Chapter 4

Adjoint tomography based on source subspace projection

Note

This chapter contains excerpts from a paper in preparation by Carl Tape, Malcolm Sambridge, and Jeroen Tromp. Each author is an equal contributor to the paper. Sambridge proposed using the subspace of sources. The concept was further developed by Tromp within the theoretical framework of *Tromp et al.* (2005). My primary contribution was to implement and test the source subspace projection method in comparison with a conjugate gradient algorithm. I will also focus on joint source-structure inversions using the source subspace method, in comparison with the conjugate gradient results in *Tape et al.* (2007).

4.1 Introduction

In adjoint tomography, for a given model m, one generally has access to the value of the objective function, $F(\mathsf{m})$, and its Fréchet derivative, $\partial F/\partial \mathsf{m}$, but not its second derivative or Hessian, $\partial^2 F/\partial \mathsf{m} \partial \mathsf{m}$. From an inverse theory perspective, this implies that one has to resort to conjugate-gradient based methods to determine the minimum of the objective function, rather than more rapidly converging and thus more desirable Gauss-Newton methods. Numerically, the gradient may be obtained based upon just two simulations for each earthquake: one calculation for the current model and a second, 'adjoint', calculation that uses time-reversed signals at the receivers as simultaneous, fictitious sources (e.g.,

Tarantola, 1984, 1986; *Akçelik et al.*, 2002, 2003; *Tromp et al.*, 2005; *Tape et al.*, 2007). The calculation of the gradient is independent of the number of receivers, components, and picks.

To increase the convergence rate of nonlinear inversion algorithms, *Sambridge et al.* (1991) proposed an improvement to the conjugate gradient algorithm advocated by *Tarantola* (1986) for the nonlinear inversion of seismic reflection data. The *Sambridge et al.* (1991) approach involves decomposing the gradient of the misfit function in terms of parts that correspond to a particular parameter type, e.g., separating the contributions to the gradient due to density, bulk-sound wave speed, shear wave speed, source location, and source mechanism. Collectively, these contributions to the gradient define a small subspace, and the algorithm proceeds by minimizing the objective function within this subspace. By solving a linearized problem within this subspace, at each iteration one only needs to invert a small matrix, which is the projection of the full Hessian onto the subspace.

In this article we introduce an alternative to the *Sambridge et al.* (1991) algorithm, which involves a projection onto the subspace spanned by the model parameters. Instead, we will consider a strategy that involves projecting the gradient and Hessian of the objective function onto the subspace spanned by the earthquakes; hence the phrase 'source subspace projection'. By performing projections in the data space the new approach differs from all earlier applications of subspace methods in seismology, which carried out projections in the model space (e.g. *Kennett et al.*, 1988; *Sambridge*, 1990; *Rawlinson et al.*, 2001). The resulting source-projected Hessian is still manageable, having the dimension of the number of earthquakes, which will be in the hundreds to thousands.

We compare the convergence rate of the classical conjugate gradient method with that of the source subspace projection algorithm by repeating some of the 2D experiments presented by *Tape et al.* (2007). We demonstrate that the source subspace projection algorithm involves only minor modifications of the classical conjugate gradient method, but that the source subspace projection algorithm converges two to three times faster. The conjugate gradient approach involves the determination of a trial model in the (conjugate) gradient direction, followed by quadratic or cubic interpolation to determine the minimum of the misfit function in the search direction. The calculation and associated storage and evaluation of this trial model is avoided in the source subspace projection algorithm, thus saving considerable compute time, I/O, and storage.

4.2 Classical least-squares solutions

To set the stage, and to introduce the necessary notation, we begin by considering the classical least-squares solution to an inverse problem (*Tarantola*, 2005). Let m_0 denote a reference *a priori* model and m a new model; these are M-dimensional vectors. The prior $M \times M$ symmetric, positive-definite model covariance matrix is denoted by $\mathsf{C}_{\mathsf{m}_0}$. The N-dimensional data vector is denoted by d, and the associated $N \times N$ symmetric, positive-definite data covariance matrix is denoted by C_d. The prediction for the current model is represented by the N-dimensional vector $\mathsf{g}(\mathsf{m})$.

Following *Tarantola* (2005), consider the *a posteriori* probability density function in the model space:

$$\sigma_\mathsf{m} = \text{const. } \exp[-F(\mathsf{m})], \tag{4.1}$$

where

$$2F(\mathsf{m}) = [\mathsf{g}(\mathsf{m}) - \mathsf{d}]^\mathsf{T} \mathsf{C}_\mathsf{d}^{-1} [\mathsf{g}(\mathsf{m}) - \mathsf{d}] + (\mathsf{m} - \mathsf{m}_0)^\mathsf{T} \mathsf{C}_{\mathsf{m}_0}^{-1} (\mathsf{m} - \mathsf{m}_0). \tag{4.2}$$

If the function $\mathsf{g}(\mathsf{m})$ can be linearized around m_0, we may write

$$\mathsf{g}(\mathsf{m}) \approx \mathsf{g}(\mathsf{m}_0) + \mathsf{G}(\mathsf{m} - \mathsf{m}_0), \tag{4.3}$$

where G denotes the $N \times M$ partial derivative matrix

$$\mathsf{G} = \frac{\partial \mathsf{g}}{\partial \mathsf{m}}. \tag{4.4}$$

Now let us introduce the notation

$$\Delta\mathsf{m} = \mathsf{m} - \mathsf{m}_0, \tag{4.5}$$

$$\Delta\mathsf{d} = \mathsf{d} - \mathsf{g}(\mathsf{m}_0). \tag{4.6}$$

Then to first order in $\Delta\mathsf{m}$ and $\Delta\mathsf{d}$ (4.2) becomes

$$2F \approx (\mathsf{G}\Delta\mathsf{m} - \Delta\mathsf{d})^\mathsf{T} \mathsf{C}_\mathsf{d}^{-1} (\mathsf{G}\Delta\mathsf{m} - \Delta\mathsf{d}) + \Delta\mathsf{m}^\mathsf{T} \mathsf{C}_{\mathsf{m}_0}^{-1} \Delta\mathsf{m}, \tag{4.7}$$

and thus the *a posteriori* probability density function is approximately Gaussian, such that (*Tarantola*, 2005)

$$\Delta \mathsf{m} = (\mathsf{G}^\mathsf{T}\mathsf{C}_\mathsf{d}^{-1}\mathsf{G} + \mathsf{C}_{\mathsf{m}_0}^{-1})^{-1}\mathsf{G}^\mathsf{T}\mathsf{C}_\mathsf{d}^{-1}\Delta \mathsf{d} = \mathsf{C}_{\mathsf{m}_0}\mathsf{G}^\mathsf{T}(\mathsf{G}\mathsf{C}_{\mathsf{m}_0}\mathsf{G}^\mathsf{T} + \mathsf{C}_\mathsf{d})^{-1}\Delta \mathsf{d}. \tag{4.8}$$

Using a quasi-Newton method, we may use the iterative algorithm (*Tarantola*, 2005)

$$\begin{aligned} \mathsf{m}_{n+1} &= \mathsf{m}_n + \lambda_n(\mathsf{G}_n^\mathsf{T}\mathsf{C}_\mathsf{d}^{-1}\mathsf{G}_n + \mathsf{C}_{\mathsf{m}_0}^{-1})^{-1}(\mathsf{G}_n^\mathsf{T}\mathsf{C}_\mathsf{d}^{-1}\Delta \mathsf{d}_n - \mathsf{C}_{\mathsf{m}_0}^{-1}\Delta \mathsf{m}_n) \\ &= \mathsf{m}_n - \lambda_n\Delta \mathsf{m}_n + \lambda_n\mathsf{C}_{\mathsf{m}_0}\mathsf{G}_n^\mathsf{T}(\mathsf{G}_n\mathsf{C}_{\mathsf{m}_0}\mathsf{G}_n^\mathsf{T} + \mathsf{C}_\mathsf{d})^{-1}(\Delta \mathsf{d}_n + \mathsf{G}_n\Delta \mathsf{m}_n), \end{aligned} \tag{4.9}$$

where $\lambda_n \approx 1$, $\Delta \mathsf{d}_n = \mathsf{d} - \mathsf{g}(\mathsf{m}_n)$, and $\Delta \mathsf{m}_n = \mathsf{m}_n - \mathsf{m}_0$.

4.3 Source subspace projection method

In the source subspace projection approach, we project the problem onto the subspace spanned by the sources as follows. We partition the data vector as follows:

$$\Delta \mathsf{d} = (\Delta d_i,\ i = 1, N) = \{(\Delta d_{sp},\ p = 1, N_s)\,,\ s = 1, S\}\,, \tag{4.10}$$

where S is the total number of sources, N_s is the number of measurements per source, and Δd_{sp} is the pth measurement for source s. Then the total number of data, N, is defined in terms of the number of sources, S, and the number of picks per source, N_s, by

$$N = \sum_{s=1}^{S} N_s. \tag{4.11}$$

Because the data covariance matrix C_d is symmetric and positive-definite we can define its square root, which will be denoted by $\mathsf{C}_\mathsf{d}^{1/2}$, and its inverse by $\mathsf{C}_\mathsf{d}^{-1/2}$. We will assume that the data covariance matrix C_d is block diagonal, with S symmetric positive-definite blocks $\mathsf{C}_{\mathsf{d}s}$ of size $N_s \times N_s$. We define a set of S orthonormal N-dimensional vectors

$$\mathsf{p}_s^\mathsf{T} = (0 \cdots 0\, \Delta\overline{d}_{s1} \cdots \Delta\overline{d}_{sN_s}\, 0 \cdots 0), \tag{4.12}$$

where the N_s-dimensional vector $\Delta\bar{\mathsf{d}}_s^{\mathsf{T}} = (\Delta\overline{d}_{s1} \cdots \Delta\overline{d}_{sN_s})$ is determined by

$$\Delta\bar{\mathsf{d}}_s = \mathsf{C}_{\mathsf{d}s}^{-1/2}\Delta\mathsf{d}_s. \tag{4.13}$$

If we further assume that the data covariance matrix $\mathsf{C}_{\mathsf{d}s}$ associated with source s is diagonal with elements σ_{sp}^2, which implies that $\mathsf{C}_{\mathsf{d}s}^{1/2}$ is also diagonal with elements σ_{sp}, then (4.13) implies

$$\Delta\overline{d}_{sp} = \Delta d_{sp}/\sigma_{sp}. \tag{4.14}$$

It is easily shown that

$$\mathsf{p}_s^{\mathsf{T}}\mathsf{p}_{s'} = \delta_{ss'}\sum_{p=1}^{N_s}(\Delta d_{sp}/\sigma_{sp})^2. \tag{4.15}$$

We now define the $S \times N$ projection operator P by

$$\mathsf{P}^{\mathsf{T}} = (\mathsf{p}_1 \cdots \mathsf{p}_S). \tag{4.16}$$

We will see in what follows that this choice of projection operator fits beautifully with the adjoint approach to calculating the gradient of an objective function.

In the source subspace projection method, we consider the *a posteriori* model space probability density function

$$\tilde{\sigma}_{\mathsf{m}} = \text{const. } \exp[-\tilde{F}(\tilde{\mathsf{m}})], \tag{4.17}$$

where

$$2\tilde{F} \approx [\mathsf{P}\mathsf{C}_{\mathsf{d}}^{-1/2}(\mathsf{G}\Delta\tilde{\mathsf{m}} - \Delta\mathsf{d})]^{\mathsf{T}}[\mathsf{P}\mathsf{C}_{\mathsf{d}}^{-1/2}(\mathsf{G}\Delta\tilde{\mathsf{m}} - \Delta\mathsf{d})] + \Delta\tilde{\mathsf{m}}^{\mathsf{T}}\mathsf{C}_{\mathsf{m}_0}^{-1}\Delta\tilde{\mathsf{m}}. \tag{4.18}$$

Note that in comparison to the classical expression (4.2) this amounts to using an inverse data covariance matrix $\mathsf{C}_{\mathsf{d}}^{-1/2}\mathsf{P}^{\mathsf{T}}\mathsf{P}\mathsf{C}_{\mathsf{d}}^{-1/2}$, rather than $\mathsf{C}_{\mathsf{d}}^{-1}$. In terms of the $S \times M$ 'projected gradient'

$$\tilde{\mathsf{G}} \equiv \mathsf{P}\mathsf{C}_{\mathsf{d}}^{-1/2}\mathsf{G}, \tag{4.19}$$

and the 'projected data vector'

$$\Delta\tilde{\mathsf{d}} = \mathsf{P}\mathsf{C}_{\mathsf{d}}^{-1/2}\Delta\mathsf{d}, \tag{4.20}$$

we have

$$2\tilde{F} \approx (\tilde{\mathsf{G}}\Delta\tilde{\mathsf{m}} - \Delta\tilde{\mathsf{d}})^{\mathsf{T}}(\tilde{\mathsf{G}}\Delta\tilde{\mathsf{m}} - \Delta\tilde{\mathsf{d}}) + \Delta\tilde{\mathsf{m}}^{\mathsf{T}}\mathsf{C}_{\mathsf{m}_0}^{-1}\Delta\tilde{\mathsf{m}}. \tag{4.21}$$

Again the *a posteriori* probability density function is approximately Gaussian, such that

$$\Delta\tilde{\mathsf{m}} = (\tilde{\mathsf{G}}^{\mathsf{T}}\tilde{\mathsf{G}} + \mathsf{C}_{\mathsf{m}_0}^{-1})^{-1}\tilde{\mathsf{G}}^{\mathsf{T}}\Delta\tilde{\mathsf{d}} = \mathsf{C}_{\mathsf{m}_0}\tilde{\mathsf{G}}^{\mathsf{T}}(\tilde{\mathsf{G}}\mathsf{C}_{\mathsf{m}_0}\tilde{\mathsf{G}}^{\mathsf{T}} + \mathsf{I})^{-1}\Delta\tilde{\mathsf{d}}. \tag{4.22}$$

Note that compared to (4.8) the projected data covariance matrix

$$\tilde{\mathsf{C}}_{\mathsf{d}} = (\mathsf{P}\mathsf{C}_{\mathsf{d}}^{-1/2})\mathsf{C}_{\mathsf{d}}(\mathsf{P}\mathsf{C}_{\mathsf{d}}^{-1/2})^{\mathsf{T}} = \mathsf{I}, \tag{4.23}$$

has become the $S \times S$ identity matrix, because the data covariance matrix C_{d} is absorbed in the definition (4.19) of G. Note also that the model update (4.22) only requires the inversion of a positive-definite $S \times S$ matrix.

Using a quasi-Newton method, we may use the iterative algorithm (*Tarantola*, 2005)

$$\begin{aligned}\tilde{\mathsf{m}}_{n+1} &= \tilde{\mathsf{m}}_n + \lambda_n(\tilde{\mathsf{G}}^{\mathsf{T}}\tilde{\mathsf{G}} + \mathsf{C}_{\mathsf{m}_0}^{-1})^{-1}(\tilde{\mathsf{G}}^{\mathsf{T}}\Delta\tilde{\mathsf{d}}_n - \mathsf{C}_{\mathsf{m}_0}^{-1}\Delta\tilde{\mathsf{m}}_n) \\ &= \tilde{\mathsf{m}}_n - \lambda_n\Delta\tilde{\mathsf{m}}_n + \lambda_n\mathsf{C}_{\mathsf{m}_0}\tilde{\mathsf{G}}_n^{\mathsf{T}}(\tilde{\mathsf{G}}_n\mathsf{C}_{\mathsf{m}_0}\tilde{\mathsf{G}}_n^{\mathsf{T}} + \mathsf{I})^{-1}(\Delta\tilde{\mathsf{d}}_n + \tilde{\mathsf{G}}_n\Delta\tilde{\mathsf{m}}_n),\end{aligned} \tag{4.24}$$

where $\lambda_n \approx 1$, $\Delta\tilde{\mathsf{d}}_n = \mathsf{P}\mathsf{C}_{\mathsf{d}}^{-1/2}\Delta\mathsf{d}_n$, and $\Delta\tilde{\mathsf{m}}_n = \tilde{\mathsf{m}}_n - \mathsf{m}_0$, to determine successive model updates. Because $(\tilde{\mathsf{G}}^{\mathsf{T}}\tilde{\mathsf{G}} + \mathsf{C}_{\mathsf{m}_0}^{-1})$ is an $M \times M$ matrix and $(\tilde{\mathsf{G}}_n\mathsf{C}_{\mathsf{m}_0}\tilde{\mathsf{G}}_n^{\mathsf{T}} + \mathsf{I})$ a generally much smaller $S \times S$ matrix, in practice we use the second equality in (4.24).

4.3.1 Significance of the source-projected gradient

In this section we investigate the significance of the source-projected partial derivative matrix $\tilde{\mathsf{G}}$ given by (4.19). To make the connection between this gradient and adjoint methods, let us consider a specific problem involving N cross-correlation traveltime anomalies ΔT_i, $i = 1, \ldots, N$, with associated standard deviations σ_i, $i = 1, \ldots, N$. Let us further assume that we are dealing with a structural inversion, and that the model m is expanded in M

basis functions $B_k(\mathbf{x})$, $k = 1, \ldots, M$, such that

$$\mathsf{m}(\mathbf{x}) = \sum_{k=1}^{M} m_k B_k(\mathbf{x}). \tag{4.25}$$

In this case the partial derivative matrix G has elements

$$G_{ik} = \frac{\partial T_i}{\partial m_k} = \int_V K_i(\mathbf{x}) B_k(\mathbf{x}) \, \mathrm{d}^3\mathbf{x}, \tag{4.26}$$

where V denotes the model volume and $K_i(\mathbf{x})$ the finite-frequency sensitivity kernel associated with observation i (e.g., *Dahlen et al.*, 2000; *Tromp et al.*, 2005).

In the particular case of cross-correlation traveltime anomalies, the source subspace projection operator P is given by (4.16), where

$$\Delta\overline{d}_{sp} = \Delta T_{sp}/\sigma_{sp}. \tag{4.27}$$

It is now straightforward to show that the source-projected data vector $\Delta\tilde{\mathsf{d}}$ (4.20) has elements

$$(\Delta\tilde{\mathsf{d}})_s = \sum_{p=1}^{N_s} (\Delta T_{sp}/\sigma_{sp})^2, \tag{4.28}$$

and that the elements of the source-projected gradient $\tilde{\mathsf{G}}$ (4.19) are given by

$$(\tilde{\mathsf{G}})_{sk} = (\mathsf{P}\mathsf{C}_{\mathsf{d}}^{-1/2}\mathsf{G})_{sk} = \int_V \left[\sum_{p=1}^{N_s} (\Delta T_{sp}/\sigma_{sp}^2) K_{sp}(\mathbf{x}) \right] B_k(\mathbf{x}) \, \mathrm{d}^3\mathbf{x} = -\int_V K_s(\mathbf{x}) B_k(\mathbf{x}) \, \mathrm{d}^3\mathbf{x}, \tag{4.29}$$

where we have defined the event kernel (*Tromp et al.*, 2005; *Tape et al.*, 2007)

$$K_s(\mathbf{x}) = -\sum_{p=1}^{N_s} (\Delta T_{sp}/\sigma_{sp}^2) K_{sp}(\mathbf{x}). \tag{4.30}$$

These kernels are calculated based upon the interaction between the regular wavefield $\mathbf{s}$ and

an adjoint wavefield $\mathbf{s}^\dagger$ that is generated by the adjoint source

$$\mathbf{f}_s^\dagger(\mathbf{x},t) = -\sum_{p=1}^{N_s}(\Delta T_{sp}/\sigma_{sp}^2)\frac{1}{M_{sp}}\,\partial_t \mathbf{s}_{sp}(T-t)]\delta(\mathbf{x}-\mathbf{x}_{sp}), \tag{4.31}$$

where M_{sp} is a normalization factor, and $\mathbf{x}_{sp}$ denotes the receiver location associated with source s and pick p. The calculation of the event kernels involves only two numerical simulations per earthquake.

4.3.2 Comparison with the conjugate gradient method

In a traditional conjugate gradient method, the first model update is in the opposite direction of the gradient of the cross-correlation traveltime misfit function (e.g., *Tarantola*, 2005; *Tape et al.*, 2007), i.e.,

$$\Delta m_k \approx -\nu \sum_{k'=1}^{M}(\mathsf{C_m})_{kk'}\sum_{s=1}^{S}\int_V K_s(\mathbf{x})B_{k'}(\mathbf{x})\,\mathrm{d}^3\mathbf{x}, \tag{4.32}$$

where the scalar ν determines the step length and thus the location of the trial model. Note how the 'metric' $\mathsf{C_m}$ turns the dual vector $\hat{\gamma}_{k'} = \int_V K_s(\mathbf{x})B_{k'}(\mathbf{x})\,\mathrm{d}^3\mathbf{x}$, i.e., the gradient, into a vector: $\gamma = \mathsf{C_m}\hat{\gamma}$ (see e.g., *Tarantola*, 2005). Upon comparing this expression with the source subspace projection result (4.22), i.e.,

$$\Delta\tilde{m}_k \approx \sum_{k'=1}^{M}\sum_{s=1}^{S}(\mathsf{C_m})_{kk'}(\tilde{\mathsf{G}})_{sk'}\Delta\mu_s = -(\mathsf{C_m})_{kk'}\sum_{s=1}^{S}\Delta\mu_s\int_V K_s(\mathbf{x})B_{k'}(\mathbf{x})\,\mathrm{d}^3\mathbf{x}, \tag{4.33}$$

where the S-dimensional vector $\Delta\mu$ is determined by

$$\Delta\mu = (\tilde{\mathsf{G}}\mathsf{C}_{\mathsf{m}_0}\tilde{\mathsf{G}}^{\mathsf{T}} + \mathsf{I})^{-1}\Delta\tilde{\mathsf{d}}, \tag{4.34}$$

we see how the source subspace projection method 'preconditions' the model update by combining the event Fréchet derivatives $\int_V K_s(\mathbf{x})B_{k'}(\mathbf{x})\,\mathrm{d}^3\mathbf{x}$ with weights $\Delta\mu_s$.

4.4 2D synthetic experiments

In Figures 4.1–4.3 we compare the source-subspace (SS) inversion with a conjugate-gradient (CG) inversion. In Figure 4.2, $\mathbf{m}_{01}$ for SS (d) is much closer to the target model that the CG version (a). This can be seen visually, as well as in the misfit values in (h). A key distinction is that the CG models require an additional evaluation of the misfit function at each step (e.g., *Tape et al.*, 2007). At $\mathbf{m}_{02}$, for example, CG has used 7 forward simulations, while SS has used only 5.

Figure 4.3 demonstrates that the SS and CG algorithms perform similarly for the case of a three-parameter inversion for location and origin time.

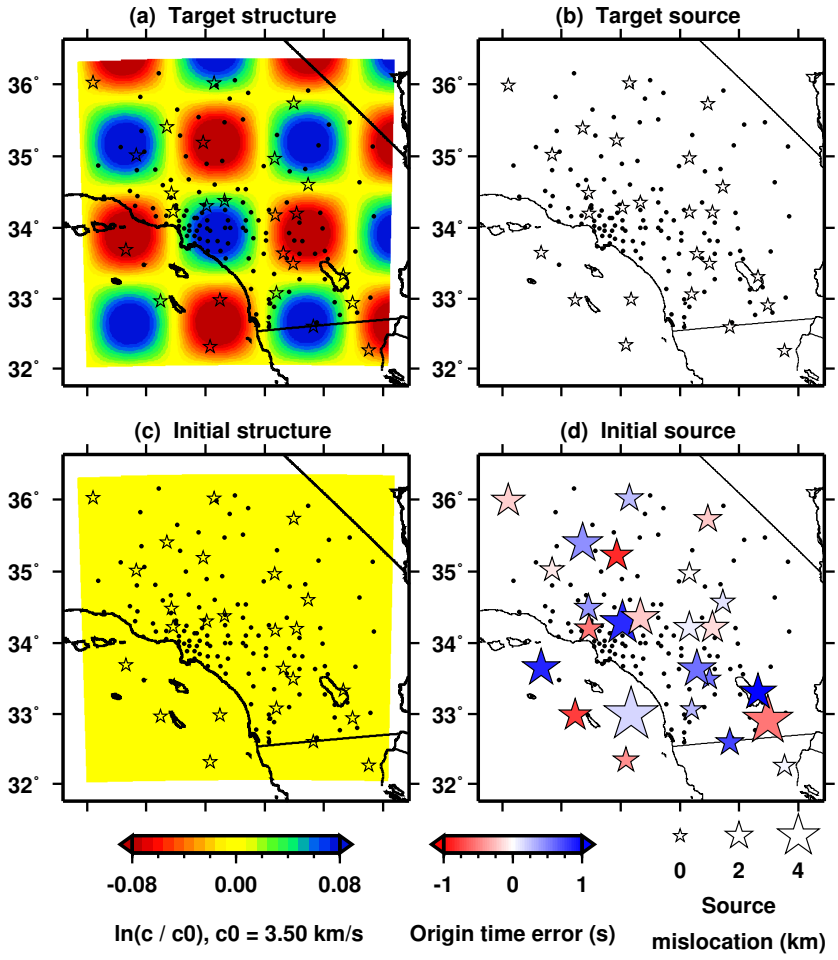

Figure 4.1: Initial and target structure and sources for subspace experiments. Target synthetic seismograms are generated using the target structure or target sources. Initial synthetics are generated using the initial structure or initial sources. Through iterative minimization of a misfit function, the initial model moves in the direction of the target model.

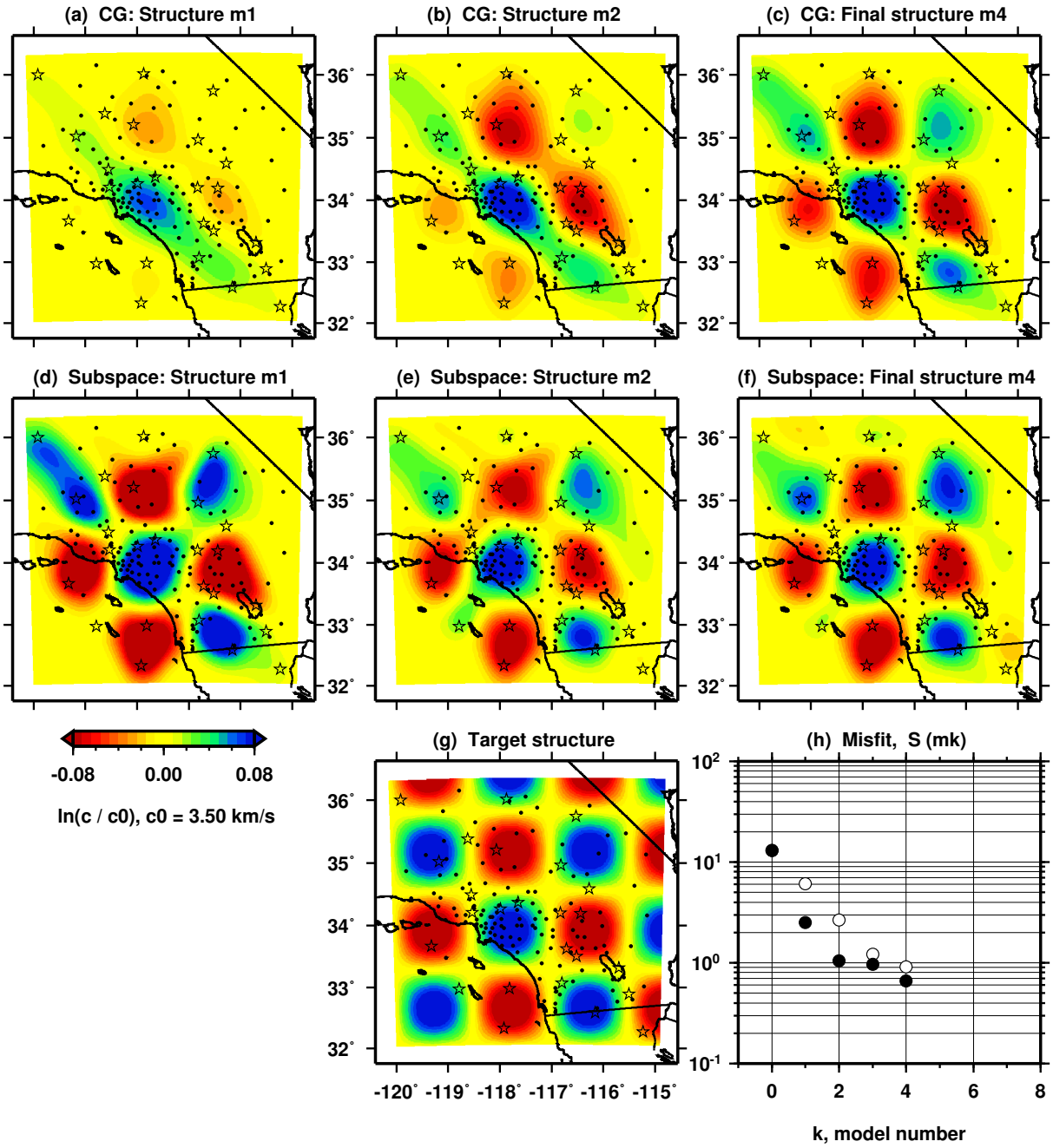

Figure 4.2: Comparison between conjugate-gradient (CG) and source-subspace synthetic inversions for structure parameters. (a)–(c) Conjugate-gradient models $\mathbf{m}_{01}$, $\mathbf{m}_{02}$, and $\mathbf{m}_{04}$. (d)–(f) Source-subspace (SS) models $\mathbf{m}_{01}$, $\mathbf{m}_{02}$, and $\mathbf{m}_{04}$. (g) Target structure. (h) Misfit function evaluations for each model. White circles are for CG models; black circles are for SS models.

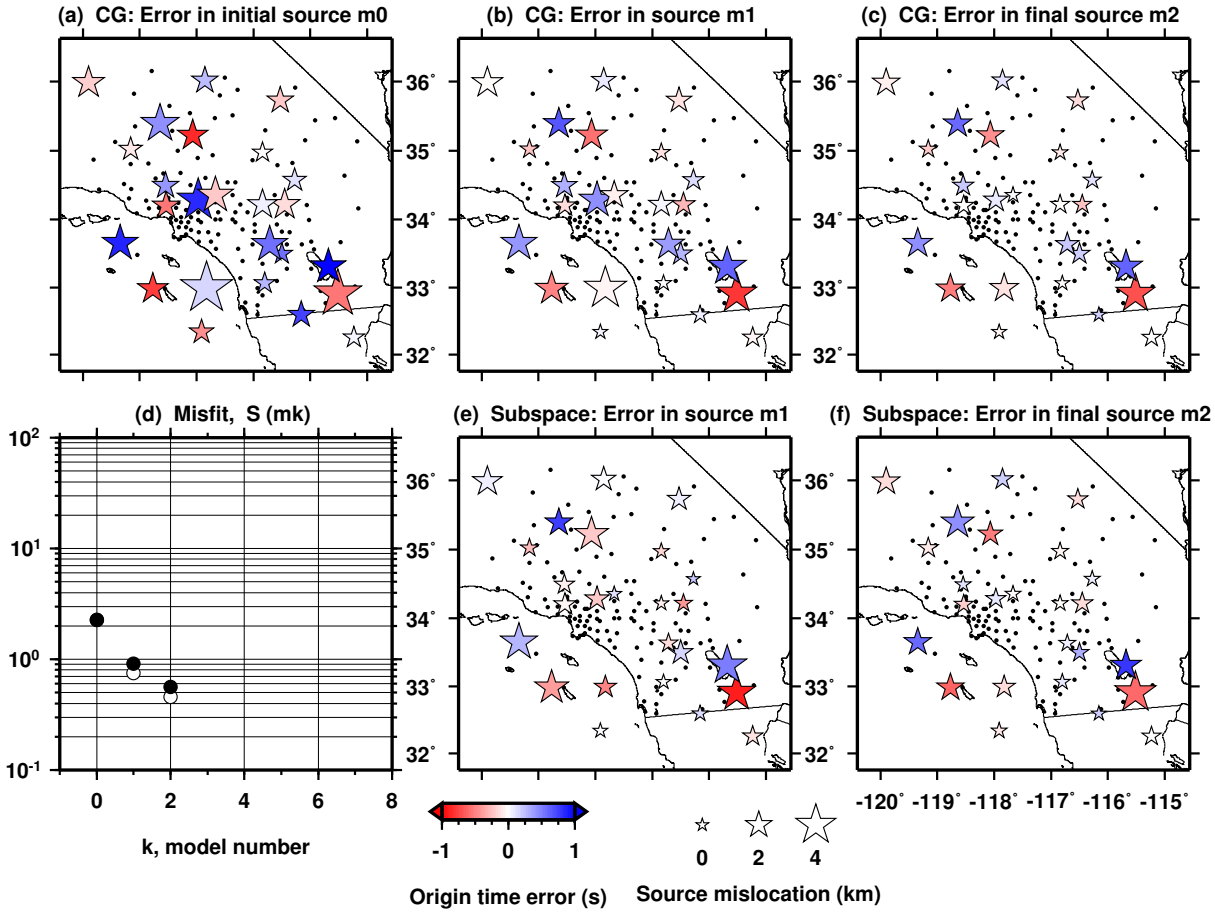

Figure 4.3: Comparison between conjugate-gradient and source-subspace synthetic inversions for source parameters. The three source parameters in the inversion experiment are (x_s, y_s) location and origin time (t_s). (a) Initial source errors for both the conjugate-gradient (CG) and source-subspace (SS) inversions. (b)–(c) CG source errors for models $\mathbf{m}_{01}$ and $\mathbf{m}_{02}$. (d) Reduction in misfit for CG (white circles) and SS (black circles). The performance of CG and SS is essentially the same for the source inversion.