Inversion of Phase Data for a Phase Velocity Map 101

Summary for CIDER12 Non-Seismologists

1. Setting up a Linear System of Equations

This is a quick-and-dirty, not-peer reviewed summary of how a dataset of phase anomalies (or group travel time anomalies, or any dataset for a 2-dimensional ray-path geometry) can be inverted for a phase velocity (anomaly) map (or any 2-dimensional model) (Figure 1). In a nutshell, to describe the problem numerically, we want to invert the linear system of equations

$$\mathbf{d} = \mathbf{A} \cdot \mathbf{m} \tag{1}$$

where vector **d** is the data vector, vector **m** is the model vector, and matrix **A** is the data kernels matrix.

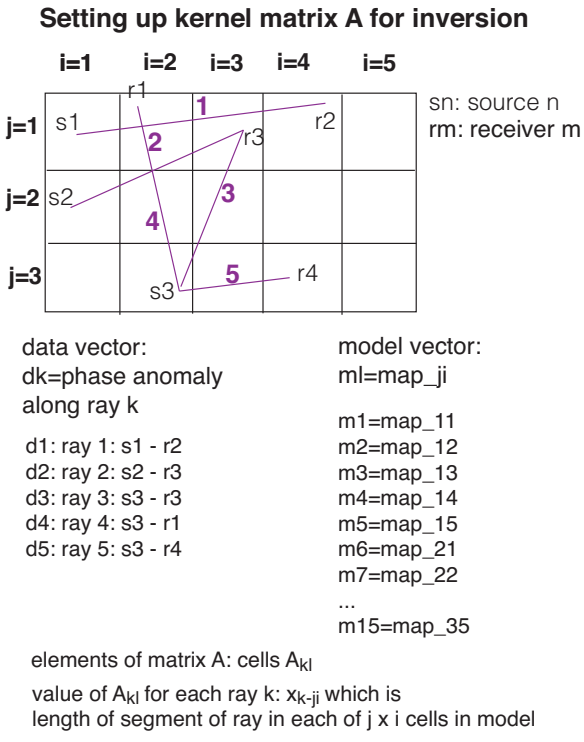

Figure 1. Top: Schematic diagram of 5 rays from n_{MAX}=3 sources to m_{MAX}=4 receivers (*not all possibilities are plotted!*) in a phase velocity map with 3 rows (latitudes) and 5 columns (longitudes). **Bottom:** scheme of how the data vector and the model vectors as well as the data kernel matrix A are set up.

At this point, we have not yet made any assumptions on how the data relate to the model, i.e. whether we use ray theory or a more sophisticated finite frequency approach. All that is assumed here is that the dependence of the data on the model is linear.

1.1 Use Ray Theory

Now we apply Fermat's Principle and assume that phase anomaly accumulates along the shortest path between a source and a receiver. On a sphere, this path is a great circle. The phase anomaly is assumed to not be influenced by structure away from the great circle. This assumption is valid in the infinite-frequency limit, i.e. when the wavelength of the wave is much smaller than the wavelength of the structure through which it travels. Since the phase accumulates as $\Phi = kx$ where k is the wave number and x the travel distance, a phase anomaly at distance $a\Delta$ is $\delta\Phi = \delta\overline{k} \cdot a\Delta$ where $\overline{k}$ is the path-averaged wave number anomaly along the path, a Earth radius and Δ the epicentral distance in radians. $\overline{k}$ is related to the path-averaged phase velocity, $\overline{c}$, through $c = \omega/k$ and $\delta k = -\omega/c \cdot \delta c/c$. So our phase datum with index k, at frequency ω becomes

$$\begin{array}{cccc} \underbrace{\delta\Phi_k} & \simeq & \underbrace{-\frac{\omega a}{c_0}\int_0^\Delta dx \cdot} & \underbrace{\frac{\delta c(\theta,\phi)}{c_0}} \\ d_k & = & \sum_k \sum_l A_{kl} & m_l \end{array} \tag{2}$$

where θ and ϕ are geographical coordinates. We integrate over $\delta c/c_0$ but the integral is rearranged to make the connection to equation (1). The $\simeq$ signifies that we replaced the unknown actual path-averaged phase velocity, $\overline{c}$, by a known, predicted reference phase velocity, c_0 (e.g. for a known model). It is ok to do this because $\delta c \ll c_0$. The index l for the model vector in the matrix notation relates to the cells in a map as illustrated in Figure 1.

1.2 Make Data Kernel Matrix

For illustration purposes on how we set up our inversion, we assume that our map has 3 rows and 5 column. This gives $3 \times 5 = 15$ model parameters. We have 5 rays, i.e. our data vector has length 5. The data dependence on the model parameters, as traced from source, s_n, to receiver, r_m, is the following

$$\begin{aligned} d_1 =& a_{11} \cdot m_1 + a_{12} \cdot m_2 + a_{13} \cdot m_3 + a_{14} \cdot m_4 \\ d_2 =& a_{26} \cdot m_6 + a_{27} \cdot m_7 + a_{22} \cdot m_2 + a_{23} \cdot m_3 \\ d_3 =& a_{3(12)} \cdot m_{(12)} + a_{3(13)} \cdot m_{(13)} + a_{28} \cdot m_8 + a_{33} \cdot m_3 \\ d_4 =& a_{4(12)} \cdot m_{(12)} + a_{47} \cdot m_7 + a_{42} \cdot m_2 \\ d_5 =& a_{5(12)} \cdot m_{(12)} + a_{5(13)} \cdot m_{(13)} + a_{5(14)} \cdot m_{(14)} \end{aligned} \tag{3}$$

$$\mathbf{A} = \begin{pmatrix} x_{1-11} & x_{1-12} & x_{1-13} & x_{1-14} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & x_{2-12} & x_{2-13} & 0 & 0 & x_{2-21} & x_{2-22} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & x_{3-13} & 0 & 0 & 0 & 0 & x_{3-23} & 0 & 0 & 0 & x_{3-32} & x_{3-33} & 0 & 0 \\ 0 & x_{4-12} & 0 & 0 & 0 & 0 & x_{4-22} & 0 & 0 & 0 & 0 & x_{4-32} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & x_{5-32} & x_{5-33} & x_{5-34} & 0 \end{pmatrix} \tag{4}$$

In a real global inversion we have many more data, e.g. 50,000 or so. Our maps are also much larger. The model vector for a 2-degree map would be $90 \times 180 = 16,000$. To accommodate the fact that a 2×2-degree cell is much smaller near the poles than near the equator but the data resolution is not latitude-dependent, we use an equal-area cell parameterization, i.e. cells are 2 degrees wide near the equator, but much wider near the poles.

Consequently, we have 180 cells near the equator but only 3 at $\pm 89^\circ$ latitude. This gives a total of 10,312 cells in the model, reduces the number of model parameters but providing a more evenly spaced model parameterization at the same time. Numerically, less independent data would be required to constrain the model.

2. Setting up the Inversion

Formally, we want to determine the model **m** by inverting equation (1), so

$$\mathbf{m} = \mathbf{A^{-1}} \cdot \mathbf{d}.$$

Numerically, it is difficult to invert a rectangular matrix, and we apply a trick by multiplying equation (1) by $\mathbf{A^T}$ where ^{T} denotes transpose. So,

$$\mathbf{A^T d} = \mathbf{A^T A} \cdot \mathbf{m} \tag{5}$$

We can do this because **A** has the same eigenvectors and eigenvalues as $\mathbf{A^T A}$. Inversion of this equation gives

$$\hat{\mathbf{m}} = (\mathbf{A^T A})^{-1} \cdot \mathbf{A^T d}. \tag{6}$$

where

$$\mathbf{G} = (\mathbf{A^T A})^{-1} \cdot \mathbf{A}^T$$

is called the generalized inverse. The hat signifies that because of the imperfection of the data, we will not be able to recover the actual model but a version of the model as "seen" (or filtered) by the data.

An inversion using (6) is a least-squares inversion, i.e. a model is retrieved that minimizes the misfit, χ^2

$$\chi^2 = \sum_k (\frac{d_k - \hat{d}_k}{\sigma_k})^2$$

were $\hat{d}_k$ are the predictions calculated with the new model $\hat{\mathbf{m}}$ and σ_k are the data errors. When $\chi^2 = 1$, the model is said to fit the data to within their errors. If $\chi^2 < 1$, the model overfits the data and contains components that are not required to fit the data, from a numerical point of view. If $\chi^2 > 1$ then the model does not fit the data, either because the model parameterization is not adequate or the data are internally inconsistent (e.g. because of noise contamination or systematic effects that are not modeled such as an erroneous earthquake event times or locations).

2.1 Model Regularization or Damping an Inversion

As seen in equation (4), many elements of **A** are zero as some cells in the model remain unsampled, resulting in a determinant for $\mathbf{A^T A}$ that is zero. Matrix **A** is singular and some eigenvalues are zero which does not allow us to retrieve the complete model vector. The matrix is ill-conditioned, with an infinite condition number (largest eigenvalue divided by lowest; a low number signifies a well-conditioned matrix). A strategy often used is that only the non-zero eigenvalues (or the large eigenvalues) and corresponding eigenvectors are used to construct parts of the model. Some workers discard this approach as it leaves unwanted "holes" in the model though, numerically, this may be the correct approach.

An alternative is to regularize an inversion by adding something to the matrix to decrease the condition number, e.g. our inversion may look like

$$\hat{\mathbf{m}} = (\mathbf{A^T A} + \mu \mathbf{I})^{-1} \cdot \mathbf{A^T d}. \tag{7}$$

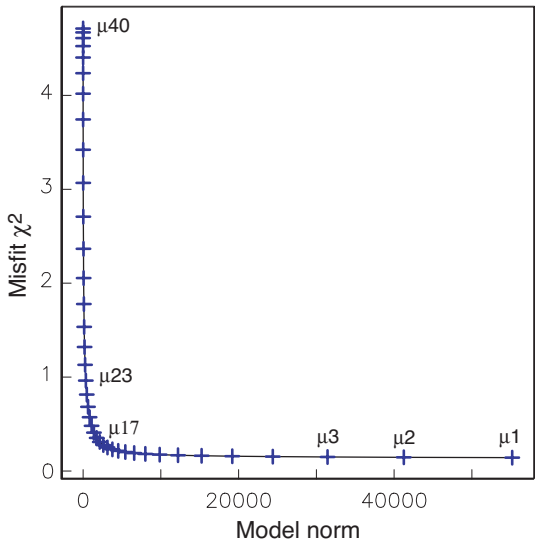

Figure 2. Trade-off curve between model norm (x-axis) and data misfit (y-axis). Increasing μ indices denote increasing values. The model norm chosen here is $|\partial\mathbf{m}|^2$. Both the model norm and the data misfit cannot be arbitrarily small at the same time. An "optimal" model is often chosen at the bend where both model norm and data misfit do not change much with changing regularization parameter μ.

where

$$\mathbf{G} = (\mathbf{A^T A} + \mu\mathbf{I})^{-1} \cdot \mathbf{A}^T$$

The regularization or damping factor μ determines how much the inversion is regularized. The identity matrix **I** in this equation limits the size of the model, where the model length decreases with increasing μ regardless of the structural wavelength in the model (i.e. all elements in the model vector are penalized in the same way). This type of inversion minimizes the weighted sum of data misfit and model norm, or the misfit function, MF,

$$MF = \chi^2 + \mu|\mathbf{m}|^2.$$

There exists a trade-off between misfit and model norm. A small μ produces a "large" model but some components may not be constrained by the data as the difference between data and predictions are much smaller than the error bars and the misfit is far below 1. With increasing μ the size of the model decreases, but such a model is less able to predict the data to within their errors, so the misfit increases (Figure 2).

In principle, we can impose any regularization on the model, e.g. we can impose a maximum value on individual elements of the model vector (e.g. phase velocity anomalies must not be larger than 10%). We can also impose constraints on gradients or curvatures in the model, in which case we replace the identity matrix in equation (7) by the first or second derivatives:

$$\hat{\mathbf{m}} = (\mathbf{A^T A} + \mu\partial^{\mathbf{T}}\partial)^{-1} \cdot \mathbf{A^T d} \tag{8}$$

and ∂ stands for either the first or second derivative and

$$\mathbf{G} = (\mathbf{A^T A} + \mu \partial^{\mathbf{T}} \partial)^{-1}.$$

The misfit function then becomes

$$MF = \chi^2 + \mu |\partial \mathbf{m}|^2.$$

The choice of the optimal model is somewhat subjective. A model a long the trade-off curve is often chosen for which both the model norm (model norm here meaning the vector length or the length of first or second derivatives thereof) and the data misfit both do not change much when changing the damping parameter. In Figure 2, this would be any model near μ_{17}. However, models in that range overfit the data and contain elements that are not required by the data. So, some people would rather choose a model near μ_{23}. Also, sometimes, models near μ_{17} are unrealistically rough or have unrealistically large values, so a model farther up the trade-off curve is chosen as most optimal model.

2.2 Quick Notes on Numerical Inversion Techniques and Other Approaches

On the computer, matrices can be inverted to use ”canned” computer tools. As mentioned in the introduction, a classical way leads through finding the eigenvalues and eigenvectors through singular value decomposition. As seen in Figure 1, many of our matrix elements are zero, so the matrix is sparse. To save computer time, iterative sparse matrix solvers are available. Without going into too much detail, such iterative techniques explore the misfit function by trying to find to fastest way toward the minimum. One such technique is LSQR or least squares OR (see, e.g., van der Sluis and van der Vorst, 1987. This technique is used in the tutorial. Here, one has to make sure to go through enough iterations so that the that the process has converged, i.e. the model no longer changes significantly.

A completely different approach to a model is by forward modeling. Here, models are compiled using a variety of strategies. Synthetic data are computed using these models and the misfit determined. The model is kept if it satisfies a certain misfit criterion or discarded if it does not. This way, a group of models can be found that may look completely different from the one obtaining through an inversion but that satisfy the data equally well. Proponents of this approach prefer this over inversions because the latter may get caught in local minima if the misfit function is complex. The art of forward modeling is to decide how to search the model space. Monte Carlo approaches do this randomly. Other approaches using, e.g. genetic algorithms or evolutionary programming, produce models that are bases on ”mutations” of models using certain rules.

3. A Few (Uncomplete) Notes on Model Evaluation

If we consider equations (1) and (8) together, we get

$$\hat{\mathbf{m}} = \mathbf{G} \cdot \mathbf{d} = \mathbf{G} \cdot \mathbf{A} \cdot \mathbf{m} \tag{9}$$

where the resolution matrix $\mathbf{R} = \mathbf{G} \cdot \mathbf{A}$ maps the true model, $\mathbf{m}$, into the version of the model, $\hat{\mathbf{m}}$, as ’”seen” by the data through filter $\mathbf{R}$:

$$\hat{\mathbf{m}} = \mathbf{R} \cdot \mathbf{m}.$$

Ideally, the resolution matrix should be the identity matrix. If the diagonal elements are less than 1 then the amplitudes of the corresponding model element is not recovered completely. Any non-zero off-diagonal elements indicate that some cross-mapping between model parameters exists. Though this strategy is somewhat contested (see below) it is useful to evaluate the resolution matrix. For example, in ”resolution test” one model

parameter in **m** is switched on, while all other elements are set to zero. The structure of the resulting $\hat{\mathbf{m}}$ then reveals how this particular element is resolved and/or smeared into other model elements.

If the model is a cell-parameterized phase velocity map, then such a test is equivalent to a spike test. If the model is a map expanded in surface spherical harmonics, then such a test is equivalent to a checker board test (Figure 3). Such tests are only meaningful if the same damping and the same error bars are used in this test as in the inversion of the real data, i.e. do the test on equation (8), not on equation (6)! The advantage of using checkerboards as input model for resolution tests in maps is that one gets a quick geographical overview over where in the map structure of a certain wavelength is resolved or not. In our example, good recovery is observed for parts of the western Pacific Ocean (lots of earthquakes) and along the coast of western North America (lots of stations). Input checker boards can have larger wavelengths (lower harmonic degree) or shorter wavelengths (higher harmonic degree).

Opponents of checkerboard tests argue that the checkerboard tests do not tell you about the real resolution capabilities of the data. E.g. when one uses ray theory instead of finite-frequency theory, then the checkerboard test does not take into account that not accounting for wavefront healing effects in ray theory degrades resolution. But this argument holds true for any input structure in tests described above, regardless whether it is a checker board, spike or some oddly-shaped structure that looks like the structure that the data imaged (as also often seen in the literature). One could argue that the latter is a most subjective test while a checker board gives a more objective view.

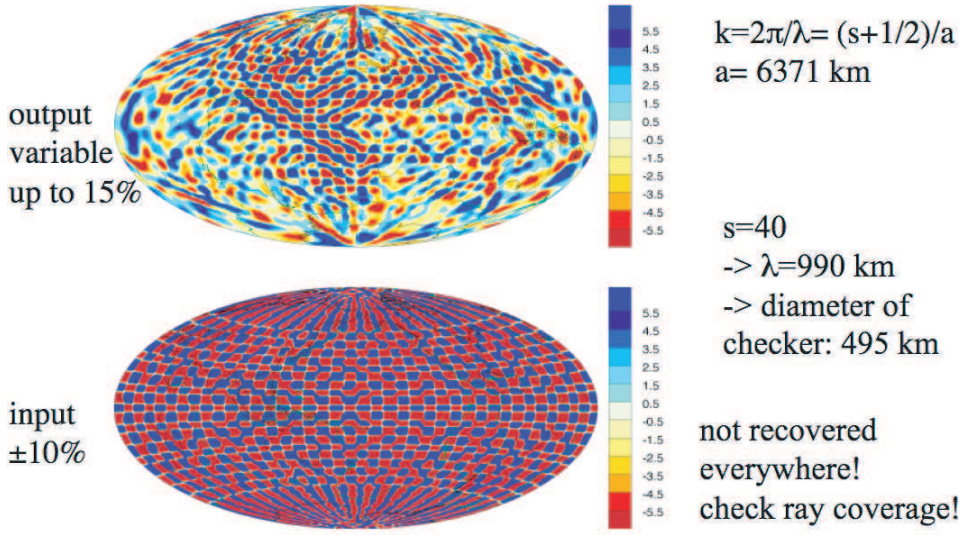

Figure 3. Input (bottom) and output (top) of a checkerboard test for a phase velocity map as obtained using the CIDER 12 dataset for 50-s Rayleigh waves. The map view is centered on the Pacific Ocean. The input model is a degree-40 checker board. Note that the inversion strategy used here was LSQR, not a classical SVD matrix inversion.

4. References

Dziewonski, A. M. and D. L. Anderson, D.L., 1981. Preliminary reference Earth model. *Phys. Earth Planet. Int*, **25**, 297–356.

Menke, W., 1989. *Geophysical Data Analysis: Discrete Inverse Theory*. Academic Press, 289 pp.

and update/MATLAB version thereof

Menke, W., 2012. *Geophysical Data Analysis: Discrete Inverse Theory, Third Edition: MATLAB Edition*. International Geophysics Series; Academic Press, 348 pp.

van der Sluis, A. and van de Vorst, H.A., 1987. Numerical solution of large sparse linear systems arising from tomographic problems. *"Seismic Tomography", G. Nolet ed, D. Reidel Publishing Company, Dordrecht,* pp. 49–83.