*Geophys. J. Int.* (2007) **170,** 359–370 doi: 10.1111/j.1365-246X.2007.03409.x

# Tomographic inversion using $\ell_1$-norm regularization of wavelet coefficients

Ignace Loris,[1,2] Guust Nolet,[3] Ingrid Daubechies[1] and F. A. Dahlen[3]

[1] *Program in Applied and Computational Mathematics, Princeton University, Princeton, NJ, USA*
[2] *Dienst Theoretische Natuurkunde, Vrije Universiteit Brussel, Brussels, Belgium. E-mail: igloris@vub.ac.be*
[3] *Department of Geosciences, Princeton University, Princeton, NJ, USA*

Accepted 2007 February 15. Received 2006 December 12; in original form 2006 July 25

## SUMMARY

We propose the use of $\ell_1$ regularization in a wavelet basis for the solution of linearized seismic tomography problems $\mathbf{Am} = \mathbf{d}$, allowing for the possibility of sharp discontinuities superimposed on a smoothly varying background. An iterative method is used to find a sparse solution $\mathbf{m}$ that contains no more fine-scale structure than is necessary to fit the data $\mathbf{d}$ to within its assigned errors.

**Key words:** inverse problem, one-norm, sparsity, tomography, wavelets.

## 1 INTRODUCTION

Like most geophysical inverse problems, the linearized problem $\mathbf{Am} = \mathbf{d}$ in seismic tomography is underdetermined, or at best offers a mix of over and underdetermined parameters. It has, therefore, long been recognized that it is important to suppress artefacts that could be falsely interpreted as 'structure' in the earth's interior. Not surprisingly, strategies that yield the smoothest solution $\mathbf{m}$ have been dominant in most global or regional tomographic applications; these strategies include seeking global models represented as a low-degree spherical harmonic expansion (Dziewonski *et al.* 1975; Dziewonski & Woodhouse 1987; Masters *et al.* 1996) as well as regularization via minimization of the gradient ($\nabla\mathbf{m}$) or second derivative ($\nabla^2\mathbf{m}$) norm of a dense local parametrization (Constable *et al.* 1987; Nolet 1987; Spakman & Nolet 1988; VanDecar & Snieder 1994; Trampert & Snieder 1996).

Smooth solutions, however, while not introducing small-scale artefacts, produce a distorted image of the earth through the strong averaging over large areas, thereby making small-scale detail difficult to see, or even hiding it. Sharp discontinuities are blurred into gradual transitions. For example, the inability of global, spherical-harmonic, tomographic models to yield as clear an image of upper-mantle subduction zones as produced by more localized studies has long been held against them. Deal *et al.* (1999) and Deal & Nolet (1999) optimize images of upper-mantle slabs to fit physical models of heat diffusion, in an effort to suppress small-scale imaging artefacts while retaining sharp boundaries. Portniaguine & Zhdanov (1999) use a conjugate-gradient method to seek the smallest possible anomalous domain by minimizing a norm based on a renormalized gradient $\nabla\mathbf{m}/(\nabla\mathbf{m}\cdot\nabla\mathbf{m}+\gamma^2)^{\frac{1}{2}}$, where $\gamma$ is a small constant. Like all methods that deviate from a least-squares type of solution, both these methods are non-linear and pose their own problems of practical implementation.

From a mathematical point of view, the linear inverse problem $\mathbf{Am} = \mathbf{d}$ requires regularization whenever $\mathbf{A}$ has either very small singular values or a null space. Even small variations in $\mathbf{d}$ can then cause large differences in $\mathbf{m}$; regularization restores stability to the problem (Bertero & Boccacci 1998). The choice among the many different regularization techniques depends on the problem, and can be governed, to some extent, by the desire to preserve or emphasize particular features of the (unknown) $\mathbf{m}$ that are expected. For instance, if the small singular values of $\mathbf{A}$ are mostly associated to highly oscillatory singular functions (as is often the case), and if one is mostly interested in large-scale features of $\mathbf{m}$, then it is natural to introduce a regularizing constraint or penalty term involving $\nabla\mathbf{m}$ or $\nabla^2\mathbf{m}$. Or, if one seeks to find solutions $\mathbf{m}$ that are 'sparse' with respect to a given basis $\{\varphi_\ell\,;\,\ell=1,\ldots,L\}$, that is, solutions of the form $\mathbf{m}=\sum_{\ell=1}^{L} c_\ell\varphi_\ell$ in which only a small fraction of the coefficients $c_1, c_2, \ldots, c_L$ are non-zero, then this can be achieved by a constraint or a penalty term involving $\|c\|_1=\sum_{\ell=1}^{L}|\langle\mathbf{m},\varphi_\ell\rangle|$. This regularization procedure selects solutions that are particularly 'simple' when expressed in terms of the basis $\{\varphi_\ell\,;\,\ell=1,\ldots,L\}$.

The notion that we should seek the 'simplest' model $\mathbf{m}$ that fits a measured set of data $\mathbf{d}$ to within the assigned errors is intuitively equivalent to the notion that the model should be describable with a small number of parameters. However, clearly, restricting the model to a few low-degree spherical-harmonic or Fourier coefficients, or a few large-scale blocks or tetrahedra, does not necessarily lead to a geophysically plausible solution. In this paper, we investigate whether a multiscale representation based upon wavelets (Daubechies 1992) has enough flexibility to represent the class of models we seek. We propose an $\ell_1$-norm regularization method which yields a model $\mathbf{m}$ that has a strong tendency to be *sparse* in a wavelet basis, meaning that it can be faithfully represented by a relatively small number of non-zero wavelet coefficients. This allows for models that vary smoothly in regions of limited coverage without sacrificing any sharp or small-scale features in well-covered regions that are required to fit the data. Our approach is different from an approach briefly suggested by de Hoop & van der Hilst (2005), in which the mapping between data and model is decomposed in

curvelets: here we are concerned with applying the principle of parsimony to the solution of the inverse problem, without any special preference for singling out linear features, for which curvelets are probably better adapted than wavelets. We choose finite-frequency modelling and not ray theory as a framework for our analysis, and focus on the effect of the regularization itself on seismic reconstructions. These effects will be present whatever the assumption on the underlying physics is, and hence are worthy of study in their own right.

In Section 2, we give a short description of the mathematical method, and in Section 3 we consider a geophysically motivated, toy 2-D application, in which the synthetic data are a small set of regional, fundamental-mode, Rayleigh-wave dispersion measurements expressed as wavenumber perturbations $\delta k(\nu)$ at various frequencies $\nu$. To enable us to concentrate on the mathematical rather than the geophysical aspects of the inverse problem, we assume that the fractional shear velocity perturbations $\delta \ln \beta = \delta\beta/\beta$ within the region are depth-independent. Finite-frequency interpretation of the surface wave dispersion data (Zhou *et al.* 2004) then yields a 2-D linearized inverse problem of the form $\mathbf{Am} = \mathbf{d}$. We compare wavelet-basis models $\mathbf{m}$ obtained using our proposed $\ell_1$-norm regularization with models obtained using more conventional $\ell_2$ regularization, both with and without wavelets, and show that the former are sparser and have fewer small-scale artefacts.

## 2 MATHEMATICAL PRINCIPLES

In any realistic tomographic problem, the linear system $\mathbf{Am} = \mathbf{d}$ is not invertible: even when the number of data exceeds the number of unknowns, the least-squares matrix $\mathbf{A}^T\mathbf{A}$ is (numerically) singular. Additional conditions always have to be imposed. The proposed regularization method is based on the fundamental assumption that the model $\mathbf{m}$ is sparse in a wavelet basis (Daubechies 1992). We believe that this is an appropriate inversion philosophy for finding a smoothly varying model while still allowing for whatever sharp or small-scale features are required to fit the data $\mathbf{d}$. An important feature of the method is that the location of the small-scale features does not have to be specified beforehand.

The properties of the wavelet basis promote the generation of a solution that we call 'parsimonious', that is, void of pre-suppositions and as economical in its description as allowed by the data. This use of the principle of parsimony is true to the philosophy of Occam's razor (e.g., Constable *et al.* 1987). However, it should not be confused with the common practice of finding the 'smoothest' model that fits the data, because that choice pre-supposes that the true earth tends to be smooth. For an extensive discussion of the principle of parsimony in modern science, see Gauch (2003).

A wavelet decomposition is a special kind of basis transformation that can be computed efficiently (the number of operations is proportional to the number of components in the input). At each step the algorithm strips off detail belonging to the finest scale present—this detail is encoded in wavelet coefficients, broadly corresponding to local differences—and calculates a coarse version—encoded in scaling coefficients, broadly corresponding to local averages—that is only half the size of the original in 1-D and only one quarter the size in 2-D. This procedure is repeated on the successive coarse versions. The resulting wavelet coefficients (at the different scales) and scaling coefficients (at the final coarsest scale only) are called the wavelet decomposition of the input. By this construction each wavelet coefficient carries information belonging to a certain scale (by virtue of the decimation) and a certain position (use of *local* differences). The final few scaling coefficients represent a (very) coarse average.

The mathematical relation between the wavelet-basis expansion coefficients $\mathbf{w}$ and the model $\mathbf{m}$ is the wavelet transform $\mathbf{W}$ (a linear operator): $\mathbf{m} = \mathbf{W}^T\mathbf{w}$. By choosing the local differences and averages carefully (corresponding to a choice among many different so-called wavelet families), the inverse transformation from $\mathbf{w}$ back to $\mathbf{m}$ can be made equally efficient. In our application we will use a special kind of 2-D wavelet basis that is overcomplete: it contains six different wavelets corresponding to different directions. Because of this overcompleteness, the wavelet transform $\mathbf{W}$ has a left inverse (namely $\mathbf{W}^T$): $\mathbf{W}^T\mathbf{W} = \mathbf{I}$, but no right inverse, $\mathbf{WW}^T \neq \mathbf{I}$. Appendix B contains a short overview of this particular construction. In short, our wavelet and scaling coefficients $\mathbf{w}$ contain information on scale, position and direction.

For the tomographic reconstruction, we will require a sparse set of wavelet-basis coefficients: the vast majority of these represent differences and will only be present around non-smooth features. In this way we regularize the inversion by adapting ourselves to the model rather than to the operator. As a measure of sparsity we will use the $\ell_1$-norm of the wavelet representation $\mathbf{w}$ of the model $\mathbf{m}$, that is, we will look for a solution of the linear equations $\mathbf{Am} = \mathbf{d}$ that has a small $\|\mathbf{w}\|_1 = \sum_i |w_i|$. Since $|w_i| > |w_i|^2$ for small $w_i$ and $|w_i| < |w_i|^2$ for large $w_i$, this type of penalization will favor a small number of large coefficients over a large number of small coefficients in the reconstruction (whereas a traditional $\ell_2$ penalization might do the opposite—see the right-hand panel in Fig. 1). The left-hand panel in Fig. 1 illustrates heuristically how $\ell_1$ penalization leads to sparse solutions, in a model space with only two degrees of freedom, in which the collection of $\mathbf{w}$ that satisfy $\mathbf{d} = \mathbf{AW}^T\mathbf{w}$ constitutes a whole line. To find the solution with smallest $\ell_2$-norm, one can imagine taking a small circle around the origin and 'growing' its radius until it first touches the solution line: the tangent point is the minimum $\ell_2$-norm solution. Likewise, to find the solution with smallest $\ell_1$-norm, one 'grows' a small $\ell_1$-ball around the origin until it first touches the solution line: the touching point is the minimum $\ell_1$-norm solution, which is indeed sparser, since it has only one coordinate different from zero. Sparsity of minimum $\ell_1$-norm solutions follows from the 'pointed' nature of the $\ell_1$-balls, as compared to the 'rounder' $\ell_2$-balls.

Even though $\ell_1$ penalized solutions will always be sparse, they need not be unique in general (in much higher dimensions than in Fig. 1), and they need not lead to the very sparsest solution. Penalizations by $\ell_p$-norms with $p < 1$ have also been proposed, but they have the inconvenience of leading to non-convex problems, for which finding global minimizers is much more tricky. For $p = 0$, the penalty term is simply the number of non-zero coefficients, and arguably the 'best' measure for sparsity; finding the optimal solution with this penalty is not computationally tractable, however, for the number of coefficients used for seismic tomography reconstruction (many thousands). For this reason, $\ell_1$ penalization remains a preferred method to promote sparsity. In addition, recent work has shown (Donoho 2004; Candes *et al.* 2006) that for some very large classes of matrices $\mathbf{A}$, and when sufficiently sparse solutions exist, the $\ell_1$ minimizer is unique, and leads to the sparsest solution.

In particular, our strategy will consist of finding the minimizer of the functional

$$I_1(\mathbf{w}) = \|\mathbf{d} - \mathbf{Am}\|_2^2 + 2\tau\|\mathbf{w}\|_1 = \|\mathbf{d} - \mathbf{AW}^T\mathbf{w}\|_2^2 + 2\tau\|\mathbf{w}\|_1, \quad (1)$$

where $\tau$ is an adjustable parameter at our disposal. Here, the first (quadratic) term corresponds to the conventional statistical measure of misfit to the data, $\chi^2 = \sum_i [d_i - (\mathbf{Am})_i]^2$, and the second

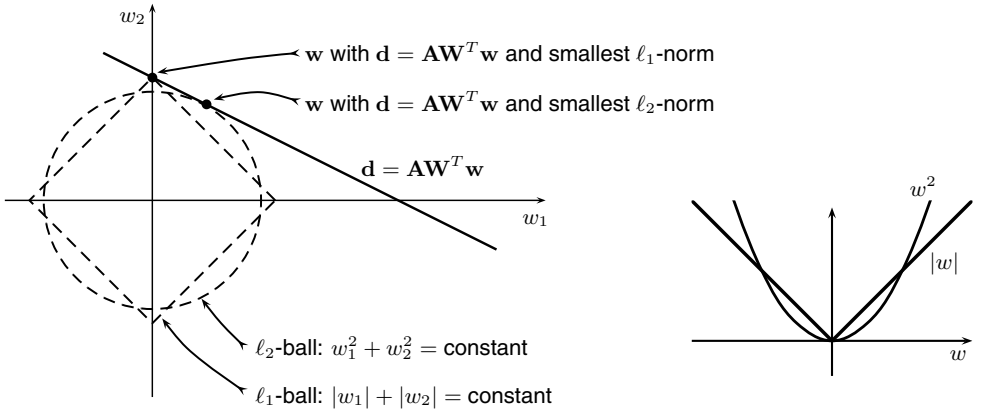


**Figure 1.** Sparsity, $\ell_1$ minimization and $\ell_2$ minimization. Left: schematic representation of the space of wavelet coefficients (only 2 coefficients out of 16 384 are shown). The set of all solutions to the underdetermined linear equations $\mathbf{d} = \mathbf{A}\mathbf{W}^T\mathbf{w}$ is represented by a straight line. The aim of regularization is to choose one point of this line that we will be considered as the actual solution. Because the $\ell_1$-ball has no bulge, the solution with smallest $\ell_1$-norm is sparser (only one non-zero component) than the solution with smallest $\ell_2$-norm (two non-zero components). Right: a $|w|$ penalization effects small coefficients more and large coefficients less than the (traditional) $w^2$ penalization.

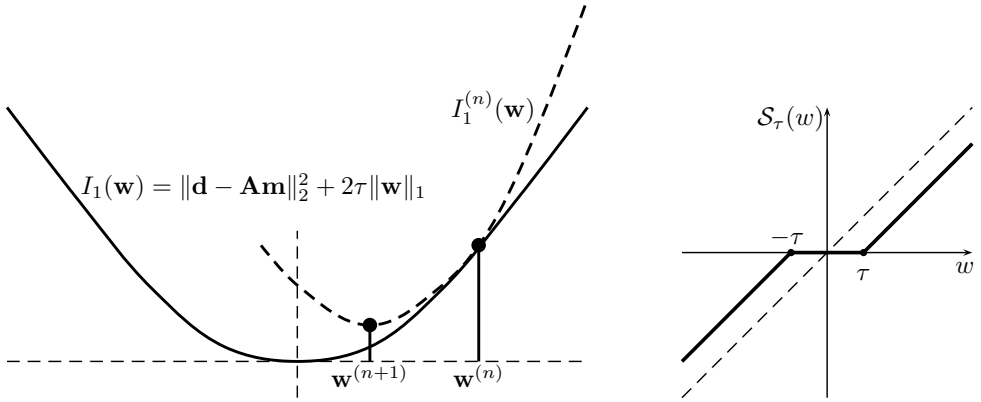


**Figure 2.** Left: the functional $I_1(\mathbf{w})$ is approximated in the vicinity of $\mathbf{w}^{(n)}$ by a surrogate functional $I_1^{(n)}(\mathbf{w})$, constructed in such a way that its minimum is easy to find (eq. 5). This defines the next step in the iteration. Right: soft thresholding function $\mathcal{S}_\tau(w)$.

($\ell_1$-norm) term $2\tau\|\mathbf{w}\|_1$ is introduced to regularize the inversion. In writing $\chi^2$ in this form, we have made the simplifying assumption that the noisy data $\mathbf{d}$ are uncorrelated with unit variance. More generally, the misfit portion of the functional (1) is $\chi^2 = (\mathbf{d} - \mathbf{A}\mathbf{m})^T \Sigma^{-1} (\mathbf{d} - \mathbf{A}\mathbf{m})$, where $\mathbf{\Sigma}$ is the data covariance matrix. In the 2-D toy problem considered in Section 3, we invert synthetic data $\mathbf{d}$ having a constant (but non-unit) variance, $\Sigma = \sigma^2\mathbf{I}$. The threshold $\tau$ is a regularization parameter. It measures the relative importance of data fit and $\ell_1$ norm (proxy for sparsity) of a candidate solution. A higher threshold will correspond to a sparser solution (but with a worse data fit); a lower threshold will correspond to a better data fit, but less sparse solution. This parameter will remain present in the reconstruction algorithm below. The exact choice of $\tau$ depends on the level of noise present in the data, as discussed in Section 3.

The minimizer of the functional (1) can be found by iteration (Daubechies *et al.* 2004): starting with the present approximation $\mathbf{w}^{(n)}$ one constructs an $n$th-iterate surrogate functional

$$I_1^{(n)}(\mathbf{w}) = I_1(\mathbf{w}) - \left\|\mathbf{A}\mathbf{W}^T\left(\mathbf{w} - \mathbf{w}^{(n)}\right)\right\|_2^2 + \left\|\mathbf{w} - \mathbf{w}^{(n)}\right\|_2^2 \qquad (2)$$

that has the same value and the same derivative at the point $\mathbf{w} = \mathbf{w}^{(n)}$ as the original functional (see Fig. 2). This surrogate functional can be rewritten as

$$I_1^{(n)}(w) = \left\|\mathbf{w} - \left(\mathbf{W}\mathbf{A}^T\mathbf{d} + (\mathbf{I} - \mathbf{W}\mathbf{A}^T\mathbf{A}\mathbf{W}^T)\mathbf{w}^{(n)}\right)\right\|_2^2 + 2\tau\|\mathbf{w}\|_1 + c^{(n)}, \qquad (3)$$

where $c^{(n)}$ is independent of $\mathbf{w}$. This functional has a much simpler form than the original $I_1(\mathbf{w})$ because there is no operator $\mathbf{A}\mathbf{W}^T$ mixing different components of $\mathbf{w}$. The next approximation $\mathbf{w}^{(n+1)}$ is defined by the minimizer of this new functional. By calculating the derivative of expression (3) with respect to a specific wavelet or scaling coefficient $w_i$, one finds the following set of component-by-component equations:

$$w_i - \left(\mathbf{W}\mathbf{A}^T\mathbf{d} + (\mathbf{I} - \mathbf{W}\mathbf{A}^T\mathbf{A}\mathbf{W}^T)\mathbf{w}^{(n)}\right)_i + \tau\,\mathrm{sign}(w_i) = 0, \qquad (4)$$

valid whenever $w_i \neq 0$. These equations are solved by distinguishing the two cases $w_i > 0$ and $w_i < 0$; the solution —corresponding to the minimizer of the surrogate functional $I_1^{(n)}(w)$, and denoted by $\mathbf{w}^{(n+1)}$—is then found to equal

$$\mathbf{w}^{(n+1)} = \mathcal{S}_\tau\left[\mathbf{W}\mathbf{A}^T\mathbf{d} + (\mathbf{I} - \mathbf{W}\mathbf{A}^T\mathbf{A}\mathbf{W}^T)\mathbf{w}^{(n)}\right], \qquad (5)$$

where $S_\tau$ is the so-called soft-thresholding operation, that is,

$$\mathcal{S}_\tau(w) = \begin{cases} w - \tau & w \geq \tau \\ 0 & |w| \leq \tau \\ w + \tau & w \leq -\tau, \end{cases} \qquad (6)$$

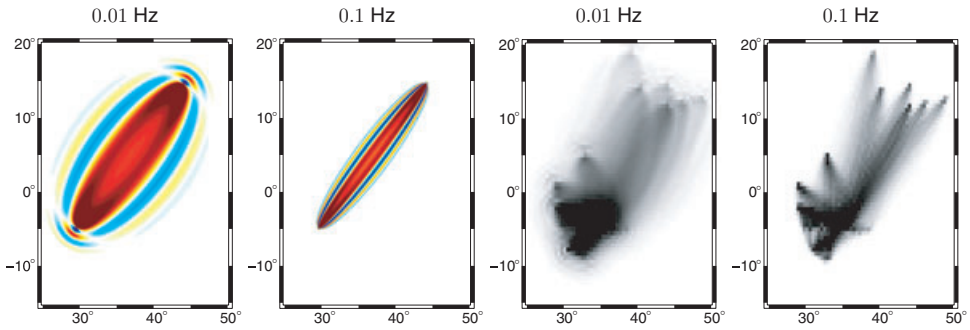


**Figure 3.** Left: Map view of a typical 2-D sensitivity kernel $K_{2-\mathrm{D}}(x, y, \nu)$ at the lowest-frequency considered, $\nu \approx 0.01$ Hz. Second from left: Highest-frequency ($\nu \approx 0.1$ Hz) kernel for the same source–receiver path. Both kernels exhibit structure on a much finer scale than the resolution of the model, necessitating the $n_x \times n_y$ numerical integration to compute the matrix $\mathbf{A}$ in eq. (15). Red denotes negative values, $K_{2-\mathrm{D}}(x, y, \nu) < 0$, and blue denotes positive values, $K_{2-\mathrm{D}}(x, y, \nu) > 0$. The cross-path tapering of the kernels as a result of the finite time-domain taper, eqs (A6)–(A7), is clearly visible. Second from right: The sum (over all source–receiver pairs) of the absolute value of the lowest-frequency ($\nu \approx 0.01$ Hz) integrated kernels (as computed in eq. 15). Far right: The sum (over all source–receiver pairs) of the absolute value of the highest-frequency ($\nu \approx 0.1$ Hz) integrated kernels (as computed in eq. 15). The coverage is adequate in the vicinity of the East African rift (by design of the original seismic deployment) but poor elsewhere.

performed on each wavelet or scaling coefficient $w_i$ individually. The starting point of the iteration procedure is arbitrary, for example, $\mathbf{w}^{(0)} = \mathbf{0}$. Because of the component-wise character of the tresholding, it is straightforward to use different thresholds $\tau_i$ for different components $w_i$ if desired, and in fact, we shall use different thresholds $\tau_{\mathrm{w}}$ and $\tau_{\mathrm{s}}$ for the wavelet and scaling coefficients in our application. A schematic representation of the idea behind the iteration (5) is given in Fig. 2. We realize that this iteration converges slowly for ill-conditioned matrices, but we use it here because it is proven to converge to the solution (Daubechies *et al.* 2004).

An improvement in convergence can be gained by rescaling the operator $\mathbf{A}$ (and rescaling the data $\mathbf{d}$ at the same time) in such a way that the largest eigenvalue of $\alpha^2\mathbf{A}^T\mathbf{A}$ is close to (but smaller than) unity. The iteration corresponding to the minimization of this new, rescaled functional is

$$\mathbf{w}^{(n+1)} = \mathcal{S}_{\tau\alpha^2}\big[\alpha^2\mathbf{W}\mathbf{A}^T\mathbf{d} + (\mathbf{I} - \alpha^2\mathbf{W}\mathbf{A}^T\mathbf{A}\mathbf{W}^T)\mathbf{w}^{(n)}\big]. \tag{7}$$

We will also make use of the following two-step procedure: from the outcome $\overline{\mathbf{m}} = \mathbf{W}^T\overline{\mathbf{w}}$ of the iteration (7), we define new, linearly shifted data $\mathbf{d}' = 2\mathbf{d} - \mathbf{A}\overline{\mathbf{m}}$ and restart the same iteration with this new data:

$$\mathbf{w}^{(n+1)} = \mathcal{S}_{\tau\alpha^2}\big[\alpha^2\mathbf{W}\mathbf{A}^T\mathbf{d}' + (\mathbf{I} - \alpha^2\mathbf{W}\mathbf{A}^T\mathbf{A}\mathbf{W}^T)\mathbf{w}^{(n)}\big], \quad \mathbf{w}^{(0)} = \overline{\mathbf{w}}. \tag{8}$$

The outcome $\overline{\overline{\mathbf{m}}} = \mathbf{W}^T\overline{\overline{\mathbf{w}}}$ of this second iteration is then the final, regularized reconstruction of the model. For the same value of the regularization parameter $\tau$, the second step improves the data fit considerably, $\|\mathbf{d} - \mathbf{A}\overline{\overline{\mathbf{m}}}\|_2^2 < \|\mathbf{d} - \mathbf{A}\overline{\mathbf{m}}\|_2^2$; hence a given level of final data fit $\chi^2$ will, in the two-step procedure, correspond to a higher value of $\tau$. Because $\tau$ specifies the threshold level, a higher value will lead to more aggressive thresholding and thus faster convergence to a sparse solution.

The above method will be demonstrated in the next section and compared to a conventional $\ell_2$-regularization method, in which the functional

$$I_2(\mathbf{m}) = \big\|\mathbf{d} - \mathbf{A}\mathbf{m}\big\|_2^2 + \tau\big\|\mathbf{m}\big\|_2^2 \tag{9}$$

is minimized (the crucial difference with $I_1(\mathbf{w})$ being the second term). This gives rise to the familiar system of damped normal equations

$$(\mathbf{A}^T\mathbf{A} + \tau\mathbf{I})\mathbf{m} = \mathbf{A}^T\mathbf{d}, \tag{10}$$

whose solution $\mathbf{m} = (\mathbf{A}^T\mathbf{A} + \tau\mathbf{I})^{-1}\,\mathbf{A}^T\,\mathbf{d}$ can be found using a linear solver of choice, since $\mathbf{A}^T\mathbf{A} + \tau\mathbf{I}$ is a regular matrix. To emphasize the similarities and differences with the $\ell_1$ method, we adopt the classical Landweber iteration (Landweber 1951) that can be (but in modern applications seldom is) used for solving the linear eq. (10):

$$\mathbf{m}^{(n+1)} = \mathbf{A}^T\mathbf{d} + \big[\mathbf{I} - (\mathbf{A}^T\mathbf{A} + \tau\mathbf{I})\big]\mathbf{m}^{(n)}, \qquad \mathbf{m}^{(0)} = \mathbf{0}. \tag{11}$$

No thresholding is employed here. Rescaling of the operator and the data again improves the rate of convergence:

$$\mathbf{m}^{(n+1)} = \alpha^2\mathbf{A}^T\mathbf{d} + \big[\mathbf{I} - (\alpha^2\mathbf{A}^T\mathbf{A} + \tau\alpha^2\mathbf{I})\big]\mathbf{m}^{(n)}, \quad \mathbf{m}^{(0)} = \mathbf{0}. \tag{12}$$

Of course it is also possible to solve the linear system (10) using a conjugate-gradient or similar algorithm in much less time.

A third option is to use an $\ell_2$ penalization on the wavelet coefficients. This allows us to penalize the scaling coefficients differently than the wavelet coefficients (with the help of different penalization parameters $\tau_{\mathrm{s}}$ and $\tau_{\mathrm{w}}$). We can use the following iteration, similar to formula (12), but now in the wavelet domain:

$$\mathbf{w}^{(n+1)} = \alpha^2\mathbf{W}\mathbf{A}^T\mathbf{d} + \big[\mathbf{I} - (\alpha^2\mathbf{W}\mathbf{A}^T\mathbf{A}\mathbf{W}^T + \alpha^2\tilde{\mathbf{I}})\big]\mathbf{w}^{(n)}, \quad \mathbf{w}^{(0)} = \mathbf{0}, \tag{13}$$

where $\tilde{\mathbf{I}}$ acts as the $\tau_{\mathrm{w}} \times$ identity on wavelet coefficients and as the $\tau_{\mathrm{s}} \times$ identity on scaling coefficients. If we were to use an orthonormal wavelet basis ($\mathbf{W}^T\,\mathbf{W} = \mathbf{W}\mathbf{W}^T = \mathbf{I}$) for our expansions, and if we penalized every coefficient the same, $\tau_{\mathrm{w}} = \tau_{\mathrm{s}} = \tau$, then this method would be identical to the previous $\ell_2$ method.

In the following, we consider both one- and two-step $\ell_1$ wavelet penalization as well as conventional $\ell_2$ penalization, both without and with wavelets, using the rescaled iterative schemes (7), (8), (12) and (13) for the purposes of comparison.

## 3 IMPLEMENTATION

To test the above ideas, we devised a dramatically simplified, 2-D, synthetic surface wave inversion problem very loosely modelled after an actual PASSCAL deployment in Tanzania (Owens *et al.* 1995). Since our primary aim is to test the feasibility of the method outlined in the previous section, the geological realism of the toy problem was not a primary concern. We opt for a surface wave data set because it easily allows for a reduction to 2-D, enabling a simple visual comparison of models obtained using various inversion strategies,

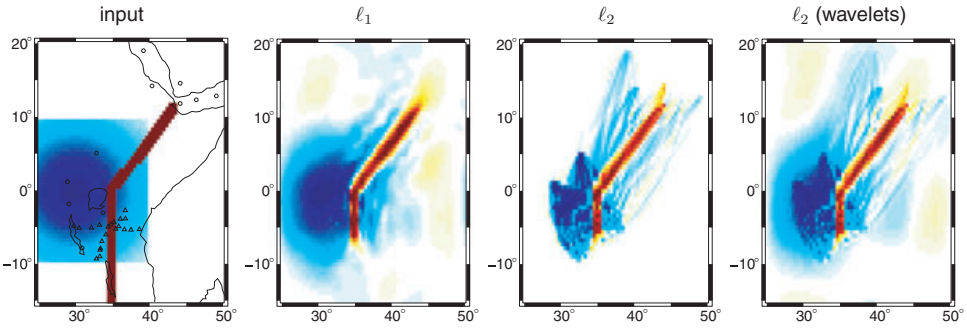


**Figure 4.** From left to right: toy 2-D velocity model for the East African rift and adjacent continental craton, showing the seismic stations (triangles) and earthquake events (circles); reconstructed model using the two-step $\ell_1$-penalization method; reconstruction using the spatial-domain $\ell_2$ method; reconstruction using the wavelet-domain $\ell_2$ method. The two-step $\ell_1$ model is that obtained after 1000 + 1000 iterations, whereas both $\ell_2$ models are after 2000 Landweber iterations. Red denotes low anomalous velocity, $\delta \ln \beta(x, y) < 0$, and blue denotes high velocity, $\delta \ln \beta(x, y) > 0$. The absolute magnitude $|\delta \ln \beta(x, y)|$ is irrelevant, since the inverse problem $\mathbf{Am} = \mathbf{d}$ is linear and the synthetic data are constructed from the input model $\mathbf{m}^{\text{input}}$ (leftmost map) via $\mathbf{d} = \mathbf{Am}^{\text{input}} + \mathbf{e}$.

and avoiding the added complexity of 3-D wavelets, which we intend to address in a separate paper. We adopt the more accurate finite-frequency modelling of the dispersion because the true sensitivity is spread over a wide geographical area (see Fig. 3); note also that a ray-theoretical sensitivity with an infinitely narrow path makes a standard regularization more complicated as one admits more detail into the solution, and since a wavelet basis sets no prior limit to the resolution this might introduce unwanted complications. Fig. 4 (left) shows the hypothetical experimental setup: the highly schematized input model consists of a sharp, bent, East African rift structure with low shear wave velocity, $\delta \ln \beta(x, y) < 0$, superimposed upon a smooth, circular cratonic positive anomaly, $\delta \ln \beta(x, y) > 0$. The input model takes on values between $-0.1$ and $+0.1$, that is, the model has maximum anomalies of $\pm 10$ per cent in shear velocity. Eleven earthquake events (circles) were taken from the NEIC catalogue to mimic realistic regional seismicity for the duration of a typical temporary deployment of the 21 stations (triangles). The locations of the seismic stations and events are listed in Table 1. For each of the $11 \times 21$ source–receiver paths, we assume that fundamental-mode Rayleigh-wave perturbations $\delta k(\nu)$ have been measured at eight selected frequencies between $\nu \approx 0.01$ and 0.1 Hz. These wavenumber perturbations are related to the 2-D, depth-independent velocity perturbations $\delta \ln \beta(x, y)$ via a 2-D, frequency-dependent sensitivity kernel (see Appendix A for more details):

$$\delta k(\nu) = \iint K_{2-\mathrm{D}}(x, y, \nu)\, \delta \ln \beta(x, y)\, dx\, dy. \tag{14}$$

Plots of the lowest-frequency ($\nu \approx 0.01$ Hz) and highest-frequency ($\nu \approx 0.1$ Hz) kernel $K_{2-\mathrm{D}}(x, y, \nu)$ for a typical source–receiver pair are shown in the left two panels of Fig. 3. Because finite-frequency scattering and diffraction effects are accounted for in the kernels $K_{2-\mathrm{D}}(x, y, \nu)$, there is significant off-path sensitivity of the measurements $\delta k(\nu)$ within the first one or two Fresnel zones (Zhou *et al.* 2004). All kernels $K_{2-\mathrm{D}}(x, y, \nu)$ and distances are computed in the flat-earth earth approximation.

The study region, which is 35° (north–south) by 25° (east–west), is subdivided into $N_x \times N_y = 64 \times 64 = 4096$ equal-sized rectangles, and the discretized model vector $\mathbf{m}$ consists of the unknown constant values of $\delta \ln \beta(x, y)$ within each rectangle. To compute the matrix $\mathbf{A}$, which maps the discretized model $\mathbf{m}$ onto the data $\mathbf{d}$ [consisting of multiple $\delta k(\nu)$], each kernel $K_{2-\mathrm{D}}(x, y, \nu)$ is sampled $n_x \times n_y$ times on each of the $n_x \times n_y$ model-vector rectangles and

**Table 1.** List of positions of seismic stations and earthquake events used in the synthetic inversion.

| Stations | | Events | |
|---|---|---|---|
| Longitude | Latitude | Longitude | Latitude |
| 33.3203° | −7.9073° | 29.02° | −1.86° |
| 35.1382° | −4.3238° | 49.10° | 12.85° |
| 32.7712° | −9.2958° | 44.15° | 11.80° |
| 33.2588° | −8.1060° | 30.82° | −7.84° |
| 29.6927° | −4.8392° | 46.34° | 12.33° |
| 38.6170° | −5.3018° | 39.17° | 19.02° |
| 30.3988° | −5.1168° | 32.78° | 5.06° |
| 36.5695° | −5.3223° | 44.15° | 14.57° |
| 37.4763° | −5.3775° | 28.84° | 1.16° |
| 36.7192° | −3.8422° | 40.33° | 14.20° |
| 35.7965° | −4.9040° | 33.67° | −3.05° |
| 36.6983° | −2.7252° | | |
| 34.3462° | −4.9610° | | |
| 34.0560° | −6.0192° | | |
| 35.4007° | −5.2508° | | |
| 33.2415° | −8.9835° | | |
| 33.1842° | −4.7145° | | |
| 33.5180° | −6.9372° | | |
| 34.7315° | −4.6403° | | |
| 36.0163° | −3.8892° | | |
| 32.0832° | −5.0878° | | |

a Riemann sum is used to compute the quantity

$$\iint_{\text{rectangle}(k,l)} K_{2-\mathrm{D}}(x, y, \nu)\, dx\, dy \approx \frac{\Delta x \Delta y}{n_x n_y} \sum_{m,n} K_{2-\mathrm{D}} \left[x_l - \Delta x/2 + (m - 1/2)\delta x,\ y_k - \Delta y/2 + (n - 1/2)\delta y,\ \nu\right], \tag{15}$$

where $\delta x = \Delta x / n_x$ and $\delta y = \Delta y / n_y$. A schematic representation of the $N_x \times N_y$ grid on which the spatial-domain model $\mathbf{m}$ is specified and the $N_x \times N_y$ integration subgrid is shown in Fig. 5. We choose $n_x = n_y = 32$ since we have found that doubling this to $n_x = n_y = 64$ yields a change of less than one percent in the integrated value of $\mathbf{A}$. The dimensions of the resulting matrix $\mathbf{A}$ are 1848 (number of stations × number of events × number of wavenumbers) by 4096 (number of model-vector pixels). To give an idea of the overall degree of coverage, we have plotted the sum (over all station-event pairs) of the absolute value of all of the lowest-frequency and all the highest-frequency discretized kernels in the right two panels of Fig. 3. It is clear that much of the study area,

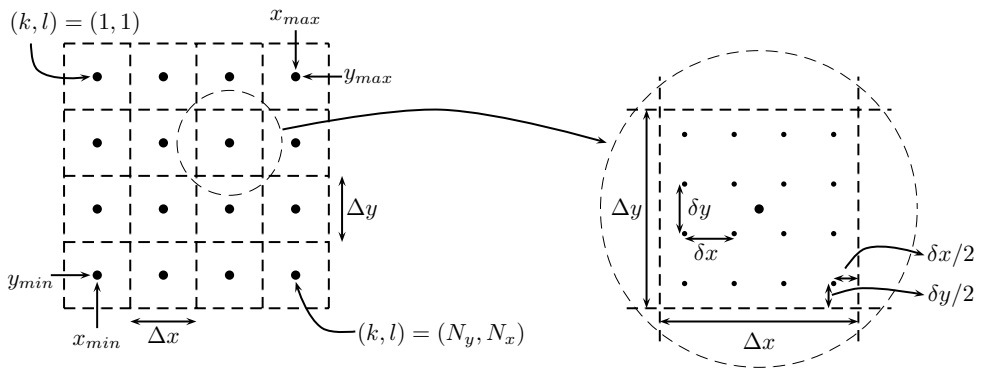


**Figure 5.** Schematic representation of the 2-D Cartesian grids used. Left: $N_x \times N_y$ grid used to specify the model $\mathbf{m}$. Right: Blowup of the finer-scale $N_x \times N_y$ grid used to compute the kernel matrix $\mathbf{A}$ via the approximate integration (15). Since the study region is rectangular in shape (see Fig. 4) and since $N_x = N_y$ and $N_x = N_y$, the actual $\Delta x \times \Delta y$ model pixels and $\delta x \times \delta y$ integration subpixels are also rectangular, rather than square as shown.

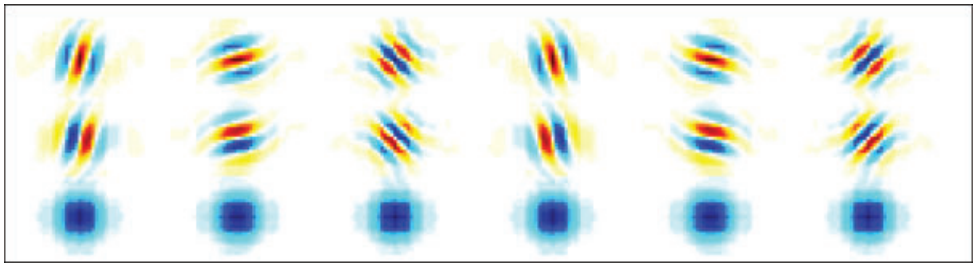

**Figure 6.** Spatial-domain structure of the 2-D dual-tree complex wavelets used in the reconstruction (figure adapted from Selesnick *et al.* (2006)). First row: real part, second row: imaginary part, third row: norm squared (i.e. sum of the squares of the top two plots). The directional character of each of the six wavelet functions is clear. Four different wavelet scales, that is, four different replicas of this picture, each a factor of two smaller than the one above it, are used in both the $\ell_1$ and $\ell_2$ wavelet-basis inversions.

particularly in the northwest and southeast, is completely uncovered (as is typical of real-world, regional seismic experiments).

Using the matrix $\mathbf{A}$ and the input model $\mathbf{m}^{\text{input}}$ with a sharp, low-velocity East African rift superimposed on a broad, high-velocity cratonic structure, we compute synthetic data $\mathbf{d} = \mathbf{A}\mathbf{m}^{\text{input}} + \mathbf{e}$, where we have added Gaussian noise $\mathbf{e}$ with zero mean and a standard deviation of $3.1 \times 10^{-7}$ km$^{-1}$ for all data, which represents a signal-to-noise ratio close to unity at the lowest frequency, but diminishes to a few per cent at the high-frequency end. Since finite-frequency inversions include the effect of scattered wave energy, a high precision of the measurement $\delta k(\nu)$ at high frequency $\nu$ is realistic. The purpose of the proposed algorithm is now to reconstruct $\mathbf{m}$ from the knowledge of the noisy data $\mathbf{d}$, the matrix $\mathbf{A}$ and the linear equations $\mathbf{A}\mathbf{m} = \mathbf{d}$.

For our purposes we will make use of the overcomplete 2-D wavelet basis described by Kingsbury (2002) and Selesnick *et al.* (2006) because of its ability to distinguish different directions (see Fig. 6). We use four wavelet scales, for a total of $4 \times 64^2 = 16\,384$ wavelet and scaling coefficients $\mathbf{w}$ (four times the number of model coefficients $\mathbf{m}$). The starting point for the iterations in both the $\ell_1$ and $\ell_2$ inversions is $\mathbf{w} = \mathbf{0}$ and $\mathbf{m} = \mathbf{0}$. As explained in the previous section we renormalize the $\ell_1$ thresholded iteration by choosing $\alpha = \lambda_{\text{max}}^{-1/2}$ (which in our case equals 4884.5) where $\lambda_{\text{max}}$ is the largest eigenvalue of $\mathbf{A}^T\mathbf{A}$. We let the iteration run for 1000 steps, adjust the data (two-step procedure) and let the second-step algorithm run for another 1000 steps. The threshold $\tau$ is chosen by hand in such a way as to arrive at a final value for the variance-adjusted misfit, $\chi^2 = \|\mathbf{d} - \mathbf{A}\mathbf{m}\|_2^2/\sigma^2$, that is approximately equal to 1848 (the number of data). The noisy data $\mathbf{d}$ are thus fit to within their standard errors $\sigma$ and no better; pushing the fit beyond this would amount to fitting the noise $\mathbf{e}$, which would lead to undesirable artefacts in the resulting model $\mathbf{m}$.

It should also be noted that the thresholding is done on pairs of wavelet coefficients: The wavelets come in pairs (at the same scale, position and orientation) that we interpret as real and imaginary part of a complex wavelet, that is, thresholding corresponds to $(w_k^{\text{re}}, w_k^{\text{im}}) \rightarrow z = w_k^{\text{re}} + i w_k^{\text{im}} \rightarrow \tilde{z} = z\mathrm{S}_\tau(|z|)/|z| \rightarrow [\text{Re}(\tilde{z}), \text{Im}(\tilde{z})]$. This particular method of thresholding is borrowed from image denoising where it is found to make a big difference in avoiding artefacts (van Spaendonck *et al.* 2003; Selesnick *et al.* 2005; Guleryuz 2006). Furthermore, the threshold for the diagonally oriented wavelets is multiplied by 1.2395 because $\|\nabla\psi_{\pm 45^\circ}\|_1 = 1.2395\|\nabla\psi_{\text{other}}\|_1$. We choose the threshold $\tau_{\text{s}}$ for the scaling coefficients to be 1/10th of the threshold $\tau_{\text{w}}$ for the wavelet coefficients; since the scaling coefficients correspond to a few large-scale averages (64 in our case versus more than 16 000 finer-scale wavelet coefficients) it is not so important that these be sparse. Likewise, in the wavelet-basis $\ell_2$ inversions, we set the penalization parameter for the scaling coefficients to 1/10th the value of the penalization parameter for the wavelet coefficients.

The two-step $\ell_1$ algorithm takes about 10 min for 1000 + 1000 iterations on a 1.5 GHz PC. The result of the $\ell_1$ inversion is compared with the outcome of both of the $\ell_2$ methods, with and without using wavelets, with the thresholding or penalization parameter $\tau$ chosen in every case to achieve the same data fit: $\chi^2 \approx 1848$ (see Fig. 7). The number of Landweber iterations is 2000, so that the total number of two-step $\ell_1$ and single-step $\ell_2$ iterations is the same. The spatial-domain $\ell_2$-regularization method yields a relative modelling error $\|\mathbf{m} - \mathbf{m}^{\text{input}}\|_2/\|\mathbf{m}^{\text{input}}\|_2$ of about 74 per cent, whereas the two-step $\ell_1$ method yields a relative modelling error of only

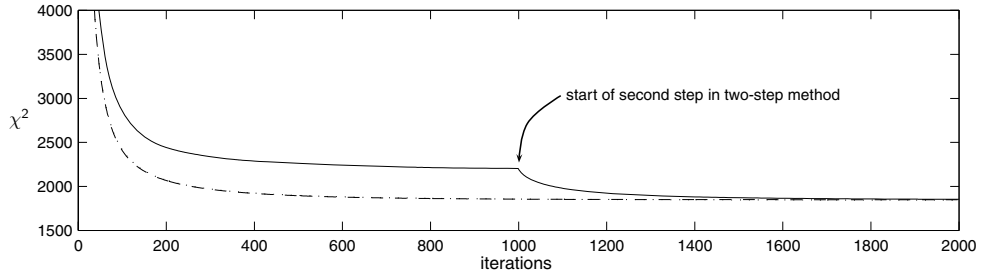


**Figure 7.** Graph of the variance-adjusted data fit $\chi^2 = \|\mathbf{d} - \mathbf{Am}\|_2^2 / \sigma^2$ versus the number of iterations: two-step $\ell_1$-regularization method (solid line), spatial-domain $\ell_2$ method (dashed line) and wavelet-basis $\ell_2$ method (dotted line, visually coincident with dashed line). The thresholding and penalization parameter $\tau$ has in each case been tailored so that the final value of $\chi^2$, after 1000 + 1000 or 2000 iterations, is equal to the number of data, namely 1848. Note the improvement in the rate of convergence toward the model with $\chi^2 = 1848$ after the implementation of the second step in the two-step $\ell_1$ iteration.

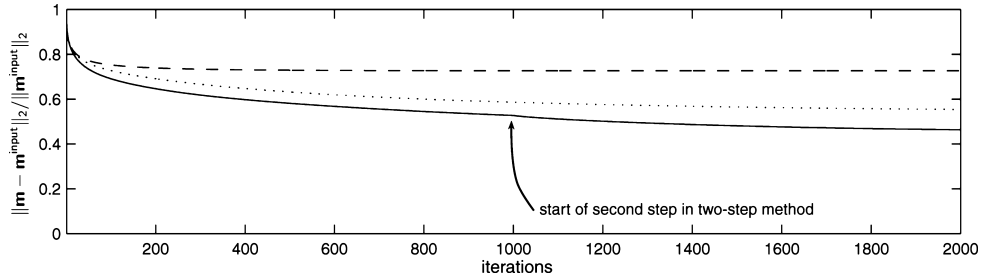


**Figure 8.** Graph showing the relative modelling error $\|\mathbf{m} - \mathbf{m}^{\text{input}}\|_2 / \|\mathbf{m}^{\text{input}}\|_2$ versus the number of iterations: two-step $\ell_1$ method (solid line), spatial-domain $\ell_2$ method (dashed line) and wavelet-basis $\ell_2$ method (dotted line). The $\ell_1$-regularization method clearly yields the most faithful reconstruction of the input model $\mathbf{m}^{\text{input}}$. Note the (slight) improvement in the rate of decrease of the modelling error following the start of the second step in the two-step $\ell_1$ iteration.

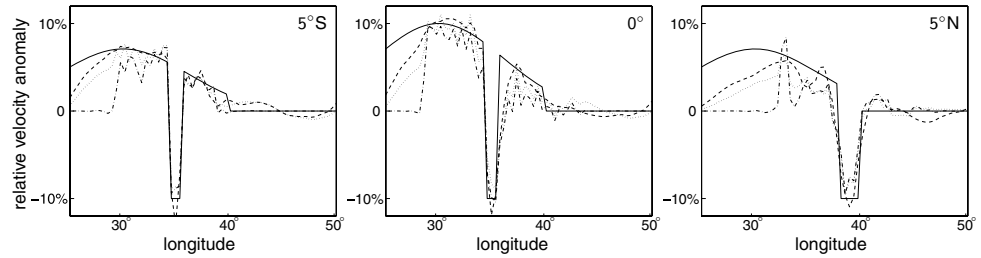


**Figure 9.** East–west slices through the input model and the various reconstructions. Left to right: −5° latitude, 0° (i.e. through the 'elbow' of the rift), 5° latitude. Solid line: input model; dashed line: $\ell_1$ reconstruction; dash–dotted line: $\ell_2$ reconstruction; dotted line: $\ell_2$ reconstruction (with wavelets).

47 per cent (see Fig. 8), and is clearly less noisy (compare the middle two maps in Fig. 4). The wavelet-basis $\ell_2$-regularized inversion (rightmost map in Fig. 4) is only slightly less noisy, with a relative modelling error of about 55 per cent (Fig. 8). One feature that can never be recovered in any of the reconstructions is the southern part of the rift, which does not lie between any station-event pair. In addition to the relative reconstruction error (which is global measure of the performance of the methods) depicted in Fig. 8, in Fig. 9 we exhibit three east–west slices of the input model and the various reconstructions. These slices give a more quantitative idea of the reconstruction quality than the colour maps; in particular, it is evident that the $\ell_1$ method best recovers the narrow low-velocity rift at 35° longitude without introducing spurious short-wavelength features elsewhere.

In Fig. 10, we compare the wavelet coefficients $\mathbf{w}$ of the input model, the two-step $\ell_1$ reconstruction and the wavelet-basis $\ell_2$ reconstruction. In accordance with our basic assumption, the $\ell_1$-regularized model is sparse in the wavelet basis. Most of the small-scale coefficients $\mathbf{w}$ are zero—in agreement with the original model on the left—indicating the effectiveness of the iterative thresholding algorithm. The wavelet coefficients of the wavelet $\ell_2$ reconstruction are clearly not sparse. Also this solution seems to suffer from large-scale artefacts (see Fig. 4, rightmost map).

In the leftmost plot in Fig. 11 we show $\chi^2$ versus $\|\mathbf{w}\|_1$ trade-off curves for the $\ell_1$ reconstruction method, both with and without using the two-step procedure. After 1000 + 1000 iterations, the $\ell_1$ wavelet norm $\|\mathbf{w}\|_1$ of the two-step reconstructed model is lower—for the same value of $\chi^2$—than the corresponding norm of the model produced by 2000 iterations of the first step, with no subsequent redefinition of the data $\mathbf{d}$ and reiteration. This is an indication that 2000 total iterations is inadequate to achieve full convergence, since the fully converged model, which minimizes the functional $I_1(\mathbf{w})$ given in eq. (1), must be the minimum-norm model for a fixed value of $\chi^2$ by definition. A much larger number of iterations seems to be required to guarantee convergence. To construct the second set of trade-off curves in Fig. 11, we employed 150 000 + 150 000 iterations in the two-step case and 300 000 in the single-step case; such a large number would be prohibitive in any larger-scale, more realistic, 3-D application. We have chosen to limit the iteration counts to 1000 + 1000 or 2000 in all of our model-space comparisons, since any changes in the spatial-domain features of the models $\mathbf{m}$ are barely discernible to the eye with further iteration. The

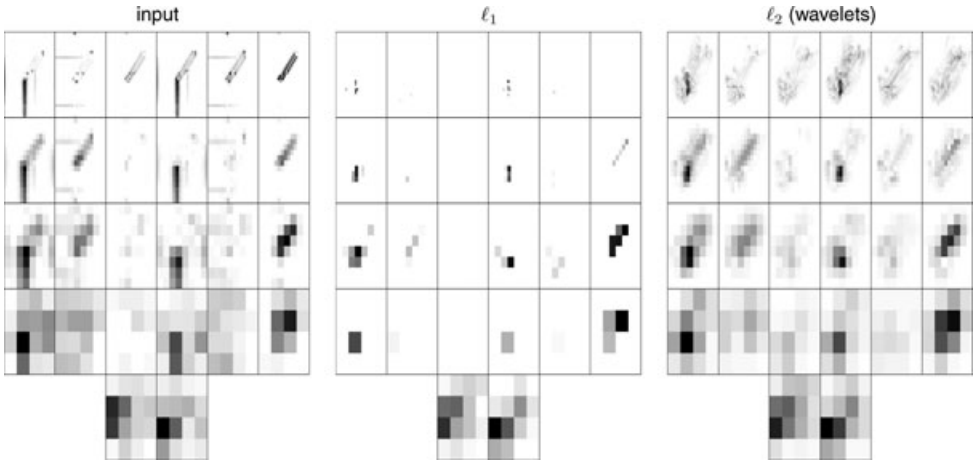


**Figure 10.** Graphical display of (the modulus of) the wavelet and scaling coefficients. Left: coefficients of the synthetic input model $\mathbf{m}^{\text{input}}$. Middle: coefficients of the two-step $\ell_1$ model after $1000 + 1000$ iterations. Right: coefficients of the wavelet-basis $\ell_2$ model after 2000 iterations. The four wavelet scales are plotted, smallest to largest, top to bottom. Each row shows the six different wavelet directions, plotted next to each other in the same left-to-right order as the wavelets plotted in Fig. 6. Scaling coefficients are plotted on the bottom row. White denotes a zero coefficient, $w_i = 0$. Each rectangle corresponds to the spatial domain 25°E–50°E × 15°S–20°N.

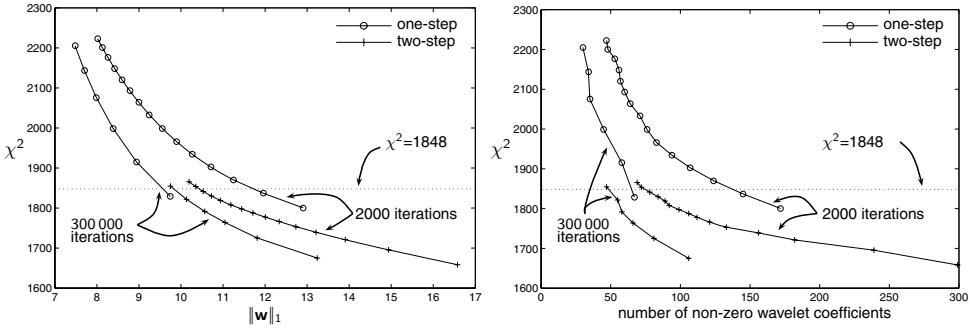


**Figure 11.** Left: Data misfit $\chi^2$ versus $\ell_1$ wavelet norm $||\mathbf{w}||_1$ trade-off curve. Right: Alternative trade-off curve showing $\chi^2$ versus the number of non-zero wavelet coefficients of the model **m**. Different values of the thresholding parameter $\tau$ were used to determine each point on the various curves. Circles: one-step method, after either 2000 or 300 000 iterations; crosses: two-step method after an equivalent number (either $1000 + 1000$ or $150\,000 + 150\,000$) of total iterations. The relative positions of the one-step and two-step curves suggests that 300 000 total iterations is sufficient to achieve full convergence. The horizontal dotted lines show the statistically meaningful value of the noisy data misfit, $\chi^2 = 1848$ (the number of data).

rightmost plot in Fig. 11 shows the principal advantage of using the two-step iteration procedure: for the same total number of iterations, either $1000 + 1000 = 2000$ or $150\,000 + 150\,000 = 300\,000$, the number of non-zero wavelet coefficients of the two-step models is always lower than the corresponding number for the single-step models. The two-step $\ell_1$ procedure, therefore, leads more quickly to a sparser wavelet-basis solution, as expected.

We also compared the single-step and two-step $\ell_1$ inversion methods with the corresponding $\ell_2$ reconstruction methods, both with and without wavelets, for a number of other input synthetic models. These include three checkerboard patterns of decreasing scale and a model similar to the geologically inspired one in Fig. 4, but with a more curvaceous low-velocity rift (see Fig. 12). Both the single-step $\ell_1$ reconstructions and the $\ell_2$ reconstructions are computed using 2000 iterations, whereas the two-step $\ell_1$ models are computed using $1000 + 1000$ iterations. In all cases, the two-step $\ell_1$ models are the most parsimonious and, therefore, to most geoscientists the most acceptable. One could consider using smoothness damping to improve the quality of the $\ell_2$ images; however, this would be done at the cost of resolving the sharpness of the rift structure. A skeptic could perhaps also argue that the 'rift' structure in the model produced by the $\ell_1$ procedure extends further northwards, albeit diminished in amplitude, whereas conventional $\ell_2$ regularization without wavelets exhibits a sharper cut-off, more like the input model. It achieves this sharp cut-off, however, at the expense of many artefacts elsewhere, especially along dominant ray directions. Perhaps the most noteworthy feature of the $\ell_1$ regularization method is its suppression of the artefacts resembling high-frequency kernel images that are streaked along surface wave ray paths in all the $\ell_2$ models, to the north of the rift and within the craton. This is one of the most serious artefacts that plague conventional seismic tomography: $\ell_2$ regularization frequently if not always seems to enhance the well-sampled regions of the model. The $\ell_1$ wavelet-basis reconstructions show no signs of this familiar deficiency.

The computational bottleneck in the present 2-D synthetic study is not the wavelet transform—which is fast, certainly on a model **m** of modest dimension $64 \times 64$—or even the number of iterations, but it is simply the size of the matrix **A**. A significant amount of time is needed to accurately pre-compute **A**, and considerable memory is needed to store the computed elements in memory; this is necessary because the product $\mathbf{A}^T\mathbf{A}$ is used in every step of the iteration. Doubling of the resolution in every direction results in a fourfold increase in size of the model **m**, and a sixteen-fold increase in the number of elements in the square matrix $\mathbf{A}^T\mathbf{A}$. All calculations were

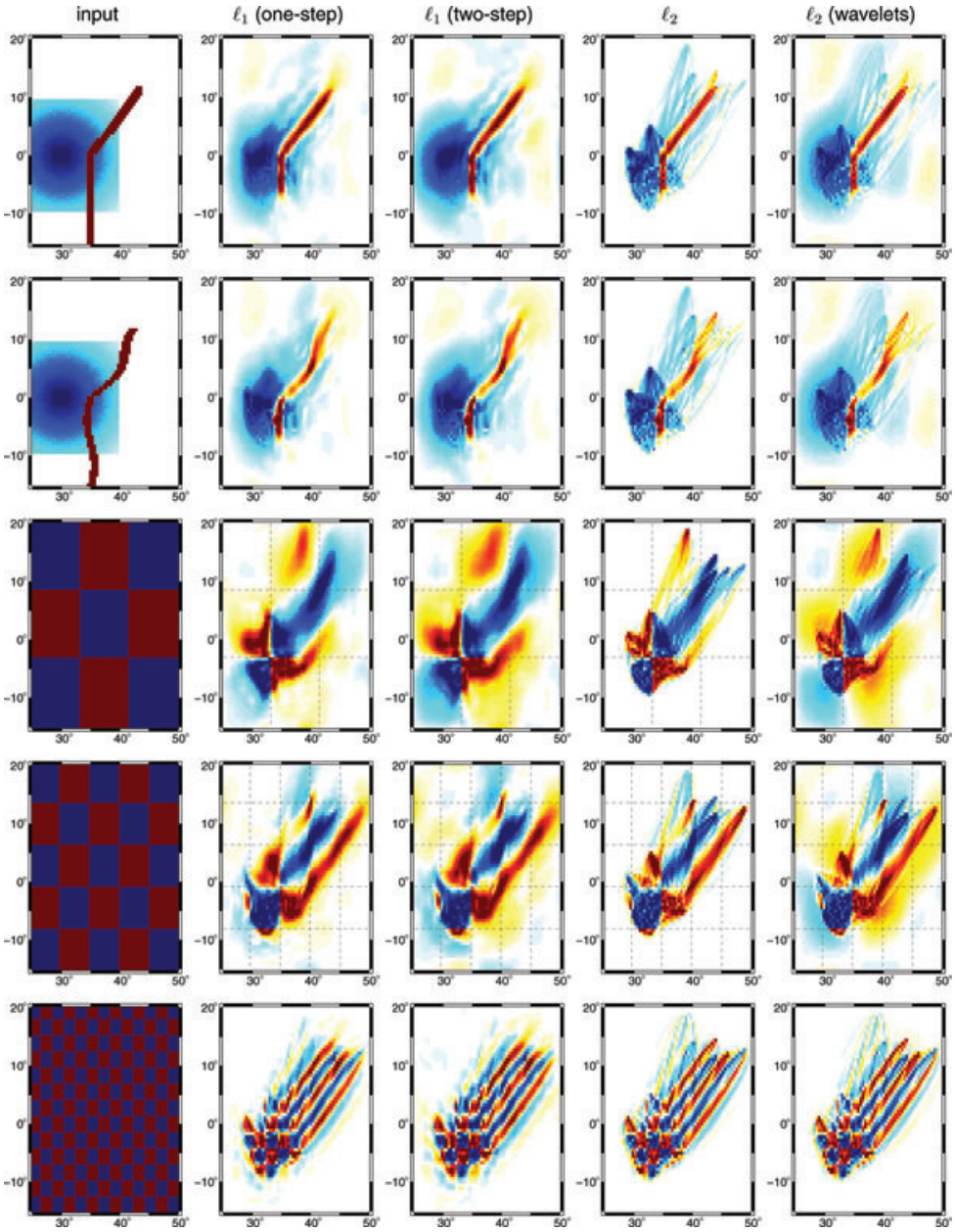


**Figure 12.** Results of applying different reconstruction techniques to a number of different 2-D toy models. From left to right: Original input model; $\ell_1$ reconstruction (2000 iterations, single-step procedure); $\ell_1$ reconstruction (1000 + 1000 iterations, two–step procedure); $\ell_2$ reconstruction (2000 iterations, without wavelets); $\ell_2$ reconstruction (2000 iterations, with wavelets).

performed using MATLAB; software for the 2-D dual-tree wavelets was downloaded from Selesnick *et al.* (2006).

In 3-D, with a much more realistic geophysical problem, nothing fundamentally changes for the regularization. The iterated soft-thresholding algorithm is still applicable (it can be used to regularize any linear problem whatever its actual origin). In this setting wavelet transform and soft-thresholding operation are still fast (i.e. the number of operations is proportional to the number of variables). One important difference however is the increase of the number of independent variables and (probably) the number of data. This leads to a significantly larger numerical matrix, which has to be computed and then stored. Another technical point that needs to be addressed is the choice of wavelet basis, in particular, if one works on the whole sphere. This is linked to the choice of grid on which the inversion problem is discretized. A suitable compromise between geophysical pre-requisites and practical considerations needs to be found.

## 4 CONCLUSIONS

We tested several new methods of regularization through wavelet decomposition of a toy 2-D tomographic problem characterized by both smooth and sharp velocity anomalies. A variety of synthetic inversion experiments show that minimization of the $\ell_1$-norm of a wavelet decomposition of the model leads to tomographic images that are parsimonious in the sense that they use only a few wavelets and still represent both smooth and sharp features well without introducing significant blurring or artefacts. The $\ell_1$-norm performs significantly better than an $\ell_2$ regularization on either the model or

its wavelet decomposition. In particular, ray path-associated artefacts are almost completely suppressed.

The choice of dual-tree complex wavelets in 2-D, representing six space directions, is sufficient to avoid directional bias, and efficient in modelling both smooth features such as the cratonic structure as well as sharp features such as the rift structure in our simplified synthetic model. Numerical comparisons between the inversion results and the input model used to generate the data confirm the superiority of the $\ell_1$-norm regularization. Though in real-world inversions such ground-truth information is not available, one can argue that the $\ell_1$ inversion method serves the principle of parsimony well and is to be preferred over more common methods. If the tomographic object (such as the real earth) is too complex to be well represented by a parsimonious expansion in wavelets, neither method is able to resolve such complexity adequately with a limited data set, as shown in the bottom rows of Fig. 12, where even the $\ell_1$ inversions begin to show the effects of ray path distribution. In this case, we expect that the principle of parsimony can be usefully applied once a richer family of building blocks is considered.

The only drawback of the method, so far, is the slow convergence of the $\ell_1$ surrogate-functional iteration procedure. Our preference for the thresholded algorithm used here arises from the fact that its convergence is guaranteed even though the $\ell_1$ problem is non-linear. We have introduced a two-step procedure that leads to a significant speedup; however, Fig. 11 indicates that even $1000 + 1000$ iterations do not suffice for complete convergence (it nevertheless produces an excellent approximation). A potentially promising approach towards further convergence improvement is to combine an efficient linear method (such as e.g. conjugate-gradient) with an *adaptive* thresholding scheme. This would then avoid the need to precompute the largest eigenvalue of $\mathbf{A}^T\mathbf{A}$ and facilitate the application of the $\ell_1$ method to a larger, 3-D, study of body-wave tomography.

## ACKNOWLEDGMENTS

Financial support for this work was provided by NSF grant DMS-0530865. I.L. is a postdoctoral fellow with the F.W.O.-Vlaanderen (Belgium). We thank the reviewers and editor for suggesting changes and clarifications that improve the readability of the paper.

## REFERENCES

Bertero, M. & Boccacci, P., 1998, *Introduction to Inverse Problems in Imaging* IOP Publishing, Bristol (UK).

Candes, E., Romberg, J. & Tao, T., 2006. Stable signal recovery from incomplete and inaccurate measurements, *Comm. Pure Appl. Math.,* in press.

Constable, S.C., Parker, R.L. & Constable, C.G., 1987. Occam's inversion: a practical algorithm for generating smooth models from electromagnetic sounding data, *Geophysics,* **52,** 289–300.

Daubechies, I., 1992. *Ten Lectures on Wavelets*, SIAM Press, Philadelphia.

Daubechies, I., Defrise, M. & De Mol, C., 2004. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Comm. Pure App. Math.,* **57,** 1413–1541, arXiv/FA/0307152.

Deal, M.M. & Nolet, G., 1999. Slab temperature and thickness from seismic tomography 2. Izu-Bonin, Japan and Kuril subduction zones, *J. Geophys. Res.,* **104,** 28 803–28 812.

Deal, M.M., Nolet, G. & van der Hilst, R.D., 1999. Slab temperature and thickness from seismic tomography, 1. Method and application to Tonga, *J. Geophys. Res.,* **104,** 28 789–28 802.

de Hoop, M.V. & van der Hilst, R.D., 2005. On sensitivity kernels for wave equation tomography, *Geophys. J. Int.,* **160,** 621–633.

Donoho, D.L., 2006. For most large underdetermined systems of linear equations the minimal $\ell_1$-norm solution is also the sparsest solution, *Comm. Pure Appl. Math.,* **59,** 797–829.

Dziewonski, A.M., Hager, B.H. & O'Connell, R.J., 1975. Large scale heterogeneities in the lower mantle, *J. Geophys. Res.,* **82,** 239–255.

Dziewonski, A.M. & Woodhouse, J.H., 1987. Global images of the Earth interior, *Science,* **236,** 37–48.

Gauch H.G., Jr, 2003. *Scientific Method in Practice*, CUP, Cambridge, 435 pp.

Guleryuz, O.G., 2003. Weighted overcomplete denoising, In *Conference Record of the Thirty-Seventh Asilomar Conference on Signals, Systems and Computers,* **2,** 1992–1996.

Kingsbury, N., 1999. Image processing with complex wavelets, *Phil. Trans. Roy. Soc. Lond.,* **A357,** 2543–2560.

Kingsbury, N.G., 2002. Complex wavelets for shift invariant analysis and filtering of signals, *Appl. Computat. Harmon. Anal.,* **10,** 234–253.

Landweber, L., 1951. An iterative formula for Fredholm integral equations of the first kind, *Am. J. Math.,* **73,** 615–624.

Masters, G., Johnson, S., Laske, G. & Bolton, H., 1996. A shear velocity model of the mantle, *Phil. Trans. Roy. Soc. Lond.,* **A354,** 1385–1410.

Nolet, G., 1987. Seismic wave propagation and seismic tomography, In G. Nolet, editor, *Seismic Tomography,* p. 1–23, Dordrecht, Reidel.

Owens, T.J., Nyblade, A.A. & Langston, C.A., 1995. The Tanzania broadband experiment, *IRIS Newsletter,* **14,** 1.

Portniaguine, O. & Zhdanov, M.S., 1999. Focusing geophysical inversion images, *Geophysics,* **64,** 874–887.

Selesnick, I.W., Baraniuk, R.G. & Kingsbury, N., 2005. The dual-tree complex wavelet transform—A coherent framework for multiscale signal and image processing, *IEEE Signal Process. Mag.,* **22,** 123–151.

Selesnick, I., Cai, S. & Li, K., 2006. MATLAB implementation of wavelet transforms, http://taco.poly.edu/WaveletSoftware/.

Spakman, W. & Nolet, G., 1988. Imaging algorithms, accuracy and resolution in delay-time tomography, in *Mathematical Geophysics*, eds Vlaar N.J., Nolet, G., Wortel, M.J.R. & Cloetingh, S.A.P.L., pp. 155–187, Hingham, Mass., Reidel.

Trampert, J. & Snieder, R., 1996. Model estimations biased by truncated expansions: possible artifacts in seismic tomography, *Science,* **271,** 1257–1260.

VanDecar, J.C. & Snieder, R., 1994. Obtaining smooth solutions to large, linear, inverse problems, *Geophysics,* **59,** 818–829.

van Spaendonck, R., Blu, T., Baraniuk, R. & Vetterli, M., 2003. Orthogonal Hilbert transform filter banks and wavelets, *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP),* **6,** 505–508.

Zhou, Y., Dahlen, F.A. & Nolet, G., 2004. Three-dimensional sensitivity kernels for surface wave observables, *Geophys. J. Int.,* **158,** 142–168.

## APPENDIX A: 2-D SENSITIVITY KERNELS

The toy linear inverse problem $\mathbf{Am} = \mathbf{d}$ used in this paper is designed to incorporate all the important characteristics of a real-world regional tomographic inversion, while at the same time being small enough to allow for repeated experimenting with reasonable CPU times on a single workstation. For this reason, we limit attention to surface wave dispersion data, specifically perturbations $\delta k(\nu)$ in the wavenumber $k(\nu)$, presumed to be measured in rad/m, of the fundamental ($n = 0$) Rayleigh mode at temporal frequency $\nu$, measured in Hz. Finite-frequency theory based upon the Born

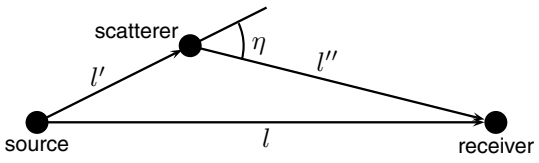


**Figure A1.** Schematic map view of the single-scattering geometry in our simplified 2-D, flat-earth, surface wave inversion problem. The quantity $l$ is the horizontal epicentral distance between the source and receiver; $l'$ and $l''$ are the lengths of the first and second legs of the detour path, respectively, and the angle $\eta$ measures the deflection of the wave at the scatterer.

**Table A1.** Parameters $\nu$, $C(\nu)$, $k(\nu)$, $E_0(\nu)$, $E_1(\nu)$ and $E_2(\nu)$ needed to compute the simplified 2-D sensitivity kernels $K_{2-D}(x, y, \nu)$. Fundamental-mode Rayleigh-wave measurements $\delta k(\nu)$ are presumed to have been made at eight frequencies ranging between $\nu \approx 0.01$ Hz (100 s period) and $\nu \approx 0.1$ Hz (10 s period).

| $\nu$ (mHz) | $C$ (m s$^{-1}$) | $k(10^{-4}$ m$^{-1})$ | $E_0(10^{-9}$ m$^{-2})$ | $E_1(10^{-9}$ m$^{-2})$ | $E_2(10^{-9}$ m$^{-2})$ |
|---|---|---|---|---|---|
| 10.742 | 3831.3 | 0.165 37 | −0.079 642 | −0.359 72 | −0.061 743 |
| 15.625 | 3829.8 | 0.245 11 | −0.126 90 | −0.776 64 | −0.129 84 |
| 20.508 | 3751.3 | 0.325 77 | −0.176 86 | −1.365 6 | −0.222 47 |
| 30.273 | 3434.4 | 0.495 75 | −0.368 58 | −3.217 7 | −0.489 28 |
| 40.039 | 3064.8 | 0.684 98 | −1.078 3 | −6.216 1 | −0.944 65 |
| 50.781 | 2861.6 | 0.914 30 | −2.788 5 | −10.769 | −1.815 0 |
| 70.313 | 2872.8 | 1.344 6 | −6.417 5 | −22.322 | −4.130 0 |
| 99.609 | 2971.5 | 1.973 3 | −11.684 | −47.879 | −8.884 8 |

approximation (Zhou *et al.* 2004) gives a linear relationship between such wavenumber perturbations and the 3-D perturbations in the fractional shear wave velocity $\delta \ln \beta(\boldsymbol{x})$ within the earth:

$$\delta k(\nu) = \iiint K_{3-D}(\boldsymbol{x}, \nu)\, \delta \ln \beta(\boldsymbol{x})\; d^3\boldsymbol{x}. \tag{A1}$$

Making use of a number of flat-earth approximations that do not fundamentally affect the nature of the inverse problem, we can write the 3-D Fréchet sensitivity kernel $K_{3-D}(\boldsymbol{x}, \nu)$, for the simplest case of an explosive source with an isotropic radiation pattern and a measurement made on the vertical component at the receiver, in the form

$$K_{3-D}(\boldsymbol{x}, \nu) = [e_0(z, \nu) + e_1(z, \nu)\cos\eta + e_2(z, \nu)\cos 2\eta] \left[\frac{1}{8\pi k(\nu) l\, l' l''}\right]^{\frac{1}{2}} \sin[k(\nu)(l' + l'' - l) + \pi/4], \tag{A2}$$

where $z$ is the depth, $l$ is the epicentral distance measured in m on the surface of the earth, and $l'$ and $l''$ are the horizontal distances of the scatterer $\boldsymbol{x} = (x, y, z)$ from the source and receiver, respectively. The quantity $\eta$ is the scattering angle, measured at the surface projection $(x, y)$ of $\boldsymbol{x}$, as shown in Fig. A1. Expressions for the depth-dependent functions $e_0(z, \nu)$, $e_1(z, \nu)$ and $e_2(z, \nu)$ can be found in the appendix of Zhou *et al.* (2004).To simplify matters even further, we assume that the velocity perturbation $\delta \ln \beta(\boldsymbol{x})$ is independent of depth $z$ and dependent only upon the horizontal Cartesian coordinates $x$ and $y$. Upon integrating the factors $e_0(z, \nu)$, $e_1(z, \nu)$ and $e_2(z, \nu)$ over depth,

$$E_0(\nu) = \int_0^\infty e_0(z, \nu)\, dz, \qquad E_1(\nu) = \int_0^\infty e_1(z, \nu)\, dz,$$
$$E_2(\nu) = \int_0^\infty e_2(z, \nu)\, dz, \tag{A3}$$

we may then relate $\delta k(\nu)$ to $\delta \ln \beta(x, y)$ via a 2-D sensitivity kernel:

$$\delta k(\nu) = \iint K_{2-D}(x, y, \nu)\, \delta \ln \beta(x, y)\, dx\, dy, \tag{A4}$$

where

$$K_{2-D}(x, y, \nu) = [E_0(\nu) + E_1(\nu)\cos\eta + E_2(\nu)\cos 2\eta] \left[\frac{1}{8\pi k(\nu) l\, l' l''}\right]^{\frac{1}{2}} \sin[k(\nu)(l' + l'' - l) + \pi/4]. \tag{A5}$$

The rapidly oscillating sinusoidal function $\sin[k(\nu)(l' + l'' - l) + \pi/4]$ in eq. (A5) is constant on ellipses, $l' + l'' =$ constant, having the surface projections of the source and receiver as foci. The $\cos\eta$ and $\cos 2\eta$ dependence and the term involving the integrable singularity $1/\sqrt{l' l''}$ act to slowly modulate this dominant elliptical dependence.

Eqs (A4) and (A5) are valid, subject to the already noted approximations, for a monochromatic wavenumber perturbation $\delta k(\nu)$, whereas actual surface wave dispersion measurements must of necessity be made on a portion of a seismogram of finite length, typically multiplied by a time-domain taper $h(t)$. Zhou *et al.* (2004) show that the effect of such a finite-length taper can be accounted for by modifying the taper as follows:

$$K_{2-D}(x, y, \nu) \rightarrow K_{2-D}(x, y, \nu)\, h[(l' + l'')/C(\nu)], \tag{A6}$$

where $C(\nu)$ is the the group velocity at frequency $\nu$ measured in m/s. This modification has the effect of limiting the cross-path width of the Fréchet kernel $K(x, y, \nu)$, since $h(t) = 0$ for large detour times. We assume the data $\delta k(\nu)$ have been measured using a Hann or cosine taper, of duration five wave periods centred on the group arrival time:

$$h(t) = \begin{cases} 0 & \text{for} \quad t \le t_{\text{arrival}} - 2.5/\nu \\ \frac{1}{2}[1 - \cos 2\pi\nu(t - t_{\text{arrival}} - 2.5/\nu)] & \text{for} \quad t_{\text{arrival}} - 2.5/\nu \le t \le t_{\text{arrival}} + 2.5/\nu \\ 0 & \text{for} \quad t \ge t_{\text{arrival}} + 2.5/\nu \end{cases}, \tag{A7}$$

where $t_{\text{arrival}} = l/C(\nu)$. Since $l' + l'' \ge l$ only the $t \ge t_{\text{arrival}}$ portion of the taper (A7) contributes to the finite-record-length sensitivity kernel (A6).The group velocity $C(\nu)$, unperturbed wavenumber $k(\nu)$ and auxiliary variables $E_0(\nu)$, $E_1(\nu)$ and $E_2(\nu)$ for fundamental-mode Rayleigh waves are listed in Table A1 at the eight selected frequencies $\nu$; the corresponding wave periods vary roughly between 100 and 10 s. Since $E_0(\nu)$, $E_1(\nu)$ and $E_2(\nu)$ are all negative, a positive velocity perturbation, $\delta \ln \beta(x, y) > 0$, gives rise to a negative wavenumber perturbation, $\delta k(\nu) < 0$, that is, an apparently longer wavelength wave, as expected. See Fig. 3 for two examples of sensitivity kernels $K_{2-D}(x, y, \nu)$ computed in this way. It is noteworthy that a 2-D surface wave inversion based upon eqs (A4)–(A7) differs from the common approach of inverting for a 2-D phase velocity map at a single specified frequency $\nu$: such maps are strictly incompatible with the notion of finite frequency, where no local phase velocity can be defined except when very crude approximations are made; for a discussion of this issue see Zhou *et al.* (2004)

## APPENDIX B: NOTES ON WAVELETS

The basic building block of the 1-D discrete wavelet transform (DWT) is a filter bank. It consists of a high-pass filter $g$ (i.e. a generalized difference) and a low-pass filter $h$ (i.e. a generalized

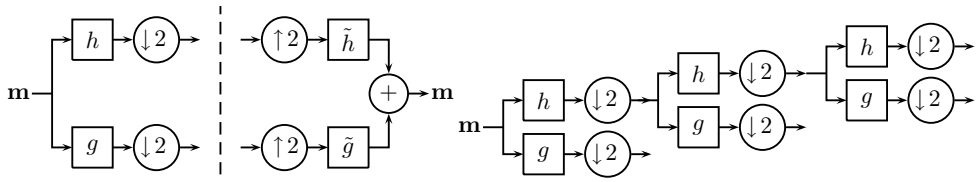


**Figure B1.** Left: schematic representation of a perfect-reconstruction filter bank that can be used to decompose or reconstruct a 1-D signal $x$. Right: a standard wavelet tree.

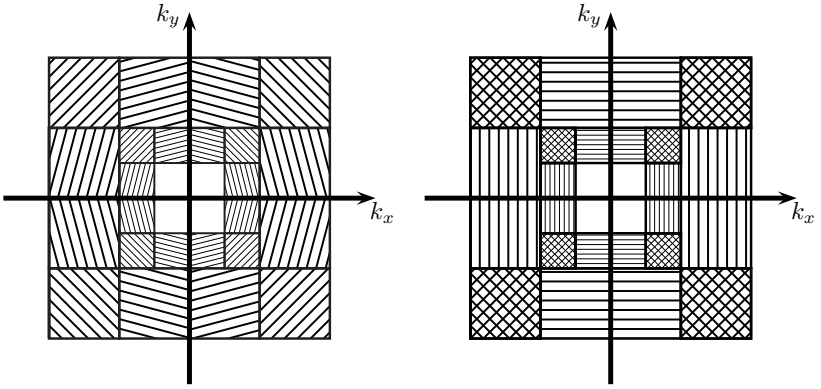


**Figure B2.** Partitioning of the 2-D Fourier domain ($k_x$, $k_y$) by the supports of the Fourier transform of wavelet functions. Only two wavelet scales are shown with the finest one on the outside. In practice the supports have smooth (overlapping) tapers. Left: With the complex 2-D wavelets used in this paper, all squares come in pairs giving rise to six dominant directions (as indicated by the hatch patterns). Right: The standard (direct product) 2-D construction only has horizontal and vertical sensitivity; the 'corner' (hi-hi) squares encode both 45° and −45° at the same time.

average) that are applied to a given signal **m** (i.e. a list of numbers) in the following way: **m** is convolved with $g$ and downsampled, **m** is convolved with $h$ and downsampled. This results in two signals, each with half the length of the original one. The process can be inverted by upsampling (inserting zeroes) the two resulting sequences and convolving each with two (carefully matched) filters $\tilde{g}$ and $\tilde{h}$ and then adding the two. A traditional way of representing this procedure is shown in the left of Fig. B1. It turns out that there exist *finite* filters that give rise to perfect reconstruction (these use finite convolutions only and lead to compactly supported wavelets); moreover in some very special cases, one can have that finite $\tilde{g}$ and $\tilde{h}$ are the reverse of $g$ and $h$ (corresponding to compactly supported *orthogonal* wavelets). The Haar wavelets have $g = (\frac{1}{2}, -\frac{1}{2})$ and $h = (\frac{1}{2}, \frac{1}{2})$, but there exist longer (perfect reconstruction) finite filters (which give rise to smoother wavelets). The so-called D4 wavelets correspond to $h = (1+\sqrt{3}, 3+\sqrt{3}, 3-\sqrt{3}, 1-\sqrt{3})/4\sqrt{2}$ and $g = (1-\sqrt{3}, -3+\sqrt{3}, 3+\sqrt{3}, -1-\sqrt{3})/4\sqrt{2}$.

The 1-D discrete wavelet transform is defined by the iteration of the analysis filter bank on the low-pass outcomes (see right-hand side of Fig. B1). In this way, successive levels of detail are stripped of the input signal **m** (and stored in wavelet coefficients), leaving a very coarse average (stored in so-called scaling coefficients). This construction is called a wavelet tree. It not only defines the DWT but also provides its practical implementation. When using finite filters, the construction automatically gives rise to a computationally efficient algorithm: as a result of the subsampling each step cost only half as much time as the previous one. The total number of operations then is $kN + kN/2 + kN/4 + kN/8 + \ldots = 2kN$, less than the $\mathcal{O}(N^2)$ for a generic linear transformation.

A standard way of generating wavelets in 2-D is to form the direct product of 1-D wavelets, that is, the filters are applied to rows and columns of an image (lo-lo, lo-hi, hi-lo and hi-hi). This, however, has the marked disadvantage of poor directional sensitivity. In this study, to obtain better directional sensitivity, we use the complex 2-D wavelets developed by Kingsbury (2002). These are constructed also by direct product but from *two* simultaneous wavelet trees (see Kingsbury 1999, for a diagram of such a dual tree). The qualitative difference between these two constructions is best seen in the Fourier domain. Fig. B2 shows a schematic representation of the supports of the Fourier transforms of the wavelet functions, both for the usual 2-D wavelet construction and for the 2-D complex wavelets. The two are fundamentally different: whereas the usual separable 2-D wavelet construction gives rise to a horizontal, a vertical and one (!) diagonal part at each scale, the complex 2-D construction has six different inherent directions per scale. A careful choice of the different filters also leads to an (almost) tight frame (i.e. the inverse wavelet transform almost coincides with the transpose).

The price to pay for these benefits is the redundancy. In 2-D the complex wavelets generate four times as many coefficients as there are pixels in the original image (two trees and real and imaginary parts of the output). For example, the $64 \times 64$ spatial-domain images we use in Section 3 give rise to $16\,320 = 2 \times 6 \times (32^2 + 16^2 + 8^2 + 4^2)$ wavelet coefficients and $64 = 2 \times 2 \times 4^2$ scaling coefficients (see e.g. Fig. 10). Together this is 16 384 which equals $4 \times 64^2$.