14

Linear inversion

In Chapter 12 we saw how the parametrization of a continuous model allows us to formulate a discrete linear relationship between data $\boldsymbol{d}$ and model $\boldsymbol{m}$. With unknown corrections added to the model vector, this linear relationship remains formally the same if we write the physical model parameters as $\boldsymbol{m}_1$ and the corrections as $\boldsymbol{m}_2$ but combine both in one vector $\boldsymbol{m}$:

$$\boldsymbol{A}_1\boldsymbol{m}_2 + \boldsymbol{A}_2\boldsymbol{m}_2 = \boldsymbol{A}\boldsymbol{m} = \boldsymbol{d} \quad (12.1) \text{ again.}$$

Assuming we have M_1 model parameters and M_2 corrections, this is a system of N equations (data) and $M = M_1 + M_2$ unknowns. For more than one reason the solution of the system is not straightforward:

- Even if we do not include multiple measurements along the same path, many of the N rows will be dependent. Since the data always contain errors, this implies we cannot solve the system exactly, but have to minimize the misfit between $\boldsymbol{A}\boldsymbol{m}$ and $\boldsymbol{d}$. For this misfit we can define different norms, and we face a choice of options.
- Despite the fact that we have (usually) many more data than unknowns (i.e. $N \gg M$), the system is almost certainly ill-posed in the sense that small errors in $\boldsymbol{d}$ can lead to large errors in $\boldsymbol{m}$; a parameter m_j may be completely undetermined ($A_{ij} = 0$ for all i) if it represents a node that is far away from any raypath. We cannot escape making a subjective choice among an infinite set of equally satisfactory solutions by imposing a *regularization* strategy.
- For large M, the numerical computation of the solution has to be done with an iterative matrix solver which is often halted when a satisfactory fit is obtained. Such efficient shortcuts interfere with the regularization strategy.

We shall deal with each of these aspects in succession. Appendix D introduces some concepts of probability theory that are needed in this chapter.

14.1 Maximum likelihood estimation and least squares

In experimental sciences, the most commonly used misfit criterion is the criterion of least squares, in which we minimize χ^2 ('chi square') as a function of the model:

$$\chi^2(\boldsymbol{m}) = \sum_{i=1}^{N} \left(\frac{|\sum_{j=1}^{M} A_{ij} m_j - d_i|^2}{\sigma_i^2} \right) = \min, \tag{14.1}$$

where σ_i is the standard deviation in datum i; χ^2 is a direct measure of the data misfit, in which we weigh the misfits inversely with their standard errors σ_i.

For uncorrelated and normally distributed errors, the principle of maximum likelihood leads naturally to the least squares definition of misfit. If there are no sources of bias, the expected value $E(d_i)$ of d_i (the average of infinitely many observations of the same observable) is equal to the 'correct' or error-free value. In practice, we have only one observation for each datum, but we usually have an educated guess at the magnitude of the errors. We almost always use a normal distribution for errors, and assume errors to be uncorrelated, such that the probability density is given by a Gaussian or 'normal' distribution of the form:

$$P(d_i) = \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(-\frac{|d_i - E(d_i)|^2}{2\sigma_i^2} \right). \tag{14.2}$$

The joint probability density for the observation of an N-tuple of data with independent errors $\boldsymbol{d} = (d_1, d_2, ..., d_N)$ is found by multiplying the individual probability densities for each datum:

$$P(\boldsymbol{d}) = \prod_{i=1}^{N} \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(-\frac{|d_i - E(d_i)|^2}{2\sigma_i^2} \right). \tag{14.3}$$

If we replace the expected values in (14.3) with the predicted values from the model parameters, we obtain again a probability, but now one that is conditional on the model parameters taking the values m_j:

$$P(\boldsymbol{d}|\boldsymbol{m}) = \prod_{i=1}^{N} \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(-\frac{|d_i - \sum_j A_{ij} m_j|^2}{2\sigma_i^2} \right). \tag{14.4}$$

We usually assume that there are no extra errors introduced by the modelling (e.g. we ignore the approximation errors introduced by linearizations, neglect of anisotropy, or the shortcomings of ray theory etc.). In fact, if such modelling errors are also uncorrelated, unbiased and normally distributed, we can take them into account by including them in σ_i – but this is a big 'if'.[†]

[†] See Tarantola [351] for a much more comprehensive discussion of this issue.

Clearly, one would like to have a model that is associated with a high probability for its predicted data vector. This leads to the definition of the likelihood function $\mathcal{L}$ for the model $\boldsymbol{m}$ given the observation of the data $\boldsymbol{d}$:

$$\mathcal{L}(\boldsymbol{m}|\boldsymbol{d}) = P(\boldsymbol{d}|\boldsymbol{m}) \propto \exp\left(-\frac{1}{2}\chi^2(\boldsymbol{m})\right).$$

Thus, maximizing the likelihood for a model involves minimizing χ^2. Since this involves minimizing the sum of squares of data misfit, the method is more generally known as the *method of least squares*. The strong point of the method of least squares is that it leads to very efficient methods of solving (12.1). Its major weakness is the reliance on a normal distribution of the errors, which may not always be the case. Because of the quadratic dependence on the misfit, outliers – misfits of several standard deviations – have an influence on the solution that may be out of proportion, which means that errors may dominate in the solution. For a truly normal distribution, large errors have such a low probability of occurrence that we would not worry about this. In practice however, many data do suffer from outliers. For picked arrival times Jeffreys [146] has already observed that the data have a tail-like distribution that deviates from the Gaussian for large deviations from the mean t_m, mainly because a later arrival is misidentified as P or S:

$$P(t) = \frac{1-\epsilon}{\sigma\sqrt{2\pi}}\mathrm{e}^{-(t-t_m)^2/2\sigma^2} + \epsilon g(t),$$

where the probability density $g(t)$ varies slowly and where $\epsilon \ll 1$. A simple method to bring the data distribution close to normal is to reject outliers with a delay that exceeds the largest delay time to be expected from reasonable effects of lateral heterogeneity. This decision can be made after a first trial inversion: for example, one may reject all data that leave a residual in excess of 3σ after a first inversion attempt.

If we divide all data – and the corresponding row of $\boldsymbol{A}$ – by their standard deviations, we end up with a data vector that is univariant, i.e. all standard deviations are equal to 1. Thus, without loss of generality, we may assume that the data are univariant, in which case we see from (14.1) that χ^2 is simply the squared length of the residual vector $|\boldsymbol{r}| = |\boldsymbol{d} - \boldsymbol{A}\boldsymbol{m}|$. From Figure 14.1 we see that $\boldsymbol{r}$ is then perpendicular to the subspace spanned by all vectors $\boldsymbol{A}\boldsymbol{y}$ (the 'range' $R(\boldsymbol{A})$ of $\boldsymbol{A}$). For if it was not, we could add a $\delta\boldsymbol{m}$ to $\boldsymbol{m}$ such that $\boldsymbol{A}\delta\boldsymbol{m}$ reduces the length of $\boldsymbol{r}$. Thus, for all $\boldsymbol{y}$ the dot product between $\boldsymbol{r}$ and $\boldsymbol{A}\boldsymbol{y}$ must be zero:

$$\boldsymbol{r} \cdot \boldsymbol{A}\boldsymbol{y} = \boldsymbol{A}^T\boldsymbol{r} \cdot \boldsymbol{y} = \boldsymbol{A}^T(\boldsymbol{d} - \boldsymbol{A}\boldsymbol{m}) \cdot \boldsymbol{y} = 0,$$

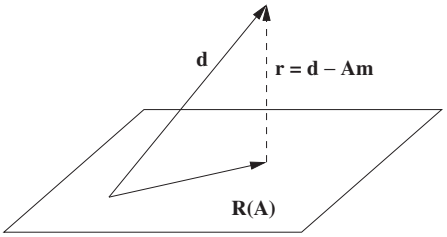

Fig. 14.1. If the data vector $\boldsymbol{d}$ does not lie in the range of $\boldsymbol{A}$, the best we can do is to minimize the length of the residual vector $\boldsymbol{r}$. This implies that $\boldsymbol{r}$ must be perpendicular to any possible vector $\boldsymbol{Ay}$.

where $\boldsymbol{A}^T$ is the transpose of $\boldsymbol{A}$ (i.e. $A^T_{ij} = A_{ji}$). Since this dot product is 0 for *all* $\boldsymbol{y}$, clearly $\boldsymbol{A}^T(\boldsymbol{d} - \boldsymbol{Am}) = 0$, or:

$$\boldsymbol{A}^T\boldsymbol{Am} = \boldsymbol{A}^T\boldsymbol{d}, \tag{14.5}$$

which is known as the set of 'normal equations' to solve the least-squares problem.

Chi square is an essential statistical measure of the goodness of fit. In the hypothetical case that we satisfy every datum with a misfit of one standard deviation we find $\chi^2 = N$; clearly values much higher than N are unwanted because the misfit is higher than could be expected from the knowledge of data errors, and values much lower than N indicate that the model is trying to fit the data errors rather than the general trend in the data. For example, if two very close rays have travel time anomalies differing by only 0.5 s and the standard deviation is estimated to be 0.7 s, we should accept that a smooth model predicts the same anomaly for each, rather than introducing a steep velocity gradient in the 3D model to try to satisfy the difference. Because we want $\chi^2 \approx N$, it is often convenient to work with the reduced χ^2 or χ^2_{red}, which is defined as χ^2/N, so that the optimum solution is found for $\chi^2_{\text{red}} \approx 1$.

But how close should χ^2 be to N? Statistical theory shows that χ^2 itself has a variance of $2N$, or a standard deviation of $\sqrt{2N}$. Thus, for 1 000 000 data the true model would with 67% confidence be found in the interval $\chi^2 = 1\,000\,000 \pm 1414$. Such theoretical bounds are almost certainly too narrow because our estimates of the standard deviations σ_i are themselves uncertain. For example, if the true σ_i are equal to 0.9 but we used 1.0 to compute χ^2, our computed χ^2 itself is in error (i.e. too low) by almost 20%, and a model satisfying this level of misfit is probably not good enough. It is therefore important to obtain accurate estimates of the standard errors, e.g. using (6.2) or (6.12). Provided one is confident that the estimated standard errors are unbiased, one should still aim for a model that brings χ^2 very close to N, say to within 20 or 30%.

An additional help in deciding how close one wishes to be to a model that fits at a level given by $\chi^2 = N$ is to plot the tradeoff between the model norm and χ^2

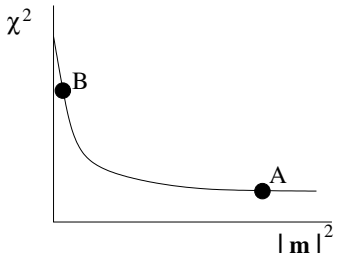

Fig. 14.2. The L- or tradeoff curve between χ^2 and model norm $|\boldsymbol{m}|^2$.

(sometimes called the L-curve), shown schematically in Figure 14.2. If the tradeoff curve shows that one could significantly reduce the norm of the model while paying only a small price in terms of an increase in χ^2 (point A in Figure 14.2), this is an indication that the standard errors in the data have been underestimated. For common data errors do not correlate between nearby stations, but the true delays should correlate – even if the Earth's properties vary erratically (because of the overlap in finite-frequency sensitivity). The badly correlating data can only be fit by significantly increasing the norm and complexity of the model, which is what we see happening on the horizontal part of the tradeoff curve. Conversely, if we notice that a significant decrease in χ^2 can be obtained at the cost of only a minor increase in model norm (point B), this indicates an overestimate of data errors and tells us we may wish to accept a model with $\chi^2 < N$. If the deviations required are unexpectedly large, this is an indication that the error estimation for the data may need to be revisited.

Depending on where on the L-curve we find that $\chi^2 = N$, we find that we do or do not have a strong constraint on the norm of the model. If the optimal data fit is obtained close to point B where the L-curve is steep, even large changes in χ^2 have little effect on the model norm. On the other hand, near point A even large changes in the model give only a small improvement of the data fit. Both A and B represent unwanted situations, since at A we are trying to fit data errors, which leads to erratic features in the model, whereas at B we are damping too strongly. In a well designed tomography experiment, $\chi^2 \approx N$ near the bend in the L-curve.

We used the term 'model norm' here in a very general sense – one may wish to inspect the Euclidean $|\boldsymbol{m}|^2$ as well as more complicated norms that we shall encounter in Section 14.5.

In many cases one inverts different data groups that have uncorrelated errors. For example Montelli et al. [215] combine travel times from the ISC catalogues with cross-correlation travel times from broadband seismometers. The ISC set, with about 10^6 data was an order of magnitude larger than the second data set (10^5), and a brute force least-squares inversion would give preference to the short period ISC data in cases where there are systematic incompatibilities. This is easily diagnosed

by computing χ^2 for the individual data groups. One would wish to weigh the data sets such that each group individually satisfies the optimal χ^2 criterion, i.e. if χ_i^2 designates the misfit for data set i with N_i data, one imposes $\chi_i^2 \approx N_i$ for each data set. This may be accomplished by giving each data set equal weight and minimizing a weighted penalty function:

$$\mathcal{P} = \sum_i \frac{1}{N_i} \chi_i^2 \,.$$

Note that this gives a solution that deviates from the maximum likelihood solution, and we should only resort to weighting if we suspect that important conditions are violated, especially those of zero mean, uncorrelated and normally distributed errors. More often, an imbalance for individual χ_i^2 simply reflects an over- or underestimation of the standard deviations for one particular group of data, and may prompt us to revisit our estimates for prior data errors.

Early tomographic studies often ignored a formal statistical appraisal of the goodness of fit, and merely quoted how much better a 3D tomographic model satisfies the data when compared to a 1D (layered or spherically symmetric) background or 'starting' model, using a quantity named 'variance reduction', essentially the reduction in the Euclidean norm of the misfit vector. This reduction is as much a function of the fit of the 1D starting model as of the data fit itself – i.e. the same 3D model can have different variance reductions depending on the starting model – and is therefore useless as a statistical measure of quality for the tomographic model.

Exercises

Exercise 14.1 Derive the normal equations by differentiating the expression for χ^2 with respect to m_k for $k = 1, ..., M$. Assume univariant data ($\sigma_i = 1$).

Exercise 14.2 Why can we not conclude from (14.5) that $\boldsymbol{Am} \equiv \boldsymbol{d}$?

14.2 Alternatives to least squares

In the parlance of mathematics, the squared Euclidean norm $\sum_i |r_i|^2$ is one of a class of Lebesgue norms defined by the power p used in the sum: $(\sum_i |r_i|^p)^{1/p}$. Thus, the Euclidean norm is also known as the 'L_2' norm because $p = 2$. Of special interest are the L_1 norm ($p = 1$) and the case $p \to \infty$ which leads to minimizing the maximum among all $|r_i|$.

Instead of simply rejecting outliers, which always requires the choice of a hard bound for acceptance, we may downweight data that show a large misfit in a

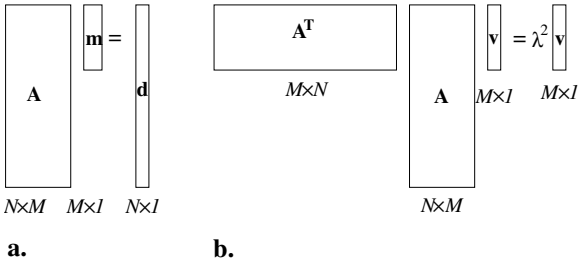

Fig. 14.3. (a) The original matrix system $\boldsymbol{Am} = \boldsymbol{d}$. (b) The eigenvalue problem for the least-squares matrix $\boldsymbol{A}^T\boldsymbol{A}$.

previous inversion attempt, and repeat the process until it converges. In 1898, the Belgian mathematician Charles Lagrange proposed such an 'iteratively weighted' least-squares solution, by iteratively solving:

$$|\boldsymbol{W}_p\boldsymbol{Am} - \boldsymbol{W}_p\boldsymbol{d}|^2 = \min\,,$$

where $\boldsymbol{W}_p$ is a diagonal matrix with elements $|r_i|^{p-2}$ and $0 \leq p < 2$, which are determined from the misfits r_i in datum i after the previous iteration. We can start with an unweighted inversion to find the first $\boldsymbol{r}$. The choice $p = 1$ leads to the minimization of the L_1 norm if it converges. The elements of the residual vector r_i vary with each iteration, and convergence is not assured, but the advantage is that the inversion makes use of the very efficient numerical tools available for linear least-squares problems. The method was introduced in geophysics by Scales et al. [304].

14.3 Singular value decomposition

Though the least squares formalism handles the incompatibility problem of data in an overdetermined system, we usually find that $\boldsymbol{A}^T\boldsymbol{A}$ has a determinant equal to zero, i.e. eigenvalues equal to zero, and its inverse does not exist. Even though in tomographic applications $\boldsymbol{A}^T\boldsymbol{A}$ is often too large to be diagonalized, we shall analyse the inverse problem using singular values ('eigenvalues' of a non-square matrix), since this formalism gives considerable insight.

Let $\boldsymbol{v}_i$ be an eigenvector of $\boldsymbol{A}^T\boldsymbol{A}$ with eigenvalue λ_i^2, so that $\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{v} = \lambda_i^2\boldsymbol{v}$. We may use squared eigenvalues because $\boldsymbol{A}^T\boldsymbol{A}$ is symmetric and has only non-negative, real eigenvalues. Its eigenvectors are orthogonal. The choice of λ^2 instead of λ as eigenvalue is for convenience: the notation λ_i^2 avoids the occurrence of $\sqrt{\lambda_i}$ later in the development. We can arrange all M eigenvectors as columns in an $M \times M$ matrix $\boldsymbol{V}$ and write (see Figure 14.3):

$$\boldsymbol{A}^T\boldsymbol{AV} = \boldsymbol{V}\boldsymbol{\Lambda}^2\,. \tag{14.6}$$

The eigenvectors are normalized such that $\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{V}\boldsymbol{V}^T = \boldsymbol{I}$.

With (14.6) we can study the underdetermined nature of the problem $\boldsymbol{Am} = \boldsymbol{d}$, of which the least-squares solution is given by the system $\boldsymbol{A}^T\boldsymbol{Am} = \boldsymbol{A}^T\boldsymbol{d}$. The eigenvectors $\boldsymbol{v}_i$ span the M-dimensional model space so $\boldsymbol{m}$ can be written as a linear combination of eigenvectors: $\boldsymbol{m} = \boldsymbol{V}\boldsymbol{y}$. Since $\boldsymbol{V}$ is orthonormal, $|\boldsymbol{m}| = |\boldsymbol{y}|$ and we can work with $\boldsymbol{y}$ instead of $\boldsymbol{m}$ if we wish to restrict the norm of the model. Using this:

$$\boldsymbol{A}^T\boldsymbol{AV}\boldsymbol{y} = \boldsymbol{V}\boldsymbol{\Lambda}^2\boldsymbol{y} = \boldsymbol{A}^T\boldsymbol{d}\,,$$

or, multiplying both on the left with $\boldsymbol{V}^T$ and using the orthogonality of $\boldsymbol{V}$:

$$\boldsymbol{\Lambda}^2\boldsymbol{y} = \boldsymbol{V}^T\boldsymbol{A}^T\boldsymbol{d}\,.$$

Since $\boldsymbol{\Lambda}$ is diagonal, this gives y_i (and with that $\boldsymbol{m} = \boldsymbol{V}\boldsymbol{y}$) simply by dividing the i-th component of the vector on the right by λ_i^2. But clearly, any y_i which is multiplied by a zero eigenvalue can take any value without affecting the data fit! We find the *minimum norm solution*, the solution with the smallest $|\boldsymbol{y}|^2$, by setting such components of $\boldsymbol{y}$ to 0. If we rank the eigenvalues $\lambda_1^2 \geq \lambda_2^2 \geq ...\lambda_K^2 > 0, 0, ..., 0$, then the last $M-K$ columns of $\boldsymbol{V}$ belong to the nullspace of $\boldsymbol{A}^T\boldsymbol{A}$. We truncate the matrices $\boldsymbol{V}$ and $\boldsymbol{\Lambda}$ to an $M \times K$ matrix $\boldsymbol{V}_K$ and a $K \times K$ diagonal matrix $\boldsymbol{\Lambda}$ to obtain the minimum norm estimate:

$$\hat{\boldsymbol{m}}_{\text{min norm}} = \boldsymbol{V}_K\boldsymbol{\Lambda}_K^{-2}\boldsymbol{V}_K^T\boldsymbol{A}^T\boldsymbol{d}\,. \tag{14.7}$$

Note that the inverse of $\boldsymbol{\Lambda}_K$ exists because we have removed the zero eigenvalues. The orthogonality of the eigenvectors still guarantees $\boldsymbol{V}_K^T\boldsymbol{V}_K = \boldsymbol{I}_K$, but now $\boldsymbol{V}_K\boldsymbol{V}_K^T \neq \boldsymbol{I}_M$.

To see how errors in the data propagate into the model, we use the fact that (14.7) represents a linear transformation of data with a covariance matrix $\boldsymbol{C}_d$. The posteriori covariance of transformed data $\boldsymbol{Td}$ is equal to $\boldsymbol{TC}_d\boldsymbol{T}^T$ (see Equation 14.39 in Appendix D). In our case we have scaled the data such that $\boldsymbol{C}_d = \boldsymbol{I}$ so that the posteriori model covariance is:

$$\begin{aligned}\boldsymbol{C}_{\hat{m}} &= \boldsymbol{V}_K\boldsymbol{\Lambda}_K^{-2}\boldsymbol{V}_K^T\boldsymbol{A}^T\boldsymbol{I}\boldsymbol{A}\boldsymbol{V}_K\boldsymbol{\Lambda}_K^{-2}\boldsymbol{V}_K^T \\ &= \boldsymbol{V}_K\boldsymbol{\Lambda}_K^{-2}\boldsymbol{\Lambda}_K^2\boldsymbol{\Lambda}_K^{-2}\boldsymbol{V}_K^T \\ &= \boldsymbol{V}_K\boldsymbol{\Lambda}_K^{-2}\boldsymbol{V}_K^T. \end{aligned} \tag{14.8}$$

Thus the posteriori variance of the estimate for parameter m_i is given by:[†]

$$\sigma_{m_i}^2 = \sum_{j=1}^{K}\frac{V_{ij}^2}{\lambda_j^2}. \tag{14.9}$$

[†] To distinguish data uncertainty from model uncertainty we denote the model standard deviation as σ_{m_i} and the data standard deviation as σ_i.

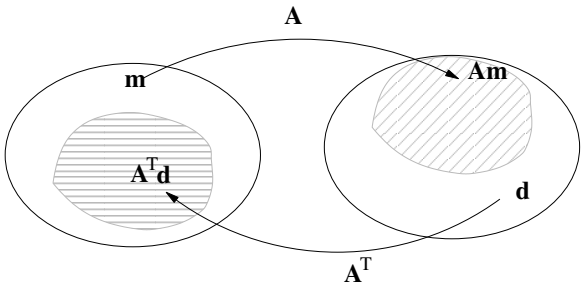

Fig. 14.4. Mappings between the model space (left) and the data space (right). The range of $\boldsymbol{A}$ is indicated by the grey area within the data space. The range of the backprojection $\boldsymbol{A}^T$ is indicated by the grey area in the model space.

This equation makes it clear that removing zero singular values is not sufficient, since the errors blow up as λ_j^{-2}, rendering the incorporation of small λ_j very dangerous. Dealing with small eigenvalues is known as *regularization* of the problem. Before we discuss this in more detail, we need to show the connection between the development given here and the theory of singular value decomposition which is more commonly found in the literature.

One way of looking at the system $\boldsymbol{Am} = \boldsymbol{d}$ is to see the components m_i as weights in a summation of the columns of $\boldsymbol{A}$ to fit the data vector $\boldsymbol{d}$. The columns make up the range of $\boldsymbol{A}$ in the data space (Figure 14.4). Similarly, the rows of $\boldsymbol{A}$ – the columns of $\boldsymbol{A}^T$ – make up the range of the backprojection $\boldsymbol{A}^T$ in the model space. The rest of the model space is the nullspace: if $\boldsymbol{m}$ is in the nullspace, $\boldsymbol{Am} = 0$. Components in the nullspace do not contribute to the data fit, but add to the norm of $\boldsymbol{m}$. We find the minimum norm solution by avoiding any components in the nullspace, in other words by selecting a model in the range of $\boldsymbol{A}^T$:

$$\hat{\boldsymbol{m}} = \boldsymbol{A}^T \boldsymbol{y}$$

and find $\boldsymbol{y}$ by solving for:

$$\boldsymbol{A}\boldsymbol{A}^T \boldsymbol{y} = \boldsymbol{d} \,.$$

The determinant of $\boldsymbol{A}\boldsymbol{A}^T$ is likely to be zero, so just as in the case of least squares we shall wish to eliminate zero eigenvalues. Let the eigenvectors of $\boldsymbol{A}\boldsymbol{A}^T$ be $\boldsymbol{u}_i$ with eigenvalues $\tilde{\lambda}_i^2$:

$$\boldsymbol{A}\boldsymbol{A}^T \boldsymbol{U} = \boldsymbol{U}\tilde{\boldsymbol{\Lambda}}^2 \,. \tag{14.10}$$

Since $\boldsymbol{A}\boldsymbol{A}^T$ is symmetric, the eigenvectors are orthogonal and we can scale them to be orthonormal, such that $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{I}$. Multiplying (14.10) on the left by

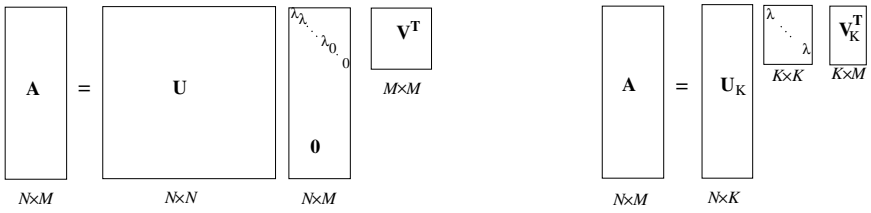

Fig. 14.5. (a) The full eigenvalue problem for $\boldsymbol{A}\boldsymbol{A}^T$ leads to a matrix with small or zero eigenvalues on the diagonal. (b) removing zero eigenvalues has no effect on $\boldsymbol{A}$.

$\boldsymbol{A}^T$ and grouping $\boldsymbol{A}^T\boldsymbol{U}$ we see that $\boldsymbol{A}^T\boldsymbol{U}$ is an eigenvector of $\boldsymbol{A}^T\boldsymbol{A}$:

$$\boldsymbol{A}^T\boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{U}) = (\boldsymbol{A}^T\boldsymbol{U})\tilde{\boldsymbol{\Lambda}}^2$$

and comparison with (14.6) shows that $\boldsymbol{A}^T\boldsymbol{u}_i$ must be a constant$\times\boldsymbol{v}_i$, and $\tilde{\lambda}_i = \lambda_i$. We choose the constant to be λ_i, so that

$$\boldsymbol{A}^T\boldsymbol{U} = \boldsymbol{V}\boldsymbol{\Lambda}. \tag{14.11}$$

Multiplying this on the left by $\boldsymbol{A}$ we obtain:

$$\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{U} = \boldsymbol{U}\boldsymbol{\Lambda}^2 = \boldsymbol{A}\boldsymbol{V}\boldsymbol{\Lambda}\,,$$

or, dividing on the right by λ_i for all $\lambda_i \neq 0$, and defining $\boldsymbol{u}_i\lambda_i = \boldsymbol{A}\boldsymbol{v}_i$ with a nullspace eigenvector $\boldsymbol{v}_i$ in case $\lambda_i = 0$:

$$\boldsymbol{A}\boldsymbol{V} = \boldsymbol{U}\boldsymbol{\Lambda}\,. \tag{14.12}$$

In the same way, by multiplying (14.12) on the right by $\boldsymbol{V}^T$ we find:

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^T \tag{14.13}$$

which is the *singular value decomposition* of $\boldsymbol{A}$. Note that in this development we have carefully avoided using the inverse of $\boldsymbol{\Lambda}$, so there is no need to truncate it to exclude zero singular values. However, because the tail of the diagonal matrix $\boldsymbol{\Lambda}$ contains only zeroes, (14.13) is equivalent to the truncated version (Figure 14.5):

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^T = \boldsymbol{U}_K\boldsymbol{\Lambda}_K\boldsymbol{V}_K^T\,. \tag{14.14}$$

Exercises

Exercise 14.3 Show that the choice (14.11) indeed implies that $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}$. Hint: use (14.12).

Exercise 14.4 Show that $\hat{\boldsymbol{m}} = \boldsymbol{V}_K\boldsymbol{\Lambda}_K^{-1}\boldsymbol{U}_K^T\boldsymbol{d}$ is equivalent to $\hat{\boldsymbol{m}}_{\text{min norm}}$.

14.4 Tikhonov regularization

The truncation to include only nonzero singular values is an example of regularization of the inverse problem. Removing zero λ_i is not sufficient however, since small singular values may give rise to large modelling errors, as shown by (14.9). This equation tells us that small errors in the data vector may cause very large excursions in model space in the direction of $\boldsymbol{v}_k$ if $\lambda_k \ll 1$. It thus seems wise to truncate $\boldsymbol{V}$ in (14.13) even further, and exclude eigenvectors belonging to small singular values. The price we pay is a small increase in χ^2, but we are rewarded by a significant reduction in the modelling error. We could apply a sharp cut-off by choosing K at some nonzero threshold level for the singular values. Less critical to the choice of threshold is a tapered cut-off. We show that the latter approach is equivalent to adding M equations of the form $\epsilon_n m_i = 0$, with ϵ_n small, to the tomographic system. Such equations act as artificial 'data' that bias the model parameters towards zero:

$$\begin{pmatrix} \boldsymbol{A} \\ \epsilon_n \boldsymbol{I} \end{pmatrix} \boldsymbol{m} = \begin{pmatrix} \boldsymbol{d} \\ \boldsymbol{0} \end{pmatrix}. \tag{14.15}$$

If the j-th column of $\boldsymbol{A}$ – associated with parameter m_j – has large elements, the addition of one additional constraint $\epsilon_n m_j = 0$ will have very little influence. But the more m_j is underdetermined by the undamped system, the more the damping will push m_j towards zero. The least squares solution of (14.15) is:

$$(\boldsymbol{A}^T \boldsymbol{A} + \epsilon_n^2 \boldsymbol{I})\boldsymbol{m} = \boldsymbol{A}^T \boldsymbol{d}\,. \tag{14.16}$$

The advantage of the formulation (14.15) is that it can easily be solved iteratively, without a need for singular value decomposition. But the solution of (14.15) does have a simple representation in terms of singular values, and it is instructive to analyse it with SVD. If $\boldsymbol{v}_k$ is an eigenvector of $\boldsymbol{A}^T \boldsymbol{A}$ with eigenvalue λ_k^2, then the damped matrix gives:

$$(\boldsymbol{A}^T \boldsymbol{A} + \epsilon_n^2 \boldsymbol{I})\boldsymbol{v}_k = (\lambda_k^2 + \epsilon_n^2)\boldsymbol{v}_k\,, \tag{14.17}$$

and we see that the damped system has the same eigenvectors but with raised eigenvalues $\lambda_k^2 + \epsilon_n^2 > 0$. The minimum norm solution (14.7) is therefore replaced by:

$$\hat{\boldsymbol{m}}_{\text{damped}} = \boldsymbol{V}_K(\boldsymbol{\Lambda}_K^2 + \epsilon_n^2 \boldsymbol{I})^{-1} \boldsymbol{V}_K^T \boldsymbol{A}^T \boldsymbol{d} \tag{14.18}$$

with the posteriori model variance given by:

$$\sigma_{m_i}^2 = \sum_{j=1}^{K} \frac{V_{ij}^2}{\lambda_j^2 + \epsilon_n^2}. \tag{14.19}$$

Since there are no zero eigenvalues, we may set $K = N$, but of course this maximizes the variance and some truncation may still be needed. For simplicity, we assumed a damping with the same ϵ_n everywhere on the diagonal. The method is often referred to as Tikhonov regularization, after its original discoverer [364]. Because one adds ϵ^2 to the diagonal of $\boldsymbol{A}^T\boldsymbol{A}$ it is also known as 'ridge regression'.

Spakman and Nolet [338] vary the damping factor ϵ_n along the diagonal. When corrections are part of the model, one should vary damping factors such that damping results in corrections that are reasonable in view of the prior uncertainty (for example, one would judge corrections as large as 100 km for hypocentral parameters usually unacceptable and increase ϵ_n for those corrections).

A comparison of (14.19) with (14.9) shows that damped model errors blow up at most by a factor ϵ_n^{-1}. Thus, damping reduces the variance of the solution. This comes at a price however: by discarding eigenvectors, we reduce our ability to shape the model. The small eigenvalues are usually associated with vectors that are strongly oscillating in space: the positive and negative parts cancel upon integration and the resulting integral (12.12) is small. Damping small eigenvalues is thus expected to lead to smoother models. However, even long-wavelength features of the model may be biased towards zero because of regularization.

The fact that biased estimations produce smaller variances is a well known phenomenon in statistical estimation, and it is easily misunderstood: one can obtain a very small model parameter m_i with a very small posteriori variance σ_i^2, yet learn nothing about the model because the bias is of the order of the true m_i. We shall come back to this in the section on resolution, but first investigate a more powerful regularization method, based on Bayesian statistics.

Exercises

Exercise 14.5 Show that the minimization of $|\boldsymbol{A}\boldsymbol{m} - \boldsymbol{d}|^2 + \epsilon^2|\boldsymbol{m}|^2$ leads to (14.16).

Exercise 14.6 In the L-curve for (14.18), indicate where $\epsilon = 0$ and where $\epsilon \to \infty$.

14.5 Bayesian inference

The simple Tikhonov regularization by norm damping we introduced in the previous section, while reducing the danger of excessive error propagation, is usually not satisfactory from a geophysical point of view. At first sight, this may seem surprising: for, when the m_i represent perturbations with respect to a background model, the damping towards 0 is defensible if we prefer the model values given by the background model in the absence of any other information. However, if the

information given by the data is unequally distributed, some parts of the model may be damped more than others, introducing an apparent structure in $\boldsymbol{m}$ that may be very misleading. The error estimate (14.19) does not represent the full modelling error because it neglects the bias. In general, we would like the model to have a minimum of unwarranted *structure*, or detail. Jackson [145] and Tarantola [349], significantly extending earlier work by Franklin [105], introduced the Bayesian method into geophysical inversion to deal with this problem, named after the Reverend Thomas Bayes (1702–1761), a British mathematician whose theorem on joint probabilities is a cornerstone of this inference method.

We shall give a brief exposé of Bayesian estimation for the case of N observations in a data vector $\boldsymbol{d}^{\text{obs}}$. Let $P(\boldsymbol{m})$ be the prior probability density for the model $\boldsymbol{m} = (m_1, m_2, ..., m_M)$, e.g. a Gaussian probability of the form:

$$P(\boldsymbol{m}) = \frac{1}{(2\pi)^{M/2}} \frac{1}{|\det \boldsymbol{C}_m|^{1/2}} \exp\left(-\frac{1}{2}\boldsymbol{m} \cdot \boldsymbol{C}_m^{-1}\boldsymbol{m}\right) . \tag{14.20}$$

Here, $\boldsymbol{C}_m$ is the *prior* covariance matrix for the model parameters. By 'prior' we mean that we generally have an idea of the allowable variations in the model values, e.g. how much the 3D Earth may differ from a 1D background model without violating more general laws of physics. We may express such knowledge as a prior probability density for the model values. The diagonal elements of $\boldsymbol{C}_m$ are the variances of that prior distribution. The off-diagonal elements reflect the correlation of model parameters – often it helps to think of them as describing the likely 'smoothness' of the model.

In a strict Bayesian philosophy such constraints may be 'subjective'. This, however, is not to say that we may impose constraints following the whim of an arbitrary person. An experienced geophysicist may often develop a very good intuition of the prior uncertainty of model parameters, perhaps because he has done experiments in the laboratory on analogue materials, or because he has experience with tomographic inversions in similar geological provinces. We shall classify such defensible subjective notions to be 'objective' after all.

The random errors in our observations make that the observed data vector $\boldsymbol{d}^{\text{obs}}$ deviates from the true (i.e. error-free) data $\boldsymbol{d}$. For the data we assume the normal distribution (14.2). Assuming the linear relationship $\boldsymbol{A}\boldsymbol{m} = \boldsymbol{d}$ has no errors (or incorporating those errors into σ_i as discussed before), we find the conditional probability density for the observed data, *given a model* $\boldsymbol{m}$:

$$P(\boldsymbol{d}|\boldsymbol{m}) = \frac{1}{(2\pi)^{N/2}} \frac{1}{|\det \boldsymbol{C}_d|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{A}\boldsymbol{m} - \boldsymbol{d}^{\text{obs}}) \cdot \boldsymbol{C}_d^{-1}(\boldsymbol{A}\boldsymbol{m} - \boldsymbol{d}^{\text{obs}})\right) , \tag{14.21}$$

where $\boldsymbol{C}_d$ is the matrix with data covariance, usually taken to be diagonal with entries σ_i^2 because we have little knowledge about data correlations.

Though we have an expression for the data probability $P(\boldsymbol{d}|\boldsymbol{m})$, for solution of the inverse problem we are more interested in the probability of the model, *given the observed data* $\boldsymbol{d}^{\text{obs}}$. This is where Bayes' theorem is useful. It starts from the recognition that the joint probability can be split up in a conditional and marginal probability in two ways, assuming the probabilities for model and data are independent:

$$P(\boldsymbol{m}, \boldsymbol{d}^{\text{obs}}) = P(\boldsymbol{m}|\boldsymbol{d}^{\text{obs}})P(\boldsymbol{d}^{\text{obs}}) = P(\boldsymbol{d}^{\text{obs}}|\boldsymbol{m})P(\boldsymbol{m}),$$

from which we find Bayes' theorem:

$$P(\boldsymbol{m}|\boldsymbol{d}^{\text{obs}}) = \frac{P(\boldsymbol{d}^{\text{obs}}|\boldsymbol{m})P(\boldsymbol{m})}{P(\boldsymbol{d}^{\text{obs}})}. \tag{14.22}$$

Using (14.20) and (14.21):

$$P(\boldsymbol{m}|\boldsymbol{d}^{\text{obs}}) \propto \exp\left[-\frac{1}{2}(\boldsymbol{A}\boldsymbol{m} - \boldsymbol{d}^{\text{obs}}) \cdot \boldsymbol{C}_d^{-1}(\boldsymbol{A}\boldsymbol{m} - \boldsymbol{d}^{\text{obs}}) - \frac{1}{2}\boldsymbol{m} \cdot \boldsymbol{C}_m^{-1}\boldsymbol{m}\right].$$

Thus, we obtain the maximum likelihood solution by minimizing:

$$(\boldsymbol{A}\boldsymbol{m} - \boldsymbol{d}^{\text{obs}}) \cdot \boldsymbol{C}_d^{-1}(\boldsymbol{A}\boldsymbol{m} - \boldsymbol{d}^{\text{obs}}) + \boldsymbol{m} \cdot \boldsymbol{C}_m^{-1}\boldsymbol{m} = \chi^2(\boldsymbol{m}) + \boldsymbol{m} \cdot \boldsymbol{C}_m^{-1}\boldsymbol{m} = \min,$$

or, differentiating with respect to m_i:

$$\boldsymbol{A}^T \boldsymbol{C}_d^{-1}(\boldsymbol{A}\boldsymbol{m} - \boldsymbol{d}^{\text{obs}}) + \boldsymbol{C}_m^{-1}\boldsymbol{m} = 0.$$

One sees that this is – again – a system of normal equations belonging to the 'damped' system:

$$\begin{pmatrix} \boldsymbol{C}_d^{-\frac{1}{2}}\boldsymbol{A} \\ \boldsymbol{C}_m^{-\frac{1}{2}} \end{pmatrix} \boldsymbol{m} = \begin{pmatrix} \boldsymbol{C}_d^{-\frac{1}{2}}\boldsymbol{d} \\ \boldsymbol{0} \end{pmatrix}. \tag{14.23}$$

Of course, if we have already scaled the data to be univariant the data covariance matrix is $\boldsymbol{C}_d = \boldsymbol{I}$. This simply shows that we are sooner or later obliged to scale the system with the data uncertainty. The prior smoothness constraint is unlikely to be a 'hard' constraint, and in practice we face again a tradeoff between the data fit and the damping of the model, much as in Figure 14.2. We obtain a manageable flexibility in the tradeoff between smoothness of the model and χ^2 by scaling $\boldsymbol{C}_d^{-\frac{1}{2}}$ with a scaling factor ϵ. Varying ϵ allows us to tweak the model damping until $\chi^2 \approx N$. Equation (14.23) is thus usually encountered in the equivalent, simplified

form:

$$\begin{pmatrix} \boldsymbol{A} \\ \epsilon \boldsymbol{C}_m^{-\frac{1}{2}} \end{pmatrix} \boldsymbol{m} = \begin{pmatrix} \boldsymbol{d} \\ \boldsymbol{0} \end{pmatrix} . \tag{14.24}$$

How should one specify $\boldsymbol{C}_m$? The model covariance essentially tells us how model parameters are correlated. Usually, such correlations are only high for nearby parameters. Thus, $\boldsymbol{C}_m$ smoothes the model when operating on $\boldsymbol{m}$. Conversely, $\boldsymbol{C}_m^{-1}$ roughens the model, and () expresses the penalization of those model elements that dominate after the roughening operation. The simplest roughening operator is the Laplacian ∇^2, which is zero when a model parameter is exactly the average of its neighbours. If we parametrize the model with tetrahedra or blocks, so that every node has well-defined nearest neighbours, we can minimize the difference between parameter m_i and the average of its neighbours (Nolet [235]):

$$\frac{1}{2} \sum_i \frac{1}{N_i} \sum_{j \in \mathcal{N}_i} (m_i - m_j)^2 = \min,$$

where $\mathcal{N}_i$ is the set of N_i nearest neighbours of mode i. Differentiating with respect to m_k gives M equations:

$$m_k - \frac{1}{N_k} \sum_{j \in \mathcal{N}_k} m_j = 0, \tag{14.25}$$

in which we recognize the k-th row of $\boldsymbol{C}_m^{-\frac{1}{2}} \boldsymbol{m}$ in (14.24).

One disadvantage of the system (14.24) is that it often converges much more slowly than the Tikhonov system (14.15) in iterative matrix solvers (VanDecar and Snieder [381]). The reason is that we are simultaneously solving a system arising from a set of integral equations, and the regularization system which involves finite-differencing. Without sacrificing the Bayesian philosophy, it is possible to transform (14.24) to a simple norm damping. Spakman and Nolet [338] introduce $\boldsymbol{m} = \boldsymbol{C}_m^{\frac{1}{2}} \boldsymbol{m}'$. Inserting this into (14.24) we find:

$$\begin{pmatrix} \boldsymbol{A} \boldsymbol{C}_m^{\frac{1}{2}} \\ \epsilon \boldsymbol{I} \end{pmatrix} \boldsymbol{m}' = \begin{pmatrix} \boldsymbol{d} \\ \boldsymbol{0} \end{pmatrix} . \tag{14.26}$$

Though it is not practical to invert the matrix $\boldsymbol{C}_m^{-\frac{1}{2}}$ that is implicit in (14.25) to find an exact expression for $\boldsymbol{C}_m^{\frac{1}{2}}$, many explicit smoothers of $\boldsymbol{m}$ may act as an appropriate 'correlation' matrix $\boldsymbol{C}_m^{\frac{1}{2}}$ for regularization purposes. After inversion for $\boldsymbol{m}'$, the tomographic model is obtained from the smoothing operation $\boldsymbol{m} = \boldsymbol{C}_m^{\frac{1}{2}} \boldsymbol{m}'$. The system (14.26) has the same form as the Tikhonov regularization (14.15).

Despite this resemblance, in my own experience the acceleration of convergence is only modest compared to inverting (14.24) directly.

14.6 Information theory

Given the lack of resolution, geophysicists are condemned to accept the fact that there are infinitely many models that all satisfy the data within the error bounds. The Earth is a laboratory, but one that is very different from those in experimental physics, where we are taught to carefully design an experiment so that we have full control. Understandably, we feel unhappy with a wide choice of regularizations, resulting in our inability to come up with a unique outcome of the experiment. The temptation is always to resort to some 'higher' – if not metaphysical – principle that allows us to choose the 'best' model among the infinite set before we start plotting tomographic cross-sections. It should be recognized that this simply replaces one subjective choice (that of a model) with another (that of a criterion). Though some tomographers religiously adhere to such metaphysical considerations, I readily confess to being an atheist. In my view, such external criteria are simply a matter of taste. As an example, the methods of regularization are related to concepts known from the field of information theory, notably to the concept of information entropy. We shall briefly look into this, but warn the reader that, in the end, there is no panacea for our fall from Paradise.

We start with a simple application of the concept of information entropy: suppose we have only one datum, a delay measured along a ray of length L. We then have a $1 \times M$ system, or just one equation:

$$d_1 = \int \boldsymbol{m}(\boldsymbol{r}) \mathrm{d}s = \sum_i m_i \mathrm{d}s,$$

As a thought experiment, assume that the segments of $\mathrm{d}s_i$ are of equal length $\mathrm{d}s$, and that we allow only one of them to cause the travel time anomaly. Which one? Information theory looks at this problem in the following way: let P_i be the probability that $m_i \neq 0$. By the law of probabilities, $\sum P_i = 1$. Intuitively, we judge that in the absence of any other information, all P_i should be equal – if not this would constitute additional information on the m_i. Formally, we may get to this conclusion by defining the information entropy:

$$I = \sum_i P_i \ln P_i, \tag{14.27}$$

which can be understood if we consider that any $P_i = 0$ will yield $I = -\infty$, thus minimizing the 'disorder' in the solution (note that if any $P_i = 1$, all others must be 0, again minimizing disorder). We express our desire to have a solution

with minimum unwarranted information as the desire to maximize I, while still satisfying $\sum P_i = 1$. Such problems are solved with the method of Lagrange multipliers. This method recognizes that the absolute maximum of I – zero for all probabilities equal to 1 – does not satisfy the constraint that $\sum P_i = 1$. So we relax the maximum condition by adding $\lambda(\sum P_i - 1)$ to I and require:

$$I + \lambda(\sum P_i - 1) = \text{Max}\,.$$

Since the added factor is required to be zero, the function to maximize has not really changed as long as we satisfy that constraint. All we have done is add another dimension, or dependent variable, the Lagrange multiplier λ. We recover the original constraint by maximizing with respect to λ. Taking the derivative with respect to P_i now gives an equation that involves λ:

$$\frac{\partial}{\partial P_i}\left(\sum_i P_i \ln P_i + \lambda \sum_i P_i\right) = 0,$$

or

$$\ln P_i = -(1+\lambda) \rightarrow P_i = \mathrm{e}^{-1-\lambda}.$$

We find the Lagrange multiplier from the constraint:

$$\sum_i P_i = N\mathrm{e}^{-1-\lambda} = 1 \rightarrow \lambda = \ln N - 1,$$

or

$$P_i = \mathrm{e}^{\ln(1/N)} = \frac{1}{N}.$$

Thus, if we impose the criterion of maximum entropy for the 'information' in our model, all m_i are equally likely to contribute. The reasoning does not change much if we allow every m_i to contribute to the anomaly and again maximize (14.27). In that case, all m_i are equally likely to contribute. In the absence of further information, there is no reason to assume that one would contribute more than any other, and all are equal: $m_i = d_1/\sum \mathrm{d}s = d_1/L$. The smoothest model is the model with the highest information entropy. Such reasoning provides a 'higher principle' to justify the damping towards smooth models.

Constable et al. [64] named the construction of the smoothest model that satisfies the data with the prescribed tolerance *Occam's inversion*, after the fourteenth century philosopher William of Occam, or Ockham, who advocated the principle that simple explanations are more likely than complicated ones and who applied what came to be known as *Occam's razor* to eliminate unnecessary presuppositions.

However, one should not assume that smooth models are free of presuppositions: in fact, if we apply (14.25) in (14.24) we arbitrarily impose that smooth structures

are more 'likely' than others. Artefacts may be suppressed, but so will sharp boundaries, e.g. the top of a subduction zone. Loris et al. [188], who invert for models that can be expanded with the fewest wavelets of a given wavelet basis, provide a variant on Occam's razor that is in principle able to preserve sharp features while eliminating unwarranted detail.

An interesting connection arises if we assume that sparse model parametrizations are a priori more probable than parametrizations with many basis functions. Assume that the prior model probability $P(\boldsymbol{m})$ is inversely proportional to the number of basis functions with nonzero coefficients in an exponential fashion:

$$P(\boldsymbol{m}) \propto \mathrm{e}^{-K},$$

where K is the number of basis functions. If we insert this into Bayes' equation, we find that the maximum likelihood equation becomes:

$$\ln \chi^2(\boldsymbol{m}) - K = \min,$$

which is Akaike's [1] criterion for the optimum selection of the number of parameters K, used in seismic tomography by Zollo et al. [422]. Note, however, that this criterion lacks a crucial element: it does not impose any restrictions on the shape of the basis functions. Presumably one could use it by ranking independently defined basis functions in order of increasing roughness, again appealing to William of Occam for his blessing.

14.7 Numerical considerations

With N often of the order of $10^5 - 10^7$ data, and M only one order of magnitude smaller than N, the matrix system $\boldsymbol{Am} = \boldsymbol{d}$ is gigantic in size. Some reduction in the number of rows N can be obtained by combining (almost) coincident raypaths into summary rays (see Section 6.1). The correct way to do this is to sum the rows of all $N_{\mathcal{S}}$ data belonging to a summary ray group $\mathcal{S}$ into one new averaged row that replaces them in the matrix:

$$\sum_{j=1}^{M} \frac{1}{N_{\mathcal{S}}} \left(\sum_{i \in \mathcal{S}} A_{ij} \right) m_j = \frac{1}{N_{\mathcal{S}}} \sum_{i \in \mathcal{S}} d_i \pm \sigma_{\mathcal{S}}\,, \tag{14.28}$$

with the variance $\sigma_{\mathcal{S}}^2$ equal to

$$\sigma_{\mathcal{S}}^2 = \frac{1}{N_{\mathcal{S}}^2} \sum_{i \in \mathcal{S}} \sigma_i^2 + \sigma_0^2\,.$$

Here, σ_0^2 is added to account for lateral variations within the summary ray that affect the variance of the sum. Gudmundsson et al. [126] analysed the relationship

between the width of a bundle and the variance in teleseismic P delay times from the ISC catalogue.

Care must be taken in defining the volume that defines the members of the summary ray. Events with a common epicentre but different depth provide important vertical resolution in the earthquake region and should often be treated separately. When using ray theory and large cells to parametrize the model we do not lose much information if we average over large volumes with size comparable to the model cells. But the Fréchet kernels of finite-frequency theory show that the sensitivity narrows down near source and receiver, and summarizing may undo some of the benefits of a finite-frequency approach.

Summary rays are sometimes applied to counteract the effect of dominant ray trajectories on the model – which may lead to strong parameter correlations along the prevailing ray direction – by ignoring the reduction of the error in the average. However, this violates statistical theory if we seek the maximum likelihood solution for normally distributed errors. The uneven distribution of sensitivity is better fought using unstructured grids with adapted resolution, and smoothness damping using a correlation matrix $\boldsymbol{C}_m$ that promotes equal parameter correlation in all directions.

If the parametrization is local, many elements of $\boldsymbol{A}$ are zero. For a least-squares solution, $\boldsymbol{A}^T\boldsymbol{A}$ has lost much of this sparseness, though, so we shall wish to avoid constructing $\boldsymbol{A}^T\boldsymbol{A}$ explicitly.[†] We can obtain a large savings in memory space by only storing the nonzero elements of $\boldsymbol{A}$. We do this row-wise – surprisingly the multiplications $\boldsymbol{Am}$ and $\boldsymbol{A}^T\boldsymbol{d}$ can both be done in row-order, using the following 'row-action' algorithms:

$\boldsymbol{p} = \boldsymbol{Am}$:
for $i = 1, N$
  for $j = 1, M$
    $p_i \leftarrow p_i + A_{ij} m_j$

$\boldsymbol{q} = \boldsymbol{A}^T\boldsymbol{d}$:
for $i = 1, N$
  for $j = 1, M$
    $q_j \leftarrow q_j + A_{ij} d_i$

where only nonzero elements of A_{ij} should take part. This often leads to complicated bookkeeping. Claerbout's [60] dot-product test: $\boldsymbol{q} \cdot \boldsymbol{Ap} = \boldsymbol{A}^T\boldsymbol{q} \cdot \boldsymbol{p}$ – for random vectors $\boldsymbol{p}$ and $\boldsymbol{q}$ – can be used as a first (though not conclusive) test to validate the coding.

Early tomographic efforts in the medical and biological sciences led to a rediscovery of row-action methods (Censor [44]). The early methods, however, had the disadvantage that they introduced an unwanted scaling into the problem that

[†] The explicit computation and use of $\boldsymbol{A}^T\boldsymbol{A}$ is also unwise from the point of view of numerical stability since its condition number – the measure of the sensitivity of the solution to data errors – is the square of that of $\boldsymbol{A}$ itself. For a discussion of this issue see *Numerical Recipes* [269].

interferes with the optimal regularization one wishes to impose (see van der Sluis and van der Vorst [377] for a detailed analysis).

Conjugate gradient methods work without implicit scaling. The stablest algorithm known today is LSQR, developed by Paige and Saunders [249] and introduced into seismic tomography by the author [233, 234]. We give a short derivation of LSQR. The main idea of the algorithm is to develop orthonormal bases $\boldsymbol{\mu}_k$ in model space, and $\boldsymbol{\rho}_k$ in data space. The first basis vector in data space, $\boldsymbol{\rho}_1$, is simply in the direction of the data vector: $\beta_1\boldsymbol{\rho}_1 = \boldsymbol{d}$, and $\boldsymbol{\mu}_1$ is the backprojection of $\boldsymbol{\rho}_1$: $\alpha_1\boldsymbol{\mu}_1 = \boldsymbol{A}^T\boldsymbol{\rho}_1$. Coefficients α_i and β_i are normalization factors such that $|\boldsymbol{\rho}_i| = |\boldsymbol{\mu}_i| = 1$. We find the second basis vector in data space by mapping $\boldsymbol{\mu}_1$ into data space, and orthogonalize to $\boldsymbol{\rho}_1$:

$$\beta_2\boldsymbol{\rho}_2 = \boldsymbol{A}\boldsymbol{\mu}_1 - (\boldsymbol{A}\boldsymbol{\mu}_1 \cdot \boldsymbol{\rho}_1)\boldsymbol{\rho}_1 = \boldsymbol{A}\boldsymbol{\mu}_1 - \alpha_1\boldsymbol{\rho}_1,$$

where we use $\boldsymbol{A}\boldsymbol{\mu}_1 \cdot \boldsymbol{\rho}_1 = \boldsymbol{\mu}_1 \cdot \boldsymbol{A}^T\boldsymbol{\rho}_1 = \boldsymbol{\mu}_1 \cdot \alpha_1\boldsymbol{\mu}_1$. Similarly:

$$\alpha_2\boldsymbol{\mu}_2 = \boldsymbol{A}^T\boldsymbol{\rho}_2 - \beta_2\boldsymbol{\mu}_1.$$

Although it would seem that we have to go through more and lengthier orthogonalizations as the basis grows, it turns out that – at least in theory, ignoring roundoff errors – the orthogonalization to the previous basis function only is sufficient. For example, for $\boldsymbol{\rho}_3$ we find $\beta_3\boldsymbol{\rho}_3 = \boldsymbol{A}\boldsymbol{\mu}_2 - \alpha_2\boldsymbol{\rho}_2$. Taking the dot product with $\boldsymbol{\rho}_1$, we find:

$$\beta_3\boldsymbol{\rho}_3 \cdot \boldsymbol{\rho}_1 = \boldsymbol{A}\boldsymbol{\mu}_2 \cdot \boldsymbol{\rho}_1 - \alpha_2\boldsymbol{\rho}_2 \cdot \boldsymbol{\rho}_1 = \boldsymbol{\mu}_2 \cdot \boldsymbol{A}^T\boldsymbol{\rho}_1 = \boldsymbol{\mu}_2 \cdot (\alpha_1\boldsymbol{\mu}_1) = 0\,,$$

and $\boldsymbol{\rho}_3$ is perpendicular to $\boldsymbol{\rho}_1$. A similar proof by induction can be made for all $\boldsymbol{\rho}_k$ and $\boldsymbol{\mu}_k$ in the iterative sequence:

$$\beta_{k+1}\boldsymbol{\rho}_{k+1} = \boldsymbol{A}\boldsymbol{\mu}_k - \alpha_k\boldsymbol{\rho}_k \tag{14.29}$$

$$\alpha_{k+1}\boldsymbol{\mu}_{k+1} = \boldsymbol{A}^T\boldsymbol{\rho}_{k+1} - \beta_{k+1}\boldsymbol{\mu}_k. \tag{14.30}$$

If we expand the solution after k iterations:

$$\boldsymbol{m}_k = \sum_{j=1}^{k} \gamma_j\boldsymbol{\mu}_j,$$

$$\sum_{j=1}^{k} \gamma_j\boldsymbol{A}\boldsymbol{\mu}_j = \boldsymbol{d},$$

and with (14.29):

$$\sum_{j=1}^{k} \gamma_j(\beta_{j+1}\boldsymbol{\rho}_{j+1} + \alpha_j\boldsymbol{\rho}_j) = \beta_1\boldsymbol{\rho}_1.$$

Taking the dot product of this with $\boldsymbol{\rho}_1$ yields $\gamma_1 = \beta_1/\alpha_1$, whereas subsequent factors are found by taking the product with $\boldsymbol{\rho}_k$ to give $\gamma_k = -\beta_k\gamma_{k-1}/\alpha_k$.

14.8 Appendix D: Some concepts of probability theory and statistics

I assume the reader is familiar with discrete probabilities, such as the probability that a flipped coin will come up with head or tail. If added up for all possible outcomes, the sum of all probabilities is 1.

This concept of probability cannot directly be applied to variables that can take any value within prescribed bounds. For such variables we use probability *density*. The probability density $P(X_0)$ for a random variable X at X_0 is equal to the probability that X is within the interval $X_0 \leq X \leq X_0 + \mathrm{d}X$, divided by $\mathrm{d}X$.

This can be extended to multiple variables. If $P(\boldsymbol{d})$ is the probability density for the data in vector $\boldsymbol{d}$, then the probability that we find the data within a small N-dimensional volume $\Delta\boldsymbol{d}$ in data space is given by $0 \leq P(\boldsymbol{d})\Delta\boldsymbol{d} \leq 1$. We only deal with normalized probability densities, i.e. the integral over all data:

$$\int P(\boldsymbol{d})\mathrm{d}^N\boldsymbol{d} = 1\,. \tag{14.31}$$

Joint probability densities give the probability that two or more random variables take a particular value, e.g. $P(\boldsymbol{m}, \boldsymbol{d})$. If the distributions for the two variables are independent, the *joint* probability density is the product of the individual densities:

$$P(\boldsymbol{m}, \boldsymbol{d}) = P(\boldsymbol{m})P(\boldsymbol{d}). \tag{14.32}$$

Conversely, one finds the *marginal* probability density of one of the variables by integrating out the second variable:

$$P(\boldsymbol{m}) = \int P(\boldsymbol{m}, \boldsymbol{d})\mathrm{d}^N\boldsymbol{d}. \tag{14.33}$$

The *conditional* probability density gives the probability of the first variable under the condition that the second variable has a given value, e.g. $P(\boldsymbol{m}|\boldsymbol{d}^{\mathrm{obs}})$ gives the probability density for model $\boldsymbol{m}$ given an observed set of data in $\boldsymbol{d}^{\mathrm{obs}}$.

The *expectation* or expected value $E(X)$ of X is defined as the average over all values of X weighted by the probability density:

$$\bar{X} \equiv E(X) = \int P(X)X\,\mathrm{d}X. \tag{14.34}$$

The expectation is a linear functional:

$$E(aX + bY) = aE(X) + bE(Y)\,, \tag{14.35}$$

and for independent variables it is separable:

$$E(XY) = E(X)E(Y)\,. \tag{14.36}$$

The *variance* is a measure of the spread of X around its expected value:

$$\sigma_X^2 = E[(X - \bar{X})^2]\,, \tag{14.37}$$

where σ_X itself is known as the standard deviation. The covariance between two random variables X and Y is defined as

$$\mathrm{Cov}(X, Y) = E[(X - \bar{X})(Y - \bar{Y})]\,. \tag{14.38}$$

In the case of an N-tuple of variables this defines an $N \times N$ covariance matrix, with the variance on the diagonal. The covariance matrix of a linear combination of variables is found by applying the linearity (14.35). Consider a linear transformation $\boldsymbol{x} = \boldsymbol{T}\boldsymbol{y}$. Since the spread of a variable does not change if we redefine the average as zero, we can assume that $E(x_i) = 0$ without loss of generality. Then:

$$\begin{aligned}\mathrm{Cov}(x_i, x_j) &= E\left(\sum_k T_{ij} y_k \sum_l T_{jl} y_l\right) = \sum_{kl} T_{ij} T_{jl} E(y_k y_l) \\ &= \sum_{kl} T_{ij} T_{jl} \mathrm{Cov}(y_k, y_l),\end{aligned}$$

or, in matrix notation:

$$\boldsymbol{C}_x = \boldsymbol{T}\boldsymbol{C}_y\boldsymbol{T}^T\,. \tag{14.39}$$