Inverse problems in a nutshell

M. Sambridge

Centre for Advanced Data Inference, Research School of Earth Sciences, Australian National University, ACT 0200, Australia

Abstract

We describe features of inverse problems and illustrate them with simple examples. The focus is on the main concepts and caveats rather than mathematical detail. Properties of linear and nonlinear inverse problems are discussed, and the effect of non-uniqueness highlighted. Iterative algorithms for nonlinear problems are briefly introduced and comments on their performance included. Fully nonlinear direct search algorithms are also mentioned. An example is given in which the performance of a direct search and an iterative optimization algorithm is compared.

Keywords: Inverse theory; Parameter estimation; Inference techniques

Introduction

Science is driven by the feedback between predictions and observations. Most of our knowledge of the Earth's interior comes from analyzing data collected at the surface. Therefore observations are almost always of an indirect nature, and there exists an 'inverse' problem to extract information about the deep interior, e.g. by building an image and "seeing" into the Earth. Our understanding of all major features within the Earth, such as the crust, mantle, liquid outer core and solid inner core, as well as the most dynamic parts of the interior such as subduction zones and mantle plumes, came about from the study of indirect measurements made at the surface. Indirect data, such as seismic, magnetic and gravity surveys are also a key tool in the search for hydrocarbon deposits as well as in understanding the contemporary plate tectonic environment, which helps quantify the risk earthquakes and other natural hazards pose to population centres and infrastructure. A key question, which is frequently asked is 'How do we extract reliable information from multi-faceted and complex geophysical data sets ?' and 'What confidence can be placed in conclusions drawn from those data sets ?'. As in many areas of the sciences, the difficulty lies as much in finding the right questions to ask, as in finding answers. Inverse theory is the name given to the study of extracting information from indirect measurements. It provides an incomplete set of mathematical, statistical and computational techniques for solving such problems. This article will explain some basic concepts, describe popular trends, and hopefully encourage the reader to look further into the subject.

Inverse problems

Inverse problems are not restricted to the geo-sciences and arise in nearly all scientific disciplines (although often under different names), indeed anywhere that data only indirectly constrain quantities of interest. An incomplete list includes medical imaging, astronomy, engineering, remote sensing, oceanography, environmental science and more recently bio-informatics. The development of practical and robust schemes for analyzing indirect data is therefore of major concern in the physical sciences. Since the earth sciences is a natural laboratory for the study of difficult data analysis problems, it forms an important crucible for the development of inversion techniques that can then be applied to other fields. A good example is helio-seismology, a technique for imaging the interior of the sun using observations in intensity fluctuations caused by sun quakes. The development of this field in the 1990s, was apparently inspired by geophysicists use of long period seismic surface waves and free oscillations to image the outer portions of the Earth, a process known as seismic tomography. (Note the geophysicists themselves stole the idea and the name from the physics who pioneered medical tomography in the early 1960s and 1970s).

Inversion techniques are often used in cases where a large amount of data are available. In many areas of the physical sciences both the quantity and quality of data has grown rapidly over the past few years. We now stand at a unique point in history where advances in instrumentation, digital storage capacity and communication speed, have increased our ability to collect and disseminate data at rates never before seen. In the earth sciences we are now able to measure the age of the smallest fragment of a rock with high precision; as well as record the precise motions of tectonic plates on a global scale. A student can sit in almost any part of the world and download high fidelity recordings of earthquake seismograms within hours of the event happening.

Linear problems

In the geosciences linear inverse problems were the first to be studied in detail. A linear inverse problem arises when the mathematical relationship between observables (e.g. electromagnetic measurements made in an aircraft) and unknowns (e.g. subsurface electrical conductivity structure of the Earth) are linear, or assumed to be linear. Pioneering work on linear inverse problems was carried out by Backus and Gilbert (1967, 1968, 1970). They considered linear inverse problems in their most general form, with the unknowns represented by continuous functions of space, rather than a discrete set of parameters. They broke inverse problems up into two parts, known as the existence problem 'Does any model exist which fits the available data ?', and the uniqueness problem 'If so, how unique is that model ?'. Backus and Gilbert showed that there exists a fundamental trade-off between the model variance (the error in unknown model value at any point in a medium) and the model resolution (the degree to which the spatial averaging or blurring occurs). In addition many inverse problems were recognized as non-unique, meaning thatan infinite class of solutions exist, each fitting the data equally well. Without extra data or introducing new assumptions there is no reason why one single model should be preferred over any other. A classic paper which explains the essential properties of linear inverse problems is Parker (1977).

A discrete inverse problem

A number of general concepts can be illustrated with a simple discrete linear inverse problem, such as seismic travel time tomography. Figure 1 shows an example. In Figure 1a we have a single seismic ray passing through a two parameter block model. The unknowns are the changes in slownesses (reciprocal of seismic velocity) in the blocks from a homogeneous slowness model $(\Delta S_1, \Delta S_2)$. The single datum, Δt_1, is the difference between the observed travel time of the ray and that calculated in the reference (homogeneous) slowness model. The question is can we find the slownesses in the blocks ? Linearization of the seismic travel time equation gives

$$\Delta t_1 = l_1 \Delta S_1 + l_2 \Delta S_2, \tag{1}$$

where l_j is the length of the ray in the j-th block. Clearly we have one linear equation and two unknowns, and hence no unique solution. In fact there is a complete trade-off between the slowness variables, and only the average slowness of the two blocks is constrained by the data. Superficially the situation in Figure 1b looks much better, but this is deceptive. Regardless of the fact that we now have many rays traversing the blocks, it's straightforward to show that each ray contributes an equation which is just a scalar multiple of (1). Hence we really still have only one independent equation, even though we now have many rays. If the data are error free the situation in Figure 1b is identical to that in Figure 1a and no extra information is present. (Note: If the data contained noise then there would be a benefit from the averaging effect of the rays, but the again only the average slowness would be constrained.) The situation in Figure 1c is entirely different. Here we have two equations that pass through different sides of the blocks. Hence the ratio of the lengths is no longer equal

$$\frac{l_1}{l_2} \neq \frac{l_3}{l_4}. \tag{2}$$

This means that the two constraint equations, are linearly independent, and hence can be solved uniquely for the slownesses $(\Delta S_1, \Delta S_2)$. The situation in Figure 1d, is an improvement of 1c, in that now many rays from different directions are present. Here we again can solve for the two unknowns and would be likely to do so with less variance in the model parameters when the data contained noise (again due to the combined effect of the many rays). In the parlance of linear algebra, Figure 1a and 1b lead to an under-determined linear system of equations; 1c to an even determined; and 1d to an overdetermined. In Figure 1e we have the same rays as in 1d but have chosen to use more blocks to represent the slowness field. Close inspection shows that all the previous cases hold simultaneously in 1e and hence this is what is known as a mixed determined problem, containing both over and under determined parameters. Comparing 1d to 1e we see that as we seek to constrain slowness variations over shorter spatial scales our model

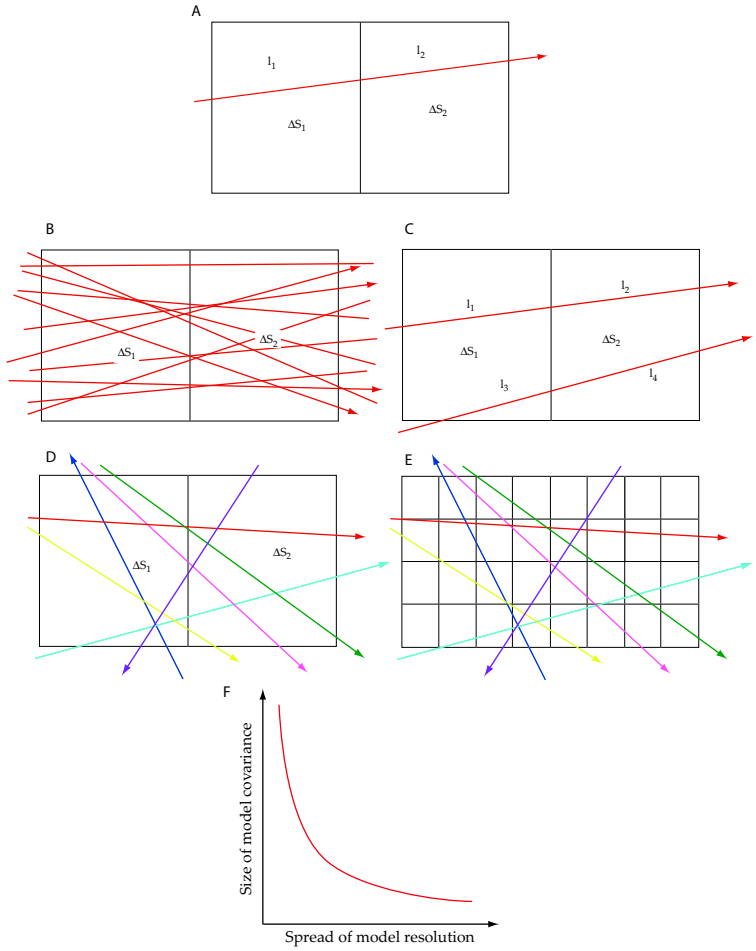

Figure 1: *A toy linear discrete inverse problem, travel time tomography. a)-d), show various numbers and orientations of rays passing through two slowness blocks. e) shows the same rays as in d) but with 32 blocks. e) illustrates the trade off between spread of model resolution (inverse cell size) and model variance (noise in the estimated slownesses). See text for a discussion.*

variance increases. This is because with noisy data even the best constrained slowness blocks in Figure 1e will have larger error than the two unknowns in 1d. Hence there is a trade-off between model variance (error) and spread of resolution (inverse of cell size). Figure 1e illustrates this trade-off, which is a general property of all linear inverse problems.

This simple linear example illustrates many important concepts of inverse problems, namely that the mere number of data is unimportant, but the (linear) independence of the data matters most (c.f. Figures 1a, 1b and 1c); that non-uniqueness is often present and parametrization is a choice; that problems are often mix-determined; and that a trade-off between resolution and variance can not be avoided. For discrete inverse problems with numbers of unknowns less than 10^3, it is practical to calculate quantities like the model covariance matrix and the model resolution matrix, which characterize these properties (see Menke, 1989; Tarantola, 2005; Aster et al., 2005, for details).

Nonlinear inverse problems

A nonlinear inverse problem arises when the mathematical relationship between unknown model and data takes the form

$$\mathbf{d} = \mathbf{g}(\mathbf{m}) \tag{3}$$

where $\mathbf{g}(\mathbf{m})$ represents the nonlinear forward model. This includes the case where no analytical expression for the forward problem exists, and $\mathbf{g}$ represents the result of an algorithm allowing data to be calculated for any given input model. To estimate the model $\mathbf{m}$ which generated the observations (3), the problem is often recast as one of

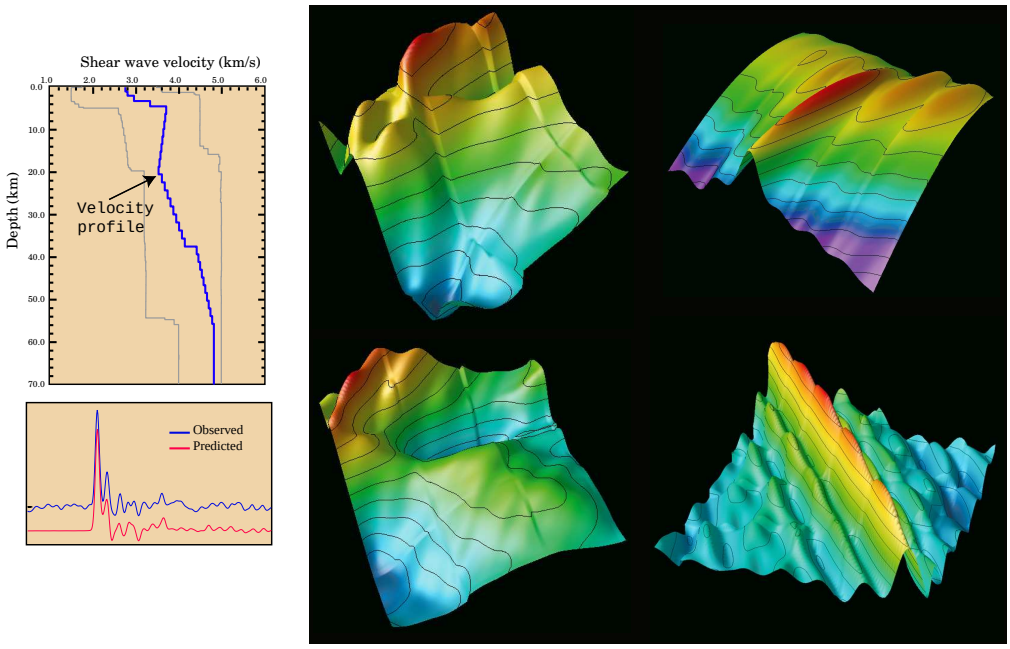

Figure 2: *A least squares data misfit for fitting of seismic receiver functions. In each of the four images the data misfit is plotted as a function of two of the twenty-four parameters defining the 1-D seismic wave-speed model. The multi-modal character of the data misfit arises because of the nonlinear relationship between model parameters and data.*

optimization. Specifically we seek the vector m which minimizes

$$\phi(\mathbf{m}) = (\mathbf{d} - \mathbf{g}(\mathbf{m}))^T C_D^{-1} (\mathbf{d} - \mathbf{g}(\mathbf{m}) + \psi(\mathbf{m}) \tag{4}$$

The first term is a least square measure of data misfit, with C_D^{-1} the covariance matrix for the data errors (Menke, 1989). The second is a regularization term designed to penalize extravagant models. The particular form of $\psi(\mathbf{m})$ is subjective. Its role is to encourage, but not always guarantee, a unique global minimum for $\phi(\mathbf{m})$. Common choices are a quadratic difference to a reference model, a derivative operator to penalize first or second spatial derivatives, (e.g. for the case when $\mathbf{m}$ represents a scalar field of some physical quantity); or a combination of all three. Many inverse problems in the geosciences are nonlinear. Figure 2 shows an example from seismic receiver function inversion (Ammon et al., 1990; Shibutani et al., 1996). The four surfaces represent the least squares data misfit between predicted and observed receiver functions (cf. the data term in (4)), plotted as a function of two model parameters. Here the model represents the shear wave-speed as a function of depth in the Earth's crust and upper mantle. Note that the surfaces are far from quadratic, indicating severe nonlinearity. They also show multi-modality making optimization difficult.

The simplest way of dealing with nonlinear inverse problems is to linearize them about some reference model $\mathbf{m}_o$ using a Taylor expansion. This leads to

$$\delta\mathbf{d} = G\delta\mathbf{m} \tag{5}$$

where $\delta\mathbf{d}$ is the difference between the observed data and the predictions from the reference model ($\mathbf{d} - g(\mathbf{m}_o)$); $\delta\mathbf{m}$ is the perturbation to the reference model, $\mathbf{m}_o$, and G is a matrix of derivatives of predictions with respect to model parameters. Once this approximation has been made, linear inverse theory can be used in an iterative fashion. For example, if the regularization term in (4) were a simple damping to a reference model we have

$$\psi(\mathbf{m}) = (\mathbf{m} - \mathbf{m}_o)^T C_M^{-1} (\mathbf{m} - \mathbf{m}_o), \tag{6}$$

where C_M^{-1} is a model covariance (weight) matrix. Then the well known damped least squares iterative algorithm results

$$\delta\mathbf{m} = (G^T C_D^{-1} G + C_M^{-1})^{-1} G^T C_D^{-1} \delta\mathbf{d}. \tag{7}$$

From the point of view of minimizing (4), the algorithm in (7) is equivalent to approximating the misfit function $\phi(\mathbf{m})$ with a local quadratic surface and moving to its minimum. Repeated iterations may converge to a minimum of ϕ or not, depending on the starting point of the process. Figure 3 shows an example. In the left panel the starting point lies in the basin of a broad minimum and repeated quadratic approximations will guide the solution to it. In the right hand panel the starting point lies outside of the basin. In this case an iterative optimization algorithm is likely to diverge

or find a secondary minimum. Iterative linearization algorithms always suffer from this type of problem. The field of gradient based optimization (e.g. Gill et al., 1981) has many techniques for damping and stabilizing such algorithms, but ultimately these methods become ineffective as the problem becomes more nonlinear, or indeed when surface derivatives (determined by the G matrix) are unavailable.

Parameter search and ensemble inference

In the examples presented we have seen two important properties of many inverse problems, namely non-uniqueness and non-linearity. These together with the presence of noise in the data tell us that it is much more important to characterize the range of acceptable solutions than finding a single best data fitting model. (Even though, as we have seen, in nonlinear cases it may be difficult to even find best data fitting models.) Figure 3 illustrates the case where multiple classes of solution exist, each fitting the data adequately. If these are disconnected as in the illustration, then no linearized iterative inversion algorithm will find more than one class of solution at any one time.

One approach in these circumstances is to avoid all linearizing approximations entirely and use only ensemble inference approaches, i.e. ones based on directly searching a parameter space (without making use of derivatives). If a 'good' sample of data fitting models can be generated then conclusions can be drawn on properties of them all, rather than the best fit model. This requires repeated testing of potential solutions, often requiring adaptive, i.e. nonuniform randomized sampling of a multi-dimensional parameter space. A convenient way to view ensemble inference is in terms of a two step approach, consisting of a search stage, where models are generated and the forward problem repeatedly solved, and an appraisal stage, where inferences are drawn from the complete ensemble of models obtained (Snieder and Trampert, 1999).

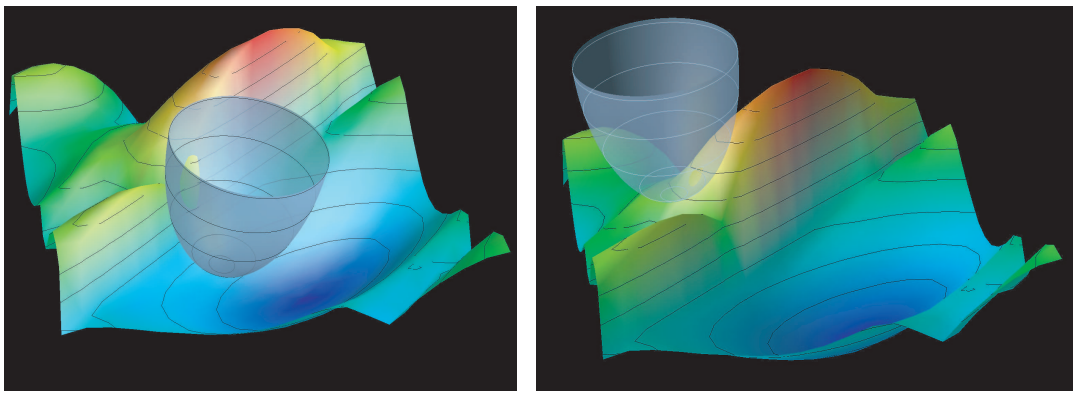

Figure 3: *Both views show a linearized approximation to an objective function arising in a nonlinear inverse problem. The quadratic (tangent) approximation in the left hand case (a) lies in the basin of the global minimum and an iterative linearized inversion algorithm is likely to converge to the deep well to the right. In the right hand case the quadratic approximation lies outside the basin of the global minimum and a linearized inversion algorithm is likely to fail. This illustrates the dependence of linearized optimization algorithms on the starting model.*

In the search stage, an algorithm (based on multi-dimensional random walks) is often used to collect samples, and the predictions of the models are compared to the data. In many cases the search process is adaptive, i.e. it makes use of samples collected to guide the search for new models. Clearly, there is a strong connection with optimization problems, and indeed many direct search (i.e. non-derivative based) optimization algorithms have been used as search algorithms in inverse problems. Examples include Simulated Annealing (Kirkpatrick et al., 1983; Mosegaard and Vestergaard, 1991; Koren et al., 1991; Sen and Stoffa, 1991), Genetic algorithms and Evolutionary Programming (Fogel et al., 1966; Holland, 1975; Sen and Stoffa, 1991; Sambridge and Drijkoningen, 1992), and the Neighbourhood algorithm (Sambridge, 1999).

In the appraisal stage the objective is to make use of the complete ensemble of parameter space samples to draw inferences from the data. Of course, the degree to which this can be sensibly done depends crucially on the type of sampling performed during the search stage. If an optimization algorithm has been used in the search stage to minimize

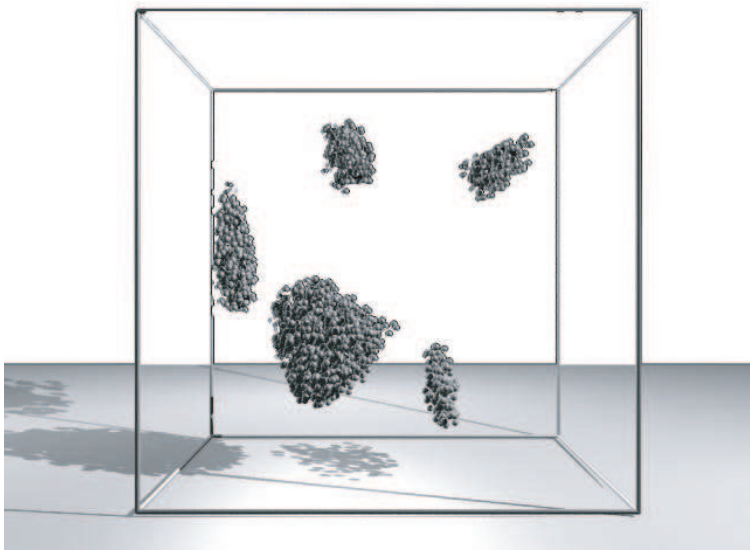

Figure 4: *An ensemble of solutions to an inverse problem. Each sphere represents a three parameter model which satisfies the corresponding data to an acceptable level (given the noise). The figure illustrates the situation in a nonlinear inverse problem where disconnected islands of solutions can exist in parameter space, each representing a different class of solution.*

a data misfit function, then the temptation is often to select the best data fitting model only and examine it in detail. However, this is almost always insufficient because of the noise in the data and the non-uniqueness of the underlying inverse problem. Even within a finite dimensional parameter space one usually finds that if one model fits the data to an acceptable level (given the noise), then an infinite number will, and so the best data fit model may well be misleading.

An alternative to taking single 'best fit' models is to try and characterize the subset of data acceptable models in the collected ensemble, which may be useful if the search algorithm has sufficiently explored the parameter space. For example one could try and detect properties, or features, which all data acceptable models share, e.g. ones which have the least structure, or can be bounded by some model property. This extremal model approach was first proposed by Parker (1977) in the context of nonlinear inverse problems, and can also be applied directly to the appraisal problem. Summaries of approaches used in the geosciences can be found in Sen and Stoffa (1995); Mosegaard and Sambridge (2002).

As the parameter spaces get large direct search techniques loose their appeal and become impractical, because the space becomes extremely large. To get an idea of how rapidly the size of a parameter space grows with dimension, it is sobering to consider a simple thought experiment. Imagine we have just a two dimensional parameter space, and want to generate at least one model in each of the four quadrants, i.e. north-east, south-east, etc. The minimum number of samples needed is obviously four. Apply the same idea to a d dimensional space, we see that there are 2^d, analogous 'quadrants'. For a three parameter space this gives the expected value of eight. For a ten parameter problem there are over a thousand corners to every unit cube, for twenty parameters this is over a million and for thirty parameters it's over a billion. The curse of dimensionality always gets you in the end !

Since the computational effort of direct search inversion will scale linearly with the cost of solving the forward problem it becomes increasingly difficult to adequately sample higher dimensional spaces. In cases where the forward problem is complex, requiring advanced numerical simulation techniques for a single solution, direct search techniques will quickly become prohibitive. Severe under-sampling always occurs, and we can never be sure that the vast oceans of unexplored parameter space might not contain data acceptable models, or even global minima. Nevertheless direct search techniques can still be useful in moderately sized problems (up to a hundred unknowns or so).

To illustrate the influence of nonlinearity and the need for derivative free search techniques, we conclude with a comparison between three algorithms on the seismic receiver function problem. Here the objective is to seek out at least one (and preferably many) models that satisfy the data to an acceptable level. Figure 5 shows the results of an iterative local optimization algorithm, Powell's direction set method (see Press et al., 1992); a direct search method, the Neighbourhood algorithm (Sambridge, 1999); and a simple uniform random search.

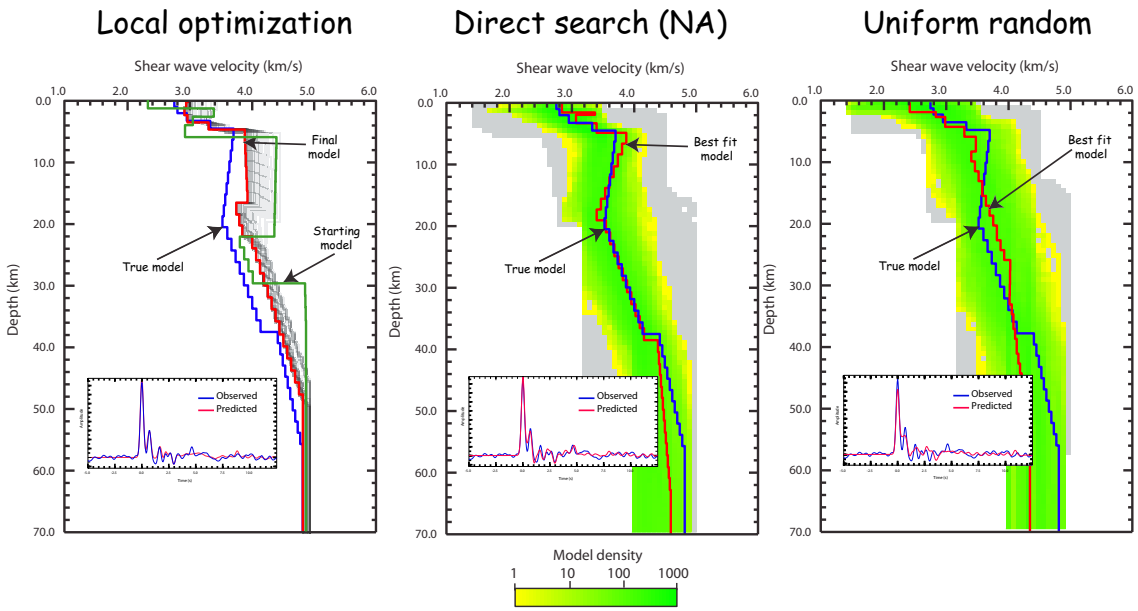

Figure 5: *A comparison between (a) an iterative optimization algorithm (Powell's method) which uses local derivative approximations, (b) a direct search technique (Neighbourhood algorithm) which uses no derivatives, and (c) a uniform random search. In each case the best fit model is shown together with its corresponding receiver function waveform. The synthetic 'observed' data were generated by the same true model shown in the figures, with the addition of correlated Gaussian noise. The spread of the models generated by each algorithm is represented by the background density plot.*

The first method is local because it generates estimates of derivative information of the misfit surface using solutions to the forward problem. It works by successively updating a starting model in a perturbative fashion, and halts when the model is no longer improving the data fit sufficiently. The use of only local information makes it descend quickly to a local minimum where it gets trapped. The third method is derivative free but makes no use of past models to guide sampling. Hence it is robust against entrapment in secondary minima, but very slow to converge. The direct search Neighbourhood algorithm also avoids use of derivatives but is able to make use of previous sampling to guide the search. This is achieved using geometrical techniques to adaptively partition the space. Like most direct search algorithms it performs better than the other two classes of approach in that it is able to adapt sampling in response to the character of the varying misfit function. Interestingly enough, although the iterative local search technique is essentially 'downhill' it still requires 22480 evaluations of the misfit surface (and hence solutions to the forward problem). This is a significant fraction of the 64128 forward solutions used by both the Neighbourhood algorithm and the uniform Monte Carlo sampling. Contrary to common perceptions, techniques using local derivative information, either by way of actual derivative calculation (matrix G in (5), or by misfit surface evaluations (as in Powell's direction set method), can be surprisingly costly.

Another potentially useful feature of modern direct search techniques is that they can be driven by just a ranking of the models with respect to complex criteria. This means that one is not forced to define a scalar objective function like (4) and then rank with respect to these values (although this is common practice). In principle any ranking criteria can be used, even non-differentiable functions and human decisions (see Boschetti and Moresi, 2001). This is true of the Neighbourhood algorithm and some of the more modern implementations of Genetic algorithms.

Conclusions

In this tutorial article we have highlighted some aspects of inverse problems and illustrated them with simple examples. Linear and nonlinear cases have been described, and pitfalls and perils have been touched upon. As the available computational power has grown, much attention has been focused on fully nonlinear problems and direct search algorithms. However it must be remembered that the issues of non-uniqueness, trade-offs, effect of data noise and influences of parametrization choices are just as relevant today as 30 years ago when they were first explained. In the author's view, inverse problems are as much about asking *the right questions of a data set* than building a model that fits it. After all,

as statistician Samuel Karlin put it 'The purpose of models is not to fit the data but to sharpen the questions'[1].

This article has not mentioned the probabilistic (Bayesian) approach to inverse problems that is popular across the sciences. The subject is discussed in detail by Tarantola (2005), and the reader is refered to that work for more information. We hope that this article is a useful summary. More extensive dicussions appear in the major books that have been published on geophysical inverse theory (see Menke, 1989; Parker, 1994; Tarantola, 2005; Aster et al., 2005), which present the subject from a range of viewpoints.

Acknowldgments

My thanks to Nick Rawlinson who read an earlier version of this manuscript and made useful suggestions.

References

C. J. Ammon, G. E. Randall, and G. Zandt. On the nonuniqueness of receiver function inversions. *J. Geophys. Res.*, 95:15,303–15,318, 1990.

R. Aster, B. Borchers, and C. H. Thurber. *Parameter estimation and inverse problems*, volume 90 of *International Geophysics Series*. Elsevier, Amsterdam, 2005.

F. Boschetti and L. Moresi. Interactive inversion in geosciences. *Geophys. Prosp.*, 64:1226–1235, 2001.

L. J. Fogel, A. J. Owens, and M. J. Walsh. *Artificial Intelligence through simulated evolution*. John Wiley and Sons, New York, 1966.

P. E. Gill, W. Murray, and M. H. Wright. *Practical optimization*. Academic Press, London, 1981.

J. H. Holland. *Adaptation in natural and artificial systems*. The university of Michigan press, Ann Arbor, 1975.

S. Kirkpatrick, C. D. Gelatt, Jr, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983.

Z. Koren, K. Mosegaard, E. Landa, P. Thore, and A. Tarantola. Monte Carlo estimation and resolution analysis of seismic background velocities. *J. Geophys. Res.*, 96(B12):20,289–20,299, 1991.

W. Menke. *Geophysical data analysis: discrete inverse theory*. Academic Press, San Diego, rev. ed edition, 1989.

K. Mosegaard and M. Sambridge. Monte Carlo analysis of inverse problems. *Inverse Problems*, 18:R29–R54, 2002.

K. Mosegaard and P. Vestergaard. A simulated annealing approach to seismic model optimization with sparse prior information. *Geophys. Prosp.*, 39:599 – 611, 1991.

R. L. Parker. Understanding inverse theory. *Ann. Rev. Earth Planet. Sci.*, 5:35 – 64, 1977.

R. L.. Parker. *Geophysical inverse theory*. Princeton University Press, Princeton, 1994.

W. H. Press, S. A. Tuekolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in FORTRAN*. Cambridge University Press, Cambridge, 1992.

M. Sambridge. Geophysical inversion with a Neighbourhood algorithm -I. Searching a parameter space. *Geophys. J. Int.*, 138:479 – 494, 1999.

M. Sambridge and G. Drijkoningen. Genetic algorithms in seismic waveform inversion. *Geophys. J. Int.*, 109:323–342, 1992.

M. K. Sen and P. L. Stoffa. Nonlinear one-dimensional seismic waveform inversion using simulated annealing. *Geophysics*, 56:1624 – 1638, 1991.

M. K. Sen and P. L. Stoffa. *Global optimization methods in geophysical inversion*, volume 4 of *Advances in Exploration Geophysics*. Elsevier, Amsterdam, 1995.

T. Shibutani, M. Sambridge, and B. L. N. Kennett. Genetic algorithm inversion for receiver functions with application to crust and uppermost mantle structure beneath eastern australia. *Geophys. Res. Lett.*, 23:1829 – 1832, 1996.

R. Snieder and J. Trampert. Inverse problems in geophysics. In A. Wirgin, editor, *Wavefield Inversion*, pages 119–190. Springer Verlag, New York, 1999.

A. Tarantola. *Inverse Problem Theory and Methods for Model Parameter Estimation*. Siam, Philadelphia, 2005.

[1] 11th R. A. Fisher Memorial Lecture, Royal Society 20, April 1983.