 250 km, whereas the S-velocity ranges from -5% to +5%. The symbol

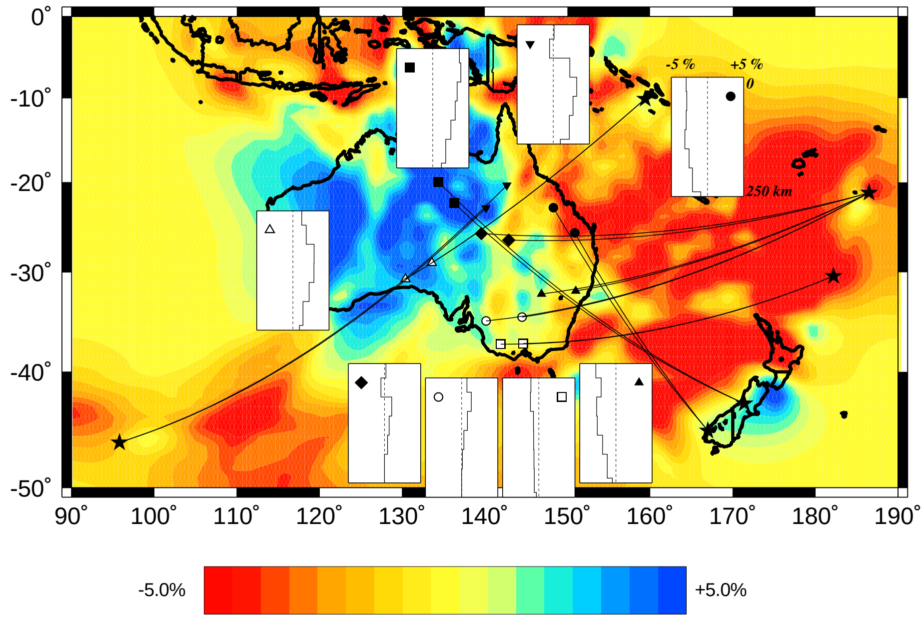

Figure 21. The S-velocity anomaly under Australia at a depth of 140 km obtained by surface wave tomography by *v*an der Hilst and Kennett [1997] where the relative velocity perturbation is shown by the colour bar at the bottom. The interstation S-velocity obtained by *P*assier et al. [1997] is shown in the panels. The S-velocity in panels ranges from -5% to +5% and is shown from the surface down to a depth of 250 km. The employed stations for each panel are indicated that corresponds with the symbol in the upper-right corner of each panel.

in each panel (circle, triangle, etc.) corresponds to the employed stations that are marked on the map with the same symbol. The depth-profile in each panel shows the S-velocity between the corresponding stations. By comparing the interstation velocity models shown in the panels with the velocity model shown in colors one can see that the locally determined S-velocity agrees well with the S-velocity obtained from the 3D tomographic inversions. Consider for example the S-velocity between the stations in the northeast marked with solid dots that is shown in the panel with the solid dot. At a depth of about 140 km the interstation shear velocity is about -2.5%. This agrees well with the S-velocity obtained from 3D tomography that is indicated with the colors.

A comparison as presented in this section is very useful for assessing the reliability of models in a qualitative way. In this sense, a healthy competition between different research groups is an important ingredient in the appraisal of models. There is one pitfall in the approach to compare models obtained from different data sets and different inversion methods; it is possible that both models are in error but that they agree with each other. The comparison of models may thus provide a false impression of their reliability. However, in practice this test often is a useful pointer to artefacts in (at least one of) the models. An example of this approach is given by Passier et al. [1997].

9. Epilogue

Linear inverse problem theory is an extremely powerful tool for solving inverse problems. Much of the information that we currently have on the Earth's interior is based on linear inverse problems. The success of modern-day oil exploration for an affordable price is to a large extent possible because of our ability to use imaging techniques that are based on single scattering theory to map oil reservoirs. One could argue that the modern world, which heavily relies on cheap access to hydrocarbons, might have a drastically different form if single scattering would not explain seismic data so well.

Despite the success of linear inverse theory, one should be aware that for many practical problems our ability to solve inverse problems is largely confined to the estimation problem. This may surprise the reader, because for linear inverse problems the resolution kernel and the statistics of linear error propagation seem to be adequate tools for carrying out the model assessment. However, as noted in section 2, in order to compute the resolution kernel one needs to explicitly define the generalized inverse $\boldsymbol{A}^{-g}$. For many problems, this entails inverting the matrix $\boldsymbol{A}^T\boldsymbol{A}$ or an equivalent matrix that contains regularization terms as well. This task is for many important problem, such as imaging seismic reflection data, not feasible.

The results of section 8 show that for nonlinear inverse problems both the estimation problem and the appraisal problem are much more difficult. In the context of a data-fitting approach, the estimation problem amounts to a problem of nonlinear optimization. Researchers can use the results of this field of mathematics to solve the estimation problem. However, the reader should be aware of the fact that there is presently no theory to describe the appraisal problem for nonlinear inverse problems. Developing a theoretical framework for the appraisal of models obtained from nonlinear inversion of data is a task of considerable complexity which urgently needs the attention of the inverse problem community.

Acknowledgments: Discussions with John Scales and Jean-Jacques Lévêque have helped to clarify a number of issues, we appreciate their input very much.

References

1. Aldridge, D.F., Linearization of the eikonal equation, *Geophysics, 59*, 1631-1632, 1994.
2. Alsina, D., R.L. Woodward, and R.K. Snieder, Shear-Wave Velocity Structure in North America from Large-Scale Waveform Inversions of Surface Waves, *J. Geophys. Res., 101*, 15969-15986, 1996.
3. Aki, K., and P.G. Richards, *Quantitative Seismology (2 volumes)*, Freeman and Co. New York, 1980.
4. Backus, G., and J.F. Gilbert, Numerical applications of a formalism for geophysical inverse problems, *Geophys. J.R. Astron. Soc., 13*, 247-276, 1967.
5. Backus, G., and J.F. Gilbert, The resolving power of gross earth data, *Geophys. J.R. Astron. Soc., 16*, 169-205, 1968.
6. Backus, G. E. and F. Gilbert, Uniqueness in the inversion of inaccurate gross earth data, *Philos. Trans. R. Soc. London, Ser. A, 266*, 123-192, 1970.
7. Ben-Menahem, A. and S.J. Singh, *Seismic waves and sources*, Springer Verlag, New York, 1981.
8. Borg, G., Eine Umkehrung der Sturm-Liouvillischen Eigenwertaufgabe, Bestimmung der Differentialgleichung durch die Eigenwerte, *Acta Math., 78*, 1-96, 1946.
9. Burridge, R., The Gel'fand-Levitan, the Marchenko and the Gopinath-Sondi integral equations of inverse scattering theory, regarded in the context of the inverse impulse response problems, *Wave Motion, 2*, 305-323, 1980.
10. Cara, M., Regional variations of higher-mode phase velocities: A spatial filtering method, *Geophys. J.R. Astron. Soc.,* 54, 439-460, 1978.
11. Claerbout, J.F., *Fundamentals of Geophysical data processing*, McGraw-Hill, New York, 1976.
12. Claerbout, J.F., *Imaging the Earth's interior*, Blackwell, Oxford, 1985.
13. Clayton, R. W. and R. P. Comer, A tomographic analysis of mantle heterogeneities from body wave travel time data, *EOS, Trans. Am. Geophys. Un., 64*, 776, 1983.
14. Constable, S.C., R.L. Parker, and C.G. Constable, Occam's inversion: a practical algorithm for generating smooth models from electromagnetic sounding data, *Geophysics, 52*, 289-300, 1987.
15. Dahlen, F.A., and J. Tromp, *Theoretical global seismology*, Princeton University Press, Princeton, 1998.
16. Dorren, H.J.S., E.J. Muyzert, and R.K. Snieder, The stability of one-dimensional inverse scattering, *Inverse Problems, 10*, 865-880, 1994.
17. Douma, H., R. Snieder, and A. Lomax, Ensemble inference in terms of Empirical Orthogonal Functions, *Geophys. J. Int., 127*, 363-378, 1996.
18. Dziewonski, A.M., and D.L. Anderson, Preliminary Reference Earth Model, *Phys. Earth. Plan. Int., 25*, 297-356, 1981.
19. Gerver, M.L. and V. Markushevitch, Determination of a seismic wave velocity from the travel time curve, *Geophys. J. Royal astro. Soc., 11* 165-173, 1966.
20. Gilbert, F., Ranking and winnowing gross Earth data for inversion and resolution, *Geophys. J. Royal astro. Soc., 23* 125-128, 1971.
21. Gouveia, W.P., and J.A. Scales, Bayesian seismic waveform inversion: parameter estimation and uncertainty analysis, *J. Geophys. Res., 103*, 2759-2779, 1998.
22. Gutenberg, B., Dispersion und Extinktion von seismischen Oberflächenwellen und der Aufbau der obersten Erdschichten, *Physikalische Zeitschrift, 25*, 377-382, 1924.
23. Herglotz, G. Über das Benndorfsche Problem des Fortpflanzungsgeschwindigkeit der Erdbebenstrahlen, *Zeitschrift fúr Geophys., 8* 145-147, 1907.
24. Keller, J.B., I. Kay, and J. Shmoys, Determination of a potential from scattering data, *Phys. Rev., 102*, 557-559, 1956.
25. Kircpatrick, S., C. Gelatt, and M.P. Vechhis, Optimization by simulated annealing, *Science, 220*, 671-680, 1983.
26. Lanczos, C., *Linear Differential Operators,* Van Nostrand, London, 1961.

27. Levenberg, K., A method for the solution of certain nonlinear problems in least squares, *Quart. Appl. Math., 2*, 164-168, 1944.
28. Lomax, A., and R. Snieder, The contrast in upper-mantle shear-wave velocity between the East European Platform and tectonic Europe obtained with genetic algorithm inversion of Rayleigh-wave group dispersion, *Geophys. J. Int., 123*, 169-182, 1995.
29. Marchenko, V.A., The construction of the potential energy from the phases of scattered waves, *Dokl. Akad. Nauk, 104*, 695-698, 1955.
30. Matsu'ura M. and N. Hirata, Generalized least-squares solutions to quasi-linear inverse problems with a priori information, *J. Phys. Earth, 30*, 451-468, 1982.
31. Mayer, K., R. Marklein, K.J. Langenberg and T.Kreutter, Three-dimensional imaging system based on Fourier transform synthetic aperture focussing technique, *Ultrasonics*, 28, 241-255, 1990.
32. Menke, W., *Geophysical data analysis: discrete inverse theory*, Academic Press, San Diego, 1984.
33. Merzbacher, E., *Quantum mechanics (2nd ed.)*, Wiley, New York, 1970.
34. Montagner, J.P., and H.C. Nataf, On the inversion of the azimuthal anisotropy of surface waves, *J. Geophys. Res., 91*, 511-520, 1986.
35. Mosegaard, K., Resolution analysis of general inverse problems through inverse Monte Carlo sampling, *Inverse Problems, 14*, 405-426, 1998.
36. Mosegaard, K., and A. Tarantola, Monte Carlo sampling of solutions to inverse problems, *J. Geophys. Res., 100*, 12431-12447, 1995.
37. Muyzert, E., and R. Snieder, An alternative parameterization for surface waves in a transverse isotropic medium, *Phys. Earth Planet Int. (submitted)*, 1999.
38. Natterer, F., H. Sielschott, and W. Derichs, Schallpyrometrie, in *Mathematik - Sclüsseltechnologie für die Zukunft*, edited by K.H. Hoffmann, W. Jäger, T. Lochmann and H. Schunk, 435-446, Springer Verlag, Berlin, 1997.
39. Newton, R.G., Inversion of reflection data for layered media: A review of exact methods, *Geophys. J.R. Astron. Soc., 65*, 191-215, 1981.
40. Newton, R.G., *Inverse Schrödinger scattering in three dimensions*, Springer Verlag, Berlin, 1989.
41. Nolet, G., The upper mantle under Western-Europe inferred from the dispersion of Rayleigh wave modes, *J. Geophys., 43*, 265-285, 1977.
42. Nolet, G., Linearized inversion of (teleseismic) data, in *The Solution of the Inverse Problem in Geophysical Interpretation*, edited by R.Cassinis, Plenum Press, New York, 1981.
43. Nolet, G., Solving or resolving inadequate and noisy tomographic systems, *J. Comp. Phys., 61*, 463-482, 1985.
44. Nolet, G., Seismic wave propagation and seismic tomography, in *Seismic Tomography*, edited by G.Nolet, pp. 1-23, Reidel, Dordrecht, 1987.
45. Nolet, G., Partitioned waveform inversion and two-dimensional structure under the Network o f Autonomously Recording Seismographs, *J. Geophys. Res., 95*, 8499-8512, 1990.
46. Nolet, G., S.P. Grand, and B.L.N. Kennett, Seismic heterogeneity in the upper mantle, *J. Geophys. Res., 99*, 23753-23766, 1994.
47. Nolet, G, and R. Snieder, Solving large linear inverse problems by projection, *Geophys. J. Int., 103*, 565-568, 1990.
48. Paige, C.G., and M.A. Saunders, LSQR: An algorithm for sparse linear equations and sparse least-squares, *ACM Trans. Math. Software, 8*, 43-71, 1982.
49. Paige, C.G., and M.A. Saunders, LSQR: Sparse linear equations and least-squares problems, *ACM Trans. Math. Software, 8*, 195-209, 1982.
50. Parker, R.L., *Geophysical Inverse Theory*, Princeton University Press, Princeton, New Jersey, 1994.
51. Passier, M.L., and R.K. Snieder, Using differential waveform data to retrieve local S-velocity structure or path-averaged S-velocity gradients, *J. Geophys. Res., 100*,

24061 - 24078, 1995.
52. Passier, M.L., and R.K. Snieder, Correlation between shear wave upper mantle structure and tectonic surface expressions: Application to central and southern Germany, *J. Geophys. Res., 101*, 25293-25304, 1996.
53. Passier, T.M., R.D. van der Hilst, and R.K. Snieder, Surface wave waveform inversions for local shear-wave velocities under eastern Australia, *Geophys. Res. Lett, 24*, 1291-1294, 1997.
54. Press, W.H., Flannery, B.P., Teukolsky, S.A. and W.T. Vetterling, *Numerical Recipies*, Cambridge University Press, Cambridge, 1989.
55. Rothman, D.H., Nonlinear inversion, statistical mechanics and residual statics estimation, *Geophysics, 50*, 2784-2796, 1985.
56. Sabatier, P.C., Discrete ambiguities and equivalent potentials, *Phys. Rev. A, 8*, 589-601, 1973.
57. Sambridge, M., Non-linear arrival time inversion: constraining velocity anomalies by seeking smooth models in 3-D, *Geophys. J.R. Astron. Soc., 102*, 653-677, 1990.
58. Sambridge, M., and G. Drijkoningen, Genetic algorithms in seismic waveform inversion, *Geophys. J. Int., 109*, 323-342, 1992.
59. Scales, J., and R. Snieder, To Bayes or not to Bayes?, *Geophysics, 62*, 1045-1046, 1997.
60. Scales, J., and R. Snieder, What is noise?, *Geophysics, 63*, 1122-1124, 1998.
61. Sen, M.K., and P.L. Stoffa, Rapid sampling of model space using genetic algorithms: examples of seismic wave from inversion, *Geophys. J. Int., 198*, 281-292, 1992.
62. Sluis, A. van der, and H.A. van der Vorst, Numerical solution of large, sparse linear algebraic systems arising from tomographic problems, in *Seismic tomography, with applications in global seismology and exploration geophysics*, edited by G. Nolet, Reidel, Dordrecht, 1987.
63. Snieder, R., 3D Linearized scattering of surface waves and a formalism for surface wave holography, *Geophys. J. R. astron. Soc., 84*, 581-605, 1986a.
64. Snieder, R., The influence of topography on the propagation and scattering of surface waves, *Phys. Earth Planet. Inter., 44*, 226-241, 1986b.
65. Snieder, R., Surface wave holography, in *Seismic tomography, with applications in global seismology and exploration geophysics*, edited by G. Nolet, pp. 323-337, Reidel, Dordrecht, 1987.
66. Snieder, R., Large-Scale Waveform Inversions of Surface Waves for Lateral Heterogeneity, 1, Theory and Numerical Examples, *J. Geophys. Res., 93*, 12055-12065, 1988.
67. Snieder, R., Large-Scale Waveform Inversions of Surface Waves for Lateral Heterogeneity, 2, Application to Surface Waves in Europe and the Mediterranean, *J. Geophys. Res., 93*, 12067-12080, 1988.
68. Snieder, R., A perturbative analysis of nonlinear inversion, *Geophys. J. Int., 101*, 545-556, 1990.
69. Snieder, R., The role of the Born-approximation in nonlinear inversion, *Inverse Problems, 6*, 247-266, 1990.
70. Snieder, R., An extension of Backus-Gilbert theory to nonlinear inverse problems, *Inverse Problems, 7*, 409-433, 1991.
71. Snieder, R., Global inversions using normal modes and long-period surface waves, in *Seismic tomography*, edited by H.M. Iyer and K. Hirahara, pp. 23-63, Prentice-Hall, London, 1993.
72. Snieder, R., and D.F. Aldridge, Perturbation theory for travel times, *J. Acoust. Soc. Am., 98*, 1565-1569, 1995.
73. Snieder, R.K., J. Beckers, and F. Neele, The effect of small-scale structure on normal mode frequencies and global inversions, *J. Geophys. Res., 96*, 501-515, 1991.
74. Snieder, R., and A. Lomax, Wavefield smoothing and the effect of rough velocity perturbations on arrival times and amplitudes, *Geophys. J. Int., 125*, 796-812, 1996.
75. Snieder, R., and G. Nolet, Linearized scattering of surface waves on a spherical

Earth, *J. Geophys., 61*, 55-63, 1987.
76. Snieder, R., and M. Sambridge, The ambiguity in ray perturbation theory, *J. Geophys. Res., 98*, 22021-22034, 1993.
77. Spakman, W., S. Van der Lee, and R.D. van der Hilst, Travel-time tomography of the European-Mediterranean mantle down to 1400 km, *Phys. Earth Planet. Int., 79*, 3-74, 1993.
78. Strang, *Linear algebra and its applications*, Harbourt Brace Jovanovisch Publishers, Fort Worth, 1988.
79. Takeuchi, H. and M. Saito, Seismic surface waves, in *Seismology: Surface waves and earth oscillations,* (Methods in computational physics, 11), Ed. B.A. Bolt, Academic Press, New York, 1972.
80. Tams, E., 1921. Über Fortplanzungsgeschwindigkeit der seismischen Oberflächenwellen längs kontinentaler und ozeanischer Wege, *Centralblatt für Mineralogie, Geologie und Paläontologie, 2-3*, 44-52, 1921.
81. Tanimoto, T., Free oscillations in a slighly anisotropic earth, *Geophys. J.R. Astron. Soc., 87*, 493-517, 1986.
82. Tarantola, A., Linearized inversion of seismic reflection data, *Geophys. Prosp., 32*, 998-1015, 1984.
83. Tarantola, A., *Inverse problem theory*, Elsevier, Amsterdam, 1987.
84. Tarantola, A. and B. Valette, Inverse problems = quest for information, *J. Geophys., 50*, 159-170, 1982a.
85. Tarantola, A., and B. Valette, Generalized nonlinear inverse problems solved using the least squares criterion, *Rev. Geophys. Space Phys., 20*, 219-232, 1982b.
86. Trampert, J., Global seismic tomography: the inverse problem and beyond, *Inverse Problems, 14*, 371-385, 1998.
87. Trampert, J., and J.J. Lévêque, Simultaneous Iterative Reconstruction Technique: Physical interpretation based on the generalized least squares solution, *J. Geophys. Res.., 95*, 12553-12559, 1990.
88. Trampert, J., J.J. Lévêque, and M. Cara, Inverse problems in seismology, in *Inverse problems in scattering and imaging*, edited by M. Bertero and E.R. Pike, pp. 131-145, Adam Hilger, Bristol, 1992.
89. Trampert, J., and R. Snieder, Model estimations based on truncated expansions: Possible artifacts in seismic tomography, *Science, 271*, 1257-1260, 1996.
90. Trampert, J., and J.H. Woodhouse, Global phase velocity maps of Love and Rayleigh waves between 40 and 150 seconds, *Geophys. J. Int., 122*, 675-690, 1995.
91. Trampert, J., and J.H. Woodhouse, High resolution global phase velocity distributions, *Geophys. Res. Lett., 23*, 21-24, 1996.
92. VanDecar, J.C., and R. Snieder, Obtaining smooth solutions to large linear inverse problems, *Geophysics, 59*, 818-829, 1994.
93. van der Hilst, R.D., S. Widiyantoro, and E.R. Engdahl, Evidence for deep mantle circulation from global tomography, *Nature, 386*, 578-584, 1997.
94. van der Hilst, R.D., and B.L.N. Kennett, Upper mantle structure beneath Australia from portable array deployments, *American Geophysical Union 'Geodynamics Series', 38*, 39-57, 1998.
95. van Heijst, H.J. and J.H. Woodhouse, Measuring surface wave overtone phase velocities using a mode-branch stripping technique, *Geophys. J. Int.,* 131, 209-230, 1997.
96. Weidelt, P., The inverse problem of geomagnetic induction, *J. Geophys., 38*, 257-289, 1972.
97. Wiechert, E., Über Erdbebenwellen. I. Theoretisches über die Ausbreitung der Erdbebenwellen, *Nachr. Ges. Wiss. Göttingen*, Math.-Phys. Klasse, 415-529 1907.
98. Woodhouse, J. H. and A. M. Dziewonski, Mapping the upper mantle: Three dimensional modelling of Earth structure by inversion of seismic waveforms, *J. Geophys. Res, 89*, 5953-5986, 1984.

99. Yilmaz, O., Seismic data processing, *Investigations in geophysics, 2*, Society of Exploration Geophysicists, Tulsa, 1987.