Geophysical Journal International

Geophys. J. Int. (2014) **199,** 276–285
GJI Seismology

doi: 10.1093/gji/ggu242

Solving large tomographic linear systems: size reduction and error estimation

Sergey Voronin,* Dylan Mikesell,† Inna Slezak and Guust Nolet

Geoazur, Université de Nice/CNRS/IRD, F-06560 Sophia Antipolis, France. E-mail: sergey.voronin@colorado.edu

Accepted 2014 June 23. Received 2014 June 19; in original form 2014 May 1

SUMMARY

We present a new approach to reduce a sparse, linear system of equations associated with tomographic inverse problems. We begin by making a modification to the commonly used compressed sparse-row format, whereby our format is tailored to the sparse structure of finite-frequency (volume) sensitivity kernels in seismic tomography. Next, we cluster the sparse matrix rows to divide a large matrix into smaller subsets representing ray paths that are geographically close. Singular value decomposition of each subset allows us to project the data onto a subspace associated with the largest eigenvalues of the subset. After projection we reject those data that have a signal-to-noise ratio (SNR) below a chosen threshold. Clustering in this way assures that the sparse nature of the system is minimally affected by the projection. Moreover, our approach allows for a precise estimation of the noise affecting the data while also giving us the ability to identify outliers. We illustrate the method by reducing large matrices computed for global tomographic systems with cross-correlation body wave delays, as well as with surface wave phase velocity anomalies. For a massive matrix computed for 3.7 million Rayleigh wave phase velocity measurements, imposing a threshold of 1 for the SNR, we condensed the matrix size from 1103 to 63 Gbyte. For a global data set of multiple-frequency *P* wave delays from 60 well-distributed deep earthquakes we obtain a reduction to 5.9 per cent. This type of reduction allows one to avoid loss of information due to underparametrizing models. Alternatively, if data have to be rejected to fit the system into computer memory, it assures that the most important data are preserved.

Key words: Inverse theory; Body waves; Surface waves and free oscillations; Computational seismology.

1 INTRODUCTION

Gilbert (1971) wrote an important paper addressing the need to condense the size of linear geophysical inverse problems so as to be able to solve them with the computing power available at the time. The IBM S/360-67, introduced in 1967, had an internal memory limited to 1 Mbyte. The first personal computer, the Apple II, offered 48 Kbytes in 1977. The IBM PC, introduced in 1981, had a memory limited to 256 Kbyte. At the time Gilbert's paper was published, a megabyte was obviously considered a major storage headache.

However, Moore's law predicting an exponential growth in the number of transistors that fit on a single chip caught up with the early limitations, and the memory capacity of computers doubled roughly every 18 months. Some of the computations presented in this paper were done on a MacBook Pro with 4 Gbyte internal memory, which is now considered more or less standard, whereas many of us have access to local clusters with a Terabyte or more of memory. As a result, Gilbert's paper was soon apparently obsolete: cited 51 times in the first ten years after its publication, it was mentioned only four times since 2000 (data from Web of Knowledge).

One of the big surprises of recent times is the extremely rapid accumulation of high quality digital seismic data, a development that has caught up with Moore's law. Combined with new methods to analyse these data, such as finite frequency tomography (Dahlen *et al.* 2000) and adjoint waveform tomography (Tromp *et al.* 2005; Fichtner *et al.* 2006), this often requires significantly more computer memory than is readily available.

The adjoint approach circumvents the problem posed by memory limitations since it computes a gradient on the fly and does a search in model space to find a minimum in the data misfit rather than inverting a linear system, but this makes it labour intensive. Because the gradient is re-computed at each iteration, adjoint inversions are thought to be better positioned to handle the strong non-linearity

*Now at: Department of Applied Mathematics, University of Colorado, Boulder, CO 80309-0526, USA.
†Now at: Earth Resources Laboratory, MIT, Cambridge, MA 02139, USA.

of waveform data. Yet in practice the adjoint method is often applied to delay times measured by cross-correlation over selected time windows (Maggi *et al.* 2009; Tape *et al.* 2009, 2010), rather than to the waveforms themselves, to reduce non-abrasivelinearity. Mercerat & Nolet (2012, 2013) show that cross-correlation delays remain linear over a large range of velocity heterogeneity (up to 10 per cent). Recomputing the gradient at every step then becomes a unnecessary burden rather than an advantage. If one can avoid the gradient search altogether, and instead invert the linear system directly, this could significantly reduce the number of non-linear iterations needed, or potentially avoid iterating at all, if the problem is sufficiently linear.

Inverting delay times (or surface wave phase delays) by linear inversion, as is done in 'finite-frequency' tomography, therefore offers a very significant speed-up in computation, though at the cost of a memory requirement that easily exceeds a few Terabytes and approaches even the memory capacity of the largest machines. At the time of writing, the Titan cluster at Oak Ridge, used by Bozdag *et al.* (2013) in their pioneering attempts to do global tomography with the adjoint method, offers 584 Tbyte, but most applications work on clusters of one, or at best a few Terabytes. It thus is worthwhile to revisit Gilbert's 'ranking and winnowing' of data to determine if this leads to the significant reduction of memory needed for modern, 3-D inverse problems.

2 SPARSE, LINEAR TOMOGRAPHIC SYSTEMS

Earthquakes often occur at (almost) the same location, and seismic stations remain where they are. As a consequence, tomographic data can be very redundant, leading to matrices with rows that are highly dependent. This section explores a method to use this redundancy to reduce the size of the linear system while retaining the sensitivity to small scale structure if it is resolvable.

We consider N travel time delays d_i that are linearly (or quasi-linearly) dependent on M model parameters m_i and that are observed with errors e_i:

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

If a local parametrization is used for m (dividing the model up in volume elements or 'voxels'), the system (1) is sparse, that is most of the elements of the sensitivity matrix A are zero. Typically, in our applications, the fraction of non-zeros is of the order of a few per cent. Sparse systems can efficiently and stably be solved with linear conjugant gradient methods such as LSQR (Paige & Saunders 1982). To exploit the extra sensitivity of finite-frequency in tomographic inversions, a fine parameterization of the model is necessary (Chevrot & Zhao 2007), leading to very large model dimension M. We use the parameterization described by Charléty *et al.* (2013), in which the Earth's mantle is represented by 3.6 million voxels. Modern applications may also require the inversion of millions of data, such that $N \times M > 10^{12}$.

The first strategy to reduce the memory needed for a matrix should focus on the way it is represented in computer memory. For completely unstructured matrices one needs to specify the column number with each non-zero element. However, finite-frequency sensitivity kernels are localized in space, and exploiting the fact that non-zeroes are clustered in each row leads to a savings that may approach 50 per cent. We describe our modified representation in the Appendix. A second strategy can be to use wavelets to reduce the storage requirement for the matrix, the model or both (Chevrot & Zhao 2007; Simons *et al.* 2011; Charléty *et al.* 2013; Voronin *et al.* 2014). In this paper, we explore a third strategy based on singular value decomposition, and combine it with the modified sparse representation.

In seismic tomography the model usually represents perturbations with respect to a 'background' model, often a spherically averaged model. The expected value of the m_i is therefore assumed to be 0. We also assume, for derivations below, that data errors as well as the model perturbations are uncorrelated. Both can be transformed to diagonalize their covariance matrix if we have prior knowledge of correlations (see the discussion in Section 5.2). Finally, assume that all model parameters have the same prior variance σ_m^2 and that all errors in the observations have the same variance σ_e^2, and are on average zero. These conditions are not essential, and will be relaxed later, but they simplify the mathematical development:

$$E[m_i] = 0, \quad E[m_i m_j] = \delta_{ij}\sigma_m^2, \tag{2}$$

$$E[e_i] = 0, \quad E[e_i e_j] = \delta_{ij}\sigma_e^2, \tag{3}$$

where $E[.]$ denotes the expected value.

2.1 Summary of SVD

We assume the reader is familiar with singular value decomposition (SVD; see also Nolet 2008, chapter 14), but recall here briefly some of the major characteristics of this approach in order to establish a useful notation. We use the SVD of the $N \times M$ matrix A:

$$A = U\mathbf{\Lambda}V^T \approx U_k\mathbf{\Lambda}_k V_k^T\,, \tag{4}$$

where $(.)^T$ indicates the transpose, U and V are eigenvector matrices of AA^T and A^TA, respectively, with eigenvalues $\mathbf{\Lambda}^2 = \mathrm{diag}(\lambda_i^2)$. The subscript k on matrix symbols indicates a truncation to k columns or rows, for example U_k is an $N \times k$ matrix with the k eigenvectors u_i, $i = 1, ..., k$ belonging to the largest k singular values as columns. The two sets of eigenvectors are related by:

$$A^T u_i = \lambda_i v_i. \tag{5}$$

We project the system (1) onto the range of U_k to obtain a consistent system of equations (if vector y is in the 'range' of U it means that there is a vector x such that $Ux = y$):

$$U_k^T Am = U_k^T d + U_k^T e\,. \tag{6}$$

If $k \leq \mathrm{Min}(M, N)$ is equal to the rank of the system, the solution using (6) is the same as the least squares solution obtained from solving (1). However, we seek a 'damped' solution for the model that is minimally influenced by the errors e. As we show below, the posteriori covariance matrix of the model is proportional to $\mathbf{\Lambda}^{-2}$.

The wish to suppress error propagations (and also the need to fit the system in limited computer memory), usually motivates us to truncate at a level k such as to remove singular values that are small, but not yet zero. We note that the eigenvectors (columns of U and V) are orthonormal, thus $U_k^T U_k = I_k$, and $V_k^T V_k = I_k$ even if $k < N$, but that a transposed product such as $U_k U_k^T$ is not equal to the unit matrix I_N unless $k = N$. We develop the true Earth m into a part projected onto the first k orthonormal eigenvectors v_i ($m_k = V_k y_k$) and a residual:

$$m = V_k y_k + m_{M-k}\,, \tag{7}$$

where m_{M-k} is the 'unresolved' part of the model not in the range of V_k, i.e. $V_k m_{M-k} = 0$, and therefore $y_k = V_k^T m$. A minimum norm

solution is obtained by setting $m_{M-k} = 0$. Similarly, we project the data onto the set of eigenvectors $u_i,\ i = 1, ..., k$:

$$d = U_k \tau_k + r_k\,, \tag{8}$$

where r_k denotes the rest term, the data component not in the range of U_k, such that $\tau_k = U_k^T d$. Note that this choice reduces the projected data vector to its component in the range of U_k, which is smaller than the range of A. If our truncation is too conservative, any unmodelled components of the observed data d are considered the same way as errors. However, as we shall show, the reduction to k equations also enables us to remove data with a low SNR from the system. We must thus find a suitable compromise in our choice of k. How to do that is the major topic of the rest of this section.

2.2 Error estimation

The covariance of the data d is related to the covariances of the model and the measurement errors. Using the fact that m and e have zero expected value and are uncorrelated (eqs 2 and 3), we find for the data covariance:

$$\begin{aligned}\mathrm{Cov}(d_i, d_j) &= E\left[\sum_{k,l}(A_{ik}m_k + e_i)(A_{jl}m_l + e_j)\right] \\ &= \sum_{k,l} A_{ik}A_{jl}E[m_k m_l] + E[e_i e_j] \\ &= \sum_{k,l} A_{ik}A_{jl}\delta_{kl}\sigma_m^2 + \sigma_e^2\delta_{ij} \\ &= \sum_{k} A_{ik}A_{jk}\sigma_m^2 + \sigma_e^2\delta_{ij}.\end{aligned} \tag{9}$$

Writing $\sigma_e^2 I$ for the error covariance matrix, and $\sigma_m^2 I$ for the prior model covariance, the total data covariance in matrix notation is:

$$\mathcal{C}_d = A\mathcal{C}_m A^T + \sigma_e^2 I = \sigma_m^2 AA^T + \sigma_e^2 I\,, \tag{10}$$

and, using the (full) singular value decomposition of A:

$$\begin{aligned}\mathcal{C}_d &= \sigma_m^2 U\mathbf{\Lambda}V^T V\mathbf{\Lambda}U^T + \sigma_e^2 I \\ &= \sigma_m^2 U\mathbf{\Lambda}^2 U^T + \sigma_e^2 I.\end{aligned} \tag{11}$$

For the covariance of the projected data $\tau_k = U_k^T d$ we find with $U_k^T U = [I_k, \varnothing]$ (i.e. the last $N - k$ columns zero) in a similar fashion:

$$\begin{aligned}\mathcal{C}_\tau &= Cov(U_k^T d) = U_k^T \mathcal{C}_d U_k \\ &= \sigma_m^2 U_k^T U\mathbf{\Lambda}^2 U^T U_k + U_k^T \sigma_e^2 I_k U_k \\ &= \sigma_m^2 \mathbf{\Lambda}_k^2 + \sigma_e^2 I_k\,.\end{aligned} \tag{12}$$

The variance of the projected data is given by the diagonal of $\mathcal{C}_\tau$:

$$\sigma_{\tau_i}^2 = \sigma_m^2\lambda_i^2 + \sigma_e^2\,, \tag{13}$$

which splits the data variance into a 'signal' part due to the model and a 'noise' part σ_e^2 due to errors in the data. For the signal-to-noise ratio (SNR) of the ith projected datum we therefore have:

$$\mathrm{SNR}_i = \frac{\sigma_m\lambda_i}{\sigma_e}. \tag{14}$$

Eqs (12) and (13) tell us that the projected data are uncorrelated, with a variance σ_τ^2 increasing with the eigenvalues and approaching σ_e^2 as the eigenvalue approaches zero. One can thus estimate the data errors by inspecting the distribution of the projected data as $\lambda_i \to 0$, or fit the complete distribution with optimized values for σ_e^2 and σ_m^2.

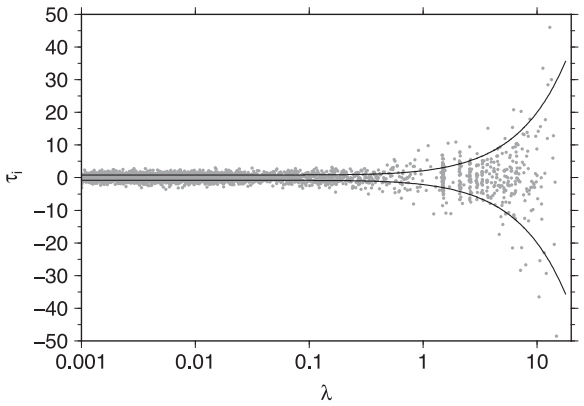

Figure 1. Projected data as a function of the eigenvalue for the largest cluster of surface wave data (see Section 4.1). The solid lines denote ± one standard deviation in the distribution of τ_i as predicted by eq. (13), with optimal σ_e^2 and σ_m^2 determined by a simple grid search. The first two data: $\tau_1 = -90.5, \tau_2 = -121.9$, fall outside the plot.

The optimization is done by assuming a normal distribution and using a grid search for σ_e^2 and σ_m^2 such that close to 68 per cent of the data falls outside $\pm\sigma_\tau$. Fig. 1 shows an example for surface wave data that we shall study in Section 4.1. The result that the projected data have the same standard error σ_e as the original data was already found by Gilbert (1971), but the derivation given here is much simplified by starting from a scaled system with uniform data error variance σ_e^2 and prior model uncertainty or variance σ_m^2.

2.3 Winnowing small eigenvalues

The data misfit χ_k^2 is found by multiplying the solution m_k with the original matrix:

$$\chi_k^2 \equiv \frac{|Am_k - d|^2}{\sigma_e^2} = \frac{|r_k|^2}{\sigma_e^2}\,, \tag{15}$$

where because of (8), r_k is the part of the data vector that remains after d is projected onto the subspace spanned by the columns of U_k. Both (14) and (15) provide convenient measures for an upper limit of k: SNRs smaller than some threshold, or χ^2 much smaller than N can be avoided by choosing k sufficiently small. Theoretically χ^2 should be equal to N for the best compromise between model resolution and error, but if the data error σ_e is uncertain, χ^2 is uncertain as well, and often a range such as $0.5N < \chi^2 < 2N$ is considered acceptable.

If the data or model averages are not zero, as we assumed, we can always redefine them by subtracting the average after a first inversion attempt. If the data errors are not uniform, we can scale the system (1) to a uniform data error, by dividing each row and its associated datum by the standard error. If we know the standard error exactly, this leads to univariant data ($\sigma_e^2 = 1$). In practice, we often assign a quality factor to the data, which represents our subjective judgement of the relative error level. For this reason we maintain an arbitrary, but uniform, error σ_e which can be different from 1, and that can be estimated using (13).

For the prior uncertainty σ_m in the model one usually has some idea of reasonable prior variations to be expected (e.g. 1 per cent for the variations in intrinsic P velocity in the lower mantle), and we scale the system such that σ_m becomes 1 for scaled parameters, even though posterior estimates for the model variance may force us to modify the prior σ_m.

Two other measures exist that may help to determine an optimal cut-off rank k, though these are in general more difficult to apply. First of all, we can solve (1) with SVD after substituting

$$m_k = V_k y_k \,, \tag{16}$$

$$A V_k y_k = U_k \mathbf{\Lambda}_k V_k^T V_k y_k = U_k \mathbf{\Lambda}_k y_k = d \,,$$

so that we find m_k after computing:

$$y_k = \mathbf{\Lambda}_k^{-1} U_k^T d = \mathbf{\Lambda}_k^{-1} \tau_k. \tag{17}$$

Since the columns of V_k are orthogonal, $m_k^T \cdot m_k = y_k^T V_k^T V_k y_k = y_k^T \cdot y_k$, so that the norms of m_k and y_k are the same and

$$|m_k|^2 = \sum_{i=1}^{k} \frac{\tau_i^2}{\lambda_i^2} \tag{18}$$

which can be used to impose a limit to the rms norm of the model variations. The problem with this measure is that, unless the full model space is resolved, the rms norm is difficult to interpret physically. As we shall see in the next section, this is certainly the case when we subdivide the matrix into clusters with geographically restricted sensitivity.

Secondly, a more physically meaningful strategy is to limit the L_∞ norm of m_k:

$$\| m_k \|_\infty = \sup(|m_i|) \,, \tag{19}$$

where we find the solution from (16) and (17):

$$m_k = V_k y_k = V_k \mathbf{\Lambda}_k^{-1} U_k^T d \,, \tag{20}$$

but this involves the non-sparse $M \times k$ matrix V_k, and thus an additional computational effort. To avoid computing V_k explicitly, we use (5) to write the first k eigenvectors in terms of U_k:

$$V_k = A^T U_k \mathbf{\Lambda}_k^{-1} \tag{21}$$

and use $A^T U_k = (U_k^T A)^T$, which we compute anyway to construct the condensed system (6). Or, combining (8), (20) and (21):

$$m_k = A^T U_k \mathbf{\Lambda}_k^{-2} \tau_k \,. \tag{22}$$

3 CLUSTERING OF SPARSE MATRIX ROWS

For large linear systems, the singular value decomposition can be accomplished using Monte Carlo techniques (Voronin *et al.* 2014). However, the projection with U is likely to destroy the sparsity of the system since many data influenced by many different geographical regions are mixed in the projected datum. In our experience, the first datum, the one with the largest eigenvalue, represents often a kind of average among all data, thus completely destroying the locally concentrated nature of the sensitivity.

To counter this disadvantage, we first perform a clustering of data such that all data within one cluster have a localized sensitivity in the same region. The basic idea is that the linear system (1) is invariant to the ordering of the data. We shall wish to group them into clusters of data that are isolated geographically, that is, that share many columns identical to zero.

To accomplish this, we find groups of rows that share the same zero columns. We define a *cluster* of rows by the set of columns that are zero in each element of the cluster, and define three measures of sparsity and overlap of non-zeros between a row and a candidate cluster:

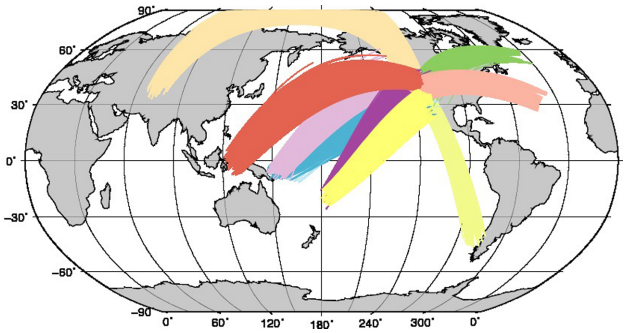

Figure 2. Ray path coverage for the first few surface wave clusters.

R_{inC}: the ratio between the number of non-zeroes in the row that overlaps with those in the cluster, and the total number of non-zero columns in the row,

C_{inR}: reversely, the fraction of non-zeroes of the cluster that overlaps with the row's non-zeroes,

N_{both}: the number of columns that has a non-zero either in the row, or in the cluster, or in both.

The algorithm that performs such clustering starts with the first row as the first cluster, and computes the overlap R_{inC} and C_{inR} of each subsequent row with all existing clusters. It selects whichever of these two is largest, then determines for which cluster this overlap is largest. If this maximum overlap is larger than a specified threshold, and if N_{both} represent an acceptable sparsity, the row is added to the cluster with largest overlap. If not, the row is the first element of a new cluster.

Though this clustering procedure can be time-consuming, several shortcuts provide a significant computational speedup. The columns in each cluster are represented by the bits in an integer array, which are set to 1 if the column is non-zero. The software was written in Fortran 90 which has convenient functions for bit manipulation and testing. Furthermore, as we create and modify clusters, we keep track of the average locations of the stations and sources constituting the endpoints of the ray paths in each cluster. If the distance between the row's station or source and that of the cluster average is larger than a specified distance $\Delta_{\max}$, the cluster is considered a non-candidate, dispensing of the need to compute the overlap. Fig. 2 shows the largest clusters for the surface wave data set discussed in the next section.

The distance parameter, $\Delta_{\max}$, not only speeds up the clustering, it also influences the width of the resulting clusters since it may be more restrictive than the minimum overlap specified. We can also set an upper limit to the sparsity allowed for a cluster, or limit the number of data in a cluster. If necessary, we can repeat the process after size reduction and cluster nearby clusters to create more populous (but wider) clusters.

Since the clustering results in submatrices with a much smaller number of rows than present in the total data set, the singular value decomposition becomes much more efficient. The column dimension of the submatrices remains the same, though, and this may be very large if one wishes to exploit the detail present in finite-frequency kernels. For example, in the wavelet-friendly parameterization advocated by Simons *et al.* (2011) and applied in full three dimensions by Charléty *et al.* (2013), the number of columns is more than 3.6×10^6 per parameter inverted (e.g. V_p, V_s), and this may still render the computation of SVD, and the storage of the (non-sparse) eigenvector matrices V difficult. An efficient way around this is to compute the non-sparse matrix AA^T, which is only

of size $N \times N$ if N is the number of rows in the subcluster, typically of order 10^2–10^3, and use (e.g. Nolet 2008):

$$AA^T U = U\mathbf{\Lambda}^2 . \tag{23}$$

Although the squaring of A leads to a loss of precision, certainly when done in single precision as we did, this is not a serious concern since all we wish to do is to project, using (6), onto a subspace of the most influential data, and inaccuracies in eigenvalues or eigenvectors do not affect the validity of this projection. The computing time needed scales as N^3, but if we limit the number of data in a cluster (we used 5000), a few hours on a single processor is sufficient for the computation of $U_k^T A$ for that cluster. Most clusters are much smaller and can be transformed in a few minutes CPU time.

The size reduction of very small clusters may not be worth the effort. We do not throw such data out; instead, we collect them in an unreduced matrix $A_{\rm rest}$. Once all large submatrices A_i have been reduced in size they may be combine with the remaining data in $A_{\rm rest}$ to formulate a linear system smaller in size but with no loss of important constraints on the model:

$$\begin{pmatrix} U_1^T A_1 \\ U_2^T A_2 \\ \dots \\ A_{\rm rest} \end{pmatrix} m = \begin{pmatrix} U_1^T d_1 \\ U_2^T d_2 \\ \dots \\ d_{\rm rest} \end{pmatrix} . \tag{24}$$

Clustering resembles the method of 'summary rays' (summing rays from nearby sources to the same or nearby stations) but is more powerful. Bolton and Masters (2001) reduce the influence of outliers by using the median of the data in a summary ray as the 'observed' delay. This assumes that there is no important variation in the delays that contribute to the summary ray, unless it is an outlier. In our approach, the variance σ_e can be used to identify outliers while taking the model influence over the cluster into account. To do this, one inverts for a model $m_k = V_k^T y_k$ using the cluster data only. Provided the model is overparametrized, 'true' data can always be fitted in this way and any remaining residuals in $r_k = d - Am_k$ must be due to data error. If this exceeds a threshold (e.g. $3\sigma_e$) one identifies (and removes) the datum as an outlier. Note that outliers cannot be removed after projection, since the transformation $U^T d$ spreads their power over all new data τ.

4 EXAMPLES

The success of the clustering SVD stands or falls with the ability to keep the decrease in sparsity of projected matrices under control. To judge our ability to do so, we investigated three different cases, one involving surface wave phase delays, the other two for body wave cross-correlation delays and delays in onsets, interpreted with finite-frequency theory and ray theory, respectively.

4.1 Surface wave phase anomalies

To determine what we accomplish in the case of data with a strong overlap, we investigate the size reduction of the sensitivity matrix for a massive data set of surface wave phase anomalies. We computed the matrix for surface wave phase velocities at five frequencies (periods 62, 88, 114, 151 and 174 s) for the fundamental Rayleigh mode phase delays that were used in the construction of tomographic model S40RTS (Van Heijst and Woodhouse 1999; Ritsema *et al.* 2011). This is only a subset of the frequencies measured; we exclude major arc data and higher modes, and we ignore intermode coupling in the computation of A. Even so, the finite frequency kernels for 3 767 043 phase anomalies fill a giant sparse matrix that occupies 1 103 139 833 531 bytes (1103 Gbyte) on disk. Application of the the optimized sparse representation described in the Appendix is powerful in the case of surface wave sensitivity: by itself it was already able to reduce the size to 610,605,684,321 bytes (611 Gbyte).

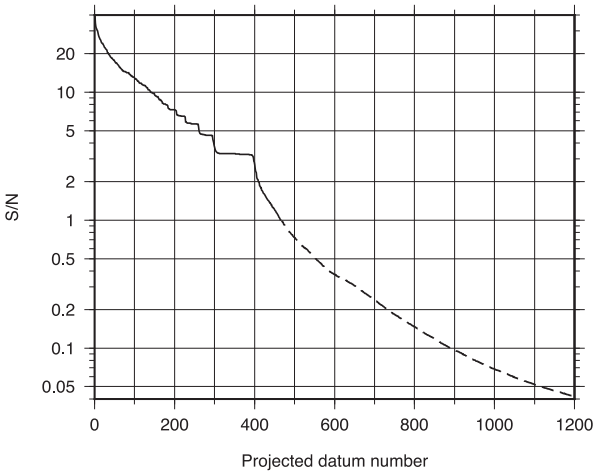

Figure 3. Signal-to-noise ratio (SNR) for the projected data τ_i of the largest cluster, with 5001 surface wave phase delays. The solid part of the curve shows the first 466 data with a SNR larger than 1.

For efficient clustering we compare only the overlap in the surface layer of the model. We started using a very restrictive clustering, setting $\Delta_{\rm max} = 700$ km, to optimize eigenvalue drop-off in the densest clusters. This yielded 1678 clusters with more than 400 data (718 of which had more than 1000 data). To avoid excessive computation times for eigenvector computations, we limit each cluster to at most 5001 data (starting a new cluster if necessary). This forced the truncation of the 176 largest clusters. On the other hand, the very strict clustering created many very small clusters, many of them probably too small to make it worthwhile to condense them by projection—the loss of sparsity is only compensated if many rows in the cluster are redundant and lead to an abundance of small eigenvalues. We therefore subjected the clusters with fewer than 400 rows to two more rounds of clustering but with increasingly relaxed $\Delta_{\rm max}$, first of 1400 km, and again clustering clusters with less than 400 data in a final round with $\Delta_{\rm max} = 2100$ km. This iterative strategy allows populous clusters to remain narrow, thereby optimizing the rate of decrease of eigenvalues in each cluster. The final result was a total of 5659 clusters with at least 10 data, 4262 of which had more than 200 data; only 5500 rows (or fewer than 0.15 per cent) were considered too isolated to fit in a cluster of at least 10 paths, and delegated to $A_{\rm rest}$ with no attempt to reduce the size.

Each of the 5659 matrices were subsequently subjected to SVD. The drop-off in λ_i or SNR_i is approximately exponential (Fig. 3), and limiting the SNR of accepted projected data τ_i to 1 allows us to reduce the number of rows by an order of magnitude. Using a cut-off at a SNR of 1, the size of the projected matrix system was reduced to 63 028 499 791 bytes (63 Gbyte), or 5.7 per cent of the original matrix size.

The rows of the transformed matrix $U^T A$ are in model space, and reflect the geographical sensitivity of the Earth to the associated data τ. For plotting purposes we equalize the Euclidean length of each row by scaling them with the associated eigenvalue (since $A^T U = V\mathbf{\Lambda}$ we have $\mathbf{\Lambda}^{-1} U^T A = V^T$ and the columns of V are eigenvectors normalized to 1). Fig. 4 plots selected rows of V^T for one of the

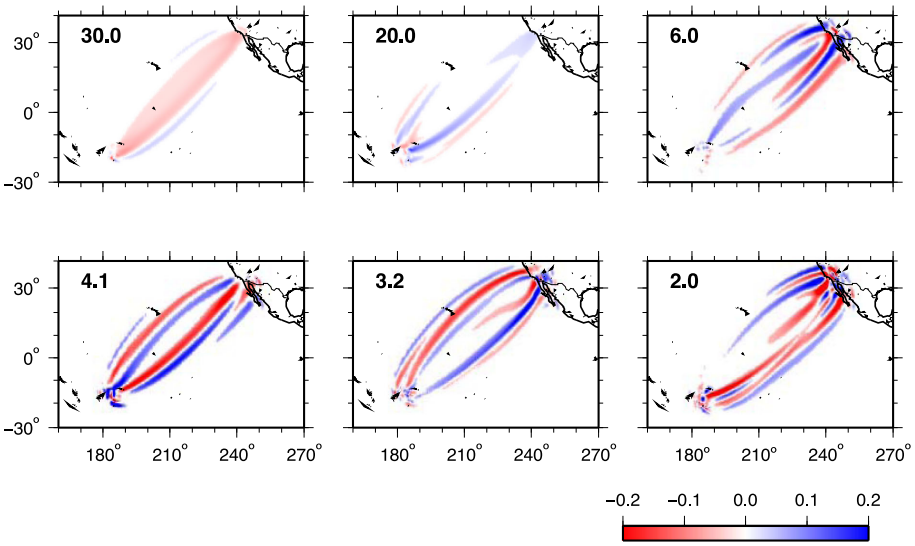

Figure 4. This figure shows horizontal cross-sections (depth 68 km) through selected rows of the matrix $V^T = \mathbf{\Lambda}^{-1}U^T A$ for one of the largest clusters of surface wave phase delays. The numbers in the upper left corner denote the signal-to-noise ratio for the projected data as determined from eq. (14).

large clusters with 5001 data. Note the increased complexity of the sensitivity as the eigenvalue decreases. The sensitivity shows structure of a scale length comparable to the voxel size (about 70 km), thus justifying the use of a model of very high dimension, as advocated by Chevrot & Zhao (2007).

4.2 *P*-wave delays by cross-correlation

We tested the clustering of body wave cross-correlation delays on a new—still incomplete—data set of multifrequency delay times. The current set has 25 046 delays measured from 60 deep earthquakes distributed evenly over the Earth. This data set is thus less redundant in path coverage than the surface wave data set discussed in Section 4.1. Also, because the sensitivity of body wave delays spreads out in the lower mantle, the need to keep sparsity under control in transformed matrices is more challenging than for surface waves.

The ray path coverage of the complete set is shown in Fig. 5. The matrix A was computed using finite-frequency sensitivity, efficiently computed with ray theory (Dahlen *et al.* 2000; Mercerat & Nolet 2012). A series of bandpass filters is applied, such that for every ray path up to five frequency-dependent arrival times are measured ('multiple-frequency tomography', Sigloch *et al.* 2008). The redundancy in our data set derives in part from lack of independence among observations in different passbands for the same source-station combination.

The original matrix occupied 6.37 Gbytes on disk, with a sparsity of 1.5 per cent. The densest row had a sparsity of 3.2 per cent. A first reduction is again obtained by optimizing the sparse representation, which reduces the size to 3.72 Gbyte.

We started with a tough clustering, specifying a narrow $\Delta_{\max} =$ 400 km. In this first run a total of 898 clusters was found, many of them too small to condense them by projection. However, the largest 43 clusters (each with more than 100 paths) have 15 092 data or 60 per cent of the total data set. The average sparsity of these 43 clusters is 2.4 per cent, with the densest matrix 3.9 per cent sparse, indicating that we are successful in retaining sparsity. The remaining clusters were then subjected to two similar rounds of clustering with $\Delta_{\max}$ increased to 1400 and 2400 km, respectively. The resulting 172 matrices have an average sparsity of 2.3 per cent (three matrices had a sparsity of 4 per cent or more). A small fraction, 132 data (0.5 per cent), was not clusterable or ended up in clusters with less than 10 data. These were not subjected to the projection procedure, but simply added as A_{rest}.

Again using a cut-off at a SNR of 1, the final size of the projected matrices is 0.38 Gbyte, a reduction to 5.9 per cent of the original size.

We inspect the largest cluster, coloured red in Fig. 5. Fig. 6 shows several rows of the matrix $\mathbf{\Lambda}^{-1}U^T A$—the sensitivity to the projected data, weighted by the eigenvalue—plotted on a vertical cross-section through the mantle, showing that the rows cluster about the ray paths from sources beneath the Sea of Japan and a dense receiver region (mostly U.S. Array stations) in North America. The top left-hand plot is representative for eigenvectors with large λ_i that have the nature of an average over a banana-shaped zone of sensitivity. As the row number increases (and the SNR decreases), the sensitivity becomes more and more complex, and extends over a wider region. Note that the complexity increases most towards the receiver end, a consequence of the dense array coverage allowing for higher resolution.

The first three eigenvectors (not plotted) are dominated by the 'correction' columns. Since origin-time and hypocentral corrections have a large weight, including them tends to dominate the first few eigenvectors. The part of these vectors in model space takes the character of an averaging kernel, while the correction terms ensure the orthogonality. This depends somewhat on the prior uncertainty used to scale the correction parameters (we used 20 km for hypocentre location, 1 s for the origin time), but the matrix entries for corrections will always dominate numerically. Note that classical techniques to render the system insensitive to source time and location (Spencer and Gubbins 1980; Masters *et al.* 2000) cannot be applied since one event may occur in more than one matrix cluster. In general one thus has to include corrections to the source parameters into the linear system, and solve for them simultaneously with the tomographic model. Since the number of them is usually much smaller than the number of data (240 corrections against 25 046 data in this case) this poses no extra burden to speak of.

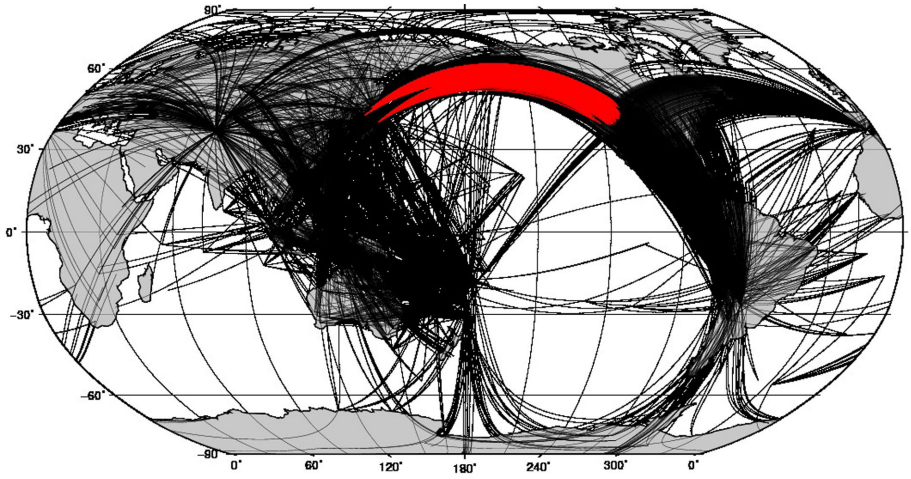

Figure 5. The ray coverage for the complete data set of *P* waves from deep earthquakes. The paths in the first cluster are indicated by the red colour.

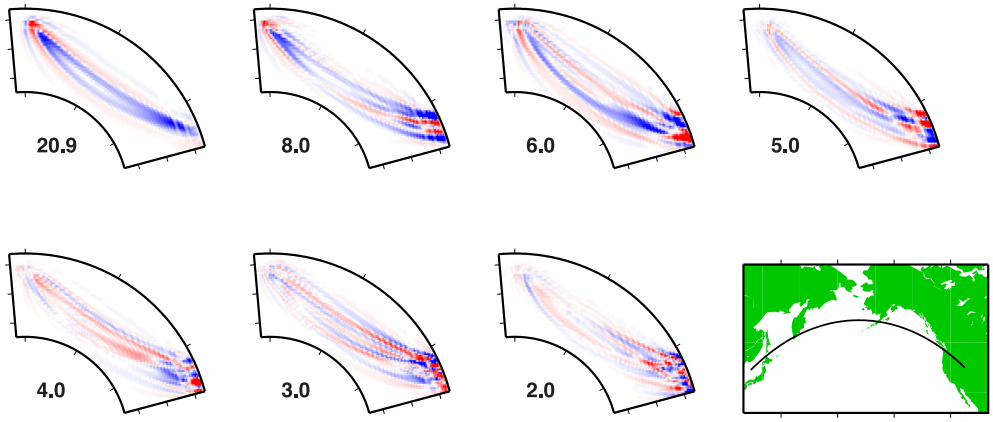

Figure 6. This figure shows a vertical cross-section through selected rows of the matrix $\mathbf{\Lambda}^{-1}U^T A$ for the largest cluster of *P*-wave delays. Numbers in the left corner denote the signal-to-noise ratio for the associated projected datum τ_i. The colour scale is between ± 1.6 in each plot. The apparent pixelation of the images is due to the finite size of the voxels used to parameterize the model. The map shows the geographical location of these cross-sections (solid line).

4.3 ISC delay times

ISC delay times have long been used in seismic tomography. The delays represent the onset of the body wave, and are therefore related to the model perturbations m with ray theory, that is the rows of A are line integrals, rather than the volume integrals used for cross-correlation delays. The matrices are thus much sparser. In the cubed Earth parametrization of Charléty *et al.* (2013), *P*-wave ray paths cover typically 100–200 voxels, equivalent to a sparsity of the order of 0.01 per cent.

While this may seem to relax the memory requirements, the fact that the ISC database contains tens of millions of delays still makes it desirable to be able to condense the linear system without loss of information. Clustering is always useful to estimate data errors with eq. (13). However, the clustering that works well for volume kernels was found to fail in the case of ray-theoretical matrices, since the decrease in sparsity tends to compensate the gains obtained by projection. In one example, a cluster of 3770 rays, truncated at a SNR of 1 for 823 data, saw the size of the matrix *increase* by a factor of more than 5.

The recommended strategy is therefore the classical remedy of summary rays, averaging rays over very closely located events to the same station. The best way to do so is to average both the delays and their associated matrix rows, rather than average the delays only and adopt some representative ray path. This involves more computation but avoids modelling errors, and has the advantage to give the rays a narrow width, which is more in accordance with the otherwise non-modelizable finite-frequency effects caused by noise (Stark and Nikolaev 1993). For cluster $\mathcal{S}$ with $N_\mathcal{S}$ members this gives one summary row:

$$\sum_{j=1}^{M} \frac{1}{N_\mathcal{S}} \left(\sum_{i \in \mathcal{S}} A_{ij} \right) m_j = \frac{1}{N_\mathcal{S}} \sum_{i \in \mathcal{S}} d_i \pm \sigma_\mathcal{S} \,. \tag{25}$$

The standard error $\sigma_\mathcal{S}$ in the averaged datum can be found from (Nolet 2008):

$$\sigma_\mathcal{S}^2 = \frac{1}{N_\mathcal{S}} \sum_{i \in \mathcal{S}} \sigma_i^2 + \sigma_m^2 \,,$$

where σ_m^2 provides a water level that accounts for the error caused by ignoring lateral variations within the summary ray. It can be estimated from the distribution of projected data using eq. (13).

5 DISCUSSION AND CONCLUSIONS

The detail visible in the sensitivity such as shown in Figs 4 and 6 for data with a significant SNR justifies the use of models with a dense parameterization, at least locally, such as to avoid any loss of information by underparametrizing. The clustering SVD mitigates,

or even removes, the computer memory problems posed by the resulting large matrices.

5.1 Cut-off criteria

There is some flexibility in the choice of cut-off rank k, such that even huge matrices might still be inverted with an acceptable loss of information, for example by setting the threshold for SNR higher than 1. For clusters, the cut-off determined by χ^2 (eq. 15) leads to a smaller system than using the SNR of the projected data, but this is misleading: usually the data assembled in a cluster can easily be fitted with an average velocity perturbation, and τ_k may be close to zero for high k. This average velocity is not likely to be sufficient to fit the total data set, for which lateral variations within the cluster may be needed. The use of (15) is thus restricted to the case where a full matrix is reduced. If the system is large enough to make clustering necessary—and this will often be the case—the cluster χ^2 has little use and (14) is the preferred diagnostic that can determine the cut-off rank k. The same applies to model perturbations over the cluster (eqs 18 and 19). In practice, eq. (14) is thus the most powerful diagnostic we have to reduce the matrix size of very large systems without the risk of losing significant information. Although we use a sharp cut-off for the eigenvalues when reducing the matrix in size, one is still free to use 'smoothing' or other regularization techniques when inverting the reduced system. Voronin *et al.* (2014) show how the main characteristics of a tomographic inversion remain preserved even with a drastic culling of eigenvectors.

5.2 Prior correlations

We assumed the data errors as well as the model perturbations to be *a priori* uncorrelated, in a Baysian sense. In principle both could be transformed to diagonalize their covariance matrix if we have prior knowledge of correlations—the difficulty is that precise information is not available and any prior covariance is at best an educated guess. For the data one therefore usually resigns oneself to use uncorrelated errors.

Whether one is justified to impose prior smoothness constraints on the model is debatable (see the discussion in Nolet 2008, p. 280). Of course, many tomographers prefer the 'smoothest possible' solution so as not to introduce unwarranted detail that might be misinterpreted. But such regularizing towards a smooth model can always be done at the time of inversion, and is not needed at the time of winnowing the data as done in this paper.

5.3 Other error estimators

It is of interest to compare the error estimation presented in this paper with earlier efforts to estimate σ_e^2. Although many efforts have been made to estimate the true variance of the errors e_i in body wave delay times (e.g. Morelli and Dziewonski 1987; Gudmundsson *et al.* 1990; Bolton and Masters 2001), considerable uncertainty exists. To the best of our knowledge no formal analysis of the errors in global surface wave delays exists, while the published estimates for the errors differ considerably even for *P*-wave delays.

Morelli & Dziewonski (1987) use summary rays to find σ_m^2 and σ_e^2 in the ISC delay time data. The assumption is that such rays have the same delay if the ray paths are very close. The variation of delays within a single bundle of summary rays then is representative for the observational error in the delays. If there are many rays in the bundle, the error in the average tends to zero: statistical theory states that the variance in an estimate over N samples decreases as $1/N$, and therefore the standard error as $1/\sqrt{N}$. By comparing the variation among delay averages of many bundles with different geographical locations, one obtains also an estimate of σ_m^2 as $N \to \infty$. Plotting the variance σ_N^2 of delay averages in bundles with N rays against N allows one to fit a curve for σ_N^2 :

$$\sigma_N^2 = \frac{\sigma_e^2}{N} + \sigma_m^2 \tag{26}$$

by optimizing σ_e^2 and σ_m^2.

The difficulty with this method is that, to obtain a sufficient number N of rays in the bundle, the source and receiver regions must be large (Morelli and Dziewonski choose $5^\circ \times 5^\circ$ areas). Gudmundsson *et al.* (1990) try to circumvent this by analysing the variance also as a function of the bundle width and investigating the variance in the limit of zero bundle width.

The similarity between (13) and (26) is deceptive. Even though the clusters apparently replace the summary rays in the earlier methods, we allow for the model to influence the distribution of observed delays over the cluster and we avoid the assumption that the true delays are the same over every ray path in the cluster. The cluster can therefore have a larger population than a typical summary ray, which improves the statistics. Since the clustering SVD allows for overparametrization there need be no danger to underestimate σ_m. We observe also that (26) represents a distribution over many ray bundles, whereas (13) refers to the distribution over one cluster only. The two approaches are thus fundamentally different.

Formal error estimates are also obtained when observing delays using the cross-correlation method of VanDecar and Crosson (1990), which is at the basis of recent data analysis strategies (Lou *et al.* 2013; Bonnin *et al.* 2014; Lou & van der Lee 2014), but in our experience these may be highly optimistic, probably because errors in the delay estimates between overlapping pairs of stations are assumed to be independent, which they clearly are not when, for example, a reverberation is present in the waveform of one particular station that influences all cross-correlations with that station.

The formal error estimates provide a rationale for the truncation of eigenvalues and is therefore essential to the matrix size reduction. Equally essential is the clustering. For body wave delays, we found that finite-frequency theory produces matrices with a sparsity of 1.5 per cent. Clustering succeeds well in keeping the loss of sparsity under control, since sparsity is raised only slightly to 2.3 per cent in the projected matrices.

For ray-theoretical matrices, appropriate for onset times such as published by the ISC, the method of summary rays is more effective to reduce matrix size than the projection method described in this paper. The disadvantage of summary rays is that one has little control over how much information is lost in the averaging. However, one could imagine applying the clustering to selected ray subsets, and investigate the eigenvalue drop-off as a function of the summary ray width. Ideally one would choose a summary ray width that results in only one significantly large eigenvalue.

5.4 Delays versus waveforms

The matrices investigated in this paper reflect delays, that is data in the phase domain. They are thus inherently more linear than full waveform data which involve harmonics like $\cos \omega \delta t$ that can only be linearized if the delay $\delta t \ll \omega^{-1}$. Abandoning waveform information for the linearity of delay times may at first sight seem unwise. However, because of the extra non-linearity, the computational demands of waveform tomography are extremely large, and

some difficulties are still encountered with non-linearity. This restricts waveform tomography to low frequencies, and delay times remain the only option for the higher frequencies.

Though the ray-based 'banana-doughnut' kernels rely on the identification of a ray path for the cross-correlated wave, one can compute the sensitivity for any part of the seismogram using finite-difference or spectral element methods (Tromp *et al.* 2005; Nissen-Meyer *et al.* 2007). Moreover, multi-frequency tomography, involving the measurement of body wave dispersion (Sigloch *et al.* 2008), recovers at least some of the information present in waveforms, as was shown by Mercerat *et al.* (2014). The large reduction in matrix size obtained then offers the perspective to forego the time-consuming gradient search now generally used with the adjoint approach. If a smooth background model is used, the kernels computed for delays in identifiable arrivals, using full waveform theory with the spectral element method, are similar to those computed with ray theory for *P* waves, and only slightly more complicated for *S* waves (Mercerat and Nolet 2012). The difference is caused by energy not modelled by ray theory, such as reverberations, mode conversions and diffractions, that may remove the 'doughnut hole' where sensitivity is small, thus reducing the sparsity of the kernels. But the sparsity of such kernels is only slightly reduced, and we suspect that the difference in sparsity for arbitrary waveforms (not associated with a simple ray path) will be similar. The difference in sparsity of *clusters*, which also tend to fill in the doughnut hole, might even be negligible. If that is the case, and if the linearity of delays holds for arbitrary waveforms, solving the reduced matrix system would have the ability to greatly speed up the adjoint approach. This will be the subject of future research.

ACKNOWLEDGEMENTS

The distribution of deep earthquakes for the *P*-wave data was contributed by Ebru Bozdag. The surface wave data matrix reflects the path coverage for a subset of surface wave data obtained from Hendrik-Jan van Heijst and Jeroen Ritsema. This research was financially supported by ERC Advanced Grant 226837.

REFERENCES

Bolton, H. & Masters, G., 2001. Travel times of P and S from global digital seismic networks: implications for the relative variation of P and S velocity in the mantle, *J. geophys. Res.,* **106,** 13 527–13 540.

Bonnin, M., Nolet, G., Villasenor, A., Gallart, J. & Thomas, C., 2014. Multiple-frequency tomography of the upper mantle beneath the African/Iberian collision zone, *Geophys. J. Int.,* **198,** 1458–1473.

Bozdag, E., Lefebvre, M., Lei, W., Peter, D.B., Smith, J.A., Zhu, H., Komatitsch, D. & Tromp, J., 2013. Global seismic imaging based on adjoint tomography, in *Proceedings of the AGU Fall Meeting,* San Francisco, Abstracts S33A-2381.

Charléty, J., Voronin, J., Nolet, G., Loris, I., Simons, F. & Daubechies, I., 2013. Global seismic tomography with sparsity constraints: comparison with smoothing and damping regularization, *J. geophys. Res.,* **118,** 4887–4899.

Chevrot, S. & Zhao, L., 2007. Multiscale finite frequency Rayleigh wave tomography of the Kaapvaal craton, *Geophys. J. Int.,* **169,** 201–215.

Dahlen, F., Hung, S.-H. & Nolet, G., 2000. Fréchet kernels for finite-frequency traveltimes—I. Theory, *Geophys. J. Int.,* **141,** 157–174.

Dutto, L., Lepage, C. & Habashi, W., 2000. Effect of the storage format of sparse linear systems on parallel CFD computations, *Comput. Methods Appl. Mech. Engrg.,* **188,** 441–453.

Fichtner, A., Bunge, H.-P. & Igel, H., 2006. The adjoint method in seismology—I. Theory, *Phys. Earth planet. Inter.,* **157,** 86–104.

Gilbert, F., 1971. Ranking and winnowing gross earth data for inversion and resolution, *Geophys. J. R. astr. Soc.,* **23,** 125–128.

Gudmundsson, O., Davies, J. & Clayton, R., 1990. Stochastic analysis of global travel time data: mantle heterogeneity and random errors in the ISC data, *Geophys. J. Int.,* **102,** 25–44.

Lou, X. & van der Lee, S., 2014. Observed and predicted North American teleseismic delay times, *Earth planet. Sci. Lett.* Available at: http://dx.doi.org/10.1016/j.epsl.2013.11.056, last accessed Date July 28 2014.

Lou, X., van der Lee, S. & Loyd, S., 2013. AIMBAT: A Python/Matplotlib tool for measuring teleseismic arrival times, *Seism. Res. Lett.,* **84,** 85–93.

Maggi, A., Tape, C., Chen, M., Chao, D. & Tromp, J., 2009. An automated time-window selection algorithm for seismic tomography, *Geophys. J. Int.,* **178,** 257–281.

Masters, G., Laske, G., Bolton, H. & Dziewonski, A., 2000. The relative behaviour of shear velocity, bulk sound speed and compressional velocity in the mantle: implications for chemical and thermal structure, in *Earth's Deep Interior,* pp. 63–88, eds Karato, S.-I., Forte, A.M., Liebermann, R.C., Masters, G. & Stixrude, L., AGU.

Mercerat, E. & Nolet, G., 2012. Comparison of ray-based and adjoint-based sensitivity kernels for body-wave seismic tomography, *Geophys. Res. Lett.,* **39,** L12301, doi:10.1029/2012GL052002.

Mercerat, E. & Nolet, G., 2013. On the linearity of cross-correlation delay times in finite-frequency tomography, *Geophys. J. Int.,* **192,** 681–687.

Mercerat, E., Nolet, G. & Zaroli, C., 2014. Cross-borehole tomography with correlation delay times, *Geophys.,* **79,** R1–R12.

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

Nissen-Meyer, T., Dahlen, F. & Fournier, A., 2007. Spherical-earth Fréchet sensitivity kernels, *Geophys. J. Int.,* **168,** 1051–1066.

Nolet, G., 2008. *A Breviary of Seismic Tomography,* Cambridge Univ. Press.

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

Ritsema, J., van Heijst, H., Deuss, A. & Woodhouse, J., 2011. S40RTS: a degree-40 shear-velocity model for the mantle from new Rayleigh wave dispersion, teleseismic traveltimes, and normal-mode splitting function measurements, *Geophys. J. Int.,* **184,** 1223–1236.

Sigloch, K., McQuarrie, N. & Nolet, G., 2008. Two-stage subduction history under North America inferred from multiple-frequency tomography, *Nat. Geosci.,* **1,** 458–462.

Simons, F. *et al.*, 2011. Solving or resolving global tomographic models with spherical wavelets, and the scale and sparsity of seismic heterogeneity, *Geophys. J. Int.,* **187,** 969–988.

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

Stark, P. & Nikolayev, D., 1993. Towards tubular tomography, *J. geophys. Res.,* **98,** 8095–8106.

Tape, C., Liu, Q., Maggi, A. & Tromp, J., 2009. Adjoint tomography of the Southern California crust, *Science,* **325,** 988–992.

Tape, C., Liu, Q., Maggi, A. & Tromp, J., 2010. Seismic tomography of the Southern California crust based on spectral-element and adjoint methods, *Geophys. J. Int.,* **180,** 433–462.

Tian, Y., Montelli, R., Nolet, G. & Dahlen, F., 2007. Computing traveltime and amplitude sensitivity kernels in finite-frequency tomography, *J. Comp. Phys.,* **226,** 2271–2288.

Tromp, J., Tape, C. & Liu, Q., 2005. Seismic tomography, adjoint methods, time reversal and banana-doughnut kernels, *Geophys. J. Int.,* **160,** 195–216.

van Heijst, H.-J. & Woodhouse, J., 1999. Global high-resolution phase velocity distributions of overtone and fundamental mode surface waves determined by mode branch stripping, *Geophys. J. Int.,* **137,** 601–620.

VanDecar, J. & Crosson, R., 1990. Determination of teleseismic arrival times using multi-channel cross-correlation and least squares, *Bull. seism. Soc. Am.,* **80,** 150–159.

Voronin, S., Nolet, G. & Mikesell, D., 2014. Compression approaches for the regularized solutions of linear systems from large-scale inverse problems, preprint (arXiv:1404.5684).

APPENDIX: SPARSE MATRIX STORAGE

If the model is described by a local parameterization, the resulting matrix A is sparse, that is most of its elements are zero (Nolet 2008). Very small elements can be set to zero with little loss of precision. The original matrix elements have been set to 0 if smaller than 3×10^{-4} times the largest element, where we used the theoretical row sum for a smooth background model (Tian *et al.* 2007) to check the accuracy and make sure that this truncation does not introduce errors in the predicted values of the delays larger than a prescribed value, usually a few per cent.

Before writing $U_k^T A$ to disk, we again perform this truncation of small elements. Since $U_k^T A = \mathbf{\Lambda}_k V_k^T$, while the rows of V_k^T are eigenvectors normalized to 1, the ith row of $U_k^T A$ is a vector of length λ_i. This property can be used to monitor the quality of the truncation. Note that the elements of the first rows are larger than those of rows belonging to the smaller eigenvalues. This is essentially why the first data have a better SNR than data associated with smaller λ_i, even though we showed that all projected data have the same standard error. Thus, the first rows decrease little in norm as a result of the truncation, but the effect is stronger when the eigenvalue, and the SNR associated with it, decreases. We keep track of the effects of truncation by checking that the length of the truncated row to that of the predicted vector length agree to better than 1 per cent. The error introduced by this truncation remains well below the observational uncertainty.

A common storage format for unstructured sparse matrices is the compressed sparse row (CSR) format, which uses an array $a(i)$ to store the non-zero elements of A, an array $ja(i)$ to store the column number of $a(i)$, and an array $na(k)$ that has either the starting index of row k in a and ja, or the number of elements in row k (Dutto *et al.* 2000). For an $N \times M$ matrix with S non-zeros, this requires $2S + N$ numbers.

Though tomographic matrices are neither band- nor block-structured (for which more powerful storage systems exist), the sensitivity kernels that form the rows of A are geographically restricted. The non-zeros in each row of A therefore occur often in groups. We found that we can obtain a significant reduction in the size of $ja(i)$ by redefining the sparse matrix format:

(i) an isolated non-zero is defined as in the classic CSR format,
(ii) the first non-zero of a group is identified by giving $ja(i)$ a negative sign,
(iii) the (positive) $ja(i + 1)$ that follows a negative $ja(i)$ indicates the last member of a non-zero group, and
(iv) $na(k)$ gives the number of $ja(i)$ in row k.

This scheme requires $\alpha S + N$ numbers with $1 + 2(N/S) \leq \alpha \leq 2$. In practice we find that this reduces the matrices for the volumetric sensitivity of finite-frequency body waves by about 30 per cent with respect to a classical CSR format, while for ray-theoretical matrices the improvement is minimal (about 5 per cent). For the very compact surface wave kernels, the memory needed to store the column numbers $ja(i)$ is an order of magnitude smaller than that needed to store the matrix elements $a(i)$, leading to a reduction of almost a factor of 2: implementing this scheme on the large matrix discussed in Section 4.1, its size was reduced to 55 per cent of the original size.