Bulletin of the Seismological Society of America, Vol. 109, No. 5, pp. 1729–1737, October 2019, doi: 10.1785/0120190079

Ⓔ

Model Misfit Minimization

by Yuanyuan Fang,* Ying Zhou, and Zhenxing Yao†

Abstract In geophysical applications, solutions to ill-posed inverse problems $\mathbf{Ax} = \mathbf{b}$ are often obtained by analyzing the trade-off between data residue $\|\mathbf{Ax} - \mathbf{b}\|^2$ and model norm $\|\mathbf{x}\|^2$. In this study, we show that the traditional L-curve analysis does not lead to solutions closest to the true models because the maximum curvature (or the corner of the L-curve) depends on the relative scaling between data residue and model norm. A Bayes approach based on empirical risk function minimization using training datasets may be designed to find a statistically optimal solution, but its success depends on the true realization of the model. To overcome this limitation, we construct training models using eigenvectors of matrix $\mathbf{A}^T\mathbf{A}$ as well as spectral coefficients calculated from the correlation between observations and eigenvector projected data. This approach accounts for data noise level but does not require it as *a priori* knowledge. Using global tomography as an example, we show that the solutions are closest to true models.

Supplemental Content: Figures showing additional scaling and L-curve analysis, Bayesian risk minimization, examples of N-point running average and model misfit (MM) minimization in seismic tomography.

Introduction

The linear inverse problem $\mathbf{Ax} = \mathbf{b}$ rises in many geophysical imaging applications. It is often ill-posed and consequently, the true model $\mathbf{x}$ is not recoverable (Jackson, 1972). A common practice is to find an approximate solution to a nearby well-posed problem by introducing regularization (Haber *et al.*, 2007; Charléty *et al.*, 2013; Fan *et al.*, 2014; Ma *et al.*, 2016). Tikhonov regularization is one of the most popular regularization methods for ill-posed problems (Tikhonov and Arsenin, 1977), and damped least-squares minimization (zero-order Tikhonov regularization) in many geophysical inverse problems is such an application (Song *et al.*, 2004; Ritsema *et al.*, 2011; Kaban *et al.*, 2016; Zhou, 2018).

The solution to the nearby well-posed zero-order Tikhonov regularization can be found by solving a minimization problem

$$\|\mathbf{Ax} - \mathbf{b}\|^2 + \alpha^2\|\mathbf{x}\|^2 = \text{minimum}, \tag{1}$$

in which $\mathbf{b}$ is the noise contaminated data $\mathbf{b} = \mathbf{b}^{\text{true}} + \epsilon$ with ϵ being the noise vector. This expression can be written equivalently as

$$(\mathbf{A}^T\mathbf{A} + \alpha^2\mathbf{I})\mathbf{x} = \mathbf{A}^T\mathbf{b}, \tag{2}$$

in which α is the Tikhonov (damping) parameter, and the ill-posedness of the inverse problem is removed by introducing the damping matrix $\alpha^2\mathbf{I}$.

In this article, we shall focus on the damped least-squares problem and develop a method for determining the Tikhonov parameter in equation (2). We point out that Tikhonov regularization does not require *a priori* knowledge of the data and model covariances. For completeness, we provide a brief review subsequently on the connection between the aforementioned Tikhonov regularization and probability inverse methods that have been used in many geophysical inverse problems. If one assumes the true model is a realization of a Gaussian distribution and the model covariance matrix $\mathbf{C}_x$ is known (*a priori*), the ill-posedness of the inverse problem $\mathbf{Ax} = \mathbf{b}$ may also be removed through an alternative approach that minimizes the posteriori probability density which solves

$$(\mathbf{A}^T\mathbf{A} + \mathbf{C}_x^{-1})\mathbf{x} = \mathbf{A}^T\mathbf{b}. \tag{3}$$

If the data have a Gaussian distribution and the covariance matrix $\mathbf{C}_b$ is also known (not an identity matrix), a more general expression of the equation (3) becomes (Jackson, 1979; Tarantola, 2005)

$$(\mathbf{A}^T\mathbf{C}_b^{-1}\mathbf{A} + \mathbf{C}_x^{-1})\mathbf{x} = \mathbf{A}^T\mathbf{C}_b^{-1}\mathbf{b} \tag{4}$$

*Also at Department of Geosciences, Virginia Tech, Blacksburg, Virginia 24061 U.S.A.; and University of Chinese Academy of Sciences, Beijing 100049, China.

†Also at University of Chinese Academy of Sciences, Beijing 100049, China.

in which the data covariance matrix $\mathbf{C}_b$ practically introduces weights on observations. One important step in solving underdetermined inverse problems is to find the regularization parameter (or matrix) by imposing additional constraints such that the matrices on the left side of equations (2–4)are invertible.

In this article, we shall focus on the Tikhonov regularization (equation 2), which does not require data and model covariance matrices as *a priori*. The linear system in equation (2) can be solved based on singular value decomposition (SVD) of the $n \times m$ matrix $\mathbf{A}$,

$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^T. \tag{5}$$

The singular value matrix $\mathbf{\Sigma}$ is a rectangular diagonal matrix with an upper $m \times m$ diagonal matrix $\mathrm{diag}(\sigma_1, ..., \sigma_m)$ and a lower zero matrix, and $\mathbf{U}$ and $\mathbf{V}$ contain data and model singular vectors $\mathbf{u}_i$ and $\mathbf{v}_i$, respectively. The model singular vectors $\mathbf{v}_i$ and σ_i^2 are eigenvectors and eigenvalues of matrix $\mathbf{A}^T\mathbf{A}$.

Tikhonov solutions to the inverse problem can be written as

$$\mathbf{x} = \sum_i \left(\frac{\sigma_i^2}{\sigma_i^2 + \alpha^2}\right) \frac{\mathbf{u}_i \cdot \mathbf{b}}{\sigma_i} \mathbf{v}_i, \tag{6}$$

and the dot product $\mathbf{u}_i \cdot \mathbf{b} = \mathbf{u}_i^T\mathbf{b}$. The most important step in solving the ill-posed inverse problem is to determine an optimal damping parameter α. In the rest of the article, without loss of generality, we shall use seismic surface-wave imaging as an example ill-posed inverse problem for discussions on choosing an optimal parameter such that the obtained model is closest to the true model.

For ill-posed inverse problems, regularizations have to be applied to find a model that approximates the true model in some way. The choice of regularization depends on the properties of the final model that one seeks. For example, the Morozov's discrepancy principle (Morozov, 1984) finds an optimal model that produces best data fit based on a known size of the noise; the generalized cross validation (Wahba, 1977) aims to find a model that best predicts each measurement as a function of the others; the quasi-optimality criterion (Tikhonov and Arsenin, 1977) finds a damping parameter in which the resulting model is least sensitive to a small change in the damping parameter; and the discrete Picard condition (Hansen, 1990) requires the projection of data onto the data singular vectors $|\mathbf{u}_i \cdot \mathbf{b}|$ to decay faster than the generalized singular values (the Picard condition). In this article, we shall propose a method that finds an optimal Tikhonov parameter to minimize model misfit (MM).

The Problem with L-Curve Analysis

L-curve was first introduced by Lawson and Hanson (1974) to investigate the relation between data residue norm $\|\mathbf{Ax} - \mathbf{b}\|^2$ and solution model norm $\|\mathbf{x}\|^2$ for regularized inverse problems. It has received wide applications in seismic tomography in determining an optimal Tikhonov (damping) parameter (Zhao, 2015). If noise level in data is *a priori*, this becomes easier because it statistically constrains data residue. The parameter α may be chosen by trying different values until the outcome model has a normalized data residue norm that equals the variance of the noise. In most cases, noise level in data is unknown, and optimal models are often chosen at the corner of the L-curve (Zhao 2015), in which the curvature (Kreyszig, 1959)

$$\kappa = \frac{h' g'' - g' h''}{(h'^2 + g'^2)^{3/2}} \tag{7}$$

is maximum. Here, $h(\alpha) = \|\mathbf{x}\|^2$ and $g(\alpha) = \|\mathbf{Ax} - \mathbf{b}\|^2$. However, it has been reported in geophysical studies that models obtained using L-curve analysis can be far away from true models (Deidda *et al.*, 2003; Zaroli *et al.*, 2013). Although it is not impossible to find an optimal model using L-curve analysis, we show in this section that this approach is very subjective. We will use surface-wave finite-frequency tomography as an example; the same applies to seismic tomography regardless of approximations in physics (e.g., ray theory).

In surface-wave tomography, efforts have been made to improve the representation of wave propagation physics by introducing finite-frequency sensitivity kernels (Snieder and Nolet, 1987; Spetzler *et al.*, 2002; Yoshizawa and Kennett, 2002; Zhou *et al.*, 2004, 2005). The linear tomographic problem can be written as a Fredholm integral of the first kind (Liu and Zhou, 2016)

$$\delta\varphi(\omega) = \iint_\Omega K_\varphi^c(\omega, \hat{\mathbf{r}}) \delta \ln c(\omega, \hat{\mathbf{r}}) d\Omega, \tag{8}$$

in which K_φ^c is the sensitivity kernel of phase delays $\delta\varphi$ to fractional perturbations in local phase velocity $\delta \ln c$, and the integration is over the surface of the unit sphere $\Omega = \{\hat{\mathbf{r}} : \|\hat{\mathbf{r}}\|^2 = 1\}$.

We use a global dataset of Rayleigh-wave phase-delay measurements from Zhou *et al.* (2006), and we calculate phase-velocity sensitivity kernels following Liu and Zhou (2016). The surface of the Earth is parameterized into 2562 triangular grid points with a lateral grid spacing of about 4° as in Zhou *et al.* (2006). The discrete form of the aforementioned equation can be written as $\mathbf{Ax} = \mathbf{b}$, in which $\mathbf{A}$ is the sensitivity kernel matrix with n number of rows and m number of columns, $\mathbf{x}$ is the vector of unknown fractional velocity perturbations on m global grid points, and $\mathbf{b}$ is the phase-delay data vector with n measurements. In this article, the unit of phase delay ($\mathbf{b}$) is radian and the fractional velocity perturbation in percentage ($\mathbf{x}$) is dimensionless. When the data vector includes both minor-arc and major-arc surface-wave observations, the number of observations ($n = 3681$) exceeds the number of unknowns ($m = 2562$). When the observations include only minor-arc surface-wave observations, $n = 1885$. The coverage of minor-arc Rayleigh wave alone illuminates the Pacific Ocean as well as the continents,

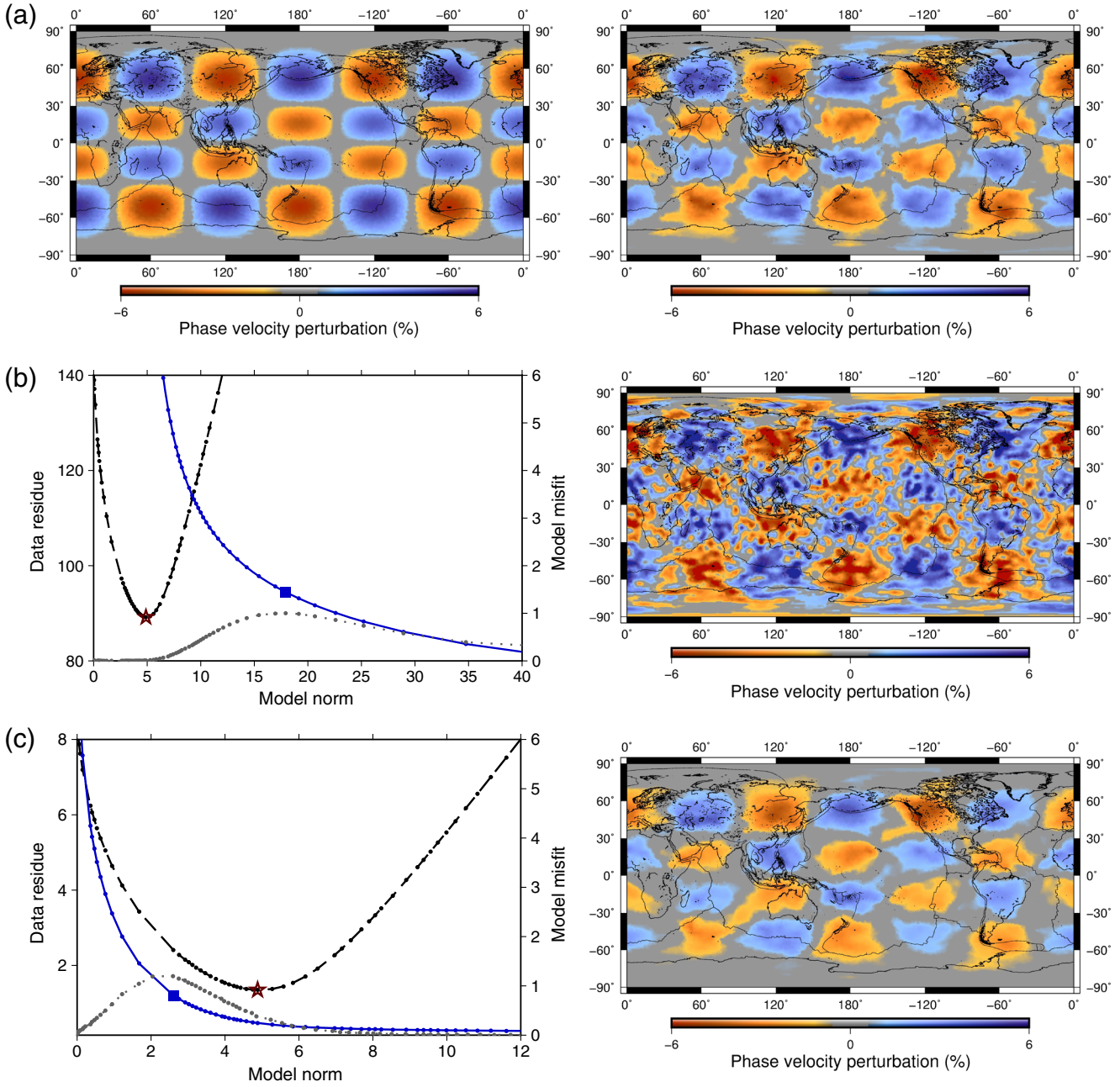

Figure 1. This figure illustrates the arbitrariness in determining an optimal model based on analysis of the L-curve (relation between data residue $\|\mathbf{Ax}-\mathbf{b}\|^2$ and model norm $\|\mathbf{x}\|^2$). (a) Input model (left) and target output model with best model fit (right). (b) L-curve (left) and maximum curvature model (right). No scaling (normalization) has been applied to the inverse problem. The solid line is the L-curve, and the curvature of the L-curve is plotted in gray dotted line. The optimal model with maximum curvature (corner of the L-curve) is indicated by the diamond, and the corresponding model is plotted in the right panel. The dotted–dashed line is the U-curve (model misfit [MM] plotted as a function of model norm), in which the star indicates the target output model. (c) Same as (b) but for a scaled inverse problem $\eta\mathbf{Ax}=\eta\mathbf{b}$, in which $\eta = 0.05$. The color version of this figure is available only in the electronic edition.

and the inclusion of major-arc measurements significantly improves global coverage, especially in the Southern Hemisphere and the Atlantic Ocean. We will first consider inversions using both minor-arc and major-arc data, and inversions using only minor-arc data will be discussed in the Model Misfit (MM) Minimization section.

In Figure 1, we show that the corner of a trade-off curve is highly subjective. For the same matrix **A** and the same data vector **b**, L-curve analysis can lead to very different optimal solutions depending on the relative scaling between data residue norm $\|\mathbf{Ax}-\mathbf{b}\|^2$ and model norm $\|\mathbf{x}\|^2$. In this experiment, the true model **x** has a spherical harmonics structure with degree $l = 6$ and order $m = 3$. Because the true model is known in this case, we may vary the damping parameter α and quantify the misfit between the true model and output model $\|\mathbf{x}-\mathbf{x}^{\text{true}}\|^2$. The U-shaped model misfit curve for

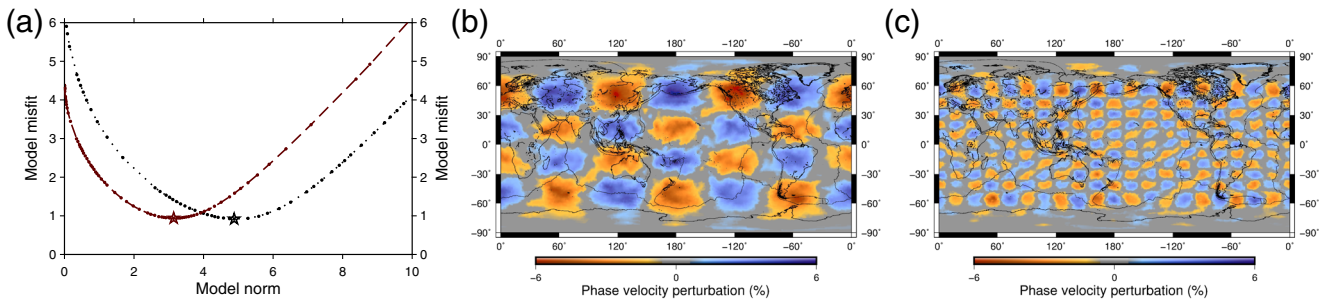

Figure 2. (a) Model misfit curves (U-curves) for input models with spherical harmonics structure $l = 6$ (dotted line) and $l = 20$ (dashed line). (b) and (c) are target models (with best model fit) for $l = 6$ and 20, respectively. The corresponding damping parameters are $\alpha = 0.14$ and 0.047, respectively. In both inversions, 20% of random noise has been added to synthetic data. The color version of this figure is available only in the electronic edition.

varying parameter α is plotted in Figure 1 in dotted-dashed lines. We added Gaussian noise to the synthetic data $\mathbf{b}$ in all inversions, and the root mean square (rms) of the noise is 20% of the rms of the synthetic data. The solution with minimum model misfit is the target model $\mathbf{x}^{\text{target}}$ (Fig. 1a) and is indicated by a star on the U-curve. The L-curves are in solid lines and their corresponding curvatures are plotted in gray dotted lines. Maximum curvature models are indicated by diamonds on the L-curves. In Figure 1b, the maximum curvature model is underdamped, leading to a large model norm as noises are amplified in the solution. L-curve normalization using extreme values of $h(\alpha)$ ($\alpha = 0$) and $g(\alpha)$ ($\alpha = \infty$) does not change the overall quality of the model at the corner of the L-curve (Ⓔ Fig. S1, available in the supplemental content to this article), consistent with the investigation made by Zaroli *et al.* (2013).

The rationale behind finding the corner (maximum curvature) of an L-curve is that the solution seems to be a fair balance between data misfit and model uncertainty (Hansen and O'Leary, 1993). However, the location of the corner is arbitrary because model norm $\|\mathbf{x}\|^2$ and data misfit $\|\mathbf{Ax} - \mathbf{b}\|^2$ have independent units. For example, if one solves a scaled inverse problem $\eta\mathbf{Ax} = \eta\mathbf{b}$, the corner of the L-curve will be different as data residue is now scaled by η^2 whereas model norm remains the same. The optimal model $\mathbf{x}$ can either be overdamped or underdamped depending on the scaling parameter η (Fig. 1 and Ⓔ Fig. S1). In seismic tomography, an optimal model is often first determined based on good estimate of noise level in data, and the L-curve plot is then scaled to have its corner at the preferred solution. This approach can lead to arbitrary solutions because estimates of noise level are subjective. We point out that L-curves are often plotted in linear scale in geophysical applications whereas a logarithmic scale has been used in other literature, including the original paper by Hansen and O'Leary (1993). This choice does not change the nature of the L-curve problem, that is, the maximum curvature of the L-curve depends on the relative scaling between data residue and model norm. In this article, we will take a different approach to minimize the mean square error and determine optimal models that are closest to the target model, that is,

$$\|\mathbf{x} - \mathbf{x}^{\text{target}}\|^2 = \text{minimum}. \tag{9}$$

We show that the target model can be approached by minimizing a risk function using training models, without having data noise level as *a priori* knowledge.

Bayesian Risk Minimization

In Figure 2, we explore optimal damping parameters for models with different structure length scales. In this experiment, we use spherical harmonic models as true models to generate synthetic data, we add 20% of Gaussian noise to the synthetic data, and then calculate solutions for different damping parameter α. A true model with large-scale structure ($l = 6$) requires more damping, whereas a model that contains only small-scale structures ($l = 20$) requires less damping.

The length scale of the true earth model is unknown in seismic tomography, but it can be considered as a sample (realization) from a set of possible earth models. It is therefore possible to find a statistically optimal regularization parameter using a finite set of training models (Chung *et al.*, 2011). The optimal parameter may be found by minimizing an empirical Bayesian risk function

$$\alpha = \arg\min \sum_i \|\mathbf{x}_i(\alpha) - \mathbf{x}_i^{\text{train}}\|^2, \tag{10}$$

in which $\mathbf{x}_i^{\text{train}}$ are training models used to generate training datasets $\mathbf{b}_i^{\text{train}}$, including different realizations of data noises. Tikhonov solutions calculated using each training dataset and parameter α is $\mathbf{x}_i(\alpha)$. The summation is over all training models.

This approach aims to minimize the misfit between the optimal model and the true model. The optimization is only in statistical sense, and the optimal parameters obtained from those training models and datasets do not necessarily minimize MM for a particular true model. The success depends on training models, noise levels in the training datasets as well as the real realization of the model and noise. In Figure 3, we use 18 spherical harmonics as training models to find the optimal Tikhonov parameter for the example inverse problem with observation contaminated by 20% of

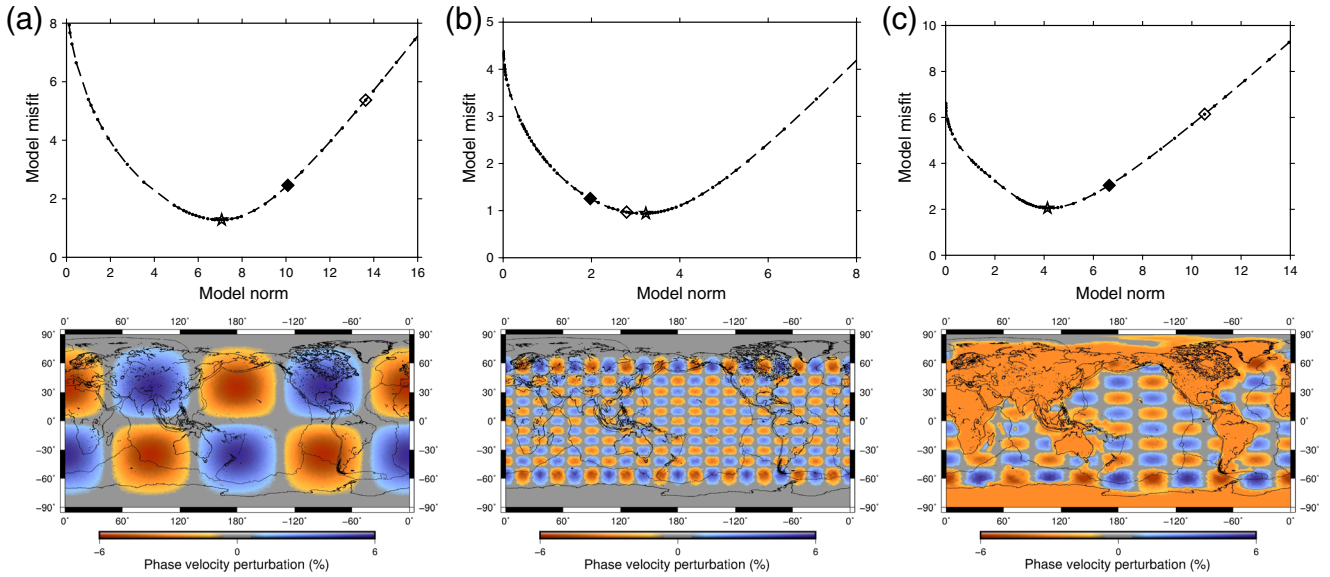

Figure 3. Model misfit curves (U-curves) for a Tikhonov inverse problem $\|\mathbf{Ax}-\mathbf{b}\|^2+\alpha\|\mathbf{x}\|^2=$ minimum with a varying damping parameter α. Synthetic data are generated for three input models (a) large scale ($l=3$), (b) small scale ($l=20$), and (c) mixed scale, all with 20% of Gaussian noise added. The mixed scale model is a spherical harmonic ($l=12$ and $m=5$) but with perturbations in continental regions replaced by a constant of −3%. The target models (with best model fit) are indicated by stars on the U-curves. The optimal models obtained based on Bayesian minimization are plotted in solid diamonds (with 20% of Gaussian noise) and open diamonds (with 10% of Gaussian noise). The training models used in this experiment are 18 spherical harmonics models with l ranging from 3 to 20. The color version of this figure is available only in the electronic edition.

Gaussian noise. The training models have a spherical harmonics degree l ranging from 3 to 20. When the true model has large-scale structure ($l=3$) and noise level in the training datasets is the same as the true realization (20%), the recovered model using the earlier Bayesian risk minimization approach is close to the target model but underdamped. When the training datasets has a noise level of 10%, it leads to an optimal model much farther away from the true model. On the other hand, when the true model has only small-scale structure ($l=20$), the optimal models are overdamped. In the case the true model has a mixed scale, the situation is somewhere in between. The success of this approach depends on prior knowledge about the true model (statistical model distribution) and true noise level. For example, if we use spherical harmonics with degree l from 3 to 40 as training models, the obtained optimal models will be different (Ⓔ Fig. S2).

Model Misfit (MM) Minimization

In the previous examples, we included the true model and models with similar length scales as training models. However, statistical distribution of the true model is not necessarily *a priori* in geophysical inverse problems; this imposes difficulties in constructing suitable calibration data. Because the target model is a linear combination of model eigenvectors

$$\mathbf{x}^{\text{target}}=\sum_i f_i^{\text{target}}\mathbf{v}_i, \tag{11}$$

the model singular vectors $\mathbf{v}_i$ can be potentially used to construct a complete set of training models. The spectral coefficients of the target model f_i^{target} are unknown, and the goal is to find spectral coefficients f_i with a similar decay rate as f_i^{target} and could be used to construct training models. We calculate the dot product (correlation) between observation $\mathbf{b}$ and eigenvector projected data $\mathbf{Av}_i$, as

$$c_i=(\mathbf{Av}_i)\cdot\mathbf{b}=(\mathbf{Av}_i)^T\mathbf{b}, \tag{12}$$

and spectral coefficients as

$$f_i=\frac{c_i}{(\sigma_i^2+\sigma_{\text{small}}^2)}. \tag{13}$$

The training models

$$\mathbf{x}_i^{\text{train}}=f_i\mathbf{v}_i \tag{14}$$

are then used to find the optimal Tikhonov parameter α in equation (10). In noise-free case, the dot product between observation and eigenvector-projected data $(\mathbf{Av}_i)\cdot\mathbf{b}=f_i\sigma_i^2$. To avoid spectral coefficients (f_i) blowing up for eigenvectors associated with small (near zero) eigenvalues, a small base eigenvalue σ_{small}^2 is added in the denominator. In determining σ_{small}^2, we make two assumptions:

1. If we order the singular values from maximum to minimum, the projection of data $\mathbf{b}$ on data singular vectors $\mathbf{u}_i$ is in general a decaying function. This condition is

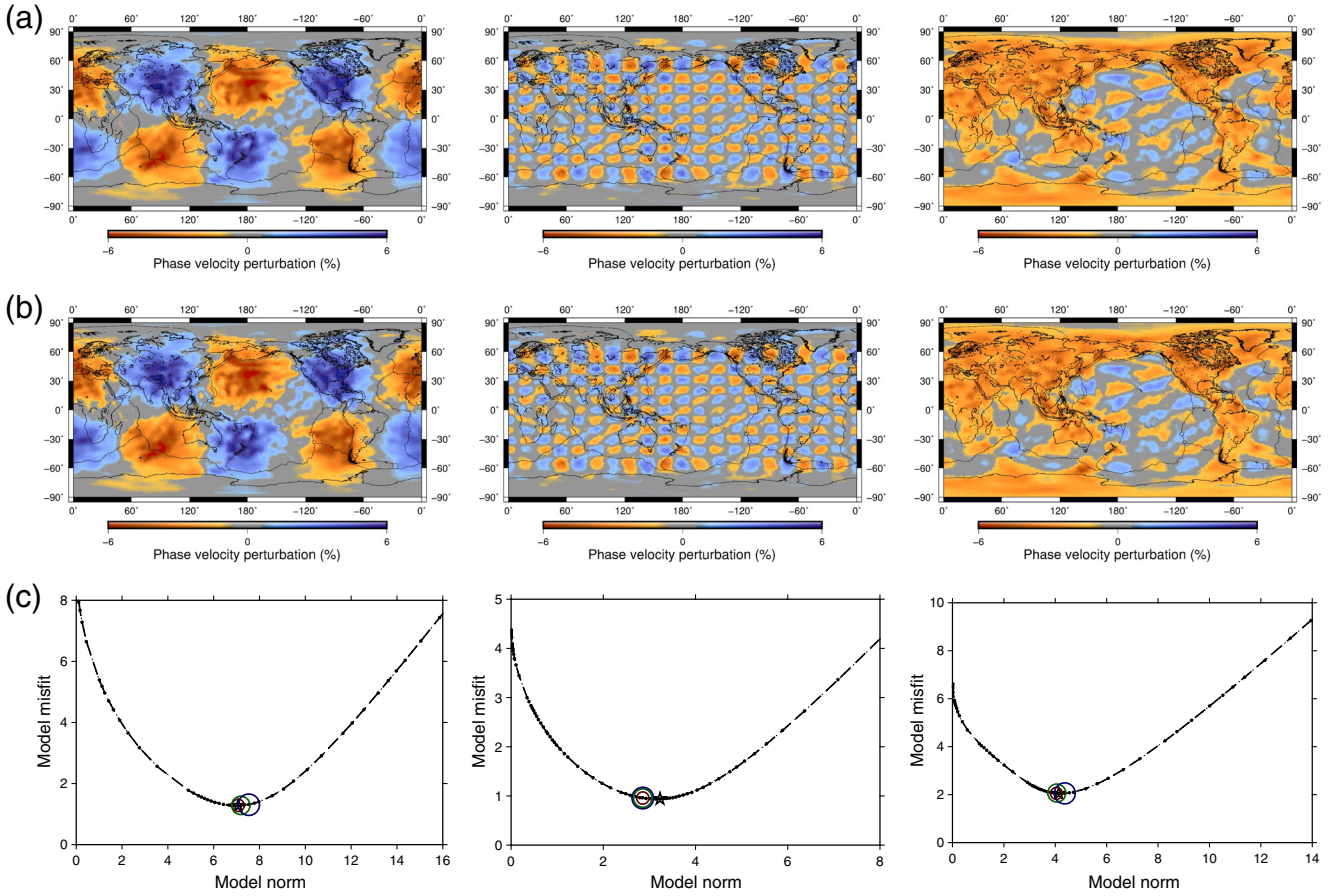

Figure 4. Tikhonov inverse problem for input models with different structure length scales: large scale $l = 3$ (left), small scale $l = 20$ (middle), and mixed scale (right). (a) Target output models that are closest to their respective input models. The target models can be determined only if the input models are known. (b) Optimal models determined using scaled eigenfunctions as training models. (c) U-curves for different damping parameters. The stars indicate the target models. Circles are optimal models determined based MM minimization with different N-point averaging noise estimates: $N = 10$ (small circles), $N = 20$ (medium circles), and $N = 50$ (large circles). The models corresponding to large circles are plotted in (b). The color version of this figure is available only in the electronic edition.

usually satisfied for forward problems that involve path integration in which σ_i decays rapidly, an indication that the inverse problem is ill-posed (Kress, 2014; Luis *et al.*, 2008).

2. The noise vector ϵ has a nearly flat spectrum in the data singular vector space, that is, $\mathbf{u}_i \cdot \epsilon \approx$ constant. The spectrum is flat in statistical sense, meaning that an N-point running average of $(\mathbf{u}_i \cdot \epsilon)$ is nearly constant.

In this case, we can order the singular values from maximum to minimum, and then calculate the N-point running average of the squared ratio between the correlation and the singular values as $\gamma_i = \|c_i/\sigma_i\|^2$. The γ_i value may increase with i at the end of its spectrum as noise in c_i blows up for small σ_i. The maximum eigenvalue associated with a γ_i value at the same level as that at the end of its spectrum is σ^2_{small} (Ⓔ Fig. S3). The basic principle behind the approach is, if the effects of noise on the coefficients (through small eigenvalues) are as large as the effects of informative signal on the coefficients, the signals are not strong enough and therefore calculations will require conditioning.

This approach is adaptive because training models will be different for different data vector $\mathbf{b}$. The noise level in real data as well as the length scale of the true model has been accounted for in the calculations of the correlation c_i and the conditioning parameter σ^2_{small}.

Because each input training model contains structure at one scale (the scale of the chosen eigenvector), summation over other eigenvectors may be dropped. The Tikhonov solutions to the training models become

$$\mathbf{x}_i(\alpha) = \left(\frac{\sigma_i^2}{\sigma_i^2 + \alpha^2}\right)\frac{\mathbf{u}_i \cdot \mathbf{b_i}}{\sigma_i}\mathbf{v}_i, \tag{15}$$

in which

$$\mathbf{b}_i = \mathbf{b}_i^{\text{train}} + \epsilon = \mathbf{A}\mathbf{x}_i^{\text{train}} + \epsilon \tag{16}$$

In Figure 4, we show that MM minimization using spectral coefficients weighted eigenvectors as training models recovers optimal models very close to the target models, regardless of the length scale of the true models. We suggest the number of points N used in estimating the noise spectrum to be around

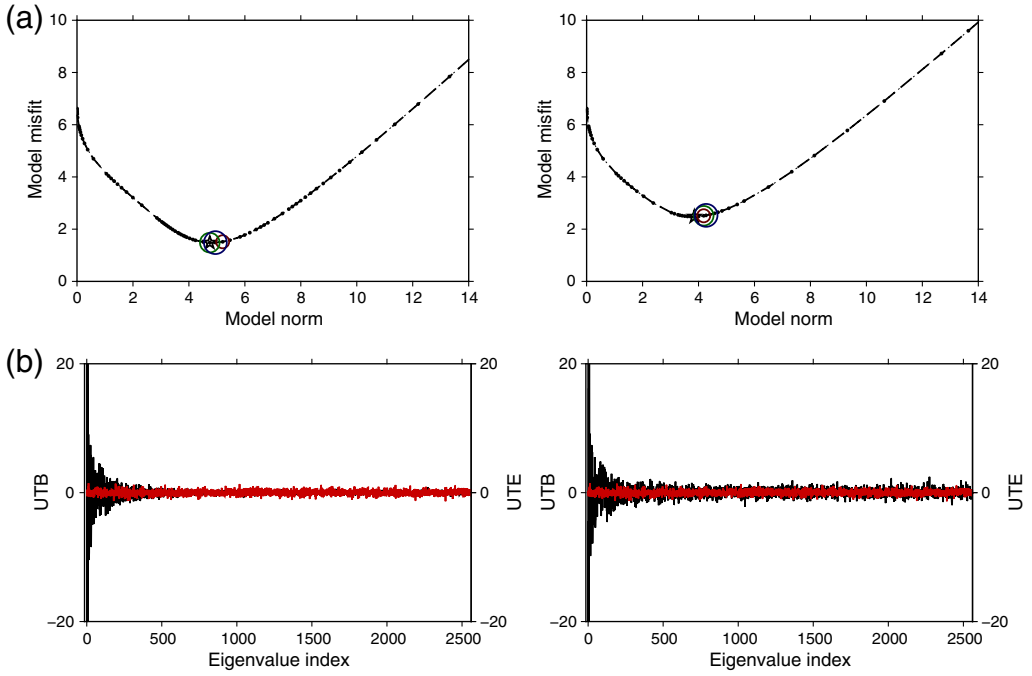

Figure 5. (a) U-curves as in Figure 4 (mixed scale) but for 10% (left) and 30% (right) Gaussian noise in data. The optimal models with $N = 10$, 20, and 50 are plotted in small, medium, and large circles, respectively. 20% of Gaussian noise has been added to training datasets in MM minimization. Optimal models do not vary significantly for different sample sizes or noise level (ϵ) used in training datasets. (b) The spectra of data $\mathbf{u}_i \cdot \mathbf{b}$ (UTB) in dark black lines and the spectra of noise $\mathbf{u}_i \cdot \epsilon$ (UTE) in bright red lines for 10% noise (left) and 30% noise (right). The spectra are in the data singular vector domain. The color version of this figure is available only in the electronic edition.

30 to make sampling statistically significant, but optimal models do not vary significantly for different sample sizes, for example, from 10 to 50. The noise vector ϵ used in training datasets $\mathbf{b}_i$ does not have a significant impact on the optimal Tikhonov parameter as long as it has a near-flat spectrum (Fig. 5). An example using a published phase velocity map as an input model and comparisons with L-curve analysis and Bayesian risk minimization are included in Ⓔ Figure S4.

Geophysical inverse problems can be very underdetermined, with the number of observations smaller than the number of unknowns. In Figure 6, we use global surface-wave tomography as an example in which the data vector $\mathbf{b}$ contains 1885 minor-arc phase-delay measurements and the unknown velocity perturbation vector $\mathbf{x}$ has 2562 elements. The optimal models obtained using MM minimization are very close to the target model. We want to point out that near-zero eigenvalues may not be computed accurately depending on the SVD code and compiler. It was the case for minor-arc inversions in which the number of unknowns exceeds the number of observations. It is important to use independent calculations to ensure the accuracy of eigenvalues. For example, the model singular vectors $\mathbf{v}_i$ can be used to calibrate the eigenvalues as $\|\mathbf{A}\mathbf{v}_i\|^2 = \sigma_i^2$.

Conclusions

Using global seismic tomography as an example, we show that optimal models obtained from traditional L-curve analysis are highly subjective, and the success of the approach mainly depends on prior knowledge about noise in data. L-curve analysis focuses on data misfit as model misfit is unknown. We develop an approach to the general inverse problem $\mathbf{A}\mathbf{x} = \mathbf{b}$ in the framework of model misfit (MM) minimization using training models. The training models are eigenvectors of the matrix $\mathbf{A}^T\mathbf{A}$, weighted by spectral coefficients calculated from the correlation between noise-contaminated observation and eigenvector-projected data. We show that this approach can be used to find optimal models very close to target models, without prior knowledge of noise in data. In the future, the same concept can be explored for higher order Tikhonov regularizations based on generalized SVD.

The dimension of the matrix $\mathbf{A}$ used in this article is 3681×2562 and the SVD of the matrix takes about 2 min on a single workstation. When the size of the matrix increases, it may become necessary to parallelize the SVD algorithm. For example, in Liu and Zhou (2016), the SVD of a matrix with a dimension of $70,000 \times 10,242$ took about 90 min on eight processors.

Data and Resources

Data used in this article are taken from published work referenced in the references.

Acknowledgments

The authors thank Editor-in-Chief Thomas L. Pratt and Delphine D. Fitzenz for their professional handling of the article and two anonymous reviewers for their careful reviews and constructive comments which

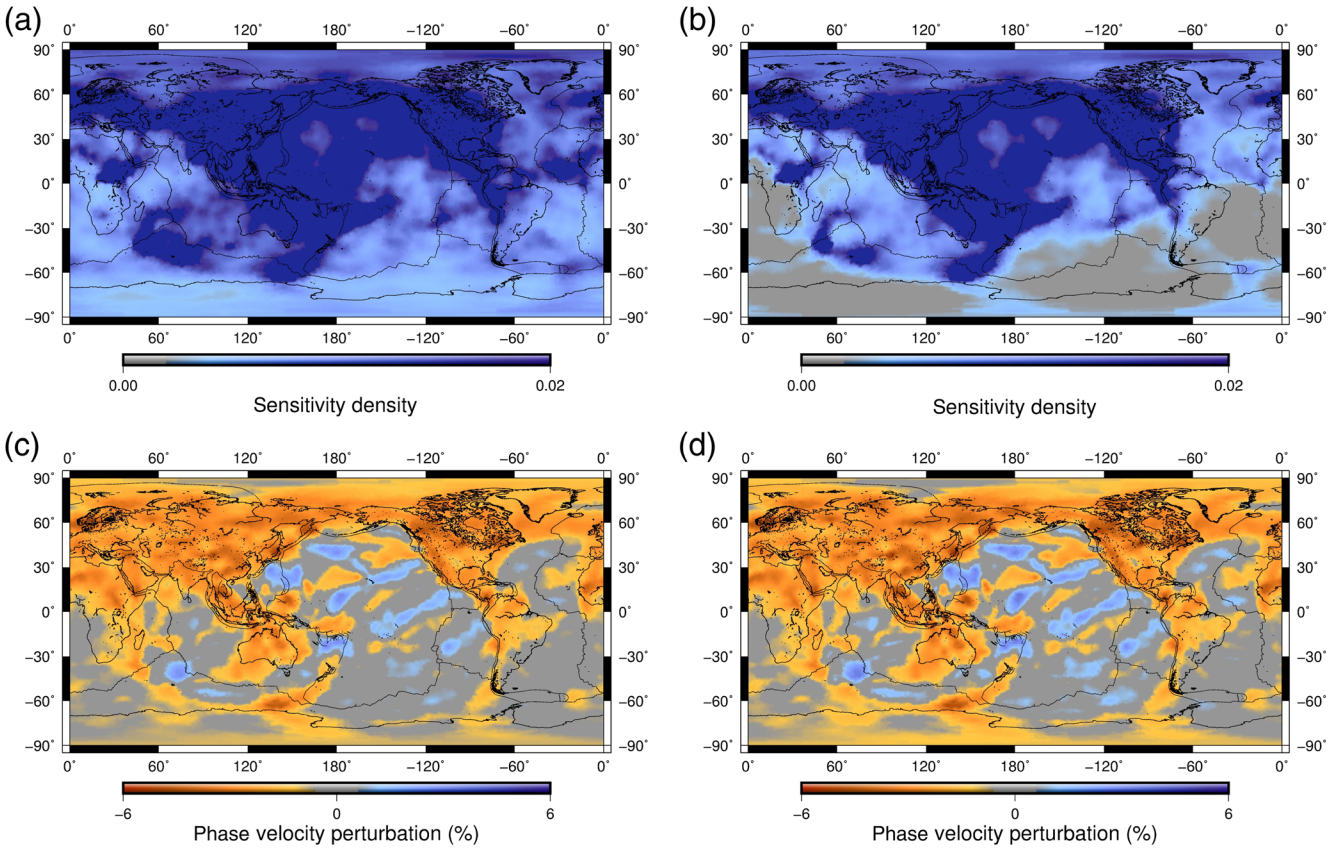

Figure 6. (a) Sensitivity density of the inverse problem using both minor-arc and major-arc surface waves. (b) Same as (a) but for minor-arc surface waves only. The sensitivity density is calculated using the diagonal elements of matrix $\mathbf{A}^T\mathbf{A}$. (c) and (d) are the target model and the optimal model obtained using MM minimization for minor-arc surface waves only. 20% noise has been added in inversions. The color version of this figure is available only in the electronic edition.

improved the article. This research was supported by the National Natural Science Foundation of China under Grant Numbers 41630210 and 41374047 and the U.S. National Science Foundation under Grant Number EAR-1737737. Advanced Research Computing at Virginia Tech provided computational resources. All figures were generated using the Generic Mapping Tools (Wessel and Smith, 1995).

References

Charléty, J., S. Voronin, G. Nolet, I. Loris, F. J. Simons, K. Sigloch, and I. C. Daubechies (2013). Global seismic tomography with sparsity constraints: Comparison with smoothing and damping regularization, *J. Geophys. Res.* **118,** 1–13.

Chung, J., M. Chung, and D. P. O'Leary (2011). Designing optimal spectral filters for inverse problems, *SIAM J. Sci. Comput.* **33,** 3132–3152.

Deidda, G. P., E. Bonomi, and C. Manzi (2003). Inversion of electrical conductivity data with Tikhonov regularization approach: Some considerations, *Ann. Geophys.* **46,** 549–558.

Fan, W., P. M. Shearer, and P. Gerstoft (2014). Kinematic earthquake rupture inversion in the frequency domain, *Geophys. J. Int.* **199,** 1138–1160.

Haber, E., D. W. Oldernburg, and R. Shekhtman (2007). Inversion of time domain three-dimensional electromagnetic data, *Geophys. J. Int.* **171,** 550–564.

Hansen, P. C. (1990). The discrete picard condition for discrete ill-posed problems, *BIT Numer. Math.* **30,** 658–672.

Hansen, P. C., and D. P. O'Leary (1993). The use of the L-curve in the regularization of discrere ill-posed problems, *SIAM J. Sci. Comput.* **14,** 1487–1503.

Jackson, D. D. (1972). Interpretation of inaccurate, insufficient and inconsistent data, *Geophys. J. Roy. Astron. Soc.* **28,** 97–109.

Jackson, D. D. (1979). The use of a priori data to resolve non-uniqueness in linear inversion, *Geophys. J. Roy. Astron. Soc.* **35,** 121–136.

Kaban, M. K., W. Stolk, M. Tesauro, S. E. Khrepy, N. Al-Arifi, F. Beekman, and S. A. P. L. Cloetingh (2016). 3d density model of the upper mantle of Asia based on inversion of gravity and seismic tomography data, *Geochem. Geophys. Geosys.* **17,** 4457–4477.

Kress, R. (2014). *Linear Integral Equations*, Springer, New York, New York.

Kreyszig, E. (1959). *Differential Geometry*, University of Toronto Press, Toronto, Canada.

Lawson, C. J., and R. J. Hanson (1974). *Solving Least Squares Problems*, Prentice-Hall, Englewood Cliffs, New Jersey.

Liu, K., and Y. Zhou (2016). Global Rayleigh wave phase-velocity maps from finite-frequency tomography, *Geophys. J. Int.* **205,** 51–66.

Luis, L. B., A. Carpio, O. Dorn, M. Moscoso, and F. Natterer (2008). *Inverse Problem and Imaging*, Springer, Berlin, Heidelberg.

Ma, Z., G. Masters, and N. Mancinelli (2016). Two-dimensional global Rayleigh wave attenuation model by accounting for finite-frequency focusing and defocusing effect, *Geophys. J. Int.* **204,** 631–649.

Morozov, V. A. (1984). *Methods for Solving Incorrectly Posed Problems*, M. Z. Nashed (Translation Editor), Springer, New York, New York.

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

Snieder, R., and G. Nolet (1987). Linearized scattering of surface waves on a spherical earth, *J. Geophys.* **61,** 55–63.

Song, L.-P., M. Koch, K. Koch, and J. Schlittenhardt (2004). 2-d anisotropic pn-velocity tomography underneath Germany using regional traveltimes, *Geophys. J. Int.* **157,** 645–663.

Spetzler, J., J. Trampert, and R. Snieder (2002). The effects of scattering in surface wave tomography, *Geophys. J. Int.* **149,** 755–767.

Tarantola, A. (2005). *Inverse Problem Theory and Methods for Model Parameter Estimation*, SIAM, Philadelphia, Pennsylvania.

Tikhonov, A. N., and V. Y. Arsenin (1977). *Solution of Ill-posed Problems*, Wiley, New York, New York.

Wahba, G. (1977). Practical approximate solutions to linear operator equations when the data are noisy, *SIAM J. Numer. Anal.* **14,** 651–667.

Wessel, P., and W. H. F. Smith (1995). New version of the generic mapping tools released, *Eos Trans. AGU* **76,** 329.

Yoshizawa, K., and B. Kennett (2002). Determination of the influence zone for surface wave paths, *Geophys. J. Int.* **149,** 440–453.

Zaroli, C., M. Sambridge, J.-J. Lévêque, E. Debayle, and G. Nolet (2013). An objective rationale for the choice of regularisation parameter with application to global multiple-frequency S-wave tomography, *Solid Earth* **4,** 357–371.

Zhao, D. (2015). *Multiscale Seismic Tomography*, Springer, New York, New York.

Zhou, Y. (2018). Anomalous mantle transition zone beneath the Yellowstone hotspot track, *Nature Geosci.* **11,** 449–453.

Zhou, Y., F. Dahlen, and G. Nolet (2004). 3-D sensitivity kernels for surface-wave observables, *Geophys. J. Int.* **158,** 142–168.

Zhou, Y., F. Dahlen, G. Nolet, and G. Laske (2005). Finite-frequency effects in global surface-wave tomography, *Geophys. J. Int.* **163,** 1087–1111.

Zhou, Y., G. Nolet, F. Dahlen, and G. Laske (2006). Global upper-mantle structure from finite-frequency surface-wave tomography, *J. Geophys. Res.* **111,** doi: 10.1029/2005JB003677.

Yuanyuan Fang
Zhenxing Yao
Key Laboratory of Earth and Planetary Physics
Institute of Geology and Geophysics
Chinese Academy of Sciences
Beijing 100029, China

Ying Zhou
Department of Geosciences
Virginia Tech
Blacksburg, Virginia 24061 U.S.A.
yingz@vt.edu

Manuscript received 28 March 2019;
Published Online 30 July 2019