3

Extensions of methods

In this chapter we extend and apply some of the methods developed in Chapter 2. The problems discussed there raise a number of issues concerning models and data that are difficult to address with the mathematical machinery already available. Among them are the introduction of left and right eigenvectors, model nonlinearity, the potential use of inequality constraints, and sampling adequacy.

3.1 The general eigenvector/eigenvalue problem

To understand some recent work on so-called pseudospectra and some surprising recent results on fluid instability, it helps to review the more general eigenvector/eigenvalue problem for arbitrary, *square,* matrices. Consider

$$\mathbf{E}\mathbf{g}_i = \lambda_i \mathbf{g}_i, \quad i = 1, 2, \ldots, N. \tag{3.1}$$

If there are no repeated eigenvalues, λ_i, then it is possible to show that there are always N independent $\mathbf{g}_i$, which are a basis, but which are *not usually orthogonal.* Because in most problems dealt with in this book, small perturbations can be made to the elements of $\mathbf{E}$ without creating any physical damage, it suffices here to assume that such perturbations can always ensure that there are N distinct λ_i. (Failure of the hypothesis leads to the Jordan form, which requires a somewhat tedious discussion.) A matrix $\mathbf{E}$ is "normal" if its eigenvectors form an orthonormal basis. Otherwise it is "non-normal." (Any matrix of forms $\mathbf{A}\mathbf{A}^{\mathrm{T}}$, $\mathbf{A}^{\mathrm{T}}\mathbf{A}$, is necessarily normal.)

Denote $\mathbf{G} = \{g_i\}$, $\mathbf{\Lambda} = \mathrm{diag}(\lambda_i)$. It follows immediately that $\mathbf{E}$ can be diagonalized:

$$\mathbf{G}^{-1}\mathbf{E}\mathbf{G} = \mathbf{\Lambda}, \tag{3.2}$$

but for a non-normal matrix, $\mathbf{G}^{-1} \neq \mathbf{G}^{\mathrm{T}}$. The general decomposition is

$$\mathbf{E} = \mathbf{G}\mathbf{\Lambda}\mathbf{G}^{-1}.$$

Matrix $\mathbf{E}^{\mathrm{T}}$ has a different set of spanning eigenvectors, but the same eigenvalues:

$$\mathbf{E}^{\mathrm{T}}\mathbf{f}_j = \lambda_j \mathbf{f}_j, \quad j = 1, 2, \ldots, N, \tag{3.3}$$

which can be written

$$\mathbf{f}_j^{\mathrm{T}}\mathbf{E} = \lambda_j \mathbf{f}_j^{\mathrm{T}}. \tag{3.4}$$

The $\mathbf{g}_i$ are hence known as the "right eigenvectors," and the $\mathbf{f}_i$ as the "left eigenvectors." Multiplying (3.1) on the left by $\mathbf{f}_j^{\mathrm{T}}$ and (3.4) on the right by $\mathbf{g}_i$ and subtracting shows

$$0 = (\lambda_i - \lambda_j)\mathbf{f}_j^{\mathrm{T}}\mathbf{g}_i, \tag{3.5}$$

or

$$\mathbf{f}_j^{\mathrm{T}}\mathbf{g}_i = 0, \quad i \neq j. \tag{3.6}$$

That is to say, the left and right eigenvectors are orthogonal for different eigenvalues, but $\mathbf{f}_j^{\mathrm{T}}\mathbf{g}_j \neq 0$. (In general the eigenvectors and eigenvalues are complex even for purely real $\mathbf{E}$. Note that some software automatically conjugates a transposed complex vector or matrix, and the derivation of (3.6) shows that it applies to the *non-conjugated* variables.)

Consider now a "model,"

$$\mathbf{A}\mathbf{x} = \mathbf{b}. \tag{3.7}$$

The norm of $\mathbf{b}$ is supposed bounded, $\|\mathbf{b}\| \leq b$, and the norm of $\mathbf{x}$ will be

$$\|\mathbf{x}\| = \|\mathbf{A}^{-1}\mathbf{b}\|. \tag{3.8}$$

What is the relationship of $\|\mathbf{x}\|$ to $\|\mathbf{b}\|$?

Let $\mathbf{g}_i$ be the right eigenvectors of $\mathbf{A}$. Write

$$\mathbf{b} = \sum_{i=1}^{N} \beta_i \mathbf{g}_i, \tag{3.9}$$

$$\mathbf{x} = \sum_{i=1}^{N} \alpha_i \mathbf{g}_i. \tag{3.10}$$

If the $\mathbf{g}_i$ were orthogonal, $|\beta_i| \leq \|\mathbf{b}\|$. But as they are not orthonormal, the β_i will need to be found through a system of simultaneous equations (2.3) (recall the discussion in Chapter 2 of the expansion of an arbitrary vector in non-orthogonal vectors) and no simple bound on the β_i is then possible; some may be very large. Substituting into (3.7),

$$\sum_{i=1}^{N} \alpha_i \lambda_i \mathbf{g}_i = \sum_{i=1}^{N} \beta_i \mathbf{g}_i. \tag{3.11}$$

A term-by-term solution is evidently no longer possible because of the lack of orthogonality. But multiplying on the left by $\mathbf{f}_j^{\mathrm{T}}$, and invoking (3.6), produces

$$\alpha_j \lambda_j \mathbf{f}_j^{\mathrm{T}} \mathbf{g}_j = \beta_j \mathbf{f}_j^{\mathrm{T}} \mathbf{g}_j, \tag{3.12}$$

or

$$\alpha_j = \beta_j / \lambda_j, \quad \lambda_j \neq 0. \tag{3.13}$$

Even if the λ_j are all of the same order, the possibility that some of the β_j are very large implies that eigenstructures in the solution may be much larger than $\|\mathbf{b}\|$. This possibility becomes very interesting when we turn to time-dependent systems. At the moment, note that partial differential equations that are self-adjoint produce discretizations that have coefficient matrices $\mathbf{A}$, such that $\mathbf{A}^{\mathrm{T}} = \mathbf{A}$. Thus self-adjoint systems have normal matrices, and the eigencomponents of the solution are all immediately bounded by $\|\mathbf{b}\| / \lambda_i$. Non-self-adjoint systems produce non-normal coefficient matrices and so can therefore unexpectedly generate very large eigenvector contributions.

Example *The equations*

$$\begin{Bmatrix} -0.32685 & 0.34133 & 0.69969 & 0.56619 \\ -4.0590 & 0.80114 & 3.0219 & 1.3683 \\ -3.3601 & 0.36789 & 2.6619 & 1.2135 \\ -5.8710 & 1.0981 & 3.9281 & 1.3676 \end{Bmatrix} \mathbf{x} = \begin{bmatrix} 0.29441 \\ -1.3362 \\ 0.71432 \\ 1.6236 \end{bmatrix},$$

are readily solved by ordinary Gaussian elimination (or matrix inversion). If one attempts to use an eigenvector expansion, it is found from Eq. (3.1) (up to rounding errors) that $\lambda_i = [2.3171, 2.2171, -0.1888, 0.1583]^{\mathrm{T}}$, *and the right eigenvectors,* $\mathbf{g}_i$, *corresponding to the eigenvalues are*

$$\mathbf{G} = \begin{Bmatrix} 0.32272 & 0.33385 & 0.20263 & 0.36466 \\ 0.58335 & 0.58478 & 0.27357 & -0.23032 \\ 0.46086 & 0.46057 & 0.53322 & 0.75938 \\ 0.58581 & 0.57832 & -0.77446 & -0.48715 \end{Bmatrix},$$

and $\mathbf{g}_{1,2}$ *are nearly parallel. If one expands the right-hand side,* $\mathbf{y}$, *of the above equations in the* $\mathbf{g}_i$, *the coefficients,* $\boldsymbol{\beta} = [25.2230, -24.7147, -3.3401, 2.9680]^{\mathrm{T}}$, *and the first two are markedly greater than any of the elements in* $\mathbf{y}$ *or in the* $\mathbf{g}_i$. *Note that, in general, such arbitrary matrices will have complex eigenvalues and eigenvectors (in conjugate pairs).*

3.2 Sampling

In Chapter 2, on p. 129, we discussed the problem of making a uniformly gridded map from irregularly spaced observations. But not just any set of observations proves adequate to the purpose. The most fundamental problem generally arises under the topic of "sampling" and "sampling error." This subject is a large and interesting one in its own right,[1] and we can only outline the basic ideas.

The simplest and most fundamental idea derives from consideration of a one-dimensional continuous function, $f(q)$, where q is an arbitrary independent variable, usually either time or space, and $f(q)$ is supposed to be sampled uniformly at intervals, Δq, an infinite number of times to produce the infinite set of sample values $\{f(n\Delta q)\}$, $-\infty \leq n \leq \infty$, n integer. The sampling theorem, or sometimes the "Shannon–Whittaker sampling theorem"[2] is a statement of the conditions under which $f(q)$ should be reconstructible from the sample values. Let the Fourier transform of $f(q)$ be defined as

$$\hat{f}(r) = \int_{-\infty}^{\infty} f(q)\mathrm{e}^{2i\pi rq}\mathrm{d}q, \tag{3.14}$$

and assumed to exist. The sampling theorem asserts that a necessary and sufficient condition to perfectly reconstruct $f(q)$ from its samples is that

$$|\hat{f}(r)| = 0, \quad |r| \geq 1/(2\Delta q). \tag{3.15}$$

It produces the Shannon–Whittaker formula for the reconstruction

$$f(q) = \sum_{n=-\infty}^{\infty} f(n\Delta q)\frac{\sin[(2\pi/2\Delta q)(q - n\Delta q)]}{(2\pi/2\Delta q)(q - n\Delta q)}. \tag{3.16}$$

Mathematically, the Shannon–Whittaker result is surprising – because it provides a condition under which a function at an uncountable infinity of points – the continuous line – can be perfectly reconstructed from information known only at a countable infinity, $n\Delta q$, of them. For present purposes, an intuitive interpretation is all we seek and this is perhaps best done by considering a special case in which the conditions of the theorem are violated.

Figure 3.1 displays an ordinary sinusoid whose Fourier transform can be represented as

$$\hat{f}(r) = \tfrac{1}{2}(\delta(r - r_0) - \delta(r + r_0)), \tag{3.17}$$

which is sampled as depicted, and in violation of the sampling theorem (δ is the Dirac delta-function). It is quite clear that there is at least one more perfect sinusoid, the one depicted with the dashed line, which is completely consistent with all the sample points and which cannot be distinguished from it using the measurements

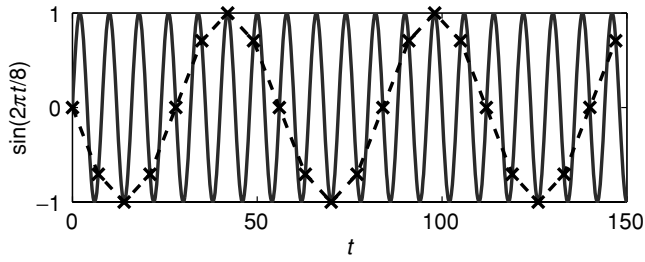

Figure 3.1 Effects of undersampling a periodic function: solid curve is $y(t) = \sin(2\pi t/8)$ sampled at time intervals of $\Delta t = 0.1$. The dashed curve is the same function, but sampled at intervals $\Delta t = 7$. With this undersampling, the curve of frequency $s = 1/8$ time units is aliased into one that appears to have a frequency $s_a = 1/8 - 1/7 = 1/56 < 1/14$. That is, the aliased curve appears to have a period of 56 time units.

alone. A little thought shows that the apparent frequency of this new sinusoid is

$$r_a = r_0 \pm \frac{n}{\Delta q}, \tag{3.18}$$

such that

$$|r_a| \leq \frac{1}{2\Delta q}. \tag{3.19}$$

The samples cannot distinguish the true high frequency sinusoid from this low frequency one, and the high frequency can be said to masquerade or "alias" as the lower frequency one.[3] The Fourier transform of a sampled function is easily seen to be periodic with period $1/\Delta q$ in the transform domain, that is, in the r space.[4] Because of this periodicity, there is no point in computing its values for frequencies outside $|r| \leq 1/2\Delta q$ (we make the convention that this "baseband," i.e., the fundamental interval for computation, is symmetric about $r = 0$, over a distance $1/2\Delta q$; see Fig. 3.2). Frequencies of absolute value larger than $1/2\Delta q$, the so-called Nyquist frequency, cannot be distinguished from those in the baseband, and alias into it. Figure 3.2 shows a densely sampled, non-periodic function and its Fourier transform compared to that obtained from the undersampled version overlain. Undersampling is a very unforgiving practice.

The consequences of aliasing range from the negligible to the disastrous. A simple example is that of the principal lunar tide, usually labeled M_2, with a period of 12.42 hours, $r = 1.932$ cycles/day. An observer measures the height of sea level at a fixed time, say 10 a.m. each day so that $\Delta q = 1$ day. Applying the formula (3.18), the apparent frequency of the tide will be 0.0676 cycles/day for a period of about 14.8 days ($n = 2$). To the extent that the observer understands what is going on, he or she will not conclude that the principal lunar tide has a period of

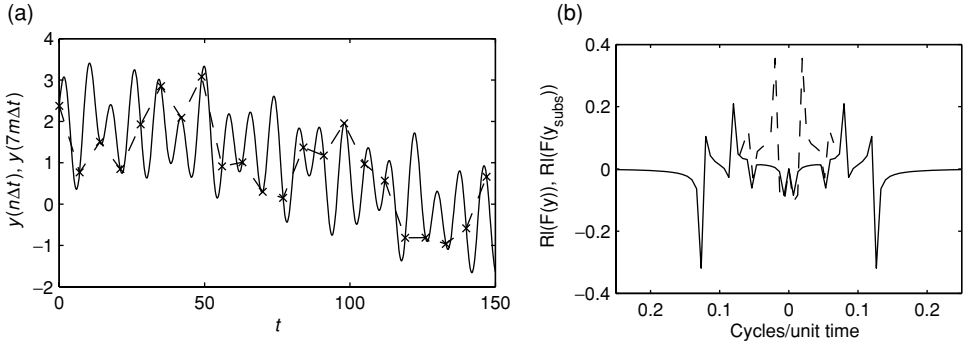

Figure 3.2 (a) A non-periodic function sampled at intervals $\Delta t = 0.1$, and the same function sampled at intervals $\Delta t = 7$ time units. (b) The real part of the Fourier components of the two functions shown in (a). The subsampled function has a Fourier transform confined to $|s| \leq 1/(2.7)$ (dashed) while that of the original, more densely sampled, function (solid) extends to $|s| \leq 1/0.1 = 10$, most of which is not displayed. The subsampled function has a very different Fourier transform from that of the original densely sampled one. Both transforms are periodic in frequency, s, with period equal to the width of their corresponding basebands. (This periodicity is suppressed in the plot.) Note in particular how erroneous an estimate of the temporal derivative of the undersampled function would be in comparison to that of the highly sampled one.

14.8 days, but will realize that the true period can be computed through (3.18) from the apparent one. But without that understanding, some bizarre theory might be produced.[5]

The reader should object that the Shannon–Whittaker theorem applies only to an infinite number of perfect samples and that one never has either perfect samples or an infinite number of them. In particular, it is true that if the duration of the data in the q domain is finite, then it is impossible for the Fourier transform to vanish over any finite interval, much less the infinite interval above the Nyquist frequency.[6] Nonetheless, the rule of thumb that results from (3.16) has been found to be quite a good one. The deviations from the assumptions of the theorem are usually dealt with by asserting that sampling should be done so that

$$\Delta q \ll 1/2r_0. \tag{3.20}$$

Many extensions and variations of the sampling theorem exist – taking account of the finite time duration , the use of "burst-sampling" and known function derivatives, etc.[7] Most of these variations are sensitive to noise. There are also extensions to multiple dimensions,[8] which are required for mapmaking purposes. Because failure to acknowledge the possibility that a signal is undersampled is so dire, one concludes that consideration of sampling is critical to any discussion of field data.

3.2.1 One-dimensional interpolation

Let there be two observations $[y_1, y_2]^{\mathrm{T}} = [x_1 + n_1, x_2 + n_2]^{\mathrm{T}}$ located at positions $[r_1, r_2]^{\mathrm{T}}$, where n_i are the observation noise. We require an estimate of $x(\tilde{r})$, where $r_1 < \tilde{r} < r_2$. The formula (3.16) is unusable – there are only two noisy observations, not an infinite number of perfect ones. We could try using linear interpolation:

$$\tilde{x}(\tilde{r}) = \frac{|r_2 - \tilde{r}|}{|r_2 - r_1|}\, y(r_1) + \frac{|r_1 - \tilde{r}|}{|r_2 - r_1|}\, y(r_2). \tag{3.21}$$

If there are N data points, r_i, $1 = 1, 2, \ldots, N$, then another possibility is Aitken–Lagrange interpolation:[9]

$$\tilde{x}(\tilde{r}) = \sum_{j=1}^{N} l_j(\tilde{r}) y_j, \tag{3.22}$$

$$l_j(\tilde{r}) = \frac{(\tilde{r} - r_1)\cdots(\tilde{r} - r_M)}{(r_j - r_1)\cdots(r_j - r_{j-1})(r_j - r_{j+1})\cdots(r_j - r_M)}. \tag{3.23}$$

Equations (3.21)– (3.23) are only two of many possible interpolation formulas. When would one be better than the other? How good are the estimates? To answer these questions, let us take a different tack, and employ the Gauss–Markov theorem, assuming we know something about the necessary covariances.

Suppose either $\langle x \rangle = \langle n \rangle = 0$ or that a known value has been removed from both (this just keeps our notation a bit simpler). Then

$$\mathbf{R}_{xy}(\tilde{r},\, r_j) \equiv \langle x(\tilde{r}) y(r_j) \rangle = \langle x(\tilde{r})(x(r_j) + n(r_j)) \rangle = \mathbf{R}_{xx}(\tilde{r},\, r_j), \tag{3.24}$$

$$\mathbf{R}_{yy}(r_i,\, r_j) \equiv \langle (x(r_i) + n(r_i))(x(r_j) + n(r_j)) \rangle \tag{3.25}$$

$$= \mathbf{R}_{xx}(r_i,\, r_j) + \mathbf{R}_{nn}(r_i,\, r_j), \tag{3.26}$$

where it has been assumed that $\langle x(r) n(q) \rangle = 0$.

From (2.396), the best linear interpolator is

$$\tilde{\mathbf{x}} = \mathbf{B}\mathbf{y}, \qquad \mathbf{B}(\tilde{r},\, \mathbf{r}) = \sum_{j=1}^{M} \mathbf{R}_{xx}(\tilde{r},\, r_j) \{\mathbf{R}_{xx} + \mathbf{R}_{nn}\}^{-1}_{ji}, \tag{3.27}$$

($\{\mathbf{R}_{xx} + \mathbf{R}_{nn}\}^{-1}_{ji}$ means the ji element of the inverse matrix) and the minimum possible error that results is

$$\mathbf{P}(\tilde{r}_\alpha,\, \tilde{r}_\beta) = \mathbf{R}_{xx}(\tilde{r}_\alpha,\, \tilde{r}_\beta) - \sum_{j}^{M} \sum_{i}^{M} \mathbf{R}_{xx}(\tilde{r}_\alpha,\, r_j) \{\mathbf{R}_{xx} + \mathbf{R}_{nn}\}^{-1}_{ji} \mathbf{R}_{xx}(r_i,\, \tilde{r}_\beta), \tag{3.28}$$

and $\tilde{\mathbf{n}} = \mathbf{y} - \tilde{\mathbf{x}}$.

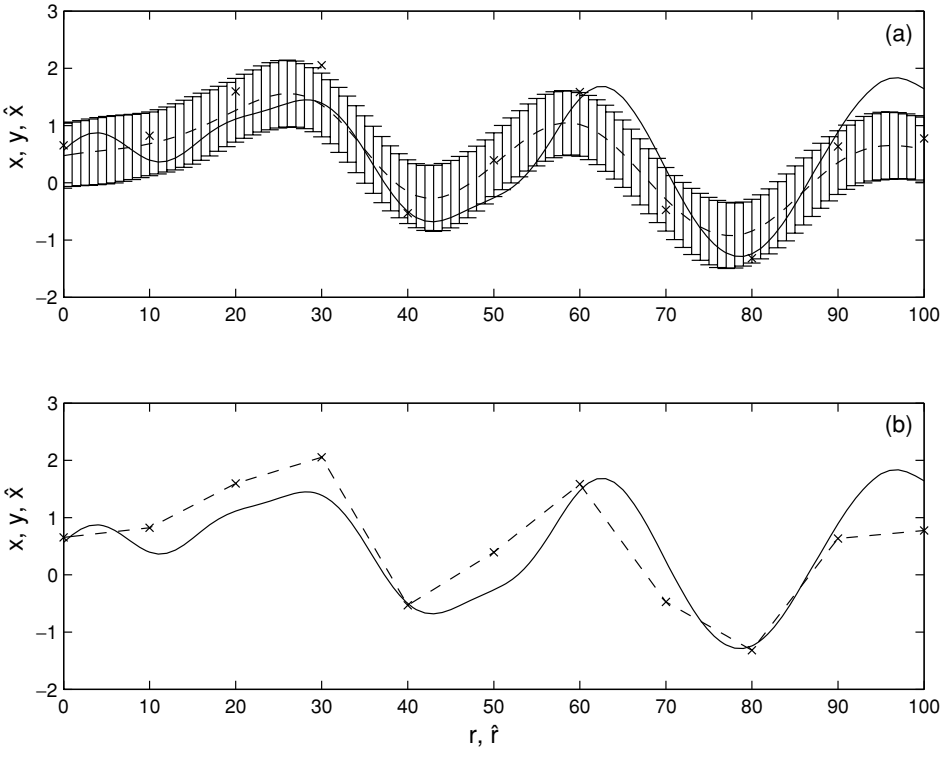

Figure 3.3 In both panels, the solid curve is the "true" curve, $x(r)$, from which noisy samples (denoted "x") have been obtained. $x\,(r)$ was generated to have a true covariance $S = \exp(-r^2/100)$, and the "data" values, $y\,(r_i) = x\,(r_i) + n_i$, where $\langle n_i \rangle = 0$, $\langle n_i n_j \rangle = (1/4)\,\delta_{ij}$, were generated from a Gaussian probability density. In (b), linear interpolation is used to generate the estimated values of $x\,(r)$ (dashed line). The estimates are identical to the observations at $r = r_i$. In (a), objective mapping was used to make the estimates (dashed line). Note that $\tilde{x}(r_i) \neq y(r_i)$, and that an error bar is available – as plotted. The true values are generally within one standard deviation of the estimated value (but about 35% of the estimated values would be expected to lie outside the error bars), and the estimated value is within two standard deviations of the correct one everywhere. The errors in the estimates, $\tilde{x}(r_i) - x(r_i)$, are clearly spatially correlated, and can be inferred from Eq. (3.28) (not shown). The values of $x(r)$ were generated to have the inferred covariance S, by forming the matrix, $\mathbf{S} =$ toeplitz $(S(r_i, r_j))$, and obtaining its symmetric factorization, $\mathbf{S} = \mathbf{U}\Lambda\mathbf{U}^{\mathrm{T}}$, $\mathbf{x}\,(r) = \mathbf{U}\Lambda\boldsymbol{\alpha}$, where the elements of $\boldsymbol{\alpha}$ are pseudo-random numbers.

Results for both linear interpolation and objective mapping are shown in Fig. 3.3. Notice that, like other interpolation methods, the optimal one is a linear combination of the data. If any other set of weights $\mathbf{B}$ is chosen, then the interpolation is not as good as it could be in the mean-square error sense; the error of any such scheme

can be obtained by substituting it into (2.394) and evaluating the result (the true covariances still need to be known).

Looking back now at the two familiar formulas in (3.21) and (3.22), it is clear what is happening: they represent a choice of $\mathbf{B}$. Unless the covariance is such as to produce one of the two sets of weights as the optimum choice, neither Aitken–Lagrange nor linear (nor any other common choice, like a spline) is the best one could do. Alternatively, if any of (3.21)–(3.23) were thought to be the best one, they are equivalent to specifying the solution and noise covariances.

If interpolation is done for two points, $\tilde{r}_\alpha, \tilde{r}_\beta$, the error of the two estimates will usually be correlated, and represented by $\mathbf{P}(\tilde{r}_\alpha, \tilde{r}_\beta)$. Knowledge of the correlations between the errors in different interpolated points is often essential – for example, if one wishes to interpolate to uniformly spaced grid points so as to make estimates of derivatives of x. Such derivatives might be numerically meaningless if the mapping errors are small scale (relative to the grid spacing) and of large amplitude. But if the mapping errors are large scale compared to the grid, the derivatives may tend to remove the error and produce better estimates than for x itself.

Both linear and Aitken–Lagrange weights will produce estimates that are exactly equal to the observed values if $\tilde{r}_\alpha = r_p$, that is, on the data points. Such a result is characteristic of "true interpolation." If no noise is present, then the observed value is the correct one to use at a data point. In contrast, the Gauss–Markov estimate will differ from the data values at the data points, because the estimator attempts to reduce the noise in the data by averaging over all observations, not just the one. The Gauss–Markov estimate is thus not a true interpolator; it is instead a "smoother." One can recover true interpolation if $\|\mathbf{R}_{nn}\| \to 0$, although the matrix being inverted in (3.27) and (3.28) can become singular. The weights $\mathbf{B}$ can be fairly complicated if there is any structure in either of $\mathbf{R}_{xx}, \mathbf{R}_{nn}$. The estimator takes explicit account of the expected spatial structure of both $\mathbf{x}, \mathbf{n}$ to weight the data in such a way as to most effectively "kill" the noise relative to the signal. One is guaranteed that no other linear filter can do better.

If $\|\mathbf{R}_{nn}\| \gg \|\mathbf{R}_{xx}\|$, $\tilde{\mathbf{x}} \to \mathbf{0}$, manifesting the bias in the estimator; this bias was deliberately introduced so as to minimize the uncertainty (minimum variance about the true value). Thus, estimated values of zero-mean processes tend toward zero, particularly far from the data points. For this reason, it is common to use expressions such as (2.413) to first remove the mean, prior to mapping the residual, and re-adding the estimated mean at the end. The interpolated values of the residuals are nearly unbiassed, because their true mean is nearly zero. Rigorous estimates of $\mathbf{P}$ for this approach require some care, as the mapped residuals contain variances owing to the uncertainty of the estimated mean,[10] but the corrections are commonly ignored.

As we have seen, the addition of small positive numbers to the diagonal of a matrix usually renders it non-singular. In the formally noise-free case, $\mathbf{R}_{nn} \to \mathbf{0}$,

and one has the prospect that $\mathbf{R}_{xx}$ by itself may be singular. To understand the meaning of this situation, consider the general case, involving both matrices. Then the symmetric form of the SVD of the sum of the two matrices is

$$\mathbf{R}_{xx} + \mathbf{R}_{nn} = \mathbf{U}\Lambda\mathbf{U}^{\mathrm{T}}. \tag{3.29}$$

If the sum covariance is positive definite, Λ will be square with $K = M$ and the inverse will exist. If the sum is not positive definite, but is only semi-definite, one or more of the singular values will vanish. The meaning is that there are *possible* structures in the data that have been assigned to neither the noise field nor the solution field. This situation is realistic only if one is truly confident that $\mathbf{y}$ does not contain such structures. In that case, the solution

$$\tilde{\mathbf{x}} = \mathbf{R}_{xx}(\mathbf{R}_{xx} + \mathbf{R}_{nn})^{-1}\mathbf{y} = \mathbf{R}_{xx}(\mathbf{U}\Lambda^{-1}\mathbf{U}^{\mathrm{T}})\mathbf{y} \tag{3.30}$$

will have components of the form $0/0$, the denominator corresponding to the zero singular values and the numerator to the absent, impossible, structures of $\mathbf{y}$. One can arrange that the ratio of these terms should be set to zero (e.g., by using the SVD). But such a delicate balance is not necessary. If one simply adds a small white noise covariance to $\mathbf{R}_{xx} + \mathbf{R}_{nn} \to \mathbf{R}_{xx} + \mathbf{R}_{nn} + \epsilon^2\mathbf{I}$, or $\mathbf{R}_{xx} \to \mathbf{R}_{xx} + \epsilon^2\mathbf{I}$, one is assured, by the discussion of tapering, that the result is no longer singular – all structures in the field are being assigned either to the noise or the solution (or in part to both).

Anyone using a Gauss–Markov estimator to make maps must check that the result is consistent with the prior estimates of $\mathbf{R}_{xx}$, $\mathbf{R}_{nn}$. Such checks include determining whether the differences between the mapped values at the data points and the observed values have numerical values consistent with the assumed noise variance; a further check involves the sample autocovariance of $\tilde{\mathbf{n}}$ and its test against $\mathbf{R}_{nn}$ (see books on regression for such tests). The mapped field should also have a variance and covariance consistent with the prior estimate $\mathbf{R}_{xx}$. If these tests are not passed, the entire result should be rejected.

3.2.2 Higher-dimensional mapping

We can now immediately write down the optimal interpolation formulas for an arbitrary distribution of data in two or more dimensions. Let the positions where data are measured be the set $\mathbf{r}_j$ with measured value $\mathbf{y}(\mathbf{r}_j)$, containing noise $\mathbf{n}$. It is assumed that aliasing errors are unimportant. The mean value of the field is first estimated and subtracted from the measurements and we proceed as though the true mean were zero.[11]

As in the case where the positions are scalars, one minimizes the expected mean-square difference between the estimated and the true field $\mathbf{x}(\tilde{\mathbf{r}}_\alpha)$. The result is (3.27)

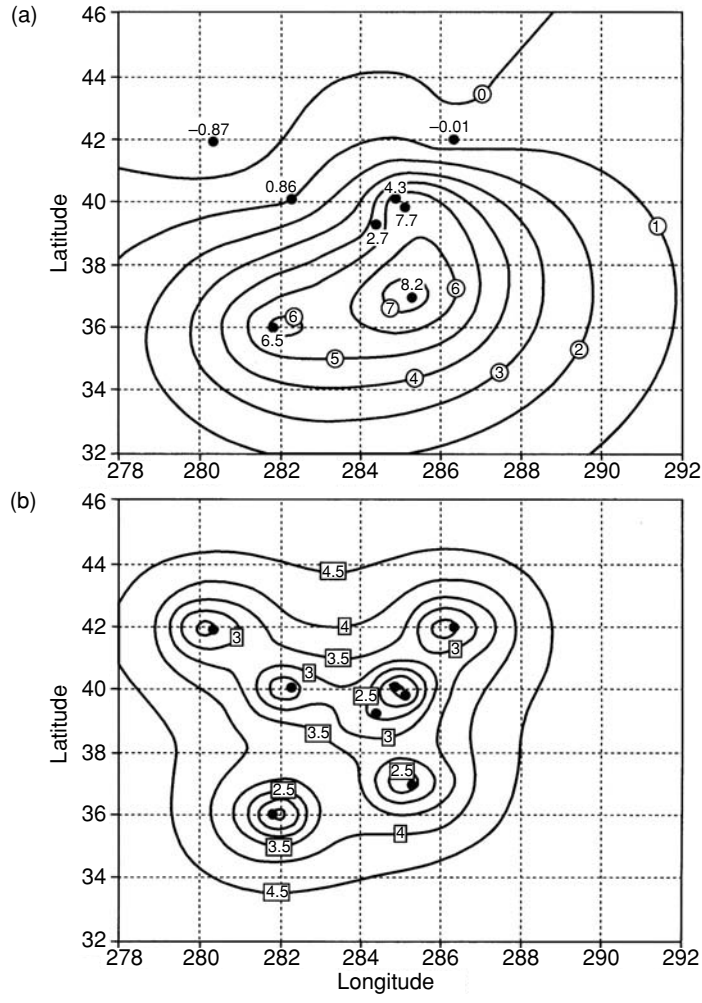

Figure 3.4 (a) Observations, shown as solid dots, from which a uniformly gridded map is desired. Contours were constructed using a fixed covariance and the Gauss–Markov estimate Eq. (3.27). Noise was assumed white with a variance of 1. (b) Expected standard error of the mapped field in (a). Values tend, far from the observations points, to a variance of 25, which was the specified field variance, and hence the largest expected error is $\sqrt{25}$. Note the minima centered on the data points.

and (3.28), except that now everything is a function of the vector positions. If the field being mapped is also a vector (e.g., two components of velocity) with known covariances between the two components, then the elements of **B** become matrices. The observations could also be vectors at each point.

An example of a two-dimensional map is shown in Fig. 3.4. The "data points," $y(\mathbf{r}_i)$, are the dots, while estimates of $\tilde{x}(\mathbf{r}_i)$ on the uniform grid were wanted.

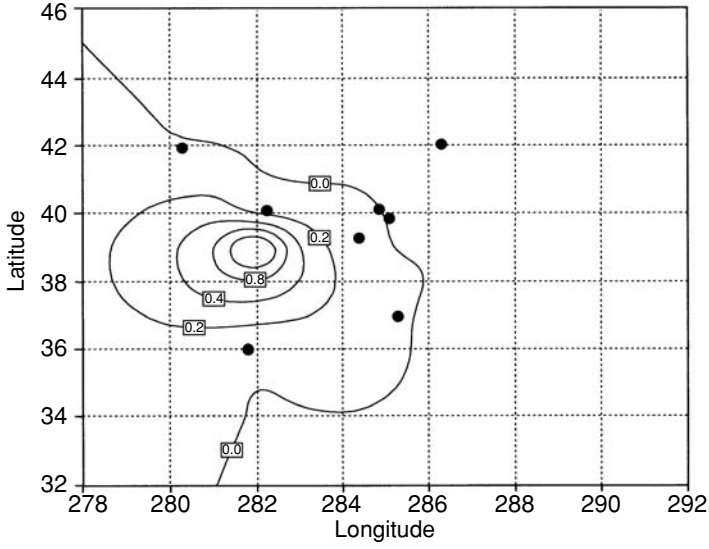

Figure 3.5 One of the rows of **P** corresponding to the grid point in Fig. 3.4 at 39°N, 282°E, displaying the expected correlations that occur in the errors of the mapped field. These errors would be important, e.g., in any use that differentiated the mapped field. For plotting purposes, the variance was normalized to 1.

The a priori noise was set to $\langle \mathbf{n} \rangle = \mathbf{0}$, $\mathbf{R}_{nn} = \langle n_i n_j \rangle = \sigma_n^2 \delta_{ij}$, $\sigma_n^2 = 1$, and the true field covariance was $\langle \mathbf{x} \rangle = \mathbf{0}$, $\mathbf{R}_{xx} = \langle \mathbf{x}(\mathbf{r}_i)\mathbf{x}(\mathbf{r}_j) \rangle = P_0 \exp -|\mathbf{r}_i - \mathbf{r}_j|^2/L_2$, $P_0 = 25$, $L_2 = 100$. Figure 3.4 also shows the estimated values and Figs. 3.4 and 3.5 show the error variance estimate of the mapped values. Notice that, far from the data points, the estimated values are 0: the mapped field goes asymptotically to the estimated true mean, with the error variance rising to the full value of 25, which cannot be exceeded. That is to say, when we are mapping far from any data point, the only real information available is provided by the prior statistics – that the mean is 0, and the variance about that mean is 25. So the expected uncertainty of the mapped field, in the absence of data, cannot exceed the prior estimate of how far from the mean the true value is likely to be. The best estimate is then the mean itself.

A complex error structure of the mapped field exists – even in the vicinity of the data points. Should a model be "driven" by this mapped field, one would need to make some provision in the model accounting for the spatial dependence in the expected errors of this forcing.

In practice, most published objective mapping (often called "OI" for "objective interpolation," although as we have seen, it is not true interpolation) has been based upon simple analytical statements of the covariances $\mathbf{R}_{xx}$, $\mathbf{R}_{nn}$ as used in the example: that is, they are commonly assumed to be spatially stationary and isotropic

(depending upon $|\mathbf{r}_i - \mathbf{r}_j|$ and not upon the two positions separately nor upon their orientation). The use of analytic forms removes the necessity for finding, storing, and computing with the potentially very large $M \times M$ data covariance matrices in which hypothetically every data or grid point has a different covariance with every other data or grid point. But the analytical convenience often distorts the solutions, as many fluid flows and other fields are neither spatially stationary nor isotropic.[12]

3.2.3 Mapping derivatives

A common problem in setting up fluid models is the need to specify the fields of quantities such as temperature, density, etc., on a regular model grid. The derivatives of these fields must be specified for use in advection–diffusion equations,

$$\frac{\partial C}{\partial t} + \mathbf{v} \cdot \nabla C = K \nabla^2 C, \tag{3.31}$$

where C is any scalar field of interest. Suppose the spatial derivative is calculated as a one-sided difference,

$$\frac{\partial C(\tilde{r}_1)}{\partial r} \sim \frac{C(\tilde{r}_1) - C(\tilde{r}_2)}{\tilde{r}_1 - \tilde{r}_2}. \tag{3.32}$$

Then it is attractive to subtract the two estimates made from Eq. (3.27) , producing

$$\frac{\partial C(\tilde{r}_1)}{\partial r} \sim \frac{1}{\Delta r}(\mathbf{R}_{xx}(\tilde{r}_1,\, r_j) - \mathbf{R}_{xx}(\tilde{r}_2,\, r_j))(\mathbf{R}_{xx} + \mathbf{R}_{nn})^{-1}\mathbf{y}. \tag{3.33}$$

Alternatively, an estimate of $\partial C/\partial r$ could be made directly from (2.392), using $\mathbf{x} = C(r_1) - C(r_2)$. $\mathbf{R}_{yy} = \mathbf{R}_{xx} + \mathbf{R}_{nn}$, which describes the data, does not change. $\mathbf{R}_{xy}$ does change:

$$\mathbf{R}_{xy} = \langle (C(\tilde{r}_1) - C(\tilde{r}_2))(C(r_j) + n(r_j)) \rangle = \mathbf{R}_{xx}(\tilde{r}_1,\, r_j) - \mathbf{R}_{xx}(\tilde{r}_2,\, r_j), \tag{3.34}$$

which when substituted into (2.396) produces (3.33). *Thus, the optimal map of the finite difference field is simply the difference of the mapped values*. More generally, the optimally mapped value of any linear combination of the values is that linear combination of the maps.[13]

3.3 Inequality constraints: non-negative least-squares

In many estimation problems, it is useful to be able to impose inequality constraints upon the solutions. Problems involving tracer concentrations, for example, usually demand that they remain positive; empirical eddy diffusion coefficients are sometimes regarded as acceptable only when non-negative; in some fluid flow

problems we may wish to impose directions, but not magnitudes, upon velocity fields.

Such needs lead to consideration of the forms

$$\mathbf{E}\mathbf{x} + \mathbf{n} = \mathbf{y}, \tag{3.35}$$

$$\mathbf{G}\mathbf{x} \geq \mathbf{h}, \tag{3.36}$$

where the use of a greater-than inequality to represent the general case is purely arbitrary; multiplication by minus 1 readily reverses it. $\mathbf{G}$ is of dimension $M_2 \times N$.

Several cases need to be distinguished. (A) Suppose $\mathbf{E}$ is full rank and fully determined; then the SVD solution to (3.35) by itself is $\tilde{\mathbf{x}}$, $\tilde{\mathbf{n}}$, and there is no solution nullspace. Substitution of the solution into (3.36) shows that the inequalities are either satisfied or that some are violated. In the first instance, the problem is solved, and the inequalities bring no new information. In the second case, the solution must be modified and, necessarily, $\|\tilde{\mathbf{n}}\|$ will increase, given the noise-minimizing nature of the SVD solution. It is also possible that the inequalities are contradictory, in which case there is no solution.

(B) Suppose that $\mathbf{E}$ is formally underdetermined – so that a solution nullspace exists. If the particular SVD solution violates one or more of the inequalities and requires modification, two subcases can be distinguished: (1) Addition of one or more nullspace vectors permits the inequalities to be satisfied. Then the solution residual norm will be unaffected, but $\|\tilde{\mathbf{x}}\|$ will increase. (2) The nullspace vectors by themselves are unable to satisfy the inequality constraints, and one or more range vectors are required to do so. Then both $\|\tilde{\mathbf{x}}\|$, $\|\tilde{\mathbf{n}}\|$ will increase.

Case (A) is the conventional one.[14] The so-called Kuhn–Tucker–Karush theorem is a requirement for a solution $\tilde{\mathbf{x}}$ to exist. Its gist is as follows: Let $M \geq N$ and $\mathbf{E}$ be full rank; there are no $\mathbf{v}_i$ in the solution nullspace. If there is a solution, there must exist a vector, $\mathbf{q}$, of dimension M_2 such that

$$\mathbf{E}^{\mathrm{T}}(\mathbf{E}\tilde{\mathbf{x}} - \mathbf{y}) = \mathbf{G}^{\mathrm{T}}\mathbf{q}, \tag{3.37}$$

$$\mathbf{G}\mathbf{x} - \mathbf{h} = \mathbf{r}, \tag{3.38}$$

where the M_2 elements of $\mathbf{q}$ are divided into two groups. For group 1, of dimension m_1,

$$r_i = 0, \quad q_i \geq 0, \tag{3.39}$$

and for group 2, of dimension $m_2 = M_2 - m_1$,

$$r_i > 0, \quad q_i = 0. \tag{3.40}$$

To understand this theorem, recall that in the solution to the ordinary overdetermined least-squares problem, the left-hand side of (3.37) vanishes identically (2.91

and 2.265), being the projection of the residuals onto the range vectors, $\mathbf{u}_i$, of $\mathbf{E}^{\mathrm{T}}$. If this solution violates one or more of the inequality constraints, structures that produce increased residuals must be introduced into the solution.

Because there are no nullspace $\mathbf{v}_i$, the rows of $\mathbf{G}$ may each be expressed exactly by an expansion in the range vectors. In the second group of indices, the corresponding inequality constraints are already satisfied by the ordinary least-squares solution, and no modification of the structure proportional to $\mathbf{v}_i$ is required. In the first group of indices, the inequality constraints are marginally satisfied, at equality, only by permitting violation of the demand (2.91) that the residuals should be orthogonal to the range vectors of $\mathbf{E}$. If the ordinary least-squares solution violates the inequality, the minimum modification required to it pushes the solution to the edge of the acceptable bound, but at the price of increasing the residuals proportional to the corresponding $\mathbf{u}_i$. The algorithm consists of finding the two sets of indices and then the smallest coefficients of the $\mathbf{v}_i$ corresponding to the group 1 indices required to just satisfy any initially violated inequality constraints. A canonical special case, to which more general problems can be reduced, is based upon the solution to $\mathbf{G} = \mathbf{I}$, $\mathbf{h} = \mathbf{0}$ – called "non-negative least-squares."[15] The requirement, $\mathbf{x} \geq 0$, is essential in many problems involving tracer concentrations, which are necessarily positive.

The algorithm can be extended to the underdetermined/rank-deficient case in which the addition, to the original basic SVD solution, of appropriate amounts of the nullspace of $\mathbf{v}_i$ is capable of satisfying any violated inequality constraints.[16] One simply chooses the smallest mean-square solution coefficients necessary to push the solution to the edge of the acceptable inequalities, producing the smallest norm. The residuals of the original problem do not increase – because only nullspace vectors are being used. $\mathbf{G}$ must have a special structure for this to be possible.

The algorithm can be further generalized[17] by considering the general case of rank-deficiency/underdeterminism where the nullspace vectors by themselves are inadequate to produce a solution satisfying the inequalities. In effect, any inequalities "left over" are satisfied by invoking the smallest perturbations necessary to the coefficients of the range vectors $\mathbf{v}_i$.

3.4 Linear programming

In a number of important geophysical fluid problems, the objective functions are linear rather than quadratic functions. Scalar property fluxes such as heat, C_i, are carried by a fluid flow at rates $\sum C_i x_i$, which are linear functions of $\mathbf{x}$. If one sought the extreme fluxes of C, it would require finding the extremal values of the corresponding linear function. Least-squares does not produce useful answers in such problems because linear objective functions achieve their minima or maxima

only at plus or minus infinity – unless the elements of $\mathbf{x}$ are bounded. The methods of *linear programming* are directed at finding extremal properties of linear objective functions subject to bounding constraints. Such problems can be written as

$$\begin{aligned} \text{minimize:}\quad & J = \mathbf{c}^{\mathrm{T}}\mathbf{x}, \\ & \mathbf{E}_1\mathbf{x} = \mathbf{y}_1, && (3.41) \\ & \mathbf{E}_2\mathbf{x} \geq \mathbf{y}_2, && (3.42) \\ & \mathbf{E}_3\mathbf{x} \leq \mathbf{y}_3, && (3.43) \\ & \mathbf{a} \leq \mathbf{x} \leq \mathbf{b}, && (3.44) \end{aligned}$$

that is, as a collection of equality and inequality constraints of both greater than or less than form, plus bounds on the individual elements of $\mathbf{x}$. In distinction to the least-squares and minimum variance equations that have been discussed so far, these constraints are hard ones; they cannot be violated even slightly in an acceptable solution.

Linear programming problems are normally reduced to what is referred to as a *canonical form*, although different authors use different definitions of what it is. But all such problems are reducible to

$$\begin{aligned} \text{minimize:}\quad & J = \mathbf{c}^{\mathrm{T}}\mathbf{x}, && (3.45) \\ & \mathbf{E}\mathbf{x} \leq \mathbf{y} && (3.46) \\ & \mathbf{x} \geq \mathbf{0}. && (3.47) \end{aligned}$$

The use of a minimum rather than a maximum is readily reversed by introducing a minus sign, and the inequality is similarly readily reversed. The last relationship, requiring purely positive elements in $\mathbf{x}$, is obtained without difficulty by translation.

Linear programming problems are widespread in many fields including, especially, financial and industrial management where they are used to maximize profits, or minimize costs, in, say, a manufacturing process. Necessarily then, the amount of a product of each type is positive, and the inequalities reflect such things as the need to consume no more than the available amounts of raw materials. In some cases, J is then literally a "cost" function. General methodologies were first developed during World War II in what became known as "operations research" ("operational research" in the UK),[18] although special cases were known much earlier. Since then, because of the economic stake in practical use of linear programming, immense effort has been devoted both to textbook discussion and efficient, easy-to-use software.[19] Given this accessible literature and software, the methodologies of solution are not described here, and only a few general points are made.

The original solution algorithm invented by G. Dantzig is usually known as the "simplex method" (a simplex is a convex geometric shape). It is a highly efficient

search method conducted along the bounding constraints of the problem. In general, it is possible to show that the outcome of a linear programming problem falls into several distinct categories: (1) The system is "infeasible," meaning that it is contradictory and there is no solution; (2) the system is unbounded, meaning that the minimum lies at negative infinity; (3) there is a unique minimizing solution; and (4) there is a unique finite minimum, but it is achieved by an infinite number of solutions $\mathbf{x}$.

The last situation is equivalent to observing that if there are two minimizing solutions, there must be an infinite number of them because then any linear combination of the two solutions is also a solution. Alternatively, if one makes up a matrix from the coefficients of $\mathbf{x}$ in Eqs. (3.45)–(3.47), one can determine if it has a nullspace. If one or more such vectors exists, it is also orthogonal to the objective function, and it can be assigned an arbitrary amplitude without changing J. One distinguishes between *feasible solutions*, meaning those that satisfy the inequality and equality constraints but which are not minimizing, and *optimal solutions*, which are both feasible and minimize the objective function.

An interesting and useful feature of a linear programming problem is that Eqs. (3.45)–(3.47) have a "dual":

$$\text{maximize:} \quad J_2 = \mathbf{y}^{\mathrm{T}}\boldsymbol{\mu}, \tag{3.48}$$

$$\mathbf{E}^{\mathrm{T}}\boldsymbol{\mu} \geq \mathbf{c}, \tag{3.49}$$

$$\boldsymbol{\mu} \geq \mathbf{0}. \tag{3.50}$$

It is possible to show that the minimum of J must equal the maximum of J_2. The reader may want to compare the structure of the original (the "primal") and dual equations with those relating the Lagrange multipliers to $\mathbf{x}$ discussed in Chapter 2. In the present case, the important relationship is

$$\frac{\partial J}{\partial y_i} = \mu_i. \tag{3.51}$$

That is, in a linear program, the dual solution provides the sensitivity of the objective function to perturbations in the constraint parameters $\mathbf{y}$. Duality theory pervades optimization problems, and the relationship to Lagrange multipliers is no accident.[20] Some simplex algorithms, called the "dual simplex," take advantage of the different dimensions of the primal and dual problems to accelerate solution. In recent years much attention has focussed upon a new, non-simplex method of solution[21] known as the "Karmarkar" or "interior point" method.

Linear programming is also valuable for solving estimation or approximation problems in which norms other than the 2-norms, which have been the focus of this book, are used. For example, suppose that one sought the solution to the

constraints $\mathbf{Ex} + \mathbf{n} = \mathbf{y}$, $M > N$, but subject not to the conventional minimum of $J = \sum_i n_i^2$, but that of $J = \sum_i |n_i|$ (a 1-norm). Such norms are less sensitive to outliers than are the 2-norms and are said to be "robust." The maximum likelihood idea connects 2-norms to Gaussian statistics, and similarly, 1-norms are related to maximum likelihood with exponential statistics.[22] Reduction of such problems to linear programming is carried out by setting $n_i = n_i^+ - n_i^-$, $n_i^+ \geq 0$, $n_i^- \geq 0$, and the objective function is

$$\min\colon \; J = \sum_i \left(n_i^+ + n_i^-\right). \tag{3.52}$$

Other norms, the most important[23] of which is the so-called infinity norm, which minimizes the maximum element of an objective function ("mini-max" optimization), are also reducible to linear programming.

3.5 Empirical orthogonal functions

Consider an arbitrary $M \times N$ matrix $\mathbf{M}$. Suppose the matrix was representable, accurately, as the product of two vectors,

$$\mathbf{M} \approx \mathbf{ab}^{\mathrm{T}},$$

where $\mathbf{a}$ was $M \times 1$, and $\mathbf{b}$ was $N \times 1$. Approximation is intended in the sense that

$$\|\mathbf{M} - \mathbf{ab}^{\mathrm{T}}\| < \varepsilon,$$

for some acceptably small ε. Then one could conclude that the MN elements of $\mathbf{A}$ contain only $M + N$ pieces of information contained in $\mathbf{a}$, $\mathbf{b}$. Such an inference has many uses, including the ability to recreate the matrix accurately from only $M + N$ numbers, to physical interpretations of the meaning of $\mathbf{a}$, $\mathbf{b}$. More generally, if one pair of vectors is inadequate, some small number might suffice:

$$\mathbf{M} \approx \mathbf{a}_1\mathbf{b}_1^{\mathrm{T}} + \mathbf{a}_2\mathbf{b}_2^{\mathrm{T}} + \cdots + \mathbf{a}_K\mathbf{b}_K^{\mathrm{T}}. \tag{3.53}$$

A general mathematical approach to finding such a representation is through the SVD in a form sometimes known as the "Eckart–Young–Mirsky theorem."[24] This theorem states that the most efficient representation of a matrix in the form

$$\mathbf{M} \approx \sum_i^K \lambda_i \mathbf{u}_i \mathbf{v}_i^{\mathrm{T}}, \tag{3.54}$$

where the $\mathbf{u}_i$, $\mathbf{v}_i$ are orthonormal, is achieved by choosing the vectors to be the singular vectors, with λ_i providing the amplitude information (recall Eq. (2.227)).

The connection to regression analysis is readily made by noticing that the sets of singular vectors are the eigenvectors of the two matrices $\mathbf{MM}^{\mathrm{T}}$, $\mathbf{M}^{\mathrm{T}}\mathbf{M}$

(Eqs. (2.253) and (2.254)). If each row of $\mathbf{M}$ is regarded as a set of observations at a fixed coordinate, then $\mathbf{M}\mathbf{M}^{\mathrm{T}}$ is just proportional to the sample second-moment matrix of all the observations, and its eigenvectors, $\mathbf{u}_i$, are the EOFs. Alternatively, if each column is regarded as the observation set for a fixed coordinate, then $\mathbf{M}^{\mathrm{T}}\mathbf{M}$ is the corresponding sample second-moment matrix, and the $\mathbf{v}_i$ are the EOFs.

A large literature provides various statistical rules for use of EOFs. For example, the rank determination in the SVD becomes a test of the statistical significance of the contribution of singular vectors to the structure of $\mathbf{M}$.[25] In the wider context, however, one is dealing with the problem of efficient relationships amongst variables known or suspected to carry mutual correlations. Because of its widespread use, this subject is plagued by multiple discovery and thus multiple jargon. In different contexts and details (e.g., how the matrix is weighted), the problem is known as that of "principal components,"[26] "empirical orthogonal functions" (EOFs), the "Karhunen–Loève" expansion (in mathematics and electrical engineering),[27] "proper orthogonal decomposition,"[28] etc. Examples of the use of EOFs will be provided in Chapter 6.

3.6 Kriging and other variants of Gauss–Markov estimation

A variant of the Gauss–Markov mapping estimators, often known as "kriging" (named for David Krige, a mining geologist), addresses the problem of a spatially varying mean field, and is a generalization of the ordinary Gauss–Markov estimator.[29]

Consider the discussion on p. 132 of the fitting of a set of functions $f_i(\mathbf{r})$ to an observed field $y(\mathbf{r}_j)$. That is, we put

$$y(\mathbf{r}_j) = \mathbf{F}\boldsymbol{\alpha} + q(\mathbf{r}_j), \tag{3.55}$$

where $\mathbf{F}(\mathbf{r}) = \{f_i(\mathbf{r})\}$ is a set of basis functions, and one seeks the expansion coefficients, $\boldsymbol{\alpha}$, and q such that the data, y, are *interpolated* (meaning reproduced exactly) at the observation points, although there is nothing to prevent further breaking up q into signal and noise components. If there is only one basis function – for example a constant – one is doing kriging, which is the determination of the mean prior to objective mapping of q, as discussed above. If several basis functions are being used, one has "universal kriging." The main issue concerns the production of an adequate statement of the expected error, given that the q are computed from a preliminary regression to determine the $\boldsymbol{\alpha}$.[30] The method is often used in situations where large-scale trends are expected in the data, and where one wishes to estimate and remove them before analyzing and mapping the q.

Because the covariances employed in objective mapping are simple to use and interpret only when the field is spatially stationary, much of the discussion

of kriging uses instead what is called the "variogram," defined as $V = \langle (y(\mathbf{r}_i) - y(\mathbf{r}_j))(y(\mathbf{r}_i) - y(\mathbf{r}_j)) \rangle$, which is related to the covariance, and which is often encountered in turbulence theory as the "structure function." Kriging is popular in geology and hydrology, and deserves wider use.

3.7 Non-linear problems

The least-squares solutions examined thus far treat the coefficient matrix $\mathbf{E}$ as given. But in many of the cases encountered in practice, the elements of $\mathbf{E}$ are computed from data and are imperfectly specified. It is well known in the regression literature that treating $\mathbf{E}$ as known, even if $\mathbf{n}$ is increased beyond the errors contained in $\mathbf{y}$, can lead to significant bias errors in the least-squares and related solutions, particularly if $\mathbf{E}$ is nearly singular.[31] The problem is known as that of "errors in regressors or errors in variables" (EIV); it manifests itself in the classical simple least-squares problem (p. 43), where a straight line is fitted to data of the form $y_i = a + bt_i$, but where the measurement positions, t_i, are partly uncertain rather than perfect.

In general terms, when $\mathbf{E}$ has errors, the model statement becomes

$$(\tilde{\mathbf{E}} + \Delta\tilde{\mathbf{E}})\tilde{\mathbf{x}} = \tilde{\mathbf{y}} + \Delta\tilde{\mathbf{y}}, \tag{3.56}$$

where one seeks estimates, $\tilde{\mathbf{x}}$, $\Delta\tilde{\mathbf{E}}$, $\Delta\tilde{\mathbf{y}}$, where the old $\mathbf{n}$ is now broken into two parts: $\Delta\tilde{\mathbf{E}}\tilde{\mathbf{x}}$ and $\Delta\tilde{\mathbf{y}}$. If such estimates can be made, the result can be used to rewrite (3.56) as

$$\tilde{\mathbf{E}}\tilde{\mathbf{x}} = \tilde{\mathbf{y}}, \tag{3.57}$$

where the relation is to be exact. That is, one seeks to modify the elements of $\mathbf{E}$ such that the observational noise in it is reduced to zero.

3.7.1 *Total least-squares*

For some problems of this form, the method of total least-squares (TLS) is a powerful and interesting method. It is worth examining briefly to understand why it is not always immediately useful, and to motivate a different approach.[32]

The SVD plays a crucial role in TLS. Consider that in Eq. (2.17) the vector $\mathbf{y}$ was written as a sum of the column vectors of $\mathbf{E}$; to the extent that the column space does not fully describe $\mathbf{y}$, a residual must be left by the solution $\tilde{\mathbf{x}}$, and ordinary least-squares can be regarded as producing a solution in which a new estimate, $\tilde{\mathbf{y}} \equiv \mathbf{E}\tilde{\mathbf{x}}$, of $\mathbf{y}$ is made; $\mathbf{y}$ is changed, but the elements of $\mathbf{E}$ are untouched. But suppose it were possible to introduce small changes in both the column vectors of $\mathbf{E}$, as well as in $\mathbf{y}$, such that the column vectors of the modified $\mathbf{E} + \Delta\mathbf{E}$ produced

a spanning vector space for $\mathbf{y} + \Delta\mathbf{y}$, where both $\|\Delta\mathbf{y}\|$, $\|\Delta\mathbf{E}\|$ were "small," then the problem as stated would be solved.

The simplest problem to analyze is the full-rank, formally overdetermined one. Let $M \geq N = K$. Then, if we form the $M \times (N + 1)$ augmented matrix

$$\mathbf{E}_a = \{\mathbf{E} \quad \mathbf{y}\},$$

the solution sought is such that

$$\{\tilde{\mathbf{E}} \quad \tilde{\mathbf{y}}\} \begin{bmatrix} \tilde{\mathbf{x}} \\ -1 \end{bmatrix} = \mathbf{0} \tag{3.58}$$

(exactly). If this solution is to exist, $[\tilde{\mathbf{x}} \; -1]^{\mathrm{T}}$ must lie in the nullspace of $\{\tilde{\mathbf{E}} \quad \tilde{\mathbf{y}}\}$. A solution is thus ensured by forming the SVD of $\{\mathbf{E} \quad \mathbf{y}\}$, setting $\lambda_{N+1} = 0$, and forming $\{\tilde{\mathbf{E}} \quad \tilde{\mathbf{y}}\}$ out of the remaining singular vectors and values. Then $[\tilde{\mathbf{x}} \quad -1]^{\mathrm{T}}$ is the nullspace of the modified augmented matrix, and must therefore be proportional to the nullspace vector $\mathbf{v}_{N+1}$. Also,

$$\{\Delta\tilde{\mathbf{E}} \quad \Delta\tilde{\mathbf{y}}\} = -\mathbf{u}_{N+1}\lambda_{N+1}\mathbf{v}_{N+1}^{\mathrm{T}}. \tag{3.59}$$

Various complications can be considered, for example, if the last element of $\mathbf{v}_{N+1} = 0$; this and other special cases are discussed in the reference (see note 24). Cases of non-uniqueness are treated by selecting the solution of minimum norm. A simple generalization applies to the underdetermined case: if the rank of the augmented matrix is p, one reduces the rank by one to $p - 1$.

The TLS solution just summarized applies only to the case in which the errors in the elements of $\mathbf{E}$ and $\mathbf{y}$ are uncorrelated and of equal variance and in which there are no required structures – for example, where certain elements of $\mathbf{E}$ must always vanish. More generally, changes in some elements of $\mathbf{E}$ require, for reasons of physics, specific corresponding changes in other elements of $\mathbf{E}$ and in $\mathbf{y}$, and vice versa. The fundamental difficulty is that the model (3.56) presents a non-linear estimation problem with correlated variables, and its solution requires modification of the linear procedures developed so far.

3.7.2 Method of total inversion

The simplest form of TLS does not readily permit the use of correlations and prior variances in the parameters appearing in the coefficient matrix and does not provide any way of maintaining the zero structure there. Methods exist that permit accounting for prior knowledge of covariances.[33] Consider a set of non-linear constraints in a vector of unknowns $\mathbf{x}$,

$$\mathbf{g}(\mathbf{x}) + \mathbf{u} = \mathbf{q}. \tag{3.60}$$

This set of equations is the generalization of the linear models hitherto used; $\mathbf{u}$ again represents any expected error in the specification of the model. An example of a scalar non-linear model is

$$8x_1^2 + x_2^2 + u = 4.$$

In general, there will be some expectations about the behavior of $\mathbf{u}$. Without loss of generality, take its expected value to be zero, and its covariance is $\mathbf{Q} = \langle \mathbf{u}\mathbf{u}^{\mathrm{T}} \rangle$. There is nothing to prevent us from combining $\mathbf{x}$, $\mathbf{u}$ into one single set of unknowns $\boldsymbol{\xi}$, and indeed if the model has some unknown parameters, $\boldsymbol{\xi}$ might as well include those as well. So (3.60) can be written as

$$\mathcal{L}(\boldsymbol{\xi}) = \mathbf{0}. \tag{3.61}$$

In addition, it is supposed that a reasonable initial estimate $\tilde{\boldsymbol{\xi}}(0)$ is available, with uncertainty $\mathbf{P}(0) \equiv \langle (\boldsymbol{\xi} - \tilde{\boldsymbol{\xi}}(0))(\boldsymbol{\xi} - \tilde{\boldsymbol{\xi}}(0))^{\mathrm{T}} \rangle$ (or the covariances of the $\mathbf{u}$, $\mathbf{x}$ could be specified separately if their uncertainties are not correlated). An objective function whose minimum is sought is written,

$$J = \mathcal{L}(\boldsymbol{\xi})^{\mathrm{T}} \mathbf{Q}^{-1} \mathcal{L}(\boldsymbol{\xi}) + (\boldsymbol{\xi} - \tilde{\boldsymbol{\xi}}(0))^{\mathrm{T}} \mathbf{P}(0)^{-1} (\boldsymbol{\xi} - \tilde{\boldsymbol{\xi}}(0)). \tag{3.62}$$

The presence of the weight matrices $\mathbf{Q}$, $\mathbf{P}(0)$ permits control of the elements most likely to change, specification of elements that should not change at all (e.g., by introducing zeros into $\mathbf{P}(0)$), as well as the stipulation of covariances. It can be regarded as a generalization of the process of minimizing objective functions, which led to least-squares in previous chapters and is sometimes known as the "method of total inversion."[34]

Consider an example for the two simultaneous equations

$$2x_1 + x_2 + n_1 = 1, \tag{3.63}$$

$$0 + 3x_2 + n_2 = 2, \tag{3.64}$$

where all the numerical values except the zero are now regarded as in error to some degree. One way to proceed is to write the coefficients of $\mathbf{E}$ in the specific perturbation form (3.56). For example, write $E_{11} = 2 + \Delta E_{11}$, and define the unknowns $\boldsymbol{\xi}$ in terms of the ΔE_{ij}. For illustration retain the full non-linear form by setting

$$\xi_1 = E_{11}, \quad \xi_2 = E_{12}, \quad \xi_3 = E_{21}, \quad \xi_4 = E_{22}, \quad \xi_5 = x_1, \quad \xi_6 = x_2,$$
$$u_1 = n_1, \quad u_2 = n_2.$$

The equations are then

$$\xi_1 \xi_5 + \xi_2 \xi_6 + u_1 - 1 = 0, \tag{3.65}$$

$$\xi_3 \xi_5 + \xi_4 \xi_6 + u_2 - 2 = 0. \tag{3.66}$$

The y_i are being treated as formally fixed, but u_1, u_2 represent their possible errors (the division into knowns and unknowns is not unique). Let there be an initial estimate,

$$\xi_1 = 2 \pm 1, \quad \xi_2 = 2 \pm 2, \quad \xi_3 = 0 \pm 0,$$
$$\xi_4 = 3.5 \pm 1, \quad \xi_5 = x_1 = 0 \pm 2, \quad \xi_6 = 0 \pm 2,$$

with no imposed correlations so that $\mathbf{P}(0) = \text{diag}\,([1, 4, 0, 1, 4, 4])$; the zero represents the requirement that E_{21} remain unchanged. Let $\mathbf{Q} = \text{diag}\,([2, 2])$. Then a useful objective function is

$$\begin{aligned} J = &(\xi_1\xi_5 + \xi_2\xi_6 - 1)^2/2 \\ &+ (\xi_3\xi_5 + \xi_4\xi_6 - 2)^2/2 + (\xi_1 - 2)^2 + (\xi_2 - 2)^2/4 \\ &+ 10^6\xi_3^2 + (\xi_4 - 3.5)^2 + \xi_5^2/4 + \xi_6^2/4. \end{aligned} \tag{3.67}$$

The 10^6 in front of the term in ξ_3^2 is a numerical approximation to the infinite value implied by a zero uncertainty in this term (an arbitrarily large value can cause numerical instability, characteristic of penalty and barrier methods).[35]

Such objective functions define surfaces in spaces of the dimension of $\boldsymbol{\xi}$. Most procedures require the investigator to make a first guess at the solution, $\tilde{\boldsymbol{\xi}}(0)$, and attempt to minimize J by going downhill from the guess. Various deterministic search algorithms have been developed and are variants of steepest descent, conjugate gradient, Newton and quasi-Newton methods. The difficulties are numerous. Some methods require computation or provision of the gradients of J with respect to $\boldsymbol{\xi}$, and the computational cost may become very great. The surfaces on which one is seeking to go downhill may become extremely tortuous, or very slowly changing. The search path can fall into local holes that are not the true minima. Non-linear optimization is something of an art. Nonetheless, existing techniques are very useful. The minimum of J corresponds to finding the solution of the non-linear normal equations that would result from setting the partial derivatives to zero.

Let the true minimum be at $\boldsymbol{\xi}^*$. Assuming that the search procedure has succeeded, the objective function is locally

$$J = \text{constant} + (\boldsymbol{\xi} - \boldsymbol{\xi}^*)^{\mathrm{T}}\mathcal{H}(\boldsymbol{\xi} - \boldsymbol{\xi}^*) + \Delta J, \tag{3.68}$$

where $\mathcal{H}$ is the Hessian and ΔJ is a correction – assumed to be small. In the linear least-squares problem (2.89), the Hessian is evidently $\mathbf{E}^{\mathrm{T}}\mathbf{E}$, the second derivative of the objective function with respect to $\mathbf{x}$. The supposition is then that near the true optimum, the objective function is locally quadratic with a small correction. To the extent that this supposition is true, the result can be analyzed in terms of the behavior of $\mathcal{H}$ as though it represented a locally defined version of $\mathbf{E}^{\mathrm{T}}\mathbf{E}$. In particular, if $\mathcal{H}$ has a nullspace, or small eigenvalues, one can expect to see all the issues arising that

we dealt with in Chapter 2, including ill-conditioning and solution variances that may become large in some elements. The machinery used in Chapter 2 (row and column scaling, nullspace suppression, etc.) thus becomes immediately relevant here and can be used to help conduct the search and to understand the solution.

Example *It remains to find the minimum of* J *in* (3.67).[36] *Most investigators are best-advised to tackle such problems by using one of the many general purpose numerical routines written by experts.*[37] *Here, a quasi-Newton method was employed to produce*

$$E_{11} = 2.0001, \quad E_{12} = 1.987, \quad E_{21} = 0.0,$$
$$E_{22} = 3.5237, \quad x_1 = -0.0461, \quad x_2 = 0.556,$$

and the minimum of $J = 0.0802$. *The inverse Hessian at the minimum is*

$$\mathcal{H}^{-1} = \begin{Bmatrix} 0.4990 & 0.0082 & -0.0000 & -0.0014 & 0.0061 & 0.0005 \\ 0.0082 & 1.9237 & 0.0000 & 0.0017 & -0.4611 & -0.0075 \\ -0.0000 & 0.0000 & 0.0000 & -0.0000 & -0.0000 & 0.0000 \\ -0.0014 & 0.0017 & -0.0000 & 0.4923 & 0.0623 & -0.0739 \\ 0.0061 & -0.4611 & -0.0000 & 0.0623 & 0.3582 & -0.0379 \\ 0.0005 & -0.0075 & 0.0000 & -0.0739 & -0.0379 & 0.0490 \end{Bmatrix}.$$

The eigenvalues and eigenvectors of $\mathcal{H}$ *are*

$$\lambda_i = \begin{bmatrix} 2.075 \times 10^6 & 30.4899 & 4.5086 & 2.0033 & 1.9252 & 0.4859 \end{bmatrix},$$

$$\mathbf{V} = \begin{Bmatrix} 0.0000 & -0.0032 & 0.0288 & 0.9993 & 0.0213 & 0.0041 \\ -0.0000 & 0.0381 & -0.2504 & 0.0020 & 0.0683 & 0.9650 \\ -1.0000 & 0.0000 & 0.0000 & -0.0000 & -0.0000 & 0.0000 \\ 0.0000 & 0.1382 & 0.2459 & -0.0271 & 0.9590 & -0.0095 \\ -0.0000 & 0.1416 & -0.9295 & 0.0237 & 0.2160 & -0.2621 \\ 0.0000 & 0.9795 & 0.1095 & 0.0035 & -0.1691 & 0.0017 \end{Bmatrix}.$$

The large jump from the first eigenvalue to the others is a reflection of the conditioning problem introduced by having one element, ξ_3, *with almost zero uncertainty. It is left to the reader to use this information about* $\mathcal{H}$ *to compute the uncertainty of the solution in the neighborhood of the optimal values – this would be the new uncertainty,* $\mathbf{P}(1)$. *A local resolution analysis follows from that of the SVD, employing knowledge of the* $\mathbf{V}$. *The particular system is too small for a proper statistical test of the result against the prior covariances, but the possibility should be clear. If* $\mathbf{P}(0)$, *etc., are simply regarded as non-statistical weights, we are free to experiment with different values until a pleasing solution is found.*

3.7.3 Variant non-linear methods, including combinatorial ones

As with the linear least-squares problems discussed in Chapter 2, many possibilities exist for objective functions that are non-linear in either data constraint terms or the model, and there are many variations on methods for searching for objective function minima.

A very interesting and useful set of methods has been developed comparatively recently, called "combinatorial optimization." Combinatorial methods do not promise that the true minimum is found – merely that it is highly probable – because they search the space of solutions in clever ways that make it unlikely that one is very far from the true optimal solution. Two such methods, simulated annealing and genetic algorithms, have recently attracted considerable attention.[38] Simulated annealing searches randomly for solutions that reduce the objective function from a present best value. Its clever addition to purely random guessing is a willingness to accept the occasional uphill solution – one that raises the value of the objective function – as a way of avoiding being trapped in purely local minima. The probability of accepting an uphill value and the size of the trial random perturbations depend upon a parameter, a temperature defined in analogy to the real temperature of a slowly cooling (annealing) solid.

Genetic algorithms, as their name would suggest, are based upon searches generated in analogy to genetic drift in biological organisms.[39] The recent literature is large and sophisticated, and this approach is not pursued here.

Notes

1 Freeman (1965), Jerri (1977), Butzer and Stens (1992), or Bracewell (2000).
2 In the Russian literature, Kotel'nikov's theorem.
3 Aliasing is familiar as the stroboscope effect. Recall the appearance of the spokes of a wagon wheel in the movies. The spokes can appear to stand still, or move slowly forward or backward, depending upon the camera shutter speed relative to the true rate at which the spokes revolve. (The terminology is apparently due to John Tukey.)
4 Hamming (1973) and Bracewell (2000) have particularly clear discussions.
5 There is a story, perhaps apocryphal, that a group of investigators was measuring the mass flux of the Gulf Stream at a fixed time each day. They were preparing to publish the exciting discovery that there was a strong 14-day periodicity to the flow, before someone pointed out that they were aliasing the tidal currents of period 12.42 hours.
6 It follows from the so-called Paley–Wiener criterion, and is usually stated in the form that "timelimited signals cannot be bandlimited."
7 Landau and Pollack (1962), Freeman (1965), Jerri (1977).
8 Petersen and Middleton (1962). An application, with discussion of the noise sensitivity, may be found in Wunsch (1989).
9 Davis and Polonsky (1965).
10 See Ripley (2004, Section 5.2).
11 Bretherton *et al.* (1976).
12 See Fukumori *et al.* (1991).
13 Luenberger (1969).
14 See Strang (1988) or Lawson and Hanson (1995); the standard full treatment is Fiacco and McCormick (1968).

15 Lawson and Hanson (1974).
16 Fu (1981).
17 Tziperman and Hecht (1987).
18 Dantzig (1963).
19 For example, Bradley *et al.* (1977), Luenberger (2003), and many others.
20 See Cacuci (1981), Hall and Cacuci (1984), Strang (1986), Rockafellar (1993), Luenberger (1997).
21 One of the few mathematical algorithms ever to be written up on the front page of the *New York Times* (November 19, 1984, story by J. Gleick) – a reflection of the huge economic importance of linear programs in industry.
22 Arthanari and Dodge (1993).
23 Wagner (1975), Arthanari and Dodge (1993).
24 Van Huffel and Vandewalle (1991).
25 The use of EOFs, with various normalizations, scalings, and in various row/column physical spaces, is widespread – for example, Wallace (1972), Wallace and Dickinson (1972), and many others.
26 Jolliffe (2002), Preisendorfer (1988), Jackson (2003).
27 Davenport and Root (1958), Wahba (1990).
28 Berkooz *et al.*(1993).
29 Armstrong (1989), David (1988), Ripley (2004).
30 Ripley (2004).
31 For example, Seber and Lee (2003).
32 Golub and van Loan (1996), Van Huffel and Vandewalle (1991).
33 Tarantola and Valette (1982), Tarantola (1987).
34 Tarantola and Valette (1982) labeled the use of similar objective functions and the determination of the minimum as the *method of total inversion*, although they considered only the case of perfect model constraints.
35 Luenberger (1984).
36 Tarantola and Valette (1982) suggested using a linearized search method, iterating from the initial estimate, which must be reasonably close to the correct answer. The method can be quite effective (e.g., Wunsch and Minster, 1982; Mercier *et al.*, 1993). In a wider context, however, their method is readily recognizable as a special case of the many known methods for minimizing a general objective function.
37 Numerical Algorithms Group (2005), Press *et al.* (1996).
38 For simulated annealing, the literature starts with Pincus (1968) and Kirkpatrick *et al.* (1983), and general discussions can be found in van Laarhoven and Aarts (1987), Ripley (2004), and Press *et al.* (1996). A simple oceanographic application to experiment design was discussed by Barth and Wunsch (1990).
39 Goldberg (1989), Holland (1992), Denning (1992).