Geophys. J. Int. (1999) **138,** 36–44

Explicit, approximate expressions for the resolution and *a posteriori* covariance of massive tomographic systems

Guust Nolet,* Raffaella Montelli and Jean Virieux
Geosciences Azur, 250 *Rue A. Einstein,* 06560 *Valbonne, France*

Accepted 1999 February 12. Received 1999 February 8; in original form 1998 March 16

SUMMARY

We present an approximate method to estimate the resolution, covariance and correlation matrix for linear tomographic systems $Ax=b$ that are too large to be solved by singular value decomposition. An explicit expression for the approximate inverse matrix A^- is found using one-step backprojections on the Penrose condition $AA^- \approx I$, from which we calculate the statistical properties of the solution. The computation of A^- can easily be parallelized, each column being constructed independently.

The method is validated on small systems for which the exact covariance can still be computed with singular value decomposition. Though A^- is not accurate enough to actually compute the solution x, the qualitative agreement obtained for resolution and covariance is sufficient for many purposes, such as rough assessment of model precision or the reparametrization of the model by the grouping of correlating parameters. We present an example for the computation of the complete covariance matrix of a very large ($69\,043 \times 9610$) system with 5.9×10^6 non-zero elements in A. Computation time is proportional to the number of non-zero elements in A. If the correlation matrix is computed for the purpose of reparametrization by combining highly correlating unknowns x_i, a further gain in efficiency can be obtained by neglecting the small elements in A, but a more accurate estimation of the correlation requires a full treatment of even the smaller A_{ij}. We finally develop a formalism to compute a damped version of A^-.

Key words: inverse theory, statistical methods, tomography.

INTRODUCTION

Seismic tomography is playing an increasingly large role in the study of the Earth and its dynamic behaviour. Tomographic images now assist us in understanding, amongst other things, the deep structure of continents, the details of the subduction process, and magma upwelling under ocean ridges and volcanoes. As the relevance of these seismological investigations grows for other Earth Science disciplines, it becomes important to deal with a fundamental shortcoming of all tomographic imaging: the non-uniqueness of the solution. The model resulting from an inversion is just one member of a subspace of models that satisfy the data equally well or better. Since the choice of the 'preferred' model in the subspace invariably involves a damping of ill-resolved aspects of the model, whereas well-resolved characteristics are more or less fixed, such damping usually reveals a strong influence of the ray path coverage in tomographic images.

* On leave from: Department of Geosciences, Princeton University, Princeton, NJ 08540, USA. E-mail: guust@geo.princeton.edu

The non-uniqueness of the solution can be characterized by its *resolution* and its *variance*, usually represented by the resolution matrix and the (*a posteriori*) covariance matrix. For small-scale problems, these matrices can be calculated conveniently using a singular value decomposition of the problem (Wiggins 1972; Jackson 1972). For larger problems this becomes impractical or downright impossible. The resolution can still be investigated using sensitivity tests (e.g. Spakman & Nolet 1988). Such tests have shortcomings (Leveque *et al.* 1993), but an even greater disadvantage is that such tests measure the sensitivity only with respect to a fixed pattern of cells (e.g. a checkerboard test), and the estimation of the resolution of single cells requires the repetition of many sensitivity tests. Furthermore, no satisfactory method exists to find the *a posteriori* covariance of the solution, other than adding random errors to the sensitivity tests and estimating the covariance matrix from the results of many such tests (Kennett & Nolet 1978), a practice too laborious to have found general acceptance. Techniques such as 'jacknifing' or 'bootstrapping' (Tichelaar & Ruff 1989) rely on the overdetermined nature of an inverse problem and should never be applied to an underdetermined system of equations. Their use on large mixed

over/underdetermined problems such as found in tomography is not only highly questionable but also computationally very expensive.

Recently, the estimation of the resolution matrix from the first few Ritz vectors (approximate eigenvectors) resulting from a Lanczos-type iteration on the linear system has been proposed (Zhang & McMechan 1995, 1996). Such schemes are seriously flawed unless the number of Ritz vectors approaches the effective rank of the matrix, a goal which is impractical for inversions with, say, the number of data and model parameters exceeding 10^5 (Deal & Nolet 1996). We can summarize the situation as follows:

(1) for models with many degrees of freedom, it becomes impossible to calculate the *a posteriori* covariance matrix of the result;

(2) resolution calculations by means of a limited number of sensitivity tests have serious shortcomings;

(3) there is no satisfactory way to suppress the influence of the uneven distribution of ray paths in the final result.

In this paper we shall develop a simple approximate algorithm to estimate the resolution *and* the *a posteriori* covariance of a tomographic solution which avoids the calculation of eigenvectors or Ritz vectors. Whilst we leave an investigation of the third problem to a future paper, we believe the influence of the ray path distribution should be reduced by an adaptive reparametrization of the model, for which an estimation of the model covariance is a necessary prerequisite.

STATEMENT OF THE PROBLEM

We consider the $n \times m$ linear inversion problem for a model x, given (exact) data b with errors ϵ:

$$Ax = b + \epsilon = \hat{b}\,, \tag{1}$$

scaled such that the covariance matrix of the data error ϵ is the $n \times n$ unit matrix,

$$C_\epsilon = I_n\,. \tag{2}$$

Without loss of generality, we assume that the expected value of the data errors as well as the model parameters is zero:

$$E[\epsilon_i] = 0\,, \quad i = 1, \ldots, n\,, \tag{3}$$

$$E[x_i] = 0\,, \quad i = 1, \ldots, m\,. \tag{4}$$

Let A^- denote the inverse of A in a generalized sense; for example, $A^-\hat{b}$ might be the minimum norm solution of the least-squares system belonging to (1). While there is considerable freedom in the choice of A^-, a generalized inverse must satisfy $AA^-A = A$, or, as paraphrased by Jackson (1972),

$$AA^- \approx I_n\,, \tag{5}$$

$$A^-A \approx I_m\,, \tag{6}$$

which we shall refer to as the two 'Penrose conditions'. We can express the error of the solution $\hat{x}$ in terms of A^- (Nolet 1987):

$$\hat{x} - x^{\text{true}} = A^-(b+\epsilon) - x^{\text{true}} = (A^-A - I_m)x^{\text{true}} + A^-\epsilon\,, \tag{7}$$

which expresses the well-known result that the error in the solution has two causes: the inadequacy of the generalized inverse to satisfy the second Penrose condition (6) exactly, and the propagation of error terms through multiplication with A^-. Using the terminology of statistics, the first term constitutes the *bias* of the solutions, the second term the statistical fluctuations (for different realizations of the observational errors) around the biased solution.

Similarly, for the data misfit we find

$$\mu \equiv A\hat{x} - b = (AA^- - I_n)b + AA^-\epsilon\,, \tag{8}$$

from which we see that the data misfit $\chi^2 = |\mu|^2$ also has a bias and a variance. If we succeed in satisfying the first Penrose condition (5) we reduce the bias. Setting $\epsilon = 0$ in (7), we find an expression for the resolution matrix:

$$\hat{x} = Rx^{\text{true}}\,, \tag{9}$$

where

$$R = A^-A\,. \tag{10}$$

If, as is usually the case, (1) is a linear approximation to a non-linear problem, we may define ϵ to include also the errors due to linearizations, or other approximations in the forward problem (Tarantola 1987). This will undoubtedly introduce some correlations between the components of the error vector ϵ, which in principle could be removed through a linear transformation. To make a reasonable *a priori* estimate of the covariance matrix of ϵ is, however, a task so daunting that we are not aware of any successful efforts to do so for the seismic tomography problem. The unscaled C_ϵ is therefore generally assumed to be diagonal, so the transformation to satisfy (2) reduces to a trivial multiplication. The *a posteriori* covariance matrix of the solution $\hat{x}$ is then given by

$$C_{\hat{x}} = A^- C_\epsilon (A^-)^{\text{T}} = A^-(A^-)^{\text{T}}\,. \tag{11}$$

As usual, this is the covariance in the 'minimum norm' solution, which may be small either because a parameter is well constrained by the data, or because it is strongly damped towards 0. For the latter, the 'bias' is large but with little uncertainty. A true indication of the model accuracy can only be obtained by inspecting both the resolution matrix R and the covariance matrix $C_{\hat{x}}$.

From (11) we can easily compute the elements of the correlation matrix, defined as

$$\rho_{ij} = \frac{C_{ij}}{[C_{ii}C_{jj}]^{1/2}}\,, \tag{12}$$

where we have suppressed the subscript $\hat{x}$. Fully unresolved parameters (for which the column in A is empty) require a special treatment: their variance, while infinite in reality, is numerically zero because the null space of A is excluded from the solution space, and the correlation is undefined. We set such $\rho_{ij} \equiv 0$.

Intuitively, one understands from (7) that the statistical error term will grow when A^- has large components. Forcing the elements of A^- to remain small will reduce the variance but increase the bias, since we also reduce our ability to satisfy (6). The early literature on geophysical inverse problems is exhaustive in its analysis of this trade-off between bias and error, or variance, of the solution, either in discrete systems such as those considered here (Wiggins 1972; Jackson 1972), or in systems where models are not discretized *a priori* (Backus & Gilbert 1970; Tarantola 1987). However, it invariably requires the inversion of large matrices, which is generally performed by the application of singular value decomposition (SVD). While the increasing capacities of large computers now allow us

to apply SVD to matrices where m and $n \approx 10^3$, large-scale traveltime inversions commonly deal with $n = 10^4$–10^6 or more data, and require 10^3–10^5 or more model elements.

Of course, the computation of the exact generalized inverse of A with SVD is not feasible for such large tomographic problems. Therefore, we can only attempt an approximate solution to our problem. We note that the computation of the solution itself does not require the computation of the inverse A^-, since we can use iterative techniques to do so. We do need the inverse, however, to characterize the resolution by means of R and $C_{\hat{x}}$.

Since we define our solution as $\hat{x} = A^- \hat{b}$, (1) implies the condition $AA^- \hat{b} = \hat{b}$, and it is obvious that the first Penrose condition (5) is the equation that we shall wish A^- to satisfy as closely as possible. We shall see later that this is not an optimum solution to (6) in our approximate analysis of the problem; that is, it does not minimize the model bias.

Define c^k as the vector equal to the kth column of A^-, and e^k as the n-dimensional unit vector in the direction k. (5) implies the following:

$$Ac^k = e^k \quad (k = 1, \ldots, n). \tag{13}$$

Nakanishi & Suetsugu (1986) have proposed solving (13) exactly for all k, a strategy which is only possible for small n. We derive a fast, approximate solution using backprojection. The backprojection direction is found by taking the negative gradient $-\nabla_c$ at location c_0^k in model space of the misfit $|Ac^k - e^k|^2$, which is equal to $-A^{\mathrm{T}}(Ac_0^k - e^k)$. In our case $c_0^k = 0$, from which we find

$$v^k \equiv A^{\mathrm{T}} e^k, \tag{14}$$

where v^k is a vector of dimension n. (14) gives simply

$$v_i^k = A_{ik}^{\mathrm{T}} = A_{ki} \quad (i = 1, \ldots, m). \tag{15}$$

We seek an approximate solution to (13) by imposing the condition that c^k is in the direction of v^k: $c^k = \alpha_k v^k$. If we impose the condition that the misfit is minimized, this implies orthogonality of the misfit vector: $(\alpha_k Av^k - e^k, \alpha_k Av^k) = 0$. Hence the coefficient

$$\alpha_k = \frac{(e^k, Av^k)}{(Av^k, Av^k)}, \tag{16}$$

or, when written out explicitly,

$$\alpha_k = \frac{\sum_{j=1}^{m} A_{kj}^2}{\sum_{i=1}^{n} \left(\sum_{j=1}^{m} A_{ij} A_{kj} \right) \left(\sum_{q=1}^{m} A_{iq} A_{kq} \right)}. \tag{17}$$

Since

$$A_{ik}^- = c_i^k = \alpha_k v_i^k = \alpha_k A_{ki}, \tag{18}$$

the generalized inverse can therefore be written as

$$A^- = A^{\mathrm{T}} D, \tag{19}$$

where D is a diagonal matrix, its diagonal elements equal to α_k:

$$D_{kk} = \frac{(AA^{\mathrm{T}})_{kk}}{\sum_{i=1}^{n} (AA^{\mathrm{T}})_{ik}^2} \quad (k = 1, \ldots, n). \tag{20}$$

Unfortunately, Penrose's second condition (6) leads to a different approximate solution. Following the same backprojection method, we find

$$A^- = D' A^{\mathrm{T}}, \tag{21}$$

with the elements of D' defined by

$$D'_{kk} = \frac{(A^{\mathrm{T}}A)_{kk}}{\sum_{i=1}^{m} (A^{\mathrm{T}}A)_{ik}^2} \quad (k = 1, \ldots, m). \tag{22}$$

Finally, we notice a difference between the last equation and the approximate inverse we would obtain by simply reducing $A^{\mathrm{T}}A$ to its diagonal. In that case we would have an inverse similar to (21): $A^- = D'' A^{\mathrm{T}}$, with $D''_{kk} = (A^{\mathrm{T}}A)_{kk}^{-1}$. We investigated this third possibility briefly and abandoned it as quickly because of its complete lack of fit to either (5) or (6).

It is well known that one iteration of a backprojection step will converge to the correct solution in the case where all singular values of A are equal. This is not even remotely the case for tomographic systems. However, backprojections often give very reasonable data fits. The reason must be sought in the sparse nature of the matrices. If there is little overlap between rays, the products involved in AA^{T} will involve multiplications with zeros, unless two rays sample the same model cell. Therefore, AA^{T} is likely to be diagonally dominant. One can easily check that (5) is satisfied as long as $(AA^{\mathrm{T}})_{ij}$ $(i \neq j)$ can be neglected with respect to $(AA^{\mathrm{T}})_{ii}$. Since cells always correlate with neighbouring cells, the diagonal of $A^{\mathrm{T}}A$ is probably less dominant, which would explain the inferior performance of the diagonal approximation $D''_{kk} = (A^{\mathrm{T}}A)_{kk}^{-1}$. This approximation may work better for systems in which $E[A_{ij}] \approx 0$, such as in diffraction tomography, but is obviously bad for systems from body wave tomography where $E[A_{ij}] \gg 0$. Such considerations are, however, far from conclusive, and in the next section we shall rely on a numerical test to justify our approach.

We can use (19) in (10) and (11) to obtain estimates of the resolution and the covariance matrix, respectively. Note that these expressions have an added advantage over the expressions for R and $C_{\hat{x}}$ as computed by SVD, apart from the saving on computer memory and CPU: they allow us to compute only part of these matrices, which is useful if our parameters are 'local' (for example, spline supports, rather than non-local parameters such as spherical harmonic coefficients). Thus, we can isolate velocity or slowness parameters from parameters designating source or station corrections, or even isolate a particular geographic region of interest. The parameter transformations inherent to SVD prohibit this with the exact computations.

Another advantage is that the computation of A^- lends itself naturally to parallel computations, since each of the columns of A^- is computed independently from the others.

VALIDATION ON A SMALL LINEAR SYSTEM

The validity of our approach depends on how well (13) is solved with only one backprojection step. Earlier experience with the iterative inversion of sparse matrices suggests that the first backprojection step almost always provides the bulk of the

variance reduction, often reducing the data misfit by more than 50 per cent of the total (converged) reduction. In this section we investigate the validity of our approach.

Although SVD is always to be preferred for matrices of small dimensions, since it allows for an exact computation of the solution statistics, we use our method on such a small system to make a comparison with the exact solution possible. We choose realistic examples. Two matrices A are taken from the *Sn* tomography study of Nolet *et al.* (1998), and are denoted by 'east' and 'west', respectively. A^{east} is a 121×115 system with a rather sparse coverage of ray paths (see Fig. 2 in Nolet *et al.* 1998). In contrast, A^{west}, 839×429, has a dense coverage with many, often overlapping ray paths. Both systems include source and station corrections in addition to unknown slowness anomalies in the vector x.

In a first test we randomly generate synthetic data vectors b that satisfy (1) exactly, and test how well $x = A^- b$ satisfies the data. This is a direct test of the first Penrose condition (5). Fig. 1 shows histograms of the fits (defined as $|Ax-b|^2/|b|^2$) for A^- computed with (19) and (21). For both east (left) and west (right) it is clear that (19) (a) yields a superior data fit, as is to be expected since (19) was constructed to satisfy (5), but the difference with (21) (b) is not large.

Although it is clear that the variance reduction is not complete, it is obvious that $A^- b$ reduces the variance by at least 50 per cent, and often by much more than that. Since this is not a small variance reduction for many tomographic inversions, and since our aim is to estimate the statistics, not to construct $\hat{x}$, we judge this outcome highly encouraging. Since backprojections work most efficiently for non-overlapping ray paths (and would result in the optimal fit if every cell was visited only once), we conjecture that the differences between east and west are due to the difference in ray path overlap, with the estimate becoming less accurate as the ray paths overlap more. This would imply that the more accurate estimate of A^- is obtained by assembling overlapping ray paths into 'summary rays' (Morelli & Dziewonski 1987).

For the actual computation of the *solution* of $Ax = b$ the application of repeated backprojections in a conjugate gradient algorithm such as LSQR is not only more accurate but also faster (Paige & Saunders 1982; Nolet 1983).

When comparing estimates of $C_{\hat{x}}$ and R with their exact counterparts we face a problem related to the damping of the SVD solution. The situation is schematically sketched in Fig. 2. In this figure, our estimated variance and resolving length (or correlation distance) is shown by the dot. The curve represents the trade-off between variance and resolving length for a truncated SVD solution as we vary the number of eigenvalues. Since our estimate A^- is not exact, our solution is not exactly *on* the curve that describes the trade-off between variance and resolving power.

We can damp the SVD solution such that we obtain the same resolving length as with the approximate inverse A^-, or the same resolving power, or make a choice in between. We shall compare variance estimates for equally resolved models (that is, point B in Fig. 2). We use the effective rank of the inverse matrix as a measure of the overall resolving power. Wiggins (1972) showed that the effective rank of the truncated SVD inverse (the number of eigenvectors used to construct the generalized inverse) is equal to the sum of the diagonal elements of R:

$$k_{\text{eff}} = \sum_{i=1}^{m} R_{ii} \,. \tag{23}$$

For R^{west} we find $k_{\text{eff}} = 21.4$. In Fig. 3 we compare the estimated and the true values of R_{ii} for 22 eigenvectors, and similarly for R^{east} for which $k_{\text{eff}} = 16.3$ we choose 17 eigenvectors. Clearly, for well-resolved parameters with $R_{ii} > 0.5$ there is broad agreement; although R_{ii} may be in error by as much as 50 per cent, only a few 'unresolved' parameters are plotted as resolved. The few that have $R_{ii}^{\text{SVD}} < 0.1$ but for which our estimate exceeds 0.1 are all event or station corrections, not slowness parameters. This conclusion does not seem to depend

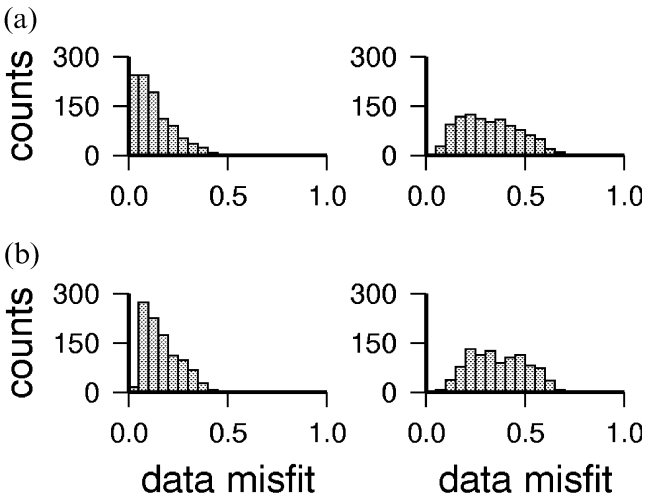

Figure 1. Histograms showing the data misfit $|Ax-b|^2/|b|^2$ for a Monte Carlo simulation using 1000 random data vectors b in the range of (left) A^{east} and (right) A^{west}. (a) A^- defined with the first Penrose condition (5); (b) A^- defined using (6).

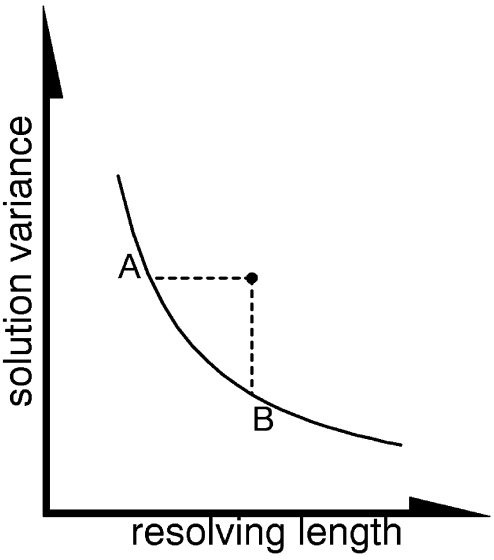

Figure 2. Schematic diagram showing the trade-off of variance versus resolving length for the SVD solution. The approximate inverse yields estimates for these quantities which are off this curve (black dot), and which may be compared either with the SVD solution with similar variance (point A) or with similar resolving power (point B), or somewhere between A and B.

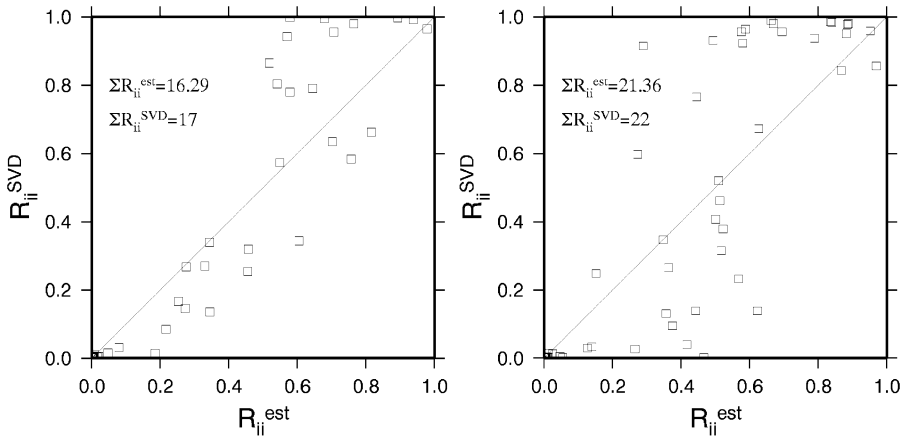

Figure 3. Comparison between the estimated diagonal elements of the resolution matrix R^{est} and the correct values R_{ii}^{SVD} for (left) A^{east} and (right) A^{west}. The SVD results were computed with 17 and 22 eigenvectors, respectively. See text for discussion.

on the exact choice of k_{eff}, since adding or subtracting an eigenvector affects only the ill-resolved parameters.

In Fig. 4 we compare the part of the covariance matrix relating to the slowness parameters with their exact counterparts for east (again calculated with 17 eigenvectors) and west (calculated with 22 eigenvectors). The colour scale is chosen to highlight parameters with a large (co)variance; that is, for which the tomographic image might be suspect.

An eyeball comparison again shows broad agreement between the estimated $C_{\hat{x}}$, denoted by 'EST', and the exact ones ('SVD'). Variances in the east, where the ray density is less than the west, are generally higher than in the west. On the diagonal, many gaps correspond to unresolved parameters for which the variance is 'numerically' zero due to the minimum norm character of the solution. Generally, the order of magnitude of the variances is well reproduced by the estimations, as are groups of covariances around the diagonal. In the off-diagonal bands corresponding to nearest-neighbour cells, the estimations seem to be biased towards somewhat larger values, but far off the diagonal (that is, for parameters located further apart) the estimated covariance is lower than the true value. A^{west} is an order of magnitude larger in size than A^{east} but no strong dependence of accuracy on matrix size is evident. If anything, the estimations for $C_{\hat{x}}^{\text{west}}$ seem to be slightly better than for $C_{\hat{x}}^{\text{east}}$. Since there is also no reason *a priori* to assume that the accuracy degrades with the size of the matrix, we are confident that, even for very large systems, the order of magnitude of the variance is estimated correctly.

APPLICATION TO A LARGE SYSTEM

We have also tested the algorithm on a much larger system. While we have no ground truth to compare the results, we investigated the efficiency of the algorithm as well as the effect of neglecting small matrix elements.

For this purpose we created a matrix A simulating a P-wave tomography experiment covering central and eastern Asia, including the subduction in the northwest Pacific, using one year of seismicity (1993). The system—formulated without source/station correction terms—has 69 043 rows and 9610 columns, and could not be handled with SVD even on a large computer. Using a linear spline parametrization (Thurber 1983), with pivots roughly 200 km apart, the matrix has 5.9×10^6 non-zero elements (0.9 per cent of the total). 44 per cent of these are smaller than 1 per cent of the largest element, 21 per cent smaller than 0.1 per cent.

The inspection of several rows of $C_{\hat{x}}$, plotted as correlations to facilitate the colour scaling, gives further confidence in the results. Fig. 5 gives these correlation coefficients for three locations, plotted in cross-sections as a function of latitude, longitude and depth. We compare the resolution in three different geographical locations plotted in Fig. 6. On the left in Fig. 5, the solution at point a, located near the surface, is clearly well constrained horizontally, but suffers from a lack of resolution in the depth direction. In the centre of Fig. 5, the solution in point b, located at 500 km depth just NE of Lake Baikal, correlates with points as far as 1000 km away. Finally, on the right of Fig. 5 one sees the effect of ray bundles for point c in the Japan slab, where the N–S cross-section evidently samples ray paths towards Australian stations, and where the lack of crossing ray paths at depth causes the elongated shape of the correlating structure.

On the Sun UltraSparc processor the computations of R and $C_{\hat{x}}$, including some overhead to calculate matrix statistics, take about 5 hr for 10^6 non-zero elements. When we neglect the smallest elements in A, computations of A^- are faster and we find that the computation time depends linearly on the number of non-zero elements of A (Fig. 7). However, the accuracy is clearly affected by truncation. We tested this by counting the number of correlation coefficients ρ_{ij} larger than a certain threshold. When we truncate A_{ij} at a level as large as 10 per cent of the maximum, we greatly increase the speed of computation (by a factor of 6), but we lose about 60 per cent of the $\rho_{ij} > 0.8$, which are now underestimated in magnitude; this is even worse for smaller ρ_{ij} (Fig. 8). Inspection of the actual covariances shows that it is mostly the smaller covariances that are affected. Since these probably belong to the ill-resolved parameters (the well-resolved parameters are associated with large elements in A), the situation shown in Fig. 7 may give a view that is too pessimistic. A modest truncation level of 1 per cent may be acceptable if we only use the ρ_{ij} for the purpose of reparametrization; this would result in a reduction of CPU time by a factor of about 2.

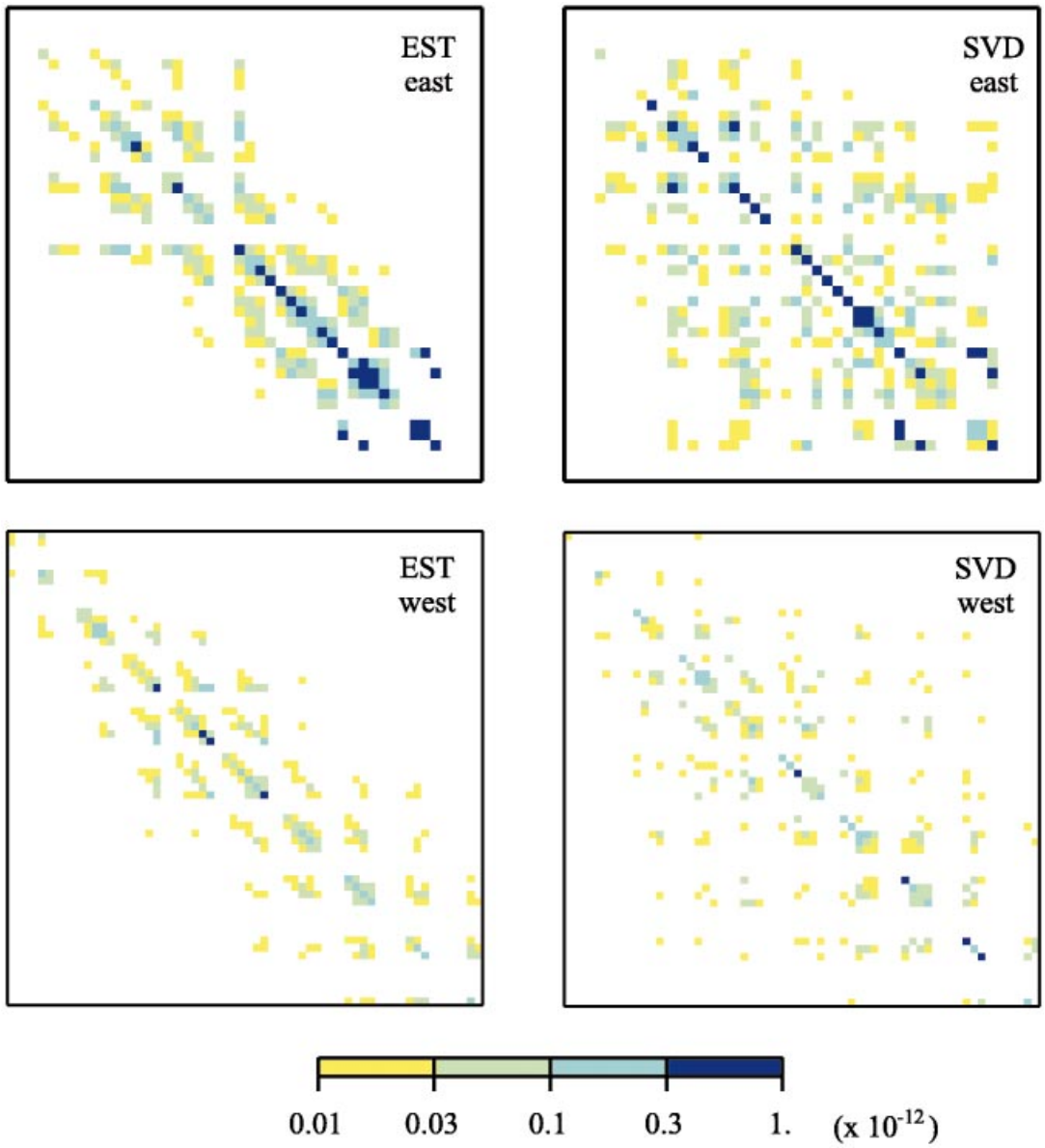

Figure 4. Comparison between the estimated covariance matrix (EST) and the correct covariance matrix (SVD) as computed for A^{east} (top) and A^{west} (bottom). Only the covariances of the slowness parameters are plotted. The scale is in 10^{-12} s^2 m^{-2} for an assumed variance in delay times of 1.0 s^2.

DISCUSSION

Programming considerations

The efficiency of the code depends strongly on some elementary programming considerations. The most commonly used scheme for the storage of non-zero elements of the matrix A is row-wise. This involves storage overhead more then double the memory required to store just the values of non-zero A_{ij}, since one has to store the column number for each element as well as the number of non-zeros in each row of A. The scheme allows for fast computation of the product of both A and A^T with a vector, by looping through the elements of A in the same order as they are stored. Whilst this strategy can still be followed for the matrix product AA^T by repeatedly multiplying A with one of its own rows, it fails for A^TA, when A^T is multiplied with *columns* of A. Although the computation of D in (20) only requires the product AA^T, reverse products occur in the computation of $R = A^TDA$ and $C_{\hat{x}} = A^TD^2A$. We have found it most efficient to store A twice: once in row and once in column order.

We also note that neither A^TA nor AA^T can be expected to be truly sparse matrices and that one should avoid storage of these products. Fortunately, for the computation of D with (20) one needs only one row of AA^T at a time. For the end-products R and $C_{\hat{x}}$ one may use mass storage to store these, generally row-wise in the form of 2-D or 3-D 'images', and for many applications a heavy truncation of smaller elements is allowed. Since the correlation matrix can be computed from $C_{\hat{x}}$ no separate storage of this is needed. If the correlation matrix is only computed to construct a sensible reparametrization of the model, an advisable strategy is to compute the diagonal elements of R first, then work from the smallest diagonal

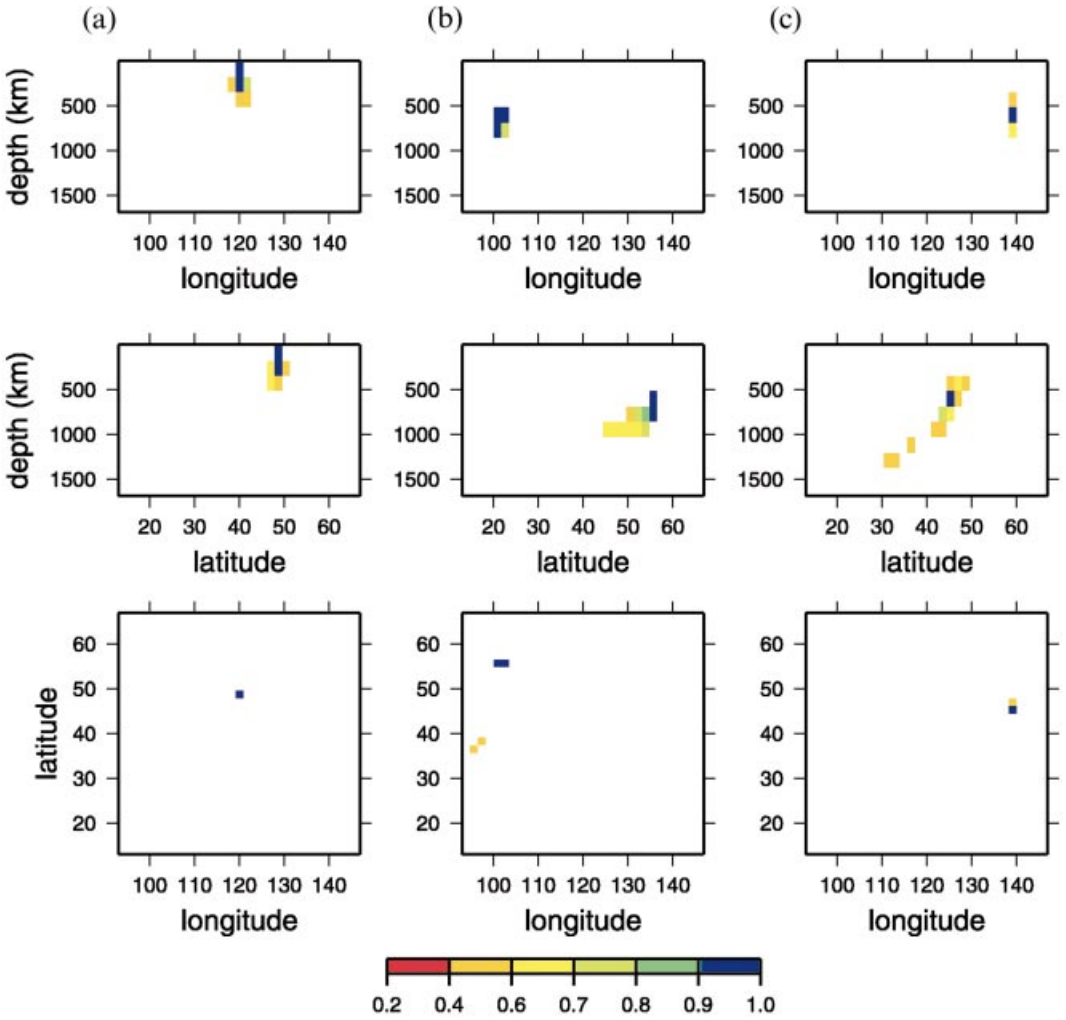

Figure 5. This figure shows three examples of rows of the correlation matrix for the large ($69\,043 \times 9610$) problem, plotted by way of cross-sections with fixed (from top to bottom) latitude, longitude and depth, respectively. (a) The P velocity near the surface below the Grand Khingan mountains in Mongolia, slowness variance 3.1×10^{-4} s^2 km^{-2}; (b) at 500 km depth NE of Lake Baikal, variance 6.1×10^{-4} s^2 km^{-2}; (c) at 500 km depth in the Japan subduction zone, variance 2.1×10^{-3} s^2 km^{-2}. The variances quoted are for an assumed variance in the delay time observations of 1.0 s^2.

elements to compute the correlations within that row and regroup parameters. This will quickly eliminate the parameters with the worst resolution and avoid unnecessary calculations.

Sensitivity tests

Our method is similar to that of sensitivity tests (Spakman & Nolet 1988; Leveque *et al.* 1993) but on n data vectors in which only one datum is equal to 1 and all others are set to zero, rather than setting one model parameter to 1 to construct a right-hand side. We also restrict the matrix solver to just one iteration. One could in fact try to forgo an analytical treatment as given here, and simply solve (13) using more iterations with a matrix solver such as LSQR. However, for large systems this will quickly saturate the available computer time. Since A^- will lose its sparse nature, this strategy may also invite storage problems, whilst truncating small elements of A^- may result in a loss of the extra precision gained by the extra iterations.

In comparison with sensitivity tests, our method gives a rough global estimate of both covariance and resolution, whereas sensitivity tests with spikes give an accurate image of the resolution, but for a few selected model parameters only, and no information on the covariances. Which is preferred depends on the application, and sometimes one may wish to use both methods, since they nicely complement each other. The main application we have in mind for our method is the reparametrization of the model by grouping of highly correlating parameters.

Lanczos iteration

Using the Ritz vectors (approximate eigenvectors) resulting from a Lanczos or conjugate gradient iteration to compute the resolution of large systems has been proposed (Zhang & McMechan 1995) as an alternative to explicit computation of the full eigensystem as in SVD. However, as pointed out by Deal & Nolet (1996), it quickly becomes infeasible to compute

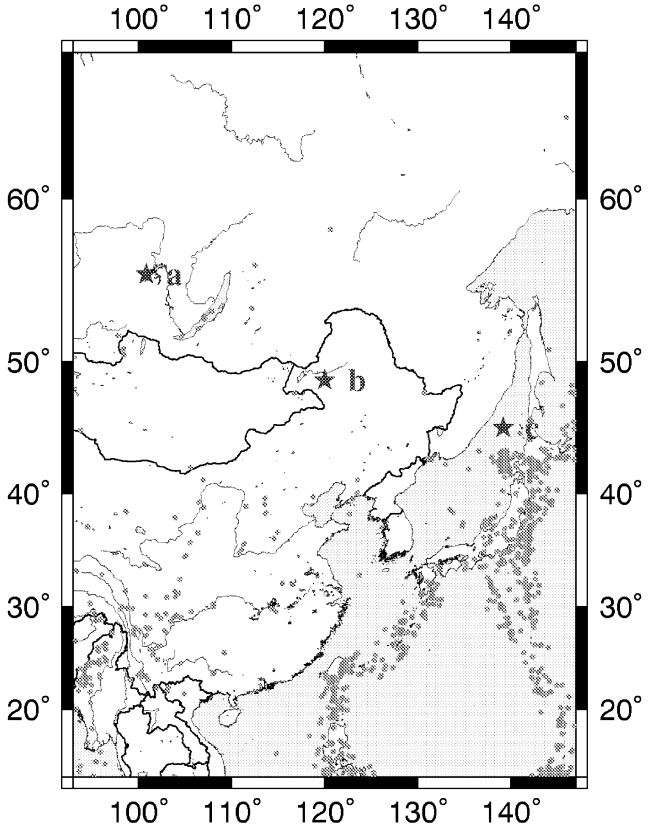

Figure 6. Geographical locations of the cells shown in Fig. 5.

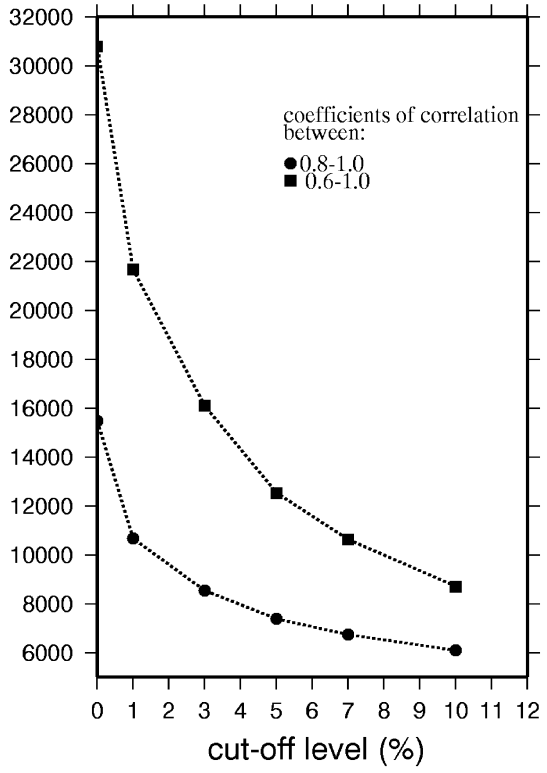

Figure 8. Test of the effect of neglecting small elements of A by measuring the number of correlation coefficients larger than 0.6 and 0.8 as a function of the cut-off threshold, defined in per cent of the largest matrix element.

all the Ritz vectors needed to span the solution space as the size of A and its effective rank grows, due to a prolific growth of duplicate vectors in the conjugate gradient scheme. For example, the effective rank k_{eff} of the large matrix used in the previous section is estimated with (23) to be 574, and to compute that many eigenvectors is very costly, and for somewhat larger problems probably even beyond the reach of iterative algorithms. We certainly do not agree with Zhang & McMechan (1996) that it is sufficient to compute the uncertainty in $\hat{x}$ by considering only a subset of Ritz vectors constructed from the data vector b: for a correct estimation of model statistics one has to allow perturbations of the model in *all* directions. Nor could one assume that the selection of a subspace spanned by an incomplete set of Ritz vectors constitutes a good basis for reparametrization (smooth models for which the statistics could then be computed). The reason is that the set of Ritz vectors is dependent on the data vector used to generate them and will ignore other directions in model space, even those that are associated with quite large eigenvalues (Deal & Nolet 1996).

This leaves the method described in this paper as the only one to estimate the resolution and covariance matrix for large systems.

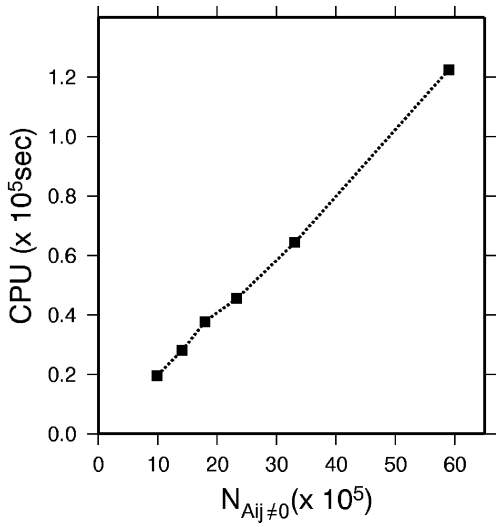

Figure 7. CPU time needed to compute R and $C_{\hat{x}}$ on a Sun UltraSparc processor as a function of the number of non-zero elements in A.

Damping

Strictly speaking, the estimated covariance and resolution is valid only for an inversion with the same damping properties as A^-. However, if the first iteration of a backprojection method such as LSQR defines the major characteristics of the solution, R and $C_{\hat{x}}$ should be useful as order of magnitude estimates. Since the variance in the data is not precisely known to begin with, attempts to increase the precision of R and $C_{\hat{x}}$ may seem futile. In principle, one could apply Newton iteration to obtain more precise versions of the inverse of A [the first correction would be $A^-(I-R)$] but the added computational effort will soon become prohibitive for really large systems. We may,

however, investigate the case that (1) needs to be damped strongly to keep the propagation of data errors under control.

Since A^- already involves a minimum degree of damping, we are limited in controlling the damping of our approximate inverse. However, in many cases the signal-to-noise ratio of the data vector $\hat{b}$ may be very small. For example, Morelli & Dziewonski (1987) estimated the variance of teleseismic P-delays at 1 s^2, which implies a signal-to-noise ratio of the order of 1. For S waves, tomographic systems are even less accurate than that. In such cases it may actually be advisable to damp the solution even further. This can be done by means of a simple modification of A^-. We may damp (1) in the same way as is done in ridge regression, adding to the system (1) m equations of the form $\lambda x_i = 0$, where λ serves to weigh these equations against the 'true' constraints:

$$\begin{pmatrix} A \\ \lambda I_m \end{pmatrix} x \equiv Bx = \begin{pmatrix} b \\ 0 \end{pmatrix}. \tag{24}$$

We then define the inverse as

$$B^- = B^{\mathrm{T}} \hat{D} = (A^{\mathrm{T}} D^{(1)} \;\; \lambda D^{(2)}), \tag{25}$$

where

$$D^{(1)}_{kk} = \frac{(AA^{\mathrm{T}})_{kk}}{\left(\sum_{i=1}^{n} (AA^{\mathrm{T}})^2_{ik} + \lambda^2 \sum_{i=1}^{m} A^2_{ki} \right)} \quad (k = 1, \ldots, n), \tag{26}$$

$$D^{(2)}_{kk} = \frac{1}{\sum_{i=1}^{n} A^2_{ik} + \lambda^2} \quad (k = 1, \ldots, m). \tag{27}$$

The definition of R now depends on a subtle interpretation of the damping. If we consider the added m equations $\lambda x = 0$ as true information on the model, that is, if we have reason to assume that the true earth model x^{true} is really 0, we would define R as before as $B^- B$ and find

$$R = A^{\mathrm{T}} D^{(1)} A + \lambda^2 D^{(2)}. \tag{28}$$

More probably, the damping equations are not reflecting true information, but are introduced to bias the model towards 0 and reduce its variance. In that case we can only say that $Ax^{\mathrm{true}} = b$, so that

$$x = (A^{\mathrm{T}} D^{(1)} \; \lambda D^{(2)}) \begin{pmatrix} Ax^{\mathrm{true}} \\ 0 \end{pmatrix} = A^{\mathrm{T}} D^{(1)} A x^{\mathrm{true}}, \tag{29}$$

which implies

$$R = A^{\mathrm{T}} D^{(1)} A. \tag{30}$$

For the covariance we find, using the same interpretation of the damping,

$$C_{\hat{x}} = A^{\mathrm{T}} D^{(1)\,2} A. \tag{31}$$

For $\lambda = 0$, this reduces to (11), and the variances behave asymptotically as λ^{-2} for $\lambda \to \infty$, as we should expect.

CONCLUSIONS

We have developed an approximate but explicit expression for the covariance and resolution of the solution of tomographic systems. In contrast to schemes based on SVD or Lanczos iteration, this can be applied to very large matrices. The CPU time required varies linearly with the number of non-zero elements in the matrix. The accuracy has been investigated with small systems and was shown to be sufficient for most purposes. Work on the application of these results in a strategy for automatic reparametrization of the model is currently in progress.

ACKNOWLEDGMENTS

We thank the Editor, Russ Evans, and Gerhard Pratt for mediating an innovative interactive review process that helped to clarify various points in this paper. We also thank Anthony Lomax for helpful discussions. GN wishes to thank CNRS for making a stay at Geosciences Azur possible, and received additional support from NSF under contract EAR-9526372. RM was supported by EEC grant ERBFMBICT.972203 attached to the European project ENV4980698.

REFERENCES

Backus, G. & Gilbert, J.F., 1970. Uniqueness on the inversion of inaccurate gross Earth data, *Phil. Trans. R. Soc. Lond.,* **A266,** 123.

Deal, M. & Nolet, G., 1996. Comment on 'Estimation of resolution and covariance for large matrix inversions' by Zhang and McMechan, *Geophys. J. Int.,* **127,** 245–250.

Jackson, D.D., 1972. Interpretation of inaccurate, insufficient and inconsistent data, *Geophys. J. R. astr. Soc.,* **28,** 97–109.

Kennett, B.L.N. & Nolet, G., 1978. Resolution analysis for discrete systems, *Geophys. J. R. astr. Soc.,* **53,** 413–425.

Leveque, J.J., Rivera, L. & Wittlinger, G., 1993. On the use of the checker-board test to assess the resolution of tomographic inversions, *Geophys. J. Int.,* **115,** 313–318.

Morelli, A. & Dziewonski, A.M., 1987. Topography of the core-mantle boundary and lateral homogeneity of the liquid core, *Nature,* **325,** 678–683.

Nakanishi, I. & Suetsugu, D., 1986. Resolution matrix calculated by a tomographic inversion method, *J. Phys. Earth,* **34,** 95–99.

Nolet, G., 1983. Inversion and resolution of linear tomographic systems, *EOS, Trans. Am. geophys. Un.,* **64,** 775–776 (abstract).

Nolet, G., 1987. Seismic wave propagation and seismic tomography, in *Seismic Tomography,* pp. 1–23, ed. Nolet, G., Reidel, Dordrecht.

Nolet, G., Coutlee, C. & Clouser, R., 1998. Sn velocities in western and eastern North America, *Geophys. Res. Lett.,* **25,** 1557–1560.

Paige, C.C. & Saunders, M.A., 1982. LSQR: an algorithm for sparse linear equations and sparse least squares, *ACM Trans. Math. Soft.,* **8,** 43–71 and 195–209.

Spakman, W. & Nolet, G., 1988. Imaging algorithms, accuracy and resolution in delay time tomography, in *Mathematical Geophysics,* pp. 155–188, eds Vlaar, N.J., Nolet, G., Wortel, M.J.R. & Cloetingh, S.A.P.L., Reidel, Dordrecht.

Tarantola, A., 1987. *Inverse Problem Theory,* Elsevier, Amsterdam.

Thurber, C.H., 1983. Earthquake location and 3D crustal structure in the Coyote Lake area, California, *J. geophys. Res.,* **88,** 8226–8236.

Tichelaar, B.W. & Ruff, L.R., 1989. How good are our best models?, *EOS, Trans. Am. geophys. Un.,* **70,** 593–606.

Wiggins, R.A., 1972. General linear inverse problem—implication of surface waves and free oscillations for Earth structure, *Rev. Geophys. Space Phys.,* **10,** 251–285.

Zhang, J. & McMechan, G.A., 1995. Estimation of resolution and covariance of large matrix inversions, *Geophys. J. Int.,* **121,** 409–426.

Zhang, J. & McMechan, G.A., 1996. Reply to comment by M. M. Deal and G. Nolet on 'Estimation of resolution and covariance of large matrix inversions', *Geophys. J. Int.,* **127,** 251–252.