3

Basic Machinery

The purpose of this chapter is to record a number of results that are essential tools for the discussion of the problems already described. Much of this material is elementary and is discussed here primarily to produce a consistent notation for later use. Reference will be made to some of the good available textbooks. But some of the material is given what may be an unfamiliar interpretation, and I urge everyone to at least skim the chapter.

Our basic tools are those of matrix and vector algebra as they relate to the solution of simultaneous equations, and some elementary statistical ideas mainly concerning covariance, correlation, and dispersion. Least squares is reviewed, with an emphasis placed upon the arbitrariness of the distinction between knowns, unknowns, and noise. The singular-value decomposition is a central building block, producing the clearest understanding of least squares and related formulations. I introduce the Gauss-Markov theorem and its use in making property maps, as an alternative method for obtaining solutions to simultaneous equations, and show its relation to and distinction from least squares. The chapter ends with a brief discussion of recursive least squares and estimation as essential background for the time-dependent methods of Chapter 6.

3.1 Matrix and Vector Algebra

This subject is very large and well developed, and it is not my intention to repeat material better found elsewhere (e.g., Noble & Daniel, 1977; Strang, 1988). Only a brief survey of central results is provided.

A matrix is an $M \times N$ array of elements of the form

$$\mathbf{A} = \{A_{ij}\},\ 1 \leq i \leq M,\ 1 \leq j \leq N\,.$$

Normally a matrix is denoted by a boldface capital letter. A vector is a special case of an $M \times 1$ matrix, written as a boldface lower-case letter, for example, $\mathbf{q}$. Corresponding capital or lower-case letters for Greek symbols are also indicated in boldface. Unless otherwise stipulated, vectors are understood to be column vectors. The transpose of a matrix interchanges its rows and columns. Transposition applied to vectors is sometimes used to save space in printing, for example, $\mathbf{q}^T = [q_1\ q_2 \ldots q_N]^T$ is the same as

$$\mathbf{q} = \begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_N \end{bmatrix} .$$

3.1.1 Matrices and Vectors

The inner, or dot, product between two $L \times 1$ vectors $\mathbf{a}$, $\mathbf{b}$ is written $\mathbf{a}^T\mathbf{b} \equiv \mathbf{a} \cdot \mathbf{b} = \sum_{i=1}^{L} a_i b_i$ and is a scalar. Such an inner product is the projection of $\mathbf{a}$ onto $\mathbf{b}$ (or vice versa). The magnitude of this projection can be measured as

$$\mathbf{a}^T\mathbf{b} = |\mathbf{a}||\mathbf{b}| \cos\phi$$

where $\cos\phi$ ranges between zero, when the vectors are orthogonal, and one, when they are parallel.

Suppose we have a collection of N vectors, $\mathbf{e}_i$, each of dimension N. If it is possible to represent perfectly an arbitrary N–dimensional vector $\mathbf{f}$ as the linear sum

$$\mathbf{f} = \sum_{i=1}^{N} \alpha_i \mathbf{e}_i \,, \tag{3.1.1}$$

then $\mathbf{e}_i$ are said to be a spanning set. A necessary and sufficient condition for them to be a spanning set is that they should be independent–that is, no one of them can be perfectly representable by the others:

$$\mathbf{e}_{j_0} - \sum_{i=1,\, i \neq j_0}^{N} \beta_i \mathbf{e}_i \neq 0, \quad 1 \leq j_0 \leq N \,, \tag{3.1.2}$$

for any choice of β_i.

The expansion coefficients α_i in (3.1.1) are obtained by taking the dot product of (3.1.1) with each of the vectors in turn:

$$\sum_{i=1}^{N} \alpha_i \mathbf{e}_k^T \mathbf{e}_i = \mathbf{e}_k^T \mathbf{f}, \quad 1 \leq k \leq N \,, \tag{3.1.3}$$

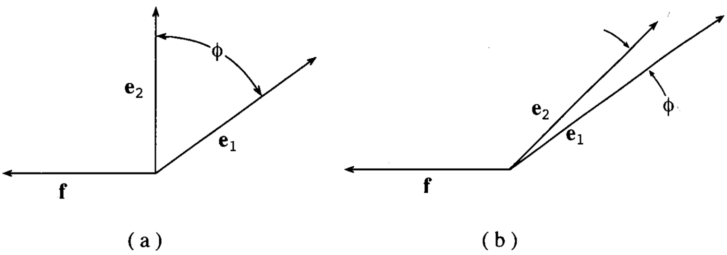

Figure 3–1. An arbitrary two-dimensional vector **f** can be expanded exactly in any two nonparallel vectors $\mathbf{e}_1$, $\mathbf{e}_2$, as in (a). In (b), the angle ϕ is not actually zero, but as it becomes arbitrarily small, it is readily confirmed that the slightest errors in knowledge of **f** render unstable calculation of the expansion coefficients.

a system of N equations in N unknowns. The α_i are most readily found if the $\mathbf{e}_i$ are a mutually orthonormal set–that is, if

$$\mathbf{e}_i^T \mathbf{e}_j = \delta_{ij} ,$$

but this requirement is not necessary for a spanning set. With a spanning set, the information contained in the set of projections, $\mathbf{e}_i^T \mathbf{f} = \mathbf{f}^T \mathbf{e}_i$, is adequate then to determine the α_i and thus all the information required to reconstruct **f**.

The concept of nearly dependent vectors is helpful and can be understood heuristically. Consider Figure 3–1a, in which the space is two-dimensional. Then the two vectors $\mathbf{e}_1$, $\mathbf{e}_2$, as depicted there, are independent and can be used to expand an arbitrary two-dimensional vector **f** in the plane. But if the vectors become nearly parallel, as in Figure 3–1b, as long as they are not exactly parallel, they can still be used mathematically to represent **f** perfectly. However, one anticipates, and we find in practice, that as the angle ϕ between them becomes very small, they are *almost dependent*, and numerical problems arise in finding the expansion coefficients α_1, α_2. The generalization to higher dimensions is left to the reader's intuition.

It has been found convenient and fruitful to define multiplication of two matrices **A**, **B** by the operation $\mathbf{C} = \mathbf{AB}$, such that

$$C_{ij} = \sum_{p=1}^{P} A_{ip} B_{pj} . \tag{3.1.4}$$

For the definition (3.1.4) to make sense, **A** must be an $M \times P$ matrix and **B** must be $P \times N$ (including the special case of $P \times 1$, a column vector). That is, the two matrices must be *conformable.* If two matrices are multiplied, or a matrix and a vector are multiplied, conformability is implied; otherwise one can be assured that an error has been made. Note that $\mathbf{AB} \neq \mathbf{BA}$ even where both products exist, except under special circumstances. For both products to exist, **A** and **B** must be square and of the same dimension.

The mathematical operation in (3.1.4) may appear arbitrary, but a physical interpretation is available: Matrix multiplication is the dot product of all of the rows of **A** with all the columns of **B**.

If we define a matrix, **E**, each of whose columns is the corresponding vector $\mathbf{e}_i$, and a vector, $\boldsymbol{\alpha} = \{\alpha_i\}$, in the same order, the expansion (3.1.1) can be written in the compact form

$$\mathbf{f} = \mathbf{E}\boldsymbol{\alpha}\,. \tag{3.1.5}$$

The transpose of a matrix **A** is written $\mathbf{A}^T$ and is defined as $\{A_{ij}\}^T = A_{ji}$, an interchange of the rows and columns of **A**. A *symmetric matrix* is one for which $\mathbf{A}^T = \mathbf{A}$. The product $\mathbf{A}^T\mathbf{A}$ represents the array of all the dot products of the columns of **A** with themselves, and similarly, $\mathbf{A}\mathbf{A}^T$ represents the set of all dot products of all the rows of **A** with themselves. It follows that $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$. Because we have $(\mathbf{A}\mathbf{A}^T)^T = \mathbf{A}\mathbf{A}^T$, $(\mathbf{A}^T\mathbf{A})^T = \mathbf{A}^T\mathbf{A}$, both these matrices are symmetric ones. [We used $(\mathbf{A}^T)^T = \mathbf{A}$.]

The *trace* of a square $M \times M$ matrix **A** is defined as $\mathrm{trace}(\mathbf{A}) = \sum_i^M A_{ii}$. A *diagonal matrix* is square and zero except for the terms along the main diagonal. The operator $\mathrm{diag}(\mathbf{q})$ makes a square diagonal matrix with **q** along the main diagonal.

The special $L \times L$ diagonal matrix $\mathbf{I}_L$, with $I_{ii} = 1$, is the *identity.* Usually, when the dimension of $\mathbf{I}_L$ is clear from the context, the subscript is omitted. If there is a matrix **B**, such that $\mathbf{BE} = \mathbf{I}$, then **B** is the *left-inverse* of **E**. If **B** is the left inverse of **E** and **E** is square, a standard result is that it must also be a right inverse: $\mathbf{EB} = \mathbf{I}$, **B** is then called *the inverse of* **E** and is usually written $\mathbf{E}^{-1}$. If **E** is not square, such an inverse cannot exist, and special inverses, like a left inverse, are sometimes written $\mathbf{E}^{+}$ and referred to as *generalized inverses.* Some of them will be encountered later. A useful result is that $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$ if the inverses exist. Square matrices with inverses are *nonsingular.*

We need the idea of the *length,* or norm, of a vector. Several choices are possible; for present purposes, the conventional l_2 norm,

$$\|\mathbf{f}\|_2 \equiv (\mathbf{f}^T\mathbf{f})^{1/2} = \left(\sum_{i=1}^{N} f_i^2\right)^{1/2} \tag{3.1.6}$$

is most useful; often the subscript will be omitted. This definition leads in turn to the measure of distance between two vectors, **a**, **b** as

$$\|\mathbf{a} - \mathbf{b}\|_2 = \sqrt{(\mathbf{a} - \mathbf{b})^T(\mathbf{a} - \mathbf{b})}\,, \tag{3.1.7}$$

the familiar Cartesian distance. Distances can also be measured in such a way that deviations of certain elements of $\mathbf{c} = \mathbf{a} - \mathbf{b}$ count for more than others–that is, a metric, or set of weights can be introduced with a definition

$$\|\mathbf{c}\| = \sqrt{\sum_i c_i W_{ii} c_i}\,, \tag{3.1.8}$$

depending upon the importance to be attached to magnitudes of different elements, stretching and shrinking various coordinates. Finally, in the most general form, distance can be measured in a coordinate system both stretched and rotated relative to the original one

$$\|\mathbf{c}\|_W = \sqrt{\mathbf{c}^T \mathbf{W} \mathbf{c}} \tag{3.1.9}$$

where $\mathbf{W}$ is an arbitrary matrix (but usually, for physical reasons, symmetric and positive definite[1]).

Consider a set of M linear equations in N unknowns,

$$\mathbf{E}\mathbf{x} = \mathbf{y}\,. \tag{3.1.10}$$

Because of the appearance of simultaneous equations in situations in which the y_i are observed, and where $\mathbf{x}$ are parameters that we wish to determine, it is often convenient to refer to (3.1.10) as a set of measurements of $\mathbf{x}$ which produced the observations or data, $\mathbf{y}$. If $M > N$, the system is said to be *overdetermined*, or *formally overdetermined.* If $M < N$, it is *underdetermined*, and if $M = N$, it is *just-determined* or *formally just-determined.* The use of the word *formally* has a purpose we will come to later. Knowledge of the matrix inverse to $\mathbf{E}$ would make it easy to solve a set of L equations in L unknowns by left-multiplying (3.1.10) by $\mathbf{E}^{-1}$. The reader is cautioned that although matrix inverses are a very powerful tool, one is usually ill advised to solve large sets of simultaneous equations by inverting the coefficient matrix (e.g., Golub & Van Loan, 1989).

There are several ways to view the meaning of any set of linear simultaneous equations. If the columns of $\mathbf{E}$ continue to be denoted $\mathbf{e}_i$, but without necessarily stipulating that they are either a spanning set or orthogonal, then (3.1.10) is of the form,

$$x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + \cdots + x_n\mathbf{e}_N = \mathbf{y}\,. \tag{3.1.11}$$

The ability to so describe an arbitrary $\mathbf{y}$, or to solve the equations, would thus depend upon whether the $M \times 1$ vector $\mathbf{y}$ can be specified by a sum of N column vectors, $\mathbf{e}_i$–that is, it would depend upon their being a spanning set. In this view, the elements of $\mathbf{x}$ are simply the corresponding expansion

[1] *Positive definite* will be defined later.

coefficients. Depending upon the ratio of M to N–that is, the number of equations to unknown elements–one faces the possibility that there are fewer expansion vectors $\mathbf{e}_i$ than elements of $\mathbf{y}$ $(M > N)$, or that there are more expansion vectors available than elements of $\mathbf{y}$ $(M < N)$. Thus, the overdetermined case corresponds to having *fewer* expansion vectors, and the underdetermined case corresponds to having *more* expansion vectors, than the dimension of $\mathbf{y}$. It is possible that in the overdetermined case, the too-few expansion vectors are not actually independent, so that there are even fewer vectors available than is first apparent. Similarly, in the underdetermined case, there is the possibility that although it appears we have more expansion vectors than required, fewer may be independent than the number of elements of $\mathbf{y}$, and the consequences of that case need to be understood as well.

Alternatively, if the rows of $\mathbf{E}$ are denoted $\mathbf{r}_i^T$, $1 \le i \le M$, (3.1.10) is a set of M-inner products,

$$\mathbf{r}_i^T \mathbf{x} = y_i, \quad 1 \le i \le M. \tag{3.1.12}$$

That is, the set of simultaneous equations is equivalent to being provided with the value of M–dot products of the N–dimensional unknown vector, $\mathbf{x}$, with M known vectors, $\mathbf{r}_i$. Whether that is sufficient information to determine $\mathbf{x}$ depends upon whether the $\mathbf{r}_i$ are a spanning set. In this view, in the overdetermined case, one has *more* dot products available than unknown elements x_i, and in the underdetermined case, there are *fewer* such values than unknowns. (These statements are particularly transparent if the rows or columns happen to be orthonormal vectors, and the reader is urged to examine the relative determinancy in that special situation.)

3.1.2 Identities, Differentiation, and So Forth

Here are some identities and matrix/vector definitions that prove useful.

A square positive definite matrix $\mathbf{A}$ is one for which the scalar quadratic form,

$$J = \mathbf{x}^T \mathbf{A} \mathbf{x}, \tag{3.1.13}$$

is positive for all vectors $\mathbf{x}$. (It suffices to consider only symmetric $\mathbf{A}$ because for a general matrix, $\mathbf{x}^T\mathbf{A}\mathbf{x} = \mathbf{x}^T[(\mathbf{A}+\mathbf{A}^T)/2]\mathbf{x}$, which follows from the scalar property of the quadratic form.) If $J \ge 0$ for all $\mathbf{x}$, $\mathbf{A}$ is positive semidefinite, or nonnegative definite. Linear algebra books show that a necessary and sufficient requirement for positive definiteness is that $\mathbf{A}$ have all positive eigenvalues and a semidefinite one must have all nonnegative eigenvalues.

Nothing has been said about actually finding the numerical values of either the matrix inverse or the eigenvectors and eigenvalues. Computational algorithms for obtaining them have been developed by experts and are discussed in many good textbooks (Lawson & Hanson, 1974; Golub & van Loan, 1989; Press, Flannery, Teukolsky, & Vetterling, 1992; etc.), and software systems like MATLAB implement them in easy-to-use form. For purposes of this book, we assume the reader has at least a rudimentary knowledge of these techniques and access to a good software implementation.

We end up doing a certain amount of differentiation and other operations with respect to matrices and vectors. A number of formulas are very helpful and save a lot of writing. They are all demonstrated by doing the derivatives term-by-term. Let $\mathbf{q}$, $\mathbf{r}$ be $N \times 1$ column vectors, and $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ be matrices. Then if s is any scalar,

$$\frac{\partial s}{\partial \mathbf{q}} = \mathbf{b} = \left\{ \begin{matrix} \frac{\partial s}{\partial q_1} \\ \cdot \\ \cdot \\ \cdot \\ \frac{\partial s}{\partial q_N} \end{matrix} \right\} \tag{3.1.14}$$

is a vector (the gradient). The second derivative of a scalar,

$$\frac{\partial^2 s}{\partial \mathbf{q}^2} = \left\{ \frac{\partial}{\partial q_i} \frac{\partial s}{\partial q_j} \right\} = \left\{ \begin{matrix} \frac{\partial^2 s}{\partial q_1^2} & \frac{\partial^2 s}{\partial q_1 q_2} & \cdot & \cdot & \frac{\partial^2 s}{\partial q_1 q_N} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \frac{\partial^2 s}{\partial q_N q_1} & \cdot & \cdot & \cdot & \frac{\partial^2 s}{\partial q_N^2} \end{matrix} \right\}, \tag{3.1.15}$$

is the *Hessian* of s.

The derivative of one vector by another is a matrix:

$$\frac{\partial \mathbf{r}}{\partial \mathbf{q}} = \left\{ \frac{\partial r_j}{\partial q_j} \right\} = \left\{ \begin{matrix} \frac{\partial r_1}{\partial q_1} & \frac{\partial r_2}{\partial q_1} & \cdot & \frac{\partial r_M}{\partial q_1} \\ \frac{\partial r_1}{\partial q_2} & \cdot & \cdot & \frac{\partial r_M}{\partial q_2} \\ \cdot & \cdot & \cdot & \cdot \\ \frac{\partial r_1}{\partial q_N} & \cdot & \cdot & \frac{\partial r_M}{\partial q_N} \end{matrix} \right\} \equiv \mathbf{P}\,. \tag{3.1.16}$$

If $\mathbf{r}$, $\mathbf{q}$ are of the same dimension, the determinant of $\mathbf{P}$ is the *Jacobian* of $\mathbf{r}$.

Assuming conformability, the inner product

$$\mathbf{r}^T \mathbf{q} = \mathbf{q}^T \mathbf{r}$$

is a scalar, and

$$\frac{\partial(\mathbf{q}^T\mathbf{r})}{\partial\mathbf{q}} = \frac{\partial(\mathbf{r}^T\mathbf{q})}{\partial\mathbf{q}} = \mathbf{r}\,, \tag{3.1.17}$$

$$\frac{\partial(\mathbf{q}^T\mathbf{q})}{\partial\mathbf{q}} = 2\mathbf{q}\,. \tag{3.1.18}$$

For a quadratic form,

$$J = \mathbf{q}^T\mathbf{A}\mathbf{q}$$

$$\frac{\partial J}{\partial\mathbf{q}} = (\mathbf{A}+\mathbf{A}^T)\mathbf{q}\,, \tag{3.1.19}$$

and its Hessian is $\mathbf{A}+\mathbf{A}^T$.

Let **A**, **B**, **C** be square. Then

$$\frac{\partial\mathrm{trace}\mathbf{A}}{\partial\mathbf{A}} = \mathbf{I}\,, \tag{3.1.20}$$

$$\frac{\partial\mathrm{trace}(\mathbf{BAC})}{\partial\mathbf{A}} = \mathbf{B}^T\mathbf{C}^T\,, \tag{3.1.21}$$

$$\frac{\partial\mathrm{trace}(\mathbf{ABA}^T)}{\partial\mathbf{A}} = \mathbf{A}(\mathbf{B}+\mathbf{B}^T)\,. \tag{3.1.22}$$

Rogers (1980) is an entire volume of matrix derivative identities, and many other useful properties are discussed by Magnus and Neudecker (1988).

There are a few, unfortunately unintuitive, matrix inversion identities that are essential to some of the later chapters. Liebelt (1967, Section 1–19) derives them by considering the square, partitioned matrix

$$\begin{Bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{C} \end{Bmatrix} \tag{3.1.23}$$

where $\mathbf{A}^T = \mathbf{A}$, $\mathbf{C}^T = \mathbf{C}$, but **B** can be rectangular of conformable dimensions in (3.1.23). The most important of the identities, sometimes called the *matrix inversion lemma* is, in one form,

$$\{\mathbf{C}-\mathbf{B}^T\mathbf{A}^{-1}\mathbf{B}\}^{-1} = \mathbf{C}^{-1} - \mathbf{C}^{-1}\mathbf{B}^T(\mathbf{B}\mathbf{C}^{-1}\mathbf{B}^T-\mathbf{A})^{-1}\mathbf{B}\mathbf{C}^{-1} \tag{3.1.24}$$

where it is assumed that the inverses exist. (The history of this not-very-obvious identity is discussed by Haykin, 1986, p. 385.) A variant (Liebelt's equation 1–51) is

$$\mathbf{A}\mathbf{B}^T(\mathbf{C}+\mathbf{B}\mathbf{A}\mathbf{B}^T)^{-1} = (\mathbf{A}^{-1}+\mathbf{B}^T\mathbf{C}^{-1}\mathbf{B})^{-1}\mathbf{B}^T\mathbf{C}^{-1}\,. \tag{3.1.25}$$

Both (3.1.24)–(3.1.25) are readily confirmed by direct multiplication, for example, by showing that $\mathbf{A}\mathbf{B}^T(\mathbf{C}+\mathbf{B}\mathbf{A}\mathbf{B}^T)^{-1}$ times the right-hand side of (3.1.25) is the identity.

Another identity, found by completing the square, is demonstrated by directly multiplying it out and requires $\mathbf{C} = \mathbf{C}^T$ ($\mathbf{A}$ is unrestricted, but the matrices must be conformable as shown):

$$\mathbf{ACA}^T - \mathbf{BA}^T - \mathbf{AB}^T = (\mathbf{A} - \mathbf{BC}^{-1})\mathbf{C}(\mathbf{A} - \mathbf{BC}^{-1})^T - \mathbf{BC}^{-1}\mathbf{B}^T \,. \quad (3.1.26)$$

A number of useful definitions of a matrix norm exist. For present purposes the so-called spectral norm or 2–norm defined as

$$\|\mathbf{A}\|_2 = \sqrt{\text{maximum eigenvalue of } (\mathbf{A}^T\mathbf{A})} \qquad (3.1.27)$$

is usually adequate. Without difficulty (e.g., Haykin, 1986, p. 61), it may be seen that this definition is equivalent to

$$\|\mathbf{A}\|_2 = \max \frac{\mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x}}{\mathbf{x}^T\mathbf{x}} = \max \frac{\|\mathbf{A}\mathbf{x}\|_2}{\|\mathbf{x}\|_2} \qquad (3.1.28)$$

where the maximum is defined over all vectors $\mathbf{x}$. Another useful measure is the Frobenius norm,

$$\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^{M} \sum_{j=1}^{N} A_{ij}^2} = \sqrt{\text{trace}(\mathbf{A}^T\mathbf{A})} \,. \qquad (3.1.29)$$

Neither definition requires $\mathbf{A}$ to be square.

These norms permit one to derive various useful results. Consider one illustration. $\mathbf{Q}$ is square, and $\|\mathbf{Q}\| < 1$, then

$$(\mathbf{I} + \mathbf{Q})^{-1} = \mathbf{I} - \mathbf{Q} + \mathbf{Q}^2 - \cdots \,, \qquad (3.1.30)$$

which may be verified by multiplying both sides by $\mathbf{I} + \mathbf{Q}$, doing term-by-term multiplication and measuring the remainders with either norm.

3.1.3 Gram-Schmidt Process

One often has a set of p-independent but nonorthonormal vectors $\mathbf{h}_i$, and it is convenient to find a new set $\mathbf{g}_i$, which are orthonormal. The *Gram-Schmidt process* operates by induction. Suppose we have orthonormalized the first k of the $\mathbf{h}_i$ to a new set, $\mathbf{g}_i$, and wish to generate the $k+1$st. Let

$$\mathbf{g}_{k+1} = \mathbf{h}_{k+1} - \sum_{j}^{k} \gamma_j \mathbf{g}_j \,. \qquad (3.1.31)$$

Because $\mathbf{g}_{k+1}$ must be orthogonal to the preceding $\mathbf{g}_i$, $i = 1, k$, we take the dot products of (3.1.31) with each of these vectors, producing a set of simultaneous equations for determining the unknown γ_j. The resulting $\mathbf{g}_{k+1}$ is easily given unit norm by division by its length.

If one has the first k of N necessary vectors, one needs an additional $N-k$ independent vectors h_i. There are several possibilities. One might simply generate the necessary vectors by filling their elements with random numbers. Or one might take a very simple trial set like $\mathbf{h}_{k+1} = [1\ \ 0\ \ 0\ \ \ldots\ \ 0]^T$, $\mathbf{h}_{k+2} = [0\ \ 1\ \ .\ \ .\ \ 0], \ldots.$ If one is unlucky, the set might prove not to be independent. But a simple numerical perturbation usually suffices to render them so. In practice, the algorithm is changed to what is usually called the *modified Gram-Schmidt process* (see Lawson & Hanson, 1974) for purposes of numerical stability.

3.2 Simple Statistics; Regression

Some statistical ideas are required, but the discussion is confined to stating some basic notions and to developing a notation. A statistics text such as Cramér (1946), or one on regression such as Seber (1977), should be consulted for real understanding.

We require the idea of a probability density for a random variable x. This subject is a very deep one–as described, for example, by Feller (1957) and Jeffreys (1961)–but our approach will be heuristic. Suppose that an arbitrarily large number of experiments can be conducted for the determination of the values of x, denoted X_i, $1 \leq i$, and a histogram of the experimental values found. The frequency function, or probability density, will be defined as the limit, supposing it exists, of the histogram per bin width of an arbitrarily large number of experiments divided into bins of arbitrarily small value ranges. Let the corresponding limiting density or frequency function be denoted $p_x(X)$. (This notation distinguishes between a random variable, x, and the numerical values it assumes, X. The distinction is not always preserved where the context prevents confusion.) The *average*, or *mean*, or *expected value* is denoted $< x >$ and defined as

$$< x > \equiv \int_{\text{all } X} X p_x(X) dX \tag{3.2.1}$$

and is the center of mass of $p_x(X)$. Using the definition of a frequency function, it is easy to show heuristically that the *sample average* or *mean*,

$$< x >_N \equiv \frac{1}{N} \sum_{i=1}^{N} X_i \, , \tag{3.2.2}$$

when it exists, will usually asymptotically approach $< x >$ in the limit as N approaches infinity. Knowledge of the true mean value of a random variable is commonly all that we are willing to assume known. If forced to forecast the numerical value of x under such circumstances, often the best we can do is to employ $< x >$. If the deviation from the true mean is denoted x' so that $x = < x > + x'$, such a forecast has the virtue that we are assured the average forecast error, $< x' >$, would be zero if many such forecasts are made. The bracket operation is very important throughout this book; it has the property that if a is a nonrandom quantity, $< ax >= a < x >$.

The idea of a frequency function generalizes easily to two or more random variables, x, y. We can in concept do an arbitrarily large number of experiments in which we count the occurrences of differing pair values, (X_i, Y_i), of x, y and make a histogram, dividing by the bin area, and taking the limit to produce a joint probability density, $p_{xy}(X, Y)$. A simple example would be the simultaneous measurement by a current meter of the two components of horizontal velocity.

An important use of joint probability densities is in what is known as *conditional probability.* Suppose that the joint probability density for x, y is known and furthermore, $y = Y$–that is, information is available concerning the actual value of y. What then is the probability density for x given that a particular value for y is known to have occurred? This new frequency function is usually written as $p_{x|y}(X|Y)$ and is read as "the probability of x, given that y has occurred with value Y." It follows immediately from the definition of the probability density that

$$p_{x|y}(X|Y) = \frac{p_{xy}(X, Y)}{p_y(Y)} . \tag{3.2.3}$$

(This equation is readily interpreted by going back to the original experimental concept and understanding the restriction on x given that y is known to lie within one of the bins.)

If one finds that $p_{xy}(X, Y) = p_x(X)p_y(Y)$, then x, y are said to be *independent.* Using the joint frequency function, define the average product as

$$< xy >= \int\int_{\text{all } X,Y} XY p_{xy}(X, Y)dXdY . \tag{3.2.4}$$

Should $< (x- < x >)(y - < y >) > \neq 0$, x, y are said to *covary* or to be *correlated.* From the definition of frequency function and the bracket operation, if x, y are independent, then $< (x- < x >)(y - < y >) > = 0$. Under these circumstances x, y are *uncorrelated* or do not covary. Independence implies lack of correlation, but the reverse is not necessarily true. If the two variables are independent, then (3.2.3) is

$$p_{x|y}(X|Y) = p_x(X) ; \tag{3.2.5}$$

that is, knowledge of the value of y does not change the probability density for x–a sensible result–and there is then no predictive power for one variable given knowledge of the other.

We need the idea of dispersion–the expected or average squared value of some quantity about some interesting value, like its mean. The most familiar measure of dispersion is the variance, already used above, the expected fluctuation of a random variable about its mean:

$$\sigma_x^2 =< (x - < x >)^2 > .$$

More generally, define the dispersion of any random variable x as

$$D^2(x) = < x^2 > .$$

Thus, $\sigma_x^2 = D^2(x - < x >)$.

Sample estimates of quantities like the mean and other properties of random variables made from observations occur throughout science. In the case of the sample mean, it is possible to show without difficulty that the expected value of $< x >_N$ is the true average–that is, $<< x >_N > = < x >$. The interpretation is that for finite N, we do not expect that the sample mean will equal the true mean, but that if we could produce sample averages from distinct groups of observations, the sample averages would themselves have an average that would fluctuate about the true mean.

The variance of the sample mean (3.2.2) is easily shown to be

$$D^2(< x >_N - << x >_N>) = \frac{\sigma_x^2}{N} , \tag{3.2.6}$$

and thus the dispersion diminishes with N.

There are many sample estimates, however, some of which we encounter, where the expected value of the sample estimate is not equal to the true estimate. Such an estimator is said to be *biased.* Otherwise, it is *unbiased.* The simplest example of a biased estimator is the sample variance if defined as

$$s^2 \equiv \frac{1}{N} \sum_i^N (X_i - < x >_N)^2 . \tag{3.2.7}$$

For simplicity, but without loss of generality, assume the true mean of x is zero, $< x > = 0$ (a nonzero mean can be removed first if necessary). Equation (3.2.7) is

$$s^2 = \frac{1}{N}\left(\sum_i^N x_i^2\right) - \frac{< x >_N^2}{N} ,$$

whose expected value is

$$< s^2 > = \sigma_x^2 - \frac{\sigma_x^2}{N} = \sigma_x^2 \frac{N-1}{N}\,, \tag{3.2.8}$$

using (3.2.6) and the zero true mean. Thus, in the definition (3.2.8) the expected value of the sample variance is not the correct value but is rather $(N-1)/N$ times it. To remove the bias, one often redefines the sample variance as

$$s^2 = \frac{1}{N-1} \sum_{i=1}^{N} (X_i - < x >_N)^2 \,. \tag{3.2.9}$$

Suppose there are two random variables x, y between which there is anticipated to be some linear relationship

$$x = ay + n \tag{3.2.10}$$

where n represents any contributions to x that remain unknown despite knowledge of y. Then

$$< x > = a < y > + < n >\,, \tag{3.2.11}$$

and (3.2.10) shows

$$x - < x > = a(y - < y >) + (n - < n >)\,,$$

or

$$x' = ay' + n', \quad x' = x - < x >, \quad \text{etc.} \tag{3.2.12}$$

From this last equation,

$$a = \frac{< x'y' >}{< y'^2 >} = \frac{< x'y' >}{(< y'^2 >< x'^2 >)^{1/2}} \frac{< x'^2 >^{1/2}}{< y'^2 >^{1/2}} = \rho \frac{< x'^2 >^{1/2}}{< y'^2 >^{1/2}} \tag{3.2.13}$$

where it was supposed $< y'n' > = 0$, thus defining n. The quantity

$$\rho \equiv \frac{< x'y' >}{(< y'^2 >< x'^2 >)^{1/2}} \tag{3.2.14}$$

is the *correlation coefficient* and is easily shown to have the property $|\rho| \leq 1$. If ρ should vanish, then so does a. If a vanishes, then knowledge of y' carries no information about the value of x'. If $\rho = \pm 1$, then it follows from the definitions that $n = 0$, and knowledge of a permits perfect prediction of x' from knowledge of y' (because probabilities are being used, rigorous usage would state "perfect prediction almost always," but this distinction will be ignored).

A measure of how well the prediction of x from y will work can be obtained in terms of the variance of x'. We have

$$< x'^2 > = a^2 < y'^2 > + < n'^2 > = \rho^2 < x'^2 > + < n'^2 >$$

or

$$(1 - \rho^2) < x'^2 > = < n'^2 > ; \tag{3.2.15}$$

that is, the fraction of the variance in x' that is unpredictable by y' is $(1 - \rho^2) < x'^2 >$ and is the *unpredictable power.* Conversely, $\rho^2 < x'^2 >$ is the *predictable power.* The limits as $\rho \to 0$, 1 are readily apparent.

Thus, we interpret the statement that two variables x', y' are correlated or covary to mean that knowledge of one permits at least a partial prediction of the other, the expected success of the prediction depending upon the size of ρ. This result represents an implementation of the statement that if two variables are not independent, then knowledge of one permits some skill in the prediction of the other. If two variables do not covary but are also not independent, a linear model like (3.2.10) would not be useful and some nonlinear one would be required. Such nonlinear methods are possible and are touched on briefly later. The idea that correlation or covariance between various physical quantities carries useful predictive skill between them is an essential ingredient of many of the methods taken up in this book.

If a sequence of pairs of values x_i, y_i is measured so that we have a set of simultaneous equations

$$ay_i + n_i = x_i\,, \tag{3.2.16}$$

we might think to use these equations to determine $< x >$, $< y >$, a, $< n >$, n', etc. This leads into the huge subject of regression analysis (see, for example, Seber, 1977; or Draper & Smith, 1982), which is necessary to understand the connection between the theoretical values of quantities like $< x >$ and their sample values computed from objects like $< x >_N$. Some more machinery is required to do so, which we will eventually obtain in part.

In the absence of other information, the Gaussian, or normal, probability density is often invoked to describe observations. Apart from its comparatively simple mathematical properties, justification for the assumption of normality lies with the so-called Central Limit Theorem (Cramér, 1946). This theorem, which can be proven under hypotheses of varying strength, shows that under general circumstances, phenomena that are the result of summing many independent stochastic phenomena will tend toward a normal distribution. But not all physical phenomena conform to the assumptions of the Central Limit Theorem; in particular, the ocean is demonstrably nonnormal in many respects, making the Gaussian hypothesis a dangerous one. Nonetheless, it is worth recalling the fundamental properties of the normal probability density. For a single variable x, it is defined as

$$p_x(X) = \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left(-\frac{(X-m_x)^2}{2\sigma_x^2}\right)$$

[sometimes abbreviated as $G(m_x,\, \sigma_x)$]. It is readily confirmed that $< x > = m_x$, $< (x- < x >)^2 > = \sigma_x^2$. Suppose that x, y are *independent* Gaussian variables $G(m_x,\, \sigma_x)$, $G(m_y,\, \sigma_y)$. Then their joint probability density is just the product of the two individual densities,

$$p_{xy}(X,Y) = \frac{1}{2\pi\sigma_x\sigma_y} \exp\left(-\frac{(X-m_x)^2}{2\sigma_x^2} - \frac{(Y-m_y)^2}{2\sigma_y^2}\right) . \qquad (3.2.17)$$

We need to consider the probability density for normal variables that are correlated. Let two new random variables, ξ_1, ξ_2, be defined as a linear combination of x, y,

$$\begin{aligned}\xi_1 &= a_{11}(x-m_x) + a_{12}(y-m_y) + m_{\xi_1}\\ \xi_2 &= a_{21}(x-m_x) + a_{22}(y-m_y) + m_{\xi_2}\end{aligned}$$

or

$$\xi = \mathbf{A}(\mathbf{x} - \mathbf{m}_x) + \mathbf{m}_\xi \qquad (3.2.18)$$

where $\mathbf{x} = \{x, y\}^T$, $\mathbf{m}_x = [m_x, m_y]^T$, $\mathbf{m}_\xi = [m_{\xi_1}, m_{\xi_2}]^T$. What is the probability density for these new variables? The general rule for changes of variable in probability densities follows from area conservation in mapping from x, y space to ξ_1, ξ_2 space–that is,

$$p_{\xi_1\xi_2}(\Xi_1, \Xi_2) = p_{xy}(X(\Xi_1,\Xi_2),\, Y(\Xi_1,\Xi_2))\frac{\partial(X,Y)}{\partial(\Xi_1,\Xi_2)} \qquad (3.2.19)$$

where $\partial(X,Y)/\partial(\Xi_1,\Xi_2)$ is the Jacobian of the transformation between the two variable sets, and the numerical values satisfy the functional relations,

$$\Xi_1 = a_{11}(X-m_x) + a_{12}(Y-m_y) + m_{\xi_1}\,,$$

etc. Suppose that the relationship (3.2.18) is invertible–that is, we can solve for

$$\begin{aligned}x &= b_{11}(\xi_1 - m_{\xi_1}) + b_{12}(\xi_2 - m_{\xi_2}) + m_x\\ y &= b_{21}(\xi_1 - m_{\xi_1}) + b_{22}(\xi_2 - m_{\xi_2}) + m_y\end{aligned}$$

or

$$\mathbf{x} = \mathbf{B}(\boldsymbol{\xi} - \mathbf{m}_\xi) + \mathbf{m}_x \,. \tag{3.2.20}$$

Then the Jacobian of the transformation is

$$\frac{\partial(X,Y)}{\partial(\Xi_1,\Xi_2)} = b_{11}b_{22} - b_{12}b_{21} = \det(\mathbf{B}) \tag{3.2.21}$$

[det(**B**) is the determinant of **B**]. Equation (3.2.18) produces

$$\begin{aligned} < \xi_1 > &= m_{\xi_1} \\ < \xi_2 > &= m_{\xi_2} \\ < (\xi_1 - < \xi_1 >)^2 > &= a_{11}^2\sigma_x^2 + a_{12}\sigma_y^2 \\ < (\xi_1 - < \xi_1 >)(\xi_2 - < \xi_2 >) > &= a_{11}a_{21}\sigma_x^2 + a_{12}a_{22}\sigma_y^2 \neq 0 \,. \end{aligned} \tag{3.2.22}$$

In the special case,

$$\mathbf{A} = \begin{Bmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{Bmatrix}, \quad \mathbf{B} = \begin{Bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{Bmatrix}, \tag{3.2.23}$$

the transformation (3.2.23) is a simple coordinate rotation through angle ϕ, and the Jacobian is 1. The second-order moments in (3.2.22) then become

$$< (\xi_1 - < \xi_1 >)^2 >= \sigma_{\xi_1}^2 = \cos^2\phi\sigma_x^2 + \sin^2\phi\sigma_y^2 \,, \tag{3.2.24}$$

$$< (\xi_2 - < \xi_2 >)^2 >= \sigma_{\xi_2}^2 = \sin^2\phi\sigma_x^2 + \cos^2\phi\sigma_y^2 \,, \tag{3.2.25}$$

$$< (\xi_1 - < \xi_1 >)(\xi_2 - < \xi_2 >) >\equiv \mu_{\xi_1\xi_2} = (\sigma_y^2 - \sigma_x^2)\cos\phi\sin\phi \,. \tag{3.2.26}$$

The new probability density is

$$p_{\xi_1\xi_2}(\Xi_1,\Xi_2) = \frac{1}{2\pi\sigma_{\xi_1}\sigma_{\xi_2}\sqrt{1-\rho_\xi^2}} \times \tag{3.2.27}$$

$$\exp\left\{ -\frac{1}{2\sqrt{1-\rho_\xi^2}} \left[\frac{(\Xi_1 - m_{\xi_1})^2}{\sigma_{\xi_1}^2} - \frac{2\rho_\xi(\Xi_1 - m_{\xi_1})(\Xi_2 - m_{\xi_2})}{\sigma_{\xi_1}\sigma_{\xi_2}} + \frac{(\Xi_2 - m_{\xi_2})^2}{\sigma_{\xi_2}^2} \right] \right\}$$

where $\rho_\xi = (\sigma_y^2 - \sigma_x^2)\sin\phi\cos\phi / \left(\sigma_{\xi_1} + \sigma_{\xi_2}\right)^{1/2} = \mu_{\xi_1\xi_2}/\sigma_{\xi_1}\sigma_{\xi_2}$ is the correlation coefficient of the new variables. A probability density derived through a linear transformation from two independent variables that are Gaussian will be said to be jointly Gaussian, and (3.2.27) is a canonical form. Because a coordinate rotation is invertible, it is important to note that if we had two random variables ξ_1, ξ_2 that were jointly Gaussian with $\rho \neq 1$, then we could find a pure rotation (3.2.23), which produces two other variables x, y that are uncorrelated and therefore *independent.* Notice that (3.2.26) shows that two such uncorrelated variables x, y will necessarily have different variances; otherwise, ξ_1, ξ_2 would have zero correlation, too.

It follows that two uncorrelated jointly Gaussian random variables are also independent. This property is one of the reasons Gaussians are so nice to work with.

3.2.1 Vector Random Processes

Simultaneous discussion of two random processes, x, y can be regarded as discussion of a vector random process $[x, y]^T$ and suggests a generalization to N dimensions. Let us label N random processes as $x(i)$ and define them as the elements of a vector $\mathbf{x}^T = [x(1), x(2), \ldots, x(N)]^T$. Then the mean is a vector: $<\mathbf{x}> = \mathbf{m}_x$, and the covariance is a matrix:

$$\mathbf{C}_{xx} = D^2(\mathbf{x}- <\mathbf{x}>) = <(\mathbf{x}- <\mathbf{x}>)(\mathbf{x}- <\mathbf{x}>)^T>, \qquad (3.2.28)$$

which is necessarily symmetric and positive semidefinite. The cross-covariance of two processes $\mathbf{x}$, $\mathbf{y}$ is

$$\mathbf{C}_{xy} = <(\mathbf{x}- <\mathbf{x}>)(\mathbf{y}- <\mathbf{y}>)^T> \qquad (3.2.29)$$

and $\mathbf{C}_{xy} = \mathbf{C}_{yx}^T$.

It proves convenient to introduce two further moment matrices in addition to the covariance matrices for which the dispersion is measured about the mean. The *second-moment* matrices will be defined as

$$\mathbf{R}_{xx} \equiv D^2(\mathbf{x}) = <\mathbf{x}\mathbf{x}^T>, \quad \mathbf{R}_{xy} = <\mathbf{x}\mathbf{y}^T>$$

($\mathbf{R}_{xy} = \mathbf{R}_{yx}^T$, etc.). Let $\tilde{\mathbf{x}}$ be an estimate of the true value, $\mathbf{x}$. Then the dispersion of $\tilde{\mathbf{x}}$ about the true value will be called the *uncertainty* (sometimes it is called the *error covariance*) and is

$$\mathbf{P} \equiv D^2(\tilde{\mathbf{x}} - \mathbf{x}) = <(\tilde{\mathbf{x}} - \mathbf{x})(\tilde{\mathbf{x}} - \mathbf{x})^T> . \qquad (3.2.30)$$

An intuitively pleasing requirement for an estimator $\mathbf{x}$ is that it should minimize the variance about the true value–that is, minimize the diagonal elements of $\mathbf{P}$. This choice is an aesthetic one, but it is the one we will use extensively later.

If there are N variables, ξ_i, $1 \le i \le N$, they will be said to have an "N-dimensional jointly normal probability density" if it is of the form

$$p_{\xi_1,\ldots,\xi_N}(\Xi_1, \ldots, \Xi_N) = \frac{\exp -\frac{1}{2}(\boldsymbol{\Xi} - \mathbf{m})^T \mathbf{C}_{\xi\xi}^{-1}(\boldsymbol{\Xi} - \mathbf{m})}{(2\pi)^{N/2}\sqrt{\det(\mathbf{C}_{\xi\xi})}} . \qquad (3.2.31)$$

It is readily demonstrated that $<\boldsymbol{\xi}>= \mathbf{m}$, $<(\boldsymbol{\xi}-\mathbf{m})(\boldsymbol{\xi}-\mathbf{m})^T> = \mathbf{C}_{\xi\xi}$. Equation (3.2.27) is a special case of (3.2.31) for $N=2$.

Positive definite symmetric matrices can be factored as

$$\mathbf{C}_{\xi\xi} = \mathbf{C}_{\xi\xi}^{T/2}\mathbf{C}_{\xi\xi}^{1/2}\,, \tag{3.2.32}$$

called the *Cholesky decomposition*, where $\mathbf{C}_{\xi\xi}^{1/2}$ is upper triangular and nonsingular. Numerical schemes for finding $\mathbf{C}_{\xi\xi}^{1/2}$ are described by Lawson and Hanson (1974) and Golub and Van Loan (1989). It follows that the transformation (a rotation and stretching),

$$\mathbf{x} = \mathbf{C}_{\xi\xi}^{-T/2}(\boldsymbol{\xi}-\mathbf{m})\,, \tag{3.2.33}$$

produces new variables $\mathbf{x}$ of zero mean, and identity covariance–that is, a probability density

$$\begin{aligned} p_{x_1,\ldots,x_N}(X_1,\ldots,X_N) &= \frac{\exp -\frac{1}{2}(X_1^2+\cdots X_N^2)}{(2\pi)^{N/2}} \\ &= \frac{\exp\left(-\frac{1}{2}X_1^2\right)}{(2\pi)^{1/2}}\cdots\frac{\exp\left(-\frac{1}{2}X_N^2\right)}{(2\pi)^{1/2}}\,, \end{aligned} \tag{3.2.34}$$

which factors into N independent, normal variates of zero mean and unit variance ($\mathbf{C}_{xx}=\mathbf{R}_{xx}=\mathbf{I}$). Such a process is often denoted *white noise*. (Cramér, 1946, discusses what happens when the determinant of $\mathbf{C}_{\xi\xi}$ vanishes–that is, if $\mathbf{C}_{\xi\xi}$ is singular.)

3.2.2 Functions of Random Variables

If the probability density of x is $p_x(x)$, then the mean of a function of x, $g(x)$ is just

$$< g(x) > = \int_{-\infty}^{\infty} g(X)p_x(X)dX\,, \tag{3.2.35}$$

which follows from the definition of the probability density as the limit of the outcome of a number of trials. The probability density for g regarded as a new random variable is given by (3.2.19) as

$$p_g(G) = p_x(X(G))\frac{dx}{dg} \tag{3.2.36}$$

where the Jacobian is just dx/dg for a one-dimensional transformation.

An important special case is $g=x^2$ where x is Gaussian of zero mean and unit variance (any Gaussian variable z of mean m and variance σ^2 can be transformed to one of zero mean and unit variance by the transformation

$$x = \frac{z-m}{\sigma},$$

whose Jacobian is very simple). Then the probability density of g is

$$p_g(G) = \frac{1}{G^{1/2}\sqrt{2\pi}} \exp(-G/2), \qquad G \geq 0, \tag{3.2.37}$$

a probability density usually denoted as χ_1^2 (chi-squared), and $< g > = 1$, $D^2(g- < g >) = 2$.

3.2.3 Sums of Random Variables

It is often helpful to be able to compute the probability density of sums of independent random variables. The procedure for doing so is based upon (3.2.35). Let x be a random variable, and consider the expected value of the function e^{ixt}:

$$< e^{ixt} > = \int_{-\infty}^{\infty} e^{iXt} p_x(X) dX \equiv \phi_x(t), \tag{3.2.38}$$

which is also the Fourier transform of $p_x(X)$; $\phi_x(t)$ is usually termed the *characteristic function* of x. Now consider the sum of two independent random variables x, y with probability densities p_x, p_y, respectively, and define a new random variable $z = x + y$. What is the probability density of z? A method for finding it is based upon first determining the characteristic function, $\phi_z(t)$ for z and then using the Fourier inversion theorem to obtain $p_x(Z)$. To obtain ϕ_z,

$$\phi_z(t) = < e^{izt} > = < e^{i(x+y)t} > = < e^{ixt} >< e^{iyt} >$$

where the last step depends upon the independence assumption. This last equation shows

$$\phi_z(t) = \phi_x(t)\phi_y(t). \tag{3.2.39}$$

That is, the characteristic function for a sum of two independent variables is the product of the characteristic functions. The *convolution theorem* (see, for example, Bracewell, 1978) asserts that the Fourier transform (forward or inverse) of a product of two functions is the convolution of the Fourier transforms of the two functions. We will not explore this relation in any detail, leaving the reader to pursue the subject in the references (e.g., Cramér, 1946). But it follows immediately that the multiplication of the characteristic functions of a sum of independent Gaussian variables produces a new variable, which is also Gaussian, with a mean equal to the sum of the means and a variance that is the sum of the variances ("sums of Gaussians are Gaussian"). It also follows immediately from the convolution theorem that if a variable ξ is defined as

$$\xi = x_1^2 + x_2^2 + \cdots + x_v^2 \qquad (3.2.40)$$

where each x_i is Gaussian of zero mean and unit variance, the probability density for ξ is

$$p_\xi(\Xi) = \frac{\Xi^{\nu/2-1}}{2^{\nu/2}\Gamma\left(\frac{\nu}{2}\right)} \exp(-\Xi/2)\,, \qquad (3.2.41)$$

known as χ_v^2, "chi-square with ν degrees of freedom." The chi-square probability density is central to the discussion of the sizes of vectors, such as $\tilde{\mathbf{n}}$, measured as $\tilde{\mathbf{n}}^T\tilde{\mathbf{n}} = \sum_i \tilde{\mathbf{n}}_i^2$ if the elements of $\tilde{\mathbf{n}}$ can be assumed to be independent and Gaussian. Equation (3.2.37) is the special case $\nu = 1$. One has,

$$< \xi > = \nu\,, \qquad D^2(\xi - < \xi >) = 2\nu\,. \qquad (3.2.42)$$

3.2.4 Degrees of Freedom

The number of independent variables described by a probability density is usually called the *number of degrees of freedom.* Thus, the densities in (3.2.31) and (3.2.34) have N degrees of freedom, and (3.2.41) has ν of them. If a sample average (3.2.2) is formed, it is said to have N degrees of freedom if each of the x_j is independent. But what if the x_j have a covariance $\mathbf{C}_{xx}$ that is nondiagonal? This question of how to interpret averages of correlated variables will be explicitly discussed in Section 3.5.

Consider for the moment only the special case of the sample variance (3.2.9), with divisor $N-1$ rather than N as might be expected. The reason is that even if the sample values $x(i)(\equiv x_i)$ are independent [we are not distinguishing here between $x(i)$ and $X(i)$], the presence of the sample average in the sample variance means that there are only $N-1$ independent terms in the sum. That this is so is most readily seen by examining the two-term case. Two samples produce a sample mean, $< x >_2 = (x_1 + x_2)/2$. A two-term sample variance is

$$s^2 = \tfrac{1}{2}((x_1 - < x >_2)^2 + (x_2 - < x >_2)^2)\,,$$

but knowledge of x_1 and the sample average permits perfect prediction of x_2 and thus of the second term in the sample variance, and there is just one independent piece of information in the two-term sample variance.

3.2.5 Stationarity

Consider a vector random variable with elements $x_i = x(i)$ where the argument i denotes a position in time or space. Then $x(i)$, $x(j)$ denote two different random variables–for example, the temperature at two different positions in the ocean, or the temperature at two different times at the same position. If the physics governing these two different random variables are independent of the parameter i (i.e., independent of time or space), then $x(i)$ is said to be *stationary*, meaning that the underlying statistics are independent of i. Specifically, $< x(i) > = < x(j) > \equiv < x >$, $D^2(x(i)) = D^2(x(j))$, etc. Furthermore, $x(i)$, $x(j)$ have a covariance

$$C_{xx}(i,j) = < (x(i) - < x(i) >)(x(j) - < x(j) >) > = C_{xx}(|i-j|)\,, \tag{3.2.43}$$

that is, independent of i, j, and depending only upon the difference $|i-j|$; $|i-j|$ is often called the *lag*. Then

$$< (x(i) - < x >)(x(j) - \ < x >)^T > = \{C_{xx}(i,j)\} = \{C_{xx}(|i-j|)\}$$

is called the *autocovariance* of $\mathbf{x}$ or just the covariance, because we now regard $x(i)$, $x(j)$ as intrinsically the same process.[2] If C_{xx} does not vanish, then by the discussion above, knowledge of the numerical value of $x(i)$ implies some predictive skill for $x(j)$ and vice versa–a result of great importance when we examine map making and objective analysis. A jointly normal stationary time series would have probability density (3.2.31) in which all the elements of $\mathbf{m}$ are identical, and the ij-th elements of $\mathbf{C}_{\xi\xi}$ depend only upon $i-j$.

3.2.6 Maximum Likelihood

Given a set of observations with known joint probability density, one can base a method for estimating various sample parameters upon a principle of maximum likelihood, which finds those parameters that render the actual observations to be the most probable ones. Consider one simple example for an uncorrelated jointly normal stationary time series,

$$< x(i) > = m, \ < (x(i) - m)(x(j) - m) > = \sigma_x^2 \delta_{ij}\,,$$

[2] If the means and variances are independent of i, j and the first cross-moment is dependent only upon $|i-j|$, the process x is said to be stationary in the *wide sense*. If all higher moments also depend only on $|i-j|$, the process is said to be stationary in the *strict sense*, or more simply, just *stationary*. A Gaussian process has the unusual property that wide-sense stationarity implies strict-sense stationarity.

with corresponding joint probability density

$$p_{\mathbf{x}}(x(1), x(2), x(3), \ldots, x(N)) = \frac{1}{(2\pi)^{N/2}\sigma_x^N} \times$$
$$\exp\left(-\frac{1}{2\sigma_x^2}\left[(x(1)-m)^2+(x(2)-m)^2+\cdots+(x(N)-m)^2\right]\right) . \quad (3.2.44)$$

Substitution of the observed values into (3.2.44)[3] permits evaluation of the probability that these particular values occurred. Denote this probability as L. One can demand those values of m, σ_x, rendering the probability a maximum of L for all possible series mean and standard deviations. The probability can be maximized by minimizing the exponent in (3.2.44)–that is, minimizing

$$\log L = \frac{1}{2\sigma_x^2}\sum (x(i)-m)^2 - \frac{1}{2}N\log(2\sigma_x) - \frac{1}{2}N\log(2\pi) , \quad (3.2.45)$$

the log-likelihood function. Maximizing log L with respect to m, σ_x produces

$$\tilde{m} = \frac{1}{N}\sum_1^N x(i),\ \tilde{\sigma}_x^2 \equiv s^2 = \frac{1}{N}\sum_1^N (x(i)-\tilde{m})^2 , \quad (3.2.46)$$

and the result is the usual sample average and the biased estimate of the sample variance (3.2.7). A likelihood function derived from (3.2.31) provides a straightforward generalization to covarying variables.

A complete methodology for most of what follows in this book can be built upon the general ideas of maximum likelihood estimation, but it is not the course I choose to follow. Extended discussions can be found in numerous places, including Van Trees (1968), who carries the idea all the way through the material found here in Chapter 6.

3.3 Least Squares

Much of what follows in this book can be described using very elegant and powerful mathematical tools. On the other hand, by restricting ourselves to discrete models and finite numbers of measurements, almost everything can also be viewed as a form of ordinary least squares. It is thus useful to go back and review what "everyone knows" about this most-familiar of all approximation methods.

[3] Strictly speaking, we should work with the conditional probability (3.2.3), $-p_{x|m,\sigma}\big(x(1), x(2)\ldots\ x(N)|m,\sigma\big)$.

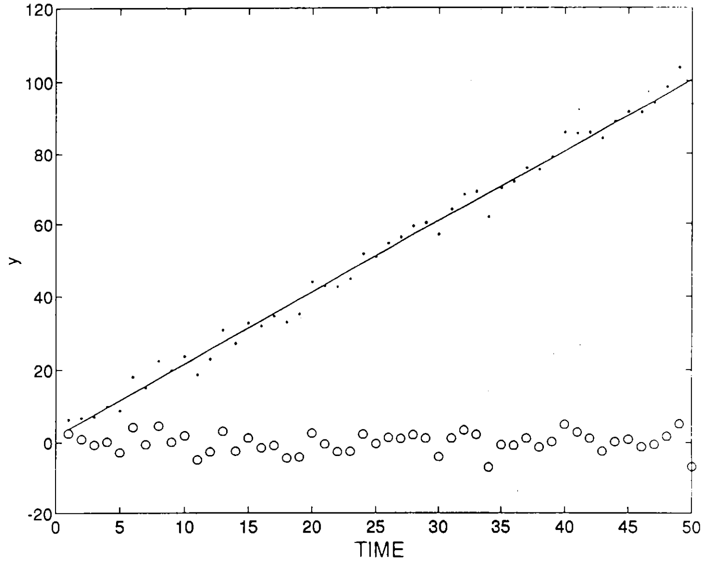

Figure 3–2. "Data" generated through the rule $y = 1 + 2t + n$, where $< n > = 0$, $< n_i n_j > = 9\delta_{ij}$, denoted by small dots. Solid line shows a least-squares fit to the data, which is $\tilde{y} = 1.9 \pm 0.8 + (1.96 \pm 0.03)t$ with $< \tilde{a}\tilde{b} > = -0.02$; open circles denote the residuals of the fit, which appear to be qualitatively white noise in character.

3.3.1 Basic Formulation

Consider the elementary problem motivated by the "data" shown in Figure 3–2; t is supposed to be an independent variable, which could be time, or a spatial coordinate or just an index. Some physical variable, call it $\theta(t)$, perhaps temperature at a point in the ocean, has been measured at times $t = t_i$, $1 \le i \le M$, as depicted in the figure.

We have reason to believe that there is a linear relationship between $\theta(t)$ and t in the form $\theta(t) = at + b$ so that the measurements are

$$y(t) = \theta(t) + n(t) = a + bt + n(t) \tag{3.3.1}$$

where $n(t)$ is the inevitable measurement noise. We want to determine a, b.

The set of observations can be written in the general standard form,

$$\mathbf{Ex} + \mathbf{n} = \mathbf{y} \tag{3.3.2}$$

where in the present special case,

$$\mathbf{E} = \left\{ \begin{matrix} 1 & t_1 \\ 1 & t_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & t_M \end{matrix} \right\}, \quad \mathbf{x} = \begin{bmatrix} a \\ b \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y(t_1) \\ y(t_2) \\ \cdot \\ \cdot \\ y(t_M) \end{bmatrix}, \tag{3.3.3}$$

and $\mathbf{n}(t)$ is the noise vector. Equation sets like (3.3.2) appear in many practical situations, including the ones described in Chapter 2.

One sometimes sees (3.3.2) written as

$$\mathbf{Ex} \sim \mathbf{y}$$

or even

$$\mathbf{Ex} = \mathbf{y}\,.$$

But Equation (3.3.2) is preferable, because it explicitly recognizes that $\mathbf{n} = \mathbf{0}$ is exceptional. Sometimes, by happenstance or arrangement, one finds that $M = N$ and that $\mathbf{E}$ has an inverse. But the obvious solution, $\mathbf{x} = \mathbf{E}^{-1}\mathbf{y}$, leads to the conclusion, $\mathbf{n} = \mathbf{0}$, probably regarded as unacceptable if the $\mathbf{y}$ are the result of measurements. We will need to return to this case, but for now, let us consider the problem where $M > N$.

Commonly, then, one sees a *best possible* solution–defined as producing the smallest possible value of $\mathbf{n}^T\mathbf{n}$–that is, the one producing the minimum of

$$J = \sum_{i=1}^{M}(a + bt_i - y(t_i))^2 \equiv \sum_{i=1}^{M} n_i^2 = \mathbf{n}^T\mathbf{n} = (\mathbf{Ex} - \mathbf{y})^T(\mathbf{Ex} - \mathbf{y})\,. \quad (3.3.4)$$

Differentiating (3.3.4) with respect to a, b or $\mathbf{x}$ [using (3.1.17) and (3.1.19)] and by setting $dJ = \sum(\partial J/\partial x_i)\,dx_i = 0$, term-by-term (anticipating a minimum rather than a maximum), leads to the system called the *normal equations*,

$$\mathbf{E}^T\mathbf{Ex} = \mathbf{E}^T\mathbf{y}\,. \quad (3.3.5)$$

Making the sometimes valid assumption that $(\mathbf{E}^T\mathbf{E})^{-1}$ exists,

$$\tilde{\mathbf{x}} = (\mathbf{E}^T\mathbf{E})^{-1}\mathbf{E}^T\mathbf{y}\,. \quad (3.3.6)$$

The solution is written as $\tilde{\mathbf{x}}$ rather than as $\mathbf{x}$ because the relationship between (3.3.6) and the correct value is not clear. The fit is displayed in Figure 3–2, as are the residuals,

$$\tilde{\mathbf{n}} = \mathbf{y} - \mathbf{E}\tilde{\mathbf{x}} = \mathbf{y} - \mathbf{E}(\mathbf{E}^T\mathbf{E})^{-1}\mathbf{E}^T\mathbf{y} = (\mathbf{I} - \mathbf{E}(\mathbf{E}^T\mathbf{E})^{-1}\mathbf{E}^T)\mathbf{y}\,. \quad (3.3.7)$$

That is, the M equations have been used to estimate N values, $\tilde{x}_i$, and M values $\tilde{n}_i$, or $M + N$ altogether.

All this is easy and familiar and applies to any set of simultaneous equations, not just the straight-line example. Before proceeding, let us apply some of the statistical machinery to understanding (3.3.6). Notice that no statistics were used in obtaining (3.3.6), but we can nonetheless ask the extent to which this value for $\tilde{\mathbf{x}}$ is affected by the random elements, the noise in $\mathbf{y}$. Let $\mathbf{y}_0$ be the value of $\mathbf{y}$ that would be obtained in the hypothetical situation for which $\mathbf{n} = 0$. Assume further that $< \mathbf{n} > = 0$ and that $\mathbf{R}_{nn} = \mathbf{C}_{nn} = < \mathbf{n}\mathbf{n}^T >$ is known. Then the expected value of $\tilde{\mathbf{x}}$ is

$$< \tilde{\mathbf{x}} > = (\mathbf{E}^T\mathbf{E})^{-1}\mathbf{E}^T\mathbf{y}_0 \,. \tag{3.3.8}$$

If the matrix inverse exists, then in many situations, including the problem of fitting a straight line to data, perfect observations would produce the correct answer, and (3.3.6) is an unbiased estimate of the true solution, $< \tilde{\mathbf{x}} > = \mathbf{x}$. On the other hand, if the data were actually produced from physics governed, for example, by a quadratic rule, $\theta(t) = a{+}ct^2$, then fitting the linear rule to such observations, even if they are perfect, could never produce the right answer, and the solution would be biased. An example of such a fit is shown in Figure 3–4. Such errors are distinguishable from the noise of observation and are properly labeled *model errors.* Assume, however, that the correct model is being used and therefore that $< \tilde{\mathbf{x}} > = \mathbf{x}$. Then the uncertainty of the solution is the same as the variance about the mean and is

$$\begin{aligned} \mathbf{P} = \mathbf{C}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}} &=< (\tilde{\mathbf{x}} - \mathbf{x})(\tilde{\mathbf{x}} - \mathbf{x})^T > \\ &= (\mathbf{E}^T\mathbf{E})^{-1}\mathbf{E}^T < \mathbf{n}\mathbf{n}^T > \mathbf{E}(\mathbf{E}^T\mathbf{E})^{-1} \\ &= (\mathbf{E}^T\mathbf{E})^{-1}\mathbf{E}^T\mathbf{R}_{nn}\mathbf{E}(\mathbf{E}^T\mathbf{E})^{-1} \,. \end{aligned} \tag{3.3.9}$$

In the special case, $\mathbf{R}_{nn} = \sigma_n^2\mathbf{I}$–that is, there is no correlation between the noise in different equations (often called *white noise*)–then (3.3.9) simplifies to

$$\mathbf{P} = \sigma_n^2(\mathbf{E}^T\mathbf{E})^{-1} \,. \tag{3.3.10}$$

If we are not confident that $< \tilde{\mathbf{x}} > = \mathbf{x}$, (3.3.9)–(3.3.10) are still interpretable but as $C_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}} = D^2(\tilde{\mathbf{x}} - < \tilde{\mathbf{x}} >)$ – the covariance of $\tilde{\mathbf{x}}$. The *standard error* of $\tilde{x}_i$ is usually defined to be $\pm\sqrt{C_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}}$ and is used to understand the adequacy of data for distinguishing different possible estimates of $\tilde{\mathbf{x}}$. If applied to the straight line fit of Figure 3–2, we obtain an estimate as $\tilde{\mathbf{x}}^T = [a \quad b]^T = [1.9 \pm 0.8,\ 1.96 \pm 0.03]^T$. If the noise in $\mathbf{y}$ is Gaussian, it follows that the probability density of $\tilde{\mathbf{x}}$ is also Gaussian, with mean $< \tilde{\mathbf{x}} >$ and covariance $\mathbf{C}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}$. Of course, if $\mathbf{n}$ is not Gaussian, then neither will be the estimate, and one must be wary of the accuracy of standard error estimates.

The uncertainty of the residual estimates is

$$\begin{aligned} \mathbf{P}_{\tilde{n}\tilde{n}} \equiv< (\tilde{\mathbf{n}} - \mathbf{n})(\tilde{\mathbf{n}} - \mathbf{n})^T > &= (\mathbf{I} - \mathbf{E}(\mathbf{E}^T\mathbf{E})^{-1}\mathbf{E}^T)\mathbf{R}_{nn}(\mathbf{I} - \mathbf{E}(\mathbf{E}^T\mathbf{E})^{-1}\mathbf{E}^T)^T \\ = \sigma_n^2(\mathbf{I} - \mathbf{E}(\mathbf{E}^T\mathbf{E})^{-1}\mathbf{E}^T)^2 &= \sigma_n^2(\mathbf{I} - \mathbf{E}(\mathbf{E}^T\mathbf{E})^{-1}\mathbf{E}^T) \end{aligned} \tag{3.3.11}$$

where the second line is valid for white-noise residuals, and where $< \mathbf{n} > = \mathbf{0}$ is assumed to be correct.

The fit of a straight line to observations demonstrates many of the issues involved in making inferences from real, noisy data that appear in more complex situations. In Figure 3–3, the correct model used to generate the data was the same as in Figure 3–2, but the noise level is very high. The parameters $[\tilde{a}, \tilde{b}] = [1.5 \pm 2.8,\ 2.0 \pm 0.1]$–that is, $\tilde{a}$ is numerically incorrect, and formally indistinguishable from zero (but consistent within one standard error with the correct value). One sometimes reads, in such situations, that "least squares failed," but such a statement represents a fundamental confusion of the methodology with the lack of data adequate to demonstrate a hypothesis. Least squares functions exactly as intended, and one could conclude legitimately either that (1) there is no evidence that a straight-line rule explains the data, or (2) the data are consistent with the hypothesis $a = 1$, $b = 2$, and there is no reason to change such a prior estimate. In Figure 3–5, the quadratic model of Figure 3–4 was used to generate the numbers, but with enough additional data supplied that the residuals now clearly fail to satisfy a hypothesis of being white noise. Modeling a quadratic field with a linear model produces a systematic or model error. In contrast, the fit of a quadratic rule $y = a + bt + ct^2$, shown in Figure 3–6, does leave small, random appearing residuals. But if the true noise were not random, one might well erroneously deduce the presence of a quadratic model; such possibilities strongly suggest that the residuals had better be examined at least as closely as the model parameter estimates–and the need to do so is a constant theme in this book.

Visual tests for randomness of residuals have obvious limitations, and elaborate statistical tests help to determine objectively whether one should accept or reject the hypothesis that no significant structure remains in a sequence of numbers. Books on regression analysis (e.g., Seber, 1977, or Box & Jenkins, 1978) should be consulted for general methodologies. As an indication of what can be done, Figures 3–7a and b show the sample autocovariance,

$$\tilde{R}_{nn}(\tau) = \frac{1}{M} \sum_{i=1}^{M-|\tau|} \tilde{n}_i \tilde{n}_{i+\tau}\,, \tag{3.3.12}$$

for the residuals of the fits shown in Figures 3–5 and 3–6. [$\tilde{R}_{nn}(\tau)$ is an estimate of $< n_i n_{i+\tau} >$.] If the residuals were truly uncorrelated, $< \tilde{n}_i\, \tilde{n}_{i+\tau} > = 0$, $\tau \neq 0$, and one expects $\tilde{R}_{nn}(\tau)$ to approach a delta function at $\tau = 0$. Tests are available to determine if the nonzero values

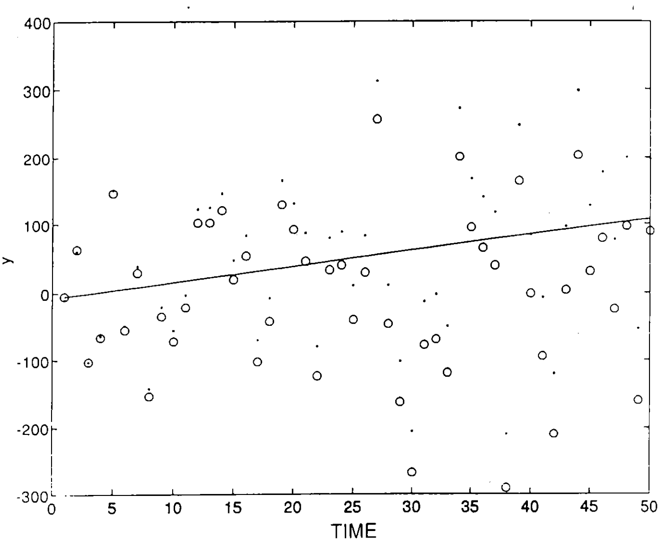

Figure 3–3. The same situation as described in Figure 3–2 except $< n > = 0$, $< n_i n_j > = 100^2\delta_{ij}$, meaning the "data" were very noisy. The fit is now $\tilde{y} = 1.5 \pm 2.75 + (2.02 \pm 0.09)t$.

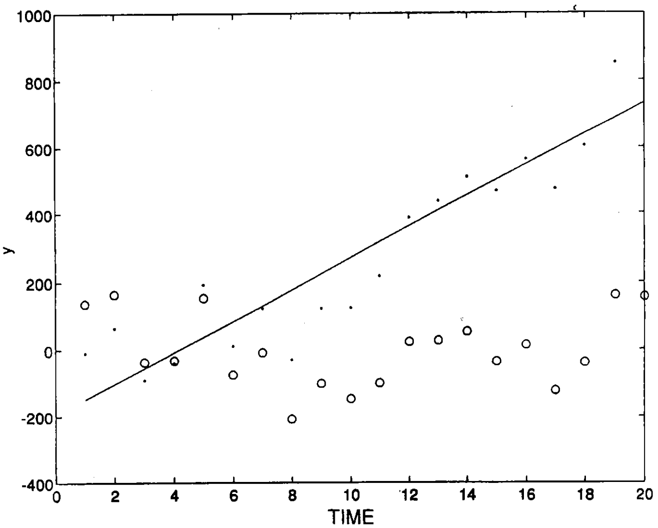

Figure 3–4. Here, the "data" were generated through a quadratic rule, $y = 1 + t^2 + n$, $< n > = 0$, $< n_i n_j > = 100^2\delta_{ij}$. But a linear fit was nonetheless made that produces $y = -194 \pm 1.6 + (46.4 \pm 0.1)t$. To the eye, at least, it is a reasonably good fit, and one might have great difficulty in rejecting the hypothesis that a straight-line rule is valid.

for $\tau \neq 0$ are significantly nonzero (Box & Jenkins, 1978, Ch. 2). Here, we merely note that the sample autocovariance behavior again confirms visually what we already know–that the fit in Figure 3–6 is adequate, and in Figure 3–5 it is not.

3.3.2 Weighted and Tapered Least Squares

The least-squares solution (3.3.6)–(3.3.7) was derived by minimizing the objective function (3.3.4), in which each residual element is given equal

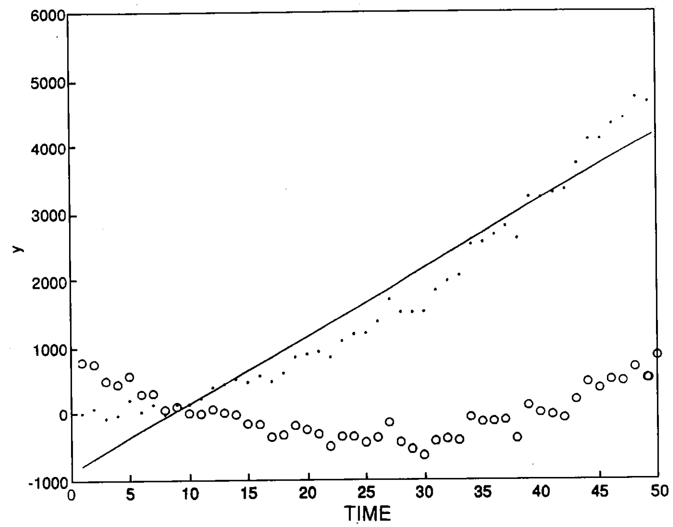

Figure 3–5. The same situation as in Figure 3–4 except that the duration was extended. Now the linear fit leaves obvious residuals that are nonrandom, producing a strong indication that a linear model is inadequate, or that the noise is not white, or both.

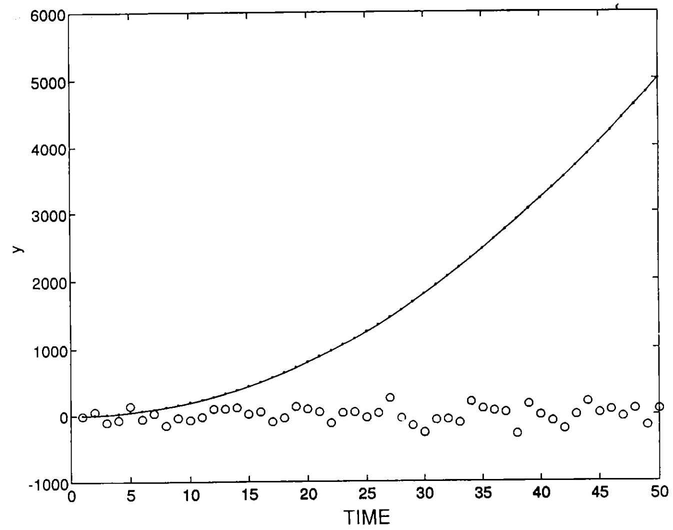

Figure 3–6. The same situation as in Figure 3–5 except that now a quadratic model, $y = a + bt + ct^2$, was fit, resulting in a solution $y = -2 \pm 5.2 + (-0.26 \pm 4.7)t + (2.0 \pm 3.9)t^2$ and leaving small, apparently random, residuals.

weight. An important feature of least squares is that we can give whatever emphasis we please to minimizing individual equation residuals, for example, by introducing an objective function

$$J = \sum_i W_{ii} n_i^2 \tag{3.3.13}$$

where W_{ii} are any numbers desired. The choice $W_{ii} = 1$ might be reasonable, but it is clearly an arbitrary one that without further justification does not produce a solution with any special claim to significance. In the least

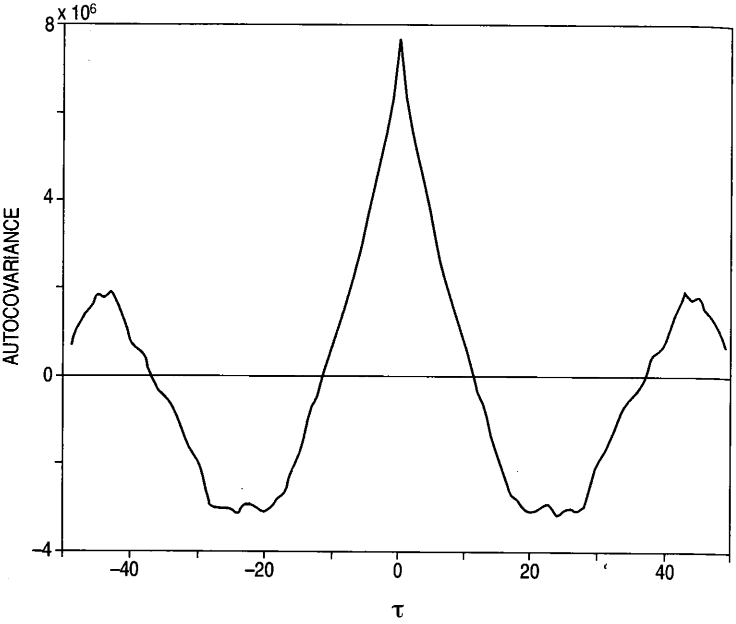

Figure 3–7a. Autocovariance of the residuals from Figure 3–5, the autocovariance of a non-white process. This is an example of a test for adequacy of a model. We would probably reject the model.

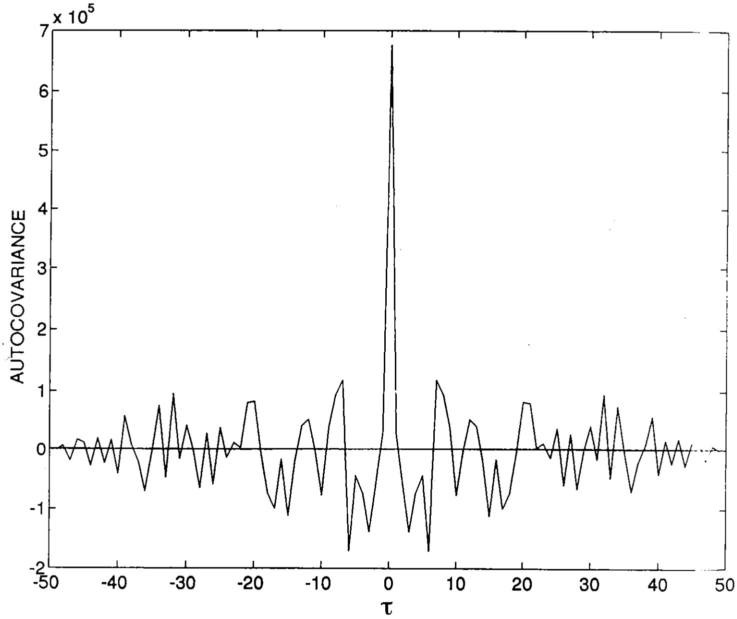

Figure 3–7b. Autocovariance of fit in Figure 3–6, when a quadratic was fit. Here, the result is indistinguishable from white noise, and the model would be acceptable.

squares context, we are free to make any other reasonable choice, including demanding that some residuals should be much larger than others, perhaps just to determine if it is possible.

A general formalism is obtained by defining a diagonal weight matrix,

$\mathbf{W} = \mathrm{diag}([W_{ii}])$.[4] Divide each equation in (3.3.2) by $\sqrt{W_{ii}}$,

$$W_{ii}^{-1/2} \sum_j E_{ij} x_j + W_{ii}^{-1/2} n_i = W_{ii}^{-1/2} y_i \tag{3.3.14}$$

or

$$\begin{gathered} \mathbf{E}'\mathbf{x} + \mathbf{n}' = \mathbf{y}' \\ \mathbf{E}' = \mathbf{W}^{-T/2}\mathbf{E},\ \mathbf{n}' = \mathbf{W}^{-T/2}\mathbf{n},\ \mathbf{y}' = \mathbf{W}^{-T/2}\mathbf{y} \end{gathered} \tag{3.3.15}$$

where we used the fact that the square root of a diagonal matrix is its element-by-element square roots. Such a matrix is its own transpose, and the purpose of writing $\mathbf{W}^{-T/2}$ will become clear below. The operation in (3.3.14) or (3.3.15) is usually called *row scaling* because it operates on the rows of $\mathbf{E}$ (as well as on $\mathbf{n}$, $\mathbf{y}$).

For the new equations (3.3.15), the objective function

$$J = \mathbf{n}'^T\mathbf{n}' = (\mathbf{y}' - \mathbf{E}'\mathbf{x})^T(\mathbf{y}' - \mathbf{E}'\mathbf{x}) = \mathbf{n}^T\mathbf{W}^{-1}\mathbf{n} = (\mathbf{y} - \mathbf{E}\mathbf{x})^T\mathbf{W}^{-1}(\mathbf{y} - \mathbf{E}\mathbf{x}) \tag{3.3.16}$$

weights the residuals as desired. If for some reason, $\mathbf{W}$ is nondiagonal but symmetric and positive-definite, then it has a Cholesky decomposition,

$$\mathbf{W} = \mathbf{W}^{T/2}\mathbf{W}^{1/2},$$

and (3.3.15) remains useful more generally.

The values $\tilde{\mathbf{x}}$, $\tilde{\mathbf{n}}$, minimizing (3.3.16), are

$$\begin{gathered} \tilde{\mathbf{x}} = (\mathbf{E}'^T\mathbf{E}')^{-1}\mathbf{E}'^T\mathbf{y}' = (\mathbf{E}^T\mathbf{W}^{-1}\mathbf{E})^{-1}\mathbf{E}^T\mathbf{W}^{-1}\mathbf{y} \\ \tilde{\mathbf{n}} = \mathbf{W}^{T/2}\mathbf{n}' = \left\{\mathbf{I} - \mathbf{E}(\mathbf{E}^T\mathbf{W}^{-1}\mathbf{E})^{-1}\mathbf{E}^T\mathbf{W}^{-1}\right\}\mathbf{y} \end{gathered} \tag{3.3.17}$$

and

$$\mathbf{P} = \mathbf{C}_{\tilde{x}\tilde{x}} = (\mathbf{E}^T\mathbf{W}^{-1}\mathbf{E})^{-1}\mathbf{E}^T\mathbf{W}^{-1}\mathbf{R}_{nn}\mathbf{W}^{-1}\mathbf{E}(\mathbf{E}^T\mathbf{W}^{-1}\mathbf{E})^{-1}. \tag{3.3.18}$$

Uniform diagonal weights are a special case. The rationale for choosing differing diagonal weights or a nondiagonal $\mathbf{W}$ is probably not very obvious to the reader. There is one common situation in which $\mathbf{W} = \mathbf{R}_{nn} = \{ < n_i n_j > \}$, that is, the weight matrix is chosen to be the expected second-moment matrix of the residuals. Then (3.3.18) simplifies to

$$\mathbf{P} = \mathbf{C}_{\tilde{x}\tilde{x}} = (\mathbf{E}^T\mathbf{R}_{nn}^{-1}\mathbf{E})^{-1}. \tag{3.3.19}$$

Here, the weighting (3.3.15) has a ready interpretation: The equations (and hence the residuals) are rotated and stretched so that in the new coordinate

[4] If $\mathbf{q}$ is a vector, the operator diag($\mathbf{q}$) forms a square diagonal matrix, whose elements are q_i.

system, the covariance of n_i is diagonal and the variances $< n_i^2 >$ are all unity. In this space, the simple, original objective functions (3.3.4) make physical sense. But we emphasize that this choice of $\mathbf{W}$ is a very special one and has confused many users of inverse methods. To emphasize again: Least squares is a deterministic process in which $\mathbf{W}$ is a set of weights wholly at the disposal of the investigator; setting $\mathbf{W} = \mathbf{R}_{nn}$ is a special case.[5]

If it is intended that $\mathbf{W}$ should reflect the actual expected noise variance, one should confirm after the fact that substitution of $\tilde{\mathbf{x}}$ into (3.3.16) produces a value of J consistent with the hypothesis. That is, because

$$< J > = < \mathbf{n}^T \mathbf{R}_{nn}^{-1} \mathbf{n} > = \sum_1^M < n_i^2 > = M - K \,, \tag{3.3.20}$$

one should obtain (3.2.42),

$$\tilde{J} = \tilde{\mathbf{n}}^T \mathbf{R}_{nn}^{-1} \tilde{\mathbf{n}} \simeq M - K \,. \tag{3.3.21}$$

$M - K$ degrees of freedom (here, $K = N$) are anticipated because the residuals are not independent but are related by (3.3.7). [That there are N degrees of freedom removed by (3.3.7) becomes obvious later on.] The degree of approximation required to $\tilde{J} = M - K$ is then readily determined from χ^2_{M-K}, assuming the $\tilde{n}_i$ are approximately Gaussian. As an illustration, 30 equations in 15 unknowns were constructed for which $(\mathbf{E}^T\mathbf{E})^{-1}$ existed. Then with $\mathbf{x}$ known, an ensemble of 50 values of $\mathbf{y}$ was generated by forming

$$\mathbf{y} = \mathbf{E}\mathbf{x} + \mathbf{n}$$

and by generating $\mathbf{n}$ with a pseudorandom number generator. The system of equations was solved for $\tilde{\mathbf{x}}$ by Equation (3.3.6), producing 50 different estimates of both $\tilde{\mathbf{x}}$, $\tilde{\mathbf{n}}$ and the resulting value of $\tilde{J}$ formed for each and plotted in Figure 3–8a. By expression (3.3.20) we would expect the mean value of $\tilde{J}$ to be near 15, as the figure suggests is approximately correct. Figure 3–8b shows the empirical frequency function of $\tilde{J}$ as compared to χ^2_{15}. The study of the deviations between the expected and computed distributions is the basis of hypothesis testing for the validity of the results, but we leave this discussion to the large literature on the subject. Similarly, the individual elements whose sum is $\tilde{J}$ should have a probability density consistent with χ^2_1.

Whether the equations are scaled or not, the previous limitations of the

[5] In maximum likelihood estimation, least-squares is used to find the likelihood function extreme, and $\mathbf{W} = \mathbf{R}_{nn}$ emerges as a natural choice (e.g., Cramér, 1946; Van Trees, 1968). But the logic of this process is distinct from least-squares as we are employing it here.

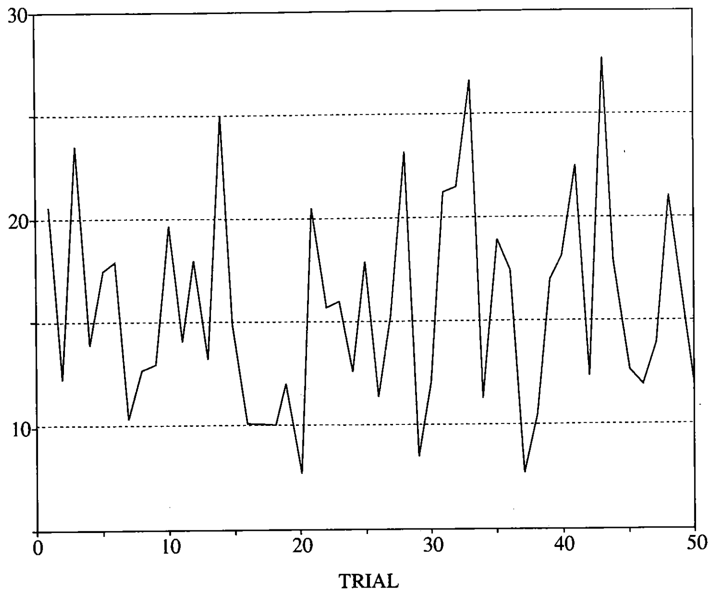

Figure 3–8a. Example in which a system of 30 equations in 15 unknowns was solved for 50 different noise realizations in $\mathbf{y}$, showing the different values of $\tilde{J}$ [equation (3.3.20)] for each resulting solution. The mean-square residuals had a value of 15.9.

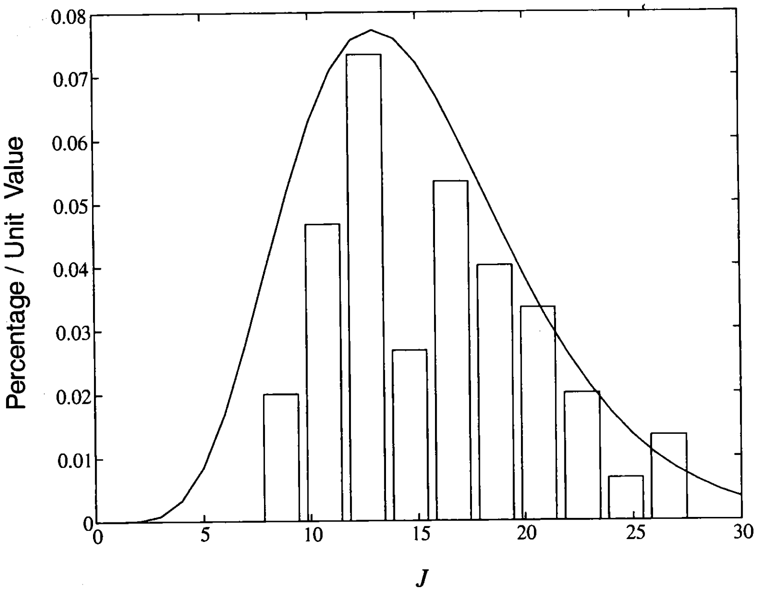

Figure 3–8b. Histogram of values of $\tilde{J}$ shown in Figure 3–8a, compared to the probability density χ^2_{15}.

simple least-squares solutions remain. In particular, we still have the problem that the solution may produce solution elements $\tilde{\mathbf{x}}$, $\tilde{\mathbf{n}}$, whose relative values are not in accord with expected or reasonable behavior, and the solution uncertainty or variances could be unusably large. These quantities are all determined, mechanically and automatically, from combinations such as $(\mathbf{E}^T\mathbf{W}^{-1}\mathbf{E})^{-1}$, an operator that is neither controllable nor very easy to understand and that may not even exist if the matrix is singular.

Suppose, without loss of generality, that any necessary weight matrix $\mathbf{W}$ has already been applied to the equations so that (3.3.4) is a reasonable objective function. It was long ago recognized that some control over the magnitudes of $\tilde{\mathbf{x}}$, $\tilde{\mathbf{n}}$, $\mathbf{C}_{\tilde{x}\tilde{x}}$ could be obtained in the simple least-squares context by modifying the objective function (3.3.4) to have an additional term:

$$J' = \mathbf{n}^T\mathbf{n} + \alpha^2\mathbf{x}^T\mathbf{x} = (\mathbf{y} - \mathbf{E}\mathbf{x})^T(\mathbf{y} - \mathbf{E}\mathbf{x}) + \alpha^2\mathbf{x}^T\mathbf{x} \qquad (3.3.22)$$

in which α^2 is a given positive constant.

If the minimum of (3.3.22) is sought by setting to zero the derivatives with respect to $\mathbf{x}$, the resulting normal equations produce

$$\boxed{\begin{aligned}
\tilde{\mathbf{x}} &= (\mathbf{E}^T\mathbf{E} + \alpha^2\mathbf{I})^{-1}\mathbf{E}^T\mathbf{y} && (3.3.23)\\
\tilde{\mathbf{n}} &= \left\{\mathbf{I} - \mathbf{E}(\mathbf{E}^T\mathbf{E} + \alpha^2\mathbf{I})^{-1}\mathbf{E}^T\right\}\mathbf{y} && (3.3.24)\\
\mathbf{C}_{\tilde{x}\tilde{x}} &= (\mathbf{E}^T\mathbf{E} + \alpha^2\mathbf{I})^{-1}\mathbf{E}^T\mathbf{R}_{nn}\mathbf{E}(\mathbf{E}^T\mathbf{E} + \alpha^2\mathbf{I})^{-1} && (3.3.25)\\
\mathbf{P}_{\tilde{n}\tilde{n}} &= \left\{\mathbf{I} - \mathbf{E}(\mathbf{E}^T\mathbf{E} + \alpha^2\mathbf{I})^{-1}\mathbf{E}^T\right\}\mathbf{R}_{nn} \times \\
&\quad \left\{\mathbf{I} - \mathbf{E}(\mathbf{E}^T\mathbf{E} + \alpha^2\mathbf{I})^{-1}\mathbf{E}^T\right\}^{-1}. && (3.3.26)
\end{aligned}}$$

By letting $\alpha^2 \to 0$, the solution (3.3.6)–(3.3.7), (3.3.9) is recovered, and if $\alpha^2 \to \infty$, $\|\tilde{\mathbf{x}}\|_2 \to 0$, $\tilde{\mathbf{n}} \to \mathbf{y}$, α^2 is called a *trade-off parameter*, because it trades the magnitude of $\tilde{\mathbf{x}}$ against that of $\tilde{\mathbf{n}}$. By varying the size of α^2, we gain some influence over the norm of the residuals relative to that of $\tilde{\mathbf{x}}$. The expected value of $\tilde{\mathbf{x}}$ is now

$$< \tilde{\mathbf{x}} > = (\mathbf{E}^T\mathbf{E} + \alpha^2\mathbf{I})^{-1}\mathbf{E}^T\mathbf{y}_0 . \qquad (3.3.27)$$

If the true solution is believed to be (3.3.8), then this new solution is biased. But the variance of $\tilde{\mathbf{x}}$ (3.3.25) has been reduced by introduction of $\alpha^2 > 0$–that is, the acceptance of a bias reduces the variance. Equations (3.3.23)–(3.3.26) are sometimes known as the *tapered least-squares* solution, a label whose implication becomes clear later.

A physical motivation for the modified objective function (3.3.22) is obtained by noticing that it would be the simplest one to use if the equations being solved consisted of (3.3.2), augmented with a second set asserting $\mathbf{x} \approx \mathbf{0}$–that is, a combined set

$$\mathbf{E}\mathbf{x} + \mathbf{n} = \mathbf{y}$$
$$\alpha^2(\mathbf{x} + \mathbf{n}_1) = \mathbf{0}$$

or

$$\mathbf{E}_1\mathbf{x} + \mathbf{n}_2 = \mathbf{y}_2$$

$$\mathbf{E}_1 = \begin{Bmatrix} \mathbf{E} \\ \alpha^2\mathbf{I} \end{Bmatrix}, \ \mathbf{n}_2^T = [\mathbf{n}^T \ \alpha^2\mathbf{n}_1^T], \ \mathbf{y}_2^T = [\mathbf{y}^T \quad \mathbf{0}^T], \tag{3.3.28}$$

and in which α^2 expresses a preference for fitting the first or second sets more accurately. It then comes as no surprise that the solution covariance depends upon the relative weight given to the second set of equations. A preference that $\mathbf{x} \approx \mathbf{x}_0$ is readily imposed instead, with an obvious change in (3.3.22).

Note the important points, to be shown later, that the matrix inverses in (3.3.23)–(3.3.26) will always exist as long as $\alpha^2 > 0$ and, furthermore, that the expressions remain valid even if $M < N$.

Tapered least squares produce some control over the sum of squares of the relative norms of $\tilde{\mathbf{x}}$, $\tilde{\mathbf{n}}$ but still does not produce control over the individual elements $\tilde{x}_i$. In analogy to the control of the elements of $\tilde{n}_i$ obtained by using a weight matrix $\mathbf{W}$, we can further generalize the objective function by introducing another $N \times N$ weight matrix, $\mathbf{S}$, and using

$$J = \mathbf{n}^T\mathbf{n} + \mathbf{x}^T\mathbf{S}^{-1}\mathbf{x} = (\mathbf{y} - \mathbf{E}\mathbf{x})^T(\mathbf{y} - \mathbf{E}\mathbf{x}) + \mathbf{x}^T\mathbf{S}^{-1}\mathbf{x}. \tag{3.3.29}$$

Setting the derivatives with respect to $\mathbf{x}$ to zero results in

$$\tilde{\mathbf{x}} = (\mathbf{E}^T\mathbf{E} + \mathbf{S}^{-1})^{-1}\mathbf{E}^T\mathbf{y} \tag{3.3.30}$$

$$\tilde{\mathbf{n}} = \left\{\mathbf{I} - \mathbf{E}(\mathbf{E}^T\mathbf{E} + \mathbf{S}^{-1})^{-1}\mathbf{E}^T\right\}\mathbf{y} \tag{3.3.31}$$

$$\mathbf{C}_{\tilde{x}\tilde{x}} = (\mathbf{E}^T\mathbf{E} + \mathbf{S}^{-1})^{-1}\mathbf{E}^T\mathbf{R}_{nn}\mathbf{E}(\mathbf{E}^T\mathbf{E} + \mathbf{S}^{-1})^{-1}. \tag{3.3.32}$$

The only restriction is that the matrix inverses must exist. Tapered least squares is a special case in which $\mathbf{S}^{-1} = \alpha^2\mathbf{I}_N$, and plain least squares further sets $\alpha^2 = 0$. Like $\mathbf{W}$, $\mathbf{S}$ is often diagonal, in which the numerical values simply assert a preference for making individual terms of $\tilde{x}_i$ large or small.

Suppose that $\mathbf{S}$ is positive definite and symmetric and thus has a Cholesky decomposition. If the equations are scaled as (sometimes called *column scaling* because it weights the columns of $\mathbf{E}$),

$$\begin{aligned}\mathbf{E}\mathbf{S}^{T/2}\mathbf{S}^{-T/2}\mathbf{x}+\mathbf{n}&=\mathbf{y}\\ \mathbf{E}'\mathbf{x}'+\mathbf{n}&=\mathbf{y}\\ \mathbf{E}'=\mathbf{E}\mathbf{S}^{T/2}\,,\qquad \mathbf{x}'&=\mathbf{S}^{-T/2}\mathbf{x}\,,\end{aligned}\tag{3.3.33}$$

then the objective function (3.3.22), with $\alpha^2 = 1$, is a plausible one in the new coordinate system of $\mathbf{x}'$. Like $\mathbf{W}$, one is completely free to choose $\mathbf{S}$ as one pleases. A common example is to write $\mathbf{S} = \mathbf{F}^T\mathbf{F}$, where $\mathbf{F}$ is $N \times N$,

$$\mathbf{F}=\alpha\left\{\begin{matrix}1 & -1 & 0 & \cdots & 0\\ 0 & 1 & -1 & \cdots & 0\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ 0 & \cdots & \cdots & 0 & 1\end{matrix}\right\}\,,\tag{3.3.34}$$

whose effect is to minimize a term $\alpha^2\sum_i(x_i-x_{i+1})^2$, which can be regarded as a smoothest solution, and using α^2 to trade smoothness against the size of $\|\tilde{\mathbf{n}}\|_2$. $\alpha\mathbf{F}$ is the Cholesky decomposition of $\mathbf{S}$. Another common choice is $\mathbf{S} = \mathbf{R}_{xx}$–that is, the second moments of the solution where known. In this special situation, the x_i' would be uncorrelated with unit variance. Usually it is assumed that both row and column scaling have been done:

$$\mathbf{W}^{-T/2}\mathbf{E}\mathbf{S}^{T/2}\mathbf{S}^{-T/2}\mathbf{x}+\mathbf{W}^{-T/2}\mathbf{n}=\mathbf{W}^{-T/2}\mathbf{y}$$

$$\mathbf{E}'\mathbf{x}'+\mathbf{n}'=\mathbf{y}'$$

$$\mathbf{E}'=\mathbf{W}^{-T/2}\mathbf{E}\mathbf{S}^{T/2},\ \mathbf{x}'=\mathbf{S}^{-T/2}\mathbf{x},\ \mathbf{n}'=\mathbf{W}^{-T/2}\mathbf{n},\ \mathbf{y}'=\mathbf{W}^{-T/2}\mathbf{y}\,,\tag{3.3.35}$$

at which point the plain objective function

$$J=\mathbf{n}'^T\mathbf{n}'+\mathbf{x}'^T\mathbf{x}'\tag{3.3.36}$$

is used (α^2 has been absorbed into $\mathbf{S}$, but it is often convenient to carry it as a separate parameter). If the primes are dropped, Equations (3.3.23)–(3.3.26) result, with $\alpha^2 = 1$. If the original variables $\mathbf{E}$, $\mathbf{x}$, $\mathbf{n}$, $\mathbf{y}$ are restored, we obtain the most general row- and column-scaled form, which is, for future reference,

$$J = \mathbf{n}\mathbf{W}^{-1}\mathbf{n} + \mathbf{x}^T\mathbf{S}^{-1}\mathbf{x} \tag{3.3.37}$$

$$\tilde{\mathbf{x}} = (\mathbf{E}^T\mathbf{W}^{-1}\mathbf{E} + \mathbf{S}^{-1})^{-1}\mathbf{E}^T\mathbf{W}^{-1}\mathbf{y} \tag{3.3.38}$$

$$\tilde{\mathbf{n}} = \left\{\mathbf{I} - \mathbf{E}(\mathbf{E}^T\mathbf{W}^{-1}\mathbf{E} + \mathbf{S}^{-1})^{-1}\mathbf{E}^T\mathbf{W}^{-1}\right\}\mathbf{y} \tag{3.3.39}$$

$$\begin{aligned}\mathbf{C}_{\tilde{x}\tilde{x}} = {} & (\mathbf{E}^T\mathbf{W}^{-1}\mathbf{E} + \mathbf{S}^{-1})^{-1}\mathbf{E}^T\mathbf{W}^{-1}\mathbf{R}_{nn} \times \\ & \mathbf{W}^{-1}\mathbf{E}(\mathbf{E}^T\mathbf{W}^{-1}\mathbf{E} + \mathbf{S}^{-1})^{-1}\end{aligned} \tag{3.3.40}$$

$$\begin{aligned}\mathbf{P}_{\tilde{n}\tilde{n}} = {} & \left\{\mathbf{I} - \mathbf{E}(\mathbf{E}^T\mathbf{W}^{-1}\mathbf{E} + \mathbf{S}^{-1})^{-1}\mathbf{E}^T\mathbf{W}^{-1}\right\}^{-1}\mathbf{R}_{nn} \times \\ & \left\{\mathbf{I} - \mathbf{E}(\mathbf{E}^T\mathbf{W}^{-1}\mathbf{E} + \mathbf{S}^{-1})\mathbf{E}^T\mathbf{W}^{-1}\right\}^T .\end{aligned} \tag{3.3.41}$$

So far, all of this is conventional. But we have made a special point of displaying explicitly not only the elements $\tilde{\mathbf{x}}$ but those of the residuals $\tilde{\mathbf{n}}$. Notice that although we have considered only the formally overdetermined system, $M > N$, we *always* determine not only the N elements of $\tilde{\mathbf{x}}$ but also the M elements of $\tilde{\mathbf{n}}$, for a total of $M + N$ values, extracted from the M equations. It is apparent that any change in any element $\tilde{n}_i$ forces changes in $\tilde{\mathbf{x}}$. In this view, to which we adhere, systems of equations involving observations *always* contain more unknowns than knowns. There is compelling reason, therefore, to rewrite (3.3.2) as

$$\mathbf{E}_1\boldsymbol{\xi} = \mathbf{y}$$

$$\mathbf{E}_1 = \{\mathbf{E} \quad \mathbf{I}_M\}, \quad \boldsymbol{\xi}^T = [\mathbf{x} \quad \mathbf{n}]^T, \tag{3.3.42}$$

which is to be solved exactly, as an underdetermined system. That even overdetermined observed systems are of necessity actually underdetermined, leads to taking a first look at formal underdetermination.

3.3.3 Undetermined Systems–A First Discussion

What does one do when the number of equations is less than the number of unknowns, and no more observations are possible? One often attempts in such a situation to reduce the number of unknowns so that the formal overdeterminism is restored. Such a parameter reduction procedure may be sensible, but there are pitfalls. Consider data produced from a law

$$y = 1 + a_M t^M + n(t)\,, \tag{3.3.43}$$

which might be deduced by fitting a parameter set $[a_0, \ldots, a_M]$. If there are fewer than M observations, an attempt to fit with fewer parameters,

$$y = \sum_{i=0}^{Q} a_i t^i\,, \quad Q < M\,, \tag{3.3.44}$$

may give a good, even perfect fit, but it would be wrong. The reduction in model parameters in such a case biases the result. One is better off retaining the underdetermined system and making inferences concerning the possible values of a_i rather than using the form (3.3.44), in which any possibility of learning something about a_M has been eliminated.

In more general terms (discussed by Wunsch & Minster, 1982), parameter reduction can lead to model errors or biases that can produce wholly illusory results. A specific example was provided by Wunsch (1988a); a two-dimensional ocean circulation model was used to calculate values for the apparent oxygen utilization rate (AOUR). But when the parameterization was made more realistic (a three-dimensional model), it was found that AOUR was indeterminate to within any useful range. The conclusions from the underparameterized model are erroneous; the second model produces the useful information that the database was inadequate to estimate AOUR, and one avoids drawing incorrect conclusions.

Another example is Munk's (1966) well-known discussion of the property fields of the abyssal Pacific Ocean (see Figure 4–29). He fit the observations to the solutions of one-dimensional vertical balance equations

$$w\frac{\partial C}{\partial z} - \kappa\frac{\partial^2 C}{\partial z^2} = \text{sinks} \tag{3.3.45}$$

where C is temperature, salinity, or radiocarbon for the vertical velocity w and vertical mixing coefficient κ. The fit was quite good and has been cherished by a generation of chemical oceanographers as showing that the ocean is one-dimensional and steady. But such an ocean circulation is impossible, and one is misled by the good fit of an underparameterized model.

A general approach to solving underdetermined problems is to render them unique by minimizing an objective function, subject to satisfaction of the linear constraints. To see how this can work, suppose that (3.3.2) are indeed formally underdetermined–that is, $M < N$–and seek the solution that exactly satisfies the equations and simultaneously renders the objective function, $J = \mathbf{x}^T\mathbf{x}$, as small as possible. Direct minimization of J leads to

$$dJ = \left(\frac{\partial J}{\partial \mathbf{x}}\right)^T d\mathbf{x} = 2\mathbf{x}^T d\mathbf{x} = 0\,, \tag{3.3.46}$$

but the coefficients of the individual dx_i can no longer be separately set to zero (i.e., $\mathbf{x} = 0$ is incorrect) because the dx_i no longer vary independently but are restricted to values satisfying $\mathbf{Ex} = \mathbf{y}$. One approach is to use the known dependencies to reduce the problem to a new one in which the differentials are independent. For example, suppose that there are general functional relationships

$$\begin{bmatrix} x_1 \\ \vdots \\ x_L \end{bmatrix} = \begin{bmatrix} \xi_1(x_{L+1}, \ldots, x_N) \\ \vdots \\ \xi_L(x_{L+1}, \ldots, x_N) \end{bmatrix}.$$

Then the first L elements of x_i may be eliminated, and the cost function becomes

$$J = \xi_1^2 + \cdots + \xi_L^2 + x_{L+1}^2 + \cdots + x_N^2$$

in which the remaining x_i, $L+1 \leq i \leq N$ are independently varying. But an explicit solution for L elements of $\mathbf{x}$ in terms of the remaining ones may be difficult to find.

When it is inconvenient to find such an explicit representation eliminating some variables in favor of others, a standard procedure for finding the constrained minimum is to introduce a new vector *Lagrange multiplier*, $\boldsymbol{\mu}$, of M unknown elements, to make a new objective function

$$J' = J - 2\boldsymbol{\mu}^T(\mathbf{Ex} - \mathbf{y}) = \mathbf{x}^T\mathbf{x} - 2\boldsymbol{\mu}^T(\mathbf{Ex} - \mathbf{y}) \qquad (3.3.47)$$

and ask for its stationary point, treating both $\boldsymbol{\mu}$ and $\mathbf{x}$ as independently varying unknowns. The numerical 2 is introduced solely for notational tidiness. The rationale for this procedure is straightforward (e.g., Morse & Feshbach, 1953, p. 238; Strang, 1986): $\mathbf{Ex} = \mathbf{y}$ requires that

$$\mathbf{E}d\mathbf{x} = \mathbf{e}_1 dx_1 + \mathbf{e}_2 dx_2 + \cdots + \mathbf{e}_N dx_N = 0$$

where the $\mathbf{e}_i$ are the column vectors of $\mathbf{E}$. A constant, $-2\boldsymbol{\mu}^T$, times this last expression can be added to $dJ = 0$ so that

$$dJ - 2\boldsymbol{\mu}^T\mathbf{E}d\mathbf{x} = \left(\frac{\partial J}{\partial x_1} - 2\boldsymbol{\mu}^T\mathbf{e}_1\right) dx_1 + \left(\frac{\partial J}{\partial x_2} - 2\boldsymbol{\mu}^T\mathbf{e}_2\right) dx_2 + \cdots$$
$$+ \left(\frac{\partial J}{\partial x_N} - 2\boldsymbol{\mu}^T\mathbf{e}_N\right) dx_N = 0\,. \qquad (3.3.48)$$

In this form, there are M elements of $\boldsymbol{\mu}$ that can be used to set any M of the coefficients of the dx_i to zero, leaving coefficients of $N-M$ of the remaining dx_i, which can be treated as independent variables. If the objective function J' is differentiated with respect to $\boldsymbol{\mu}$, $\mathbf{x}$ and is set to zero, it is readily found that the result is a set of simultaneous equations equivalent to the vanishing of each coefficient in (3.3.48), plus the model. With (3.3.47), this recipe produces

$$\tfrac{1}{2}\frac{\partial J'}{\partial \boldsymbol{\mu}} = \mathbf{E}\mathbf{x} - \mathbf{y} = \mathbf{0} \tag{3.3.49}$$

$$\tfrac{1}{2}\frac{\partial J'}{\partial \mathbf{x}} = \mathbf{x} - \mathbf{E}^T\boldsymbol{\mu} = \mathbf{0} \tag{3.3.50}$$

where the first of these are just the original equations and the second are the coefficients in (3.3.48) whose solution must therefore be equivalent to setting the individual terms of (3.3.48) to zero as required, subject to the model. Because the original equations emerge, the second term of J' will vanish at the stationary point. The convenience of being able to treat all the x_i as independently varying is offset by the increase in problem dimensions by the introduction of the unknown μ_i.

Equation (3.3.50) gives

$$\mathbf{E}^T\boldsymbol{\mu} = \mathbf{x}\,, \tag{3.3.51}$$

and substituting for $\mathbf{x}$ into (3.3.49),

$$\mathbf{E}\mathbf{E}^T\boldsymbol{\mu} = \mathbf{y}\,,$$

$$\tilde{\boldsymbol{\mu}} = (\mathbf{E}\mathbf{E}^T)^{-1}\mathbf{y}\,, \tag{3.3.52}$$

assuming the inverse exists, and

$$\tilde{\mathbf{x}} = \mathbf{E}^T(\mathbf{E}\mathbf{E}^T)^{-1}\mathbf{y} \tag{3.3.53}$$

$$\tilde{\mathbf{n}} = \mathbf{0} \tag{3.3.54}$$

$$\mathbf{C}_{\tilde{x}\tilde{x}} = \mathbf{0} \tag{3.3.55}$$

($\mathbf{C}_{\tilde{x}\tilde{x}} = 0$ because formally we estimate $\tilde{\mathbf{n}} = \mathbf{0}$).

Equations (3.3.51) for $\boldsymbol{\mu}$ in terms of $\mathbf{x}$ involves the coefficient matrix $\mathbf{E}^T$. An intimate connection exists between matrix transposes and adjoints of differential equations (see especially, Lanczos, 1961; or Morse & Feshbach, 1953), and thus $\boldsymbol{\mu}$ is sometimes called the *adjoint solution*, with $\mathbf{E}^T$ being the adjoint model.[6] The original Equations (3.3.2) were assumed formally underdetermined, and thus the adjoint model equations in (3.3.51) are necessarily formally overdetermined. The physical interpretation of $\boldsymbol{\mu}$ comes from the result

$$\frac{\partial J'}{\partial \mathbf{y}} = 2\boldsymbol{\mu}\,; \tag{3.3.56}$$

[6] But the matrix transpose should not be confused with the adjoint matrix, which is quite different.

the Lagrange multipliers represent the sensitivity of the minimum of J' to the perturbations in the data $\mathbf{y}$.

Equation (3.3.53) is the classical solution of minimum norm of $\mathbf{x}$, satisfying the constraints exactly while minimizing the solution length. That a minimum is achieved can be verified by evaluating the second derivatives of J' at the solution point. The minimum occurs at a saddle point in $\mathbf{x}$, $\boldsymbol{\mu}$ space (see Sewell, 1987, for an interesting discussion) and where the term proportional to $\boldsymbol{\mu}$ necessarily vanishes. The operator $\mathbf{E}^T(\mathbf{E}\mathbf{E}^T)^{-1}$ is sometimes called a *Moore-Penrose inverse.*

If the Equations (3.3.2) are first column-scaled using $\mathbf{S}^{-T/2}$, Equations (3.3.53)–(3.3.55) are in the primed variables, the solution in the original variables is

$$\tilde{\mathbf{x}} = \mathbf{S}\mathbf{E}^T(\mathbf{E}\mathbf{S}\mathbf{E}^T)^{-1}\mathbf{y} \tag{3.3.57}$$

$$\tilde{\mathbf{n}} = \mathbf{0} \tag{3.3.58}$$

$$\mathbf{C}_{\tilde{x}\tilde{x}} = \mathbf{0}\,, \tag{3.3.59}$$

and the result depends directly upon $\mathbf{S}$. If a row-scaling with $\mathbf{W}^{-T/2}$ is used, it is readily shown that $\mathbf{W}$ disappears from the solution and has no effect on it. Equations (3.3.57)–(3.3.59) are a valid solution, but there is a potentially fatal defect–$\tilde{\mathbf{n}} = \mathbf{0}$ is rarely acceptable when $\mathbf{y}$ are observations. Furthermore, $\|\tilde{\mathbf{x}}\|$ is again uncontrolled, and $\mathbf{E}^T\mathbf{E}$ may not have an inverse.

We have been emphasizing that $\mathbf{n}$ must be regarded as fully an element of the solution, as much as $\mathbf{x}$, that equations representing observations can always be written as (3.3.42) and can be solved exactly. Therefore, we use a modified objective function

$$J = \alpha^2\mathbf{x}^T\mathbf{x} + \mathbf{n}^T\mathbf{n} - 2\boldsymbol{\mu}^T(\mathbf{E}\mathbf{x} + \mathbf{n} - \mathbf{y})\,, \tag{3.3.60}$$

with both $\mathbf{x}$, $\mathbf{n}$ appearing in the objective function. Setting the derivatives of (3.3.60) with respect to $\mathbf{x}$, $\mathbf{n}$, $\boldsymbol{\mu}$ to zero, and solving the resulting normal equations produces

$$\begin{aligned}
\tilde{\mathbf{x}} &= \mathbf{E}^T(\mathbf{E}\mathbf{E}^T + \alpha^2\mathbf{I})^{-1}\mathbf{y} && (3.3.61)\\
\tilde{\mathbf{n}} &= \left\{\mathbf{I} - \mathbf{E}\mathbf{E}^T(\mathbf{E}\mathbf{E}^T + \alpha^2\mathbf{I})^{-1}\right\}\mathbf{y} && (3.3.62)\\
\mathbf{C}_{\tilde{x}\tilde{x}} &= \mathbf{E}^T(\mathbf{E}\mathbf{E}^T + \alpha^2\mathbf{I})^{-1}\mathbf{R}_{nn}(\mathbf{E}\mathbf{E}^T + \alpha^2\mathbf{I})^{-1}\mathbf{E} && (3.3.63)\\
\tilde{\boldsymbol{\mu}} &= \tilde{\mathbf{n}} && (3.3.64)\\
\mathbf{P}_{\tilde{n}\tilde{n}} &= \left\{\mathbf{I} - \mathbf{E}\mathbf{E}^T(\mathbf{E}\mathbf{E}^T + \alpha^2\mathbf{I})^{-1}\right\}\mathbf{R}_{nn} \times \\
&\quad \left\{\mathbf{I} - \mathbf{E}\mathbf{E}^T(\mathbf{E}\mathbf{E}^T + \alpha^2\mathbf{I})^{-1}\right\}, && (3.3.65)
\end{aligned}$$

and as before, we could employ α^2 as a means to control the relative norms of $\tilde{\mathbf{x}}$, $\tilde{\mathbf{n}}$ and the elements of $\mathbf{C}_{\tilde{x}\tilde{x}}$. If we suppose that weights $\mathbf{W}^{T/2}$, $\mathbf{S}^{T/2}$ were applied to the equations prior to forming J, then the solution (3.3.61)–(3.3.65) is in the primed variables, and in terms of the original variables is

$$\begin{aligned}
\tilde{\mathbf{x}} &= \mathbf{S}\mathbf{E}^T(\mathbf{E}\mathbf{S}\mathbf{E}^T + \mathbf{W})^{-1}\mathbf{y} && (3.3.66)\\
\tilde{\mathbf{n}} &= \left\{\mathbf{I} - \mathbf{E}\mathbf{S}\mathbf{E}^T(\mathbf{E}\mathbf{S}\mathbf{E}^T + \mathbf{W})^{-1}\right\}\mathbf{y} && (3.3.67)\\
\mathbf{C}_{\tilde{x}\tilde{x}} &= \mathbf{S}\mathbf{E}^T(\mathbf{E}\mathbf{S}\mathbf{E}^T + \mathbf{W})^{-1}\mathbf{R}_{nn}(\mathbf{E}\mathbf{S}\mathbf{E}^T + \mathbf{W})^{-1}\mathbf{E}\mathbf{S} && (3.3.68)\\
\tilde{\boldsymbol{\mu}} &= \mathbf{W}^{-T/2}\tilde{\mathbf{n}} && (3.3.69)\\
\mathbf{P}_{\tilde{n}\tilde{n}} &= \left\{\mathbf{I} - \mathbf{E}\mathbf{S}\mathbf{E}^T(\mathbf{E}\mathbf{S}\mathbf{E}^T + \mathbf{W})^{-1}\right\}\mathbf{R}_{nn} \times \\
&\quad \left\{\mathbf{I} - \mathbf{E}\mathbf{S}\mathbf{E}^T(\mathbf{E}\mathbf{S}\mathbf{E}^T + \mathbf{W})^{-1}\right\}, && (3.3.70)
\end{aligned}$$

with α^2 absorbed into $\mathbf{S}$. Despite the different form, we claim that (3.3.66)–(3.3.68) are identical to (3.3.38)–(3.3.41)–their identity is readily shown by using the matrix inversion lemma in the form (3.1.25). A choice between the two forms is often made on the basis of the dimensionality of the matrices being inverted: $\mathbf{E}^T\mathbf{W}^{-1}\mathbf{E}$ is $N \times N$ and $\mathbf{E}\mathbf{S}\mathbf{E}^T$ is $M \times M$. But even this criterion is ambiguous, for example, because $\mathbf{W}$ is $M \times M$, and if it is not actually diagonal, or its inverse otherwise known, one would have to invert it.

Equations (3.3.38)–(3.3.40) and (3.3.66)–(3.3.70) result from two very different appearing objective functions–one in which the equations are imposed in the mean square (3.3.38)–(3.3.40), and one in which they are imposed exactly (3.3.66)–(3.3.70), using Lagrange multipliers. In the terminology of Sasaki (1970) and others, exact relationships are called *strong* constraints, and those imposed in the mean-square are *weak* ones. A preferable terminology, which we will sometimes use, is *hard* and *soft* constraints. But in the present situation in particular, the distinction is illusory: Although (3.3.2) are being imposed exactly, it is only the presence of the error term, $\mathbf{n}$, that permits the equations to be written as equalities and thus as hard constraints. The hard and soft constraints here produce an identical solution. In some circumstances, which we will discuss briefly below, one may wish to impose exact constraints upon the elements of $\tilde{x}_i$; these are often model constraints, for example, that the flow should be exactly geostrophic. But it is actually rare that one's models are exactly correct, and even geostrophy is always violated slightly. The solution (3.3.53)–(3.3.55) was derived from a true hard constraint, $\mathbf{Ex} = \mathbf{y}$, but we ended by rejecting it as generally inapplicable.

It should be ever more clear that $\mathbf{n}$ is only by convention discussed separately from $\mathbf{x}$ and is fully a part of the solution. The combined form (3.3.42), which literally treats $\mathbf{x}$, $\mathbf{n}$ as the solution, is imposed through a hard constraint on the objective function,

$$J = \boldsymbol{\xi}^T\boldsymbol{\xi} - 2\boldsymbol{\mu}^T(\mathbf{E}_1\boldsymbol{\xi} - \mathbf{y}_1)\,, \tag{3.3.71}$$

which is (3.3.60) with $\alpha^2 = 1$. (There are numerical advantages, however, in working with objects in two spaces of dimensions M and N rather than a single space of dimension $M + N$.)

3.4 The Singular Vector Expansion

Least squares is a very powerful, very useful method for finding solutions of linear simultaneous equations of any dimensionality, and one might wonder why it is necessary to discuss any other form of solution. But in the simplest form of least squares, the solution is dependent upon the existence of inverses of $\mathbf{E}^T\mathbf{E}$, or $\mathbf{E}\mathbf{E}^T$. In practice, their existence cannot be guaranteed, and we need to understand first what that means, the extent to which solutions can be found when the inverses do not exist, and the effect of introducing weight matrices $\mathbf{W}$, $\mathbf{S}$. This problem is intimately related to the issue of controlling solution and residual norms. Second, the relationship between the equations and the solutions is somewhat impenetrable, in the sense that structures in the solutions are not easily related to particular elements of the data y_i. For many purposes, particularly physical insight, understanding the structure of the solution is essential.

3.4.1 Simple Vector Expansions

Consider again the elementary problem (3.1.1) of representing an L-dimensional vector $\mathbf{f}$ as a sum of a complete set of L-orthonormal vectors $\mathbf{g}_i$, $1 \leq i \leq L$, $\mathbf{g}_i^T \mathbf{g}_j = \delta_{ij}$. Without error,

$$\mathbf{f} = \sum_{j=1}^{L} a_j \mathbf{g}_j, \, a_j = \mathbf{g}_j^T \mathbf{f} \, . \tag{3.4.1}$$

But if for some reason only the first K coefficients a_j are known, we can only approximate $\mathbf{f}$ by its first K terms:

$$\begin{aligned} \tilde{\mathbf{f}} &\approx \sum_{j=1}^{K} a_j \mathbf{g}_j \\ &= \mathbf{f} + \delta \mathbf{f}_1 \, , \end{aligned} \tag{3.4.2}$$

and there is an error, $\delta \mathbf{f}_1$. From the orthogonality of the $\mathbf{g}_i$, it follows that $\delta \mathbf{f}_1$ will have minimum l_2 norm if and only if it is orthogonal to the K vectors retained in the approximation, and if and only if a_j are given by (3.4.1). The only way the error could be reduced is by increasing K.

Define an $L \times K$ matrix $\mathbf{G}_K$ whose columns are the first K of the $\mathbf{g}_j$. Then $\mathbf{a} = \mathbf{G}_K^T \mathbf{f}$ is the vector of coefficients $a_j = \mathbf{g}_j^T \mathbf{f}$, $1 \leq j \leq K$, and the finite representation (3.4.2) is (one should write it out)

$$\tilde{\mathbf{f}} = \mathbf{G}_K \mathbf{a} = \mathbf{G}_K (\mathbf{G}_K^T \mathbf{f}) = (\mathbf{G}_K \mathbf{G}_K^T) \mathbf{f} \, , \quad \mathbf{a} = \{a_i\} \tag{3.4.3}$$

where the third equality follows from the associative properties of matrix multiplication. This expression shows that a *representation of a vector in an incomplete orthonormal set produces a resulting approximation that is a simple linear combination of the elements of the correct values* (i.e., a weighted average, or *filtered* version of them).

Because the columns of $\mathbf{G}_K$ are orthonormal, $\mathbf{G}_K^T \mathbf{G}_K = \mathbf{I}_K$–that is, the $K \times K$ identity matrix; but $\mathbf{G}_K \mathbf{G}_K^T \neq \mathbf{I}_L$ unless $K = L$ (that $\mathbf{G}_L \mathbf{G}_L^T = \mathbf{I}_L$ for $K = L$ follows from the theorem for square matrices, which show a left inverse is also a right inverse; see any book on linear algebra). If $K < L$, $\mathbf{G}_K$ is semi-orthogonal. If $K = L$, it is orthogonal; in this case, $\mathbf{G}_L^{-1} = \mathbf{G}_L^T$. If it is only semi-orthogonal, $\mathbf{G}_K^T$ is a left inverse but not a right inverse.

$\mathbf{G}_K \mathbf{G}_K^T$ is known as a *resolution matrix*, with a simple interpretation. Suppose the true value of $\mathbf{f}$ were

$$\mathbf{f}_{j_0} = [0 \;\; 0 \;\; 0 \;\; . \;\; . \;\; . \;\; 0 \;\; 1 \;\; 0 \;\; . \;\; 0 \;\; .. \;\; 0]^T ,$$

that is, a Kronecker delta with unity in element j_0. Then the incomplete

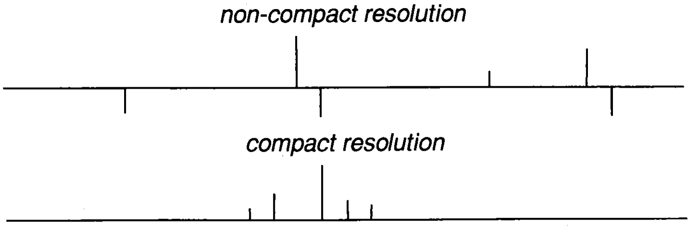

Figure 3–9. Incomplete vector expansions produce solutions that are linear combinations of the elements of correct ones. One can distinguish *compact* resolution where the linear combinations are a simple neighborhood weighted average (where *neighborhood* has a physical interpretation in time or space) from *non-compact* resolution where the averaging involves physically distant elements (although they are typically close in some other space, e.g., as measured in water mass properties). The figure shows, schematically, how the weights differ in the two cases.

expansion (3.4.2) or (3.4.3) would not reproduce the delta function but rather

$$\tilde{\mathbf{f}}_{j_0} = \mathbf{G}_K \mathbf{G}_K^T \mathbf{f}_{j_0}\,, \tag{3.4.4}$$

which is row (or column, because it is symmetric) j_0 of $\mathbf{G}_K \mathbf{G}_K^T$. The j_0-th row of the resolution matrix tells one what the corresponding form of the vector would be if its true form were a delta function at position j_0.

To form a Kronecker delta function requires a complete set of vectors. An analogous elementary result of Fourier analysis shows that a Dirac delta function demands contributions from all frequencies to arrange for a narrow, very high pulse. Removal of some of the requisite vectors (sinusoids) produces broadening and sidelobes. Here, depending upon the precise structure of the $\mathbf{g}_i$, the broadening and sidelobes can be complicated. If one is lucky, the effect could be a simple broadening (schematically shown in Figure 3–9) without distant sidelobes (Wiggins, 1972, who has a good discussion, calls this *compact resolution*), leading to the tidy interpretation of the result as a local average of the true values.

A resolution matrix has the property

$$\mathrm{trace}(\mathbf{G}_K \mathbf{G}_K^T) = K\,, \tag{3.4.5}$$

which follows from noting that

$$\mathrm{trace}(\mathbf{G}_K^T \mathbf{G}_K) = \mathrm{trace}(\mathbf{I}_K) = K\,,$$

and by direct evaluation,

$$\mathrm{trace}(\mathbf{G}_K \mathbf{G}_K^T) = \mathrm{trace}(\mathbf{G}_K^T \mathbf{G}_K)\,.$$

Orthogonal vector expansions are particularly simple to use and interpret, but their relevance to solving a set of simultaneous equations may be obscure. What we will show, however, is that we can always find sets of orthonormal vectors to simplify greatly the job of solving simultaneous equations. To do so, we digress to recall the basic elements of the eigenvector/eigenvalue problem.

Consider a *square*, $M \times M$ matrix $\mathbf{E}$ and the simultaneous equations

$$\mathbf{E}\mathbf{g}_i = \lambda_i \mathbf{g}_i, \quad 1 \le i \le M\,, \tag{3.4.6}$$

that is, the problem of finding a set of vectors $\mathbf{g}_i$ whose dot products with the rows of $\mathbf{E}$ are proportional to themselves. Such vectors are *eigenvectors*, and the constants of proportionality are the *eigenvalues.* Under special circumstances, the eigenvectors form an orthonormal spanning set. Textbooks show that if $\mathbf{E}$ is square and symmetric, such a result is guaranteed. Suppose for the moment that we have such a special case, and recall how eigenvectors can be used to solve (3.1.10). With an orthonormal, spanning set, both the known $\mathbf{y}$ and the unknown $\mathbf{x}$ can be written as

$$\mathbf{x} = \sum_{i=1}^{M} \alpha_i \mathbf{g}_i,\ \alpha_i = \mathbf{g}_i^T \mathbf{x}\,, \tag{3.4.7}$$

$$\mathbf{y} = \sum_{i=1}^{M} \beta_i \mathbf{g}_i,\ \beta_i = \mathbf{g}_i^T \mathbf{y}\,. \tag{3.4.8}$$

By convention, $\mathbf{y}$ is known, and therefore the β_i can be regarded as given. If we could find the α_i, $\mathbf{x}$ would be known.

Substitute (3.4.7) into (3.1.10), and using the eigenvector property,

$$\mathbf{E}\sum_{i=1}^{M} \alpha_i \mathbf{g}_i = \sum_{i=1}^{M} \left(\mathbf{g}_i^T \mathbf{y}\right) \mathbf{g}_i$$

or

$$\sum_{i=1}^{M} \alpha_i \lambda_i \mathbf{g}_i = \sum_{i} \left(\mathbf{g}_i^T \mathbf{y}\right) \mathbf{g}_i\,. \tag{3.4.9}$$

But the expansion vectors are orthonormal, and so

$$\lambda_i \alpha_i = \mathbf{g}_i^T \mathbf{y} \tag{3.4.10}$$

$$\alpha_i = \frac{\mathbf{g}_i^T \mathbf{y}}{\lambda_i} \tag{3.4.11}$$

$$\mathbf{x} = \sum_{i=1}^{N} \frac{\mathbf{g}_i^T \mathbf{y}}{\lambda_i} \mathbf{g}_i\,. \tag{3.4.12}$$

Apart from an evident difficulty if any eigenvalue vanishes, the problem is now completely solved. If we define a diagonal matrix, $\boldsymbol{\Lambda}$, with elements,

λ_i, ordered by convention in descending numerical value, and the matrix $\mathbf{G}$, whose columns are the corresponding $\mathbf{g}_i$ in the same order, the solution to (3.1.10) can be written [from (3.4.7), (3.4.10)–(3.4.12)] as

$$\boldsymbol{\alpha} = \boldsymbol{\Lambda}^{-1}\mathbf{G}^T\mathbf{y} \tag{3.4.13}$$

$$\mathbf{x} = \mathbf{G}\boldsymbol{\Lambda}^{-1}\mathbf{G}^T\mathbf{y} \tag{3.4.14}$$

where $\boldsymbol{\Lambda}^{-1} = \text{diag}(1/\lambda_i)$.

Vanishing eigenvalues, $i = i_0$, cause trouble, and we must consider them. Let the corresponding eigenvectors be $\mathbf{g}_{i_0}$. Then any part of the solution that is proportional to such an eigenvector is annihilated by $\mathbf{E}$–that is, $\mathbf{g}_{i_0}$ is orthogonal to all the rows of $\mathbf{E}$. Such a result means that there is no possibility that anything in $\mathbf{y}$ could provide any information about the coefficient α_{i_0}. If $\mathbf{y}$ corresponds to a set of observations (data), then $\mathbf{E}$ represents the connection (mapping) between system unknowns and observations. The existence of zero eigenvalues shows that the act of observation of $\mathbf{x}$ removes certain structures in the solution that are then indeterminate. Vectors $\mathbf{g}_{i_0}$ (and there may be many of them) are said to lie in the nullspace of $\mathbf{E}$. Eigenvectors corresponding to nonzero eigenvalues lie in its range. The simplest example is given by the observations

$$x_1 + x_2 = 3\,,$$
$$x_1 + x_2 = 3\,.$$

Any structure in $\mathbf{x}$ such that $x_1 = -x_2$ is destroyed by this observation, and by inspection, the nullspace vector must be $\mathbf{g}_2 = [1 \quad -1]^T/\sqrt{2}$ (the purpose of showing the observation twice is to produce an $\mathbf{E}$ that is square).

Suppose there are $K < M$ nonzero λ_i. Then for $i > K$, Equation (3.4.10) is

$$0\alpha_i = \mathbf{g}_i^T\mathbf{y}, \quad K+1 \le i \le M\,, \tag{3.4.15}$$

and two cases must be distinguished.

Case (1):

$$\mathbf{g}_i^T\mathbf{y} = 0\,, \qquad K+1 \le i \le M\,. \tag{3.4.16}$$

We could then put $\alpha_i = 0$, $K+1 \le i \le M$, and the solution can be written

$$\tilde{\mathbf{x}} = \sum_{i=1}^{K} \frac{\mathbf{g}_i^T\mathbf{y}}{\lambda_i}\mathbf{g}_i \tag{3.4.17}$$

and $\mathbf{E}\tilde{\mathbf{x}} = \mathbf{y}$, *exactly*. We have put a tilde over $\mathbf{x}$ because a solution of the form

$$\tilde{\mathbf{x}} = \sum_{i=1}^{K} \frac{\mathbf{g}_i^T \mathbf{y}}{\lambda_i} \mathbf{g}_i + \sum_{i=K+1}^{M} \alpha_i \mathbf{g}_i \,, \tag{3.4.18}$$

with the remaining α_i taking on arbitrary values, also satisfies the equations exactly. That is to say, the true value of $\mathbf{x}$ *could* contain structures proportional to the nullspace vectors of $\mathbf{E}$, but the equations (3.1.10) neither require their presence nor provide the information necessary to determine their amplitudes. We thus have a situation with a *solution nullspace.* If the matrix $\mathbf{G}_K$ is $M \times K$, carrying only the first K of the $\mathbf{g}_i$–that is, the range vectors–$\boldsymbol{\Lambda}_K$ is $K \times K$ with only the first K, nonzero eigenvalues, and the columns of $\mathbf{Q}_G$ are the M-K nullspace vectors [it is $M \times (M - K)$], then the solutions (3.4.17) and (3.4.18) are

$$\tilde{\mathbf{x}} = \mathbf{G}_K \boldsymbol{\Lambda}_K^{-1} \mathbf{G}_K^T \mathbf{y} \,, \tag{3.4.19}$$

$$\tilde{\mathbf{x}} = \mathbf{G}_K \boldsymbol{\Lambda}_K^{-1} \mathbf{G}_K^T \mathbf{y} + \mathbf{Q}_G \boldsymbol{\alpha}_G \tag{3.4.20}$$

where $\boldsymbol{\alpha}_G$ is the vector of unknown nullspace coefficients, respectively. Equation (3.4.16) is often known as a *solvability condition.* The solution (3.4.19) with no nullspace contribution will be called the *particular* solution.

If $\mathbf{G}$ is written as a partitioned matrix,

$$\mathbf{G} = \{\mathbf{G}_K \quad \mathbf{Q}_G\} \,,$$

it follows from the column orthonormality that

$$\mathbf{G}\mathbf{G}^T = \mathbf{I}_L = \mathbf{G}_K \mathbf{G}_K^T + \mathbf{Q}_G \mathbf{Q}_G^T \tag{3.4.21}$$

or

$$\mathbf{Q}_G \mathbf{Q}_G^T = \mathbf{I}_L - \mathbf{G}_K \mathbf{G}_K^T \,. \tag{3.4.22}$$

Case (2):

$$\mathbf{g}_i^T \mathbf{y} \neq 0 \,, \qquad i > K \,, \tag{3.4.23}$$

for one or more of the nullspace vectors. In this case, Equation (3.4.10) is the contradiction

$$0\alpha_i \neq 0 \,,$$

and Equation (3.4.9) is actually

$$\sum_{i=1}^{K} \lambda_i \alpha_i \mathbf{g}_i = \sum_{i=1}^{M} (\mathbf{g}_i^T \mathbf{y}) \mathbf{g}_i, \; K < M \,, \tag{3.4.24}$$

that is, with differing upper limits on the sums. Owing to the orthonormality of the $\mathbf{g}_i$, there is no choice of α_i, $1 \le i \le K$ on the left that can match the last M-K terms on the right. Evidently there is no solution in the conventional sense unless (3.4.16) is satisfied, hence the name *solvability condition.* What is the best we might do? Define *best* to mean that the solution $\tilde{\mathbf{x}}$ should be chosen such that

$$\mathbf{E}\tilde{\mathbf{x}} = \tilde{\mathbf{y}}$$

where the difference, $\tilde{\mathbf{n}} = \mathbf{y} - \tilde{\mathbf{y}}$, which we call the *residual,* should be as small as possible (in the l_2 norm). If this choice is made, then the orthogonality of the $\mathbf{g}_i$ shows immediately that the best choice is still (3.4.11), $1 \le i \le K$. No choice of nullspace vector coefficients, nor any other value of the coefficients of the range vectors, can reduce the norm of $\mathbf{n}$. The best solution is then also (3.4.17) or (3.4.19).

In this situation, we are no longer solving the equations (3.1.10) but rather are dealing with a set that could be written

$$\mathbf{E}\mathbf{x} \sim \mathbf{y} \tag{3.4.25}$$

where the demand is for a solution that is the best possible, in the sense just defined. Such statements of approximation are awkward, and it is more useful to always rewrite (3.4.25) as

$$\mathbf{E}\mathbf{x} + \mathbf{n} = \mathbf{y} \tag{3.4.26}$$

where $\mathbf{n}$ is the residual. If $\tilde{\mathbf{x}}$ is given by (3.4.18), then

$$\tilde{\mathbf{n}} = \sum_{i=K+1}^{M} (\mathbf{g}_i^T \mathbf{y})\mathbf{g}_i \tag{3.4.27}$$

by (3.4.24). Notice that $\tilde{\mathbf{n}}^T \tilde{\mathbf{y}} = 0$.

This situation, where we started with M equations in M unknowns, but found in practice that some structures of the solution could not actually be determined, is labeled *formally just-determined,* where the word *formally* alludes to the fact that the mere appearance of a just-determined system did not mean that the characterization was true in practice. One or more vanishing eigenvalues means that the rows and columns $\mathbf{E}$ are not spanning sets.

Some decision has to be made about the coefficients of the nullspace vectors in (3.4.18) or (3.4.20). We could use the form as it stands, regarding it as the *general solution.* The analogy with the solution of differential equations should be apparent–typically, such equations have particular and homogeneous solutions. In the present case, the homogeneous solution corresponds to the nullspace vectors. When solving a differential equation, determination of the magnitude of the homogeneous solution requires additional information, often provided by boundary or initial conditions; here, additional information is also necessary but is missing. Despite the presence of indeterminate elements in the solution, we know exactly what they are: proportional to the nullspace vectors. Depending upon the specific situation, we might conceivably be in a position to obtain more observations and would seriously consider observational strategies directed at detecting these missing structures. The reader is also reminded of the discussion of the Neumann problem in Section 1.3.

Another approach is to define a simplest solution, appealing to what is usually known as *Occam's Razor*, or the *principal of parsimony*, that in choosing between multiple explanations of a given phenomenon, the simplest one is usually the best. What is simplest can be debated, but here there is a compelling choice: The solution (3.4.17) or (3.4.19)–that is, without any nullspace contributions, is less structured than any other solution. [It is often but not always (again, recall the Neumann problem) true that the nullspace vectors are more "wiggily" than those in the range. In any case, including any vector not required by the data is arguably producing more structure than is required.] Setting all the unknown α_i to zero is thus one choice. It follows from the orthogonality of the $\mathbf{g}_i$ that this particular solution is also the one of minimum solution norm. Later, we will see some other choices for the nullspace vectors.

If the nullspace vector contributions are set to zero, the true solution has been expanded in an incomplete set of orthonormal vectors. Thus, $\mathbf{G}_K\mathbf{G}_K^T$ is the resolution matrix, and the relationship between the true solution and the particular one is just

$$\tilde{\mathbf{x}} = \mathbf{G}_K\mathbf{G}_K^T\mathbf{x} = \mathbf{x} - \mathbf{Q}_G\boldsymbol{\alpha}_G, \quad \tilde{\mathbf{y}} = \mathbf{G}_K\mathbf{G}_K^T\mathbf{y}, \quad \tilde{\mathbf{n}} = \mathbf{Q}_G\mathbf{Q}_G^T\mathbf{y}\,. \qquad (3.4.28)$$

These results are so important, we recapitulate them: (3.4.18) or (3.4.20) is the general solution. There are three vectors involved–one of them, $\mathbf{y}$, is known, and two of them, $\mathbf{x}$, $\mathbf{n}$, are unknown. Because of the assumption that $\mathbf{E}$ has a complete orthonormal set of eigenvectors, all three of these vectors can be expanded, exactly, as

$$\mathbf{x} = \sum_{i=1}^{M} \alpha_i\mathbf{g}_i, \quad \mathbf{n} = \sum_{i=1}^{M} \gamma_i\mathbf{g}_i, \quad \mathbf{y} = \sum_{i=1}^{M} (\mathbf{y}^T\mathbf{g}_i)\mathbf{g}_i\,. \qquad (3.4.29)$$

Substituting into (3.4.26), and using the eigenvector property produces

$$\sum_{i=1}^{M} \alpha_i \mathbf{E}\mathbf{g}_i + \sum_{i=1}^{M} \gamma_i \mathbf{g}_i = \sum_{i=1}^{M} (\mathbf{y}^T \mathbf{g}_i)\mathbf{g}_i$$
$$= \sum_{i=1}^{K} \lambda_i \alpha_i \mathbf{g}_i + \sum_{i=1}^{M} \gamma_i \mathbf{g}_i = \sum_{i=1}^{M} (\mathbf{y}^T \mathbf{g}_i)\mathbf{g}_i \,. \tag{3.4.30}$$

From the orthogonality property, we must have

$$\lambda_i \alpha_i + \gamma_i = \mathbf{y}^T \mathbf{g}_i, \quad 1 \leq i \leq K\,, \tag{3.4.31}$$
$$\gamma_i = \mathbf{y}^T \mathbf{g}_i, \quad K+1 \leq i \leq M\,. \tag{3.4.32}$$

In dealing with the first relationship, we must make a choice. If we set

$$\gamma_i = \mathbf{g}_i^T \mathbf{n} = 0\,, \quad 1 \leq i \leq K\,, \tag{3.4.33}$$

the residual norm is made as small as possible by completely eliminating the range vectors from the residual. This choice is motivated by the attempt to satisfy the equations as well as possible but is seen to have elements of arbitrariness. A decision about other possibilities depends upon knowing more about the system and will be the focus of considerable later attention.

It may be objected that this entire development is of little use, because the problems discussed in Chapter 2 produced **E** matrices that could not be guaranteed to have complete orthonormal sets of eigenvectors. Indeed, the problems considered produce matrices that are usually nonsquare and for which the eigenvector problem is not even defined.

For arbitrary *square* matrices, the question of when a complete orthonormal set of eigenvectors exists is not difficult to answer but becomes somewhat elaborate; it is treated in all texts on linear algebra. Brogan (1985) has a succinct discussion.

In the general situation, where an $N \times N - \mathbf{E}$ is not symmetric, one must consider cases in which there are N distinct eigenvalues and where some are repeated, and the general approach requires the so-called Jordan form. But we will find a way to avoid these intricacies and yet deal with sets of simultaneous equations of arbitrary dimensions, not just square ones. In the next several sections, a machinery is developed for doing exactly that. Although the mathematics are necessarily somewhat more complicated than is employed in solving the just-determined simultaneous linear equations using a complete orthonormal eigenvector set, this simplest problem provides full analogues to all of the issues in the more general case, and the reader will probably find it helpful to refer back to this situation for insight.

Before leaving this special case, note one more useful property of the eigenvectors/eigenvalues. For the moment, let $\mathbf{G}$ have all its columns, containing both the range and nullspace vectors, with the nullspace vectors being last. It is thus an $M \times M$ matrix. Correspondingly, let $\boldsymbol{\Lambda}$ contain all the eigenvalues on its diagonal, including the zero ones; it, too, is $M \times M$. Then the eigenvector definition (3.4.6) produces

$$\mathbf{EG} = \mathbf{G}\boldsymbol{\Lambda} . \tag{3.4.34}$$

Multiply both sides of (3.4.34) by $\mathbf{G}^T$:

$$\mathbf{G}^T\mathbf{EG} = \mathbf{G}^T\mathbf{G}\boldsymbol{\Lambda} = \boldsymbol{\Lambda} \tag{3.4.35}$$

using the orthogonality of $\mathbf{G}$; $\mathbf{G}$ is said to *diagonalize* $\mathbf{E}$. Now multiply both sides of (3.4.35) on the left by $\mathbf{G}$ and on the right by $\mathbf{G}^T$:

$$\mathbf{GG}^T\mathbf{EGG}^T = \mathbf{G}\boldsymbol{\Lambda}\mathbf{G}^T \tag{3.4.36}$$

or, using the orthogonality of $\mathbf{G}$ when it has all its columns,

$$\mathbf{E} = \mathbf{G}\boldsymbol{\Lambda}\mathbf{G}^T , \tag{3.4.37}$$

a useful decomposition of $\mathbf{G}$, consistent with the symmetry of $\mathbf{E}$. Recall that $\boldsymbol{\Lambda}$ has zeros on the diagonal corresponding to the zero eigenvalues, and the corresponding rows and columns are entirely zero. Writing out (3.4.37), these zero rows and columns multiply all the nullspace vector columns of $\mathbf{G}$ by zero, and it is found that the nullspace columns of $\mathbf{G}$ can be eliminated, $\boldsymbol{\Lambda}$ reduced to its $K \times K$ form, and the decomposition (3.4.37) is still exact in the form

$$\mathbf{E} = \mathbf{G}_K\boldsymbol{\Lambda}_K\mathbf{G}_K^T . \tag{3.4.38}$$

Then the simultaneous equations (3.4.26) are

$$\mathbf{G}_K\boldsymbol{\Lambda}_K\mathbf{G}_K^T\mathbf{x} + \mathbf{n} = \mathbf{y} . \tag{3.4.39}$$

Left multiply both sides by $\boldsymbol{\Lambda}_K^{-1}\mathbf{G}_K^T$ (existence of the inverse is guaranteed by the removal of zero eigenvalues), and

$$\mathbf{G}_K^T\mathbf{x} + \boldsymbol{\Lambda}_K^{-1}\mathbf{G}_K^T\mathbf{n} = \boldsymbol{\Lambda}_K^{-1}\mathbf{G}_K^T\mathbf{y} . \tag{3.4.40}$$

But $\mathbf{G}_K^T\mathbf{x}$ are the projection of $\mathbf{x}$ onto the range vectors of $\mathbf{E}$, and $\mathbf{G}_K^T\mathbf{n}$ is the same projection of the noise. We have agreed to regard the latter as zero, and we obtain

$$\mathbf{G}_K^T\mathbf{x} = \boldsymbol{\Lambda}_K^{-1}\mathbf{G}_K^T\mathbf{y} ,$$

the dot products of the range of $\mathbf{E}$ with the solution. Hence, it must be true, because the range vectors are orthonormal, that

$$\tilde{\mathbf{x}} \equiv \mathbf{G}_K\mathbf{G}_K^T\mathbf{x} \equiv \mathbf{G}_K\boldsymbol{\Lambda}_K^{-1}\mathbf{G}_K^T\mathbf{y}\,, \tag{3.4.41}$$

$$\tilde{\mathbf{y}} = \mathbf{E}\tilde{\mathbf{x}} = \mathbf{G}_K\mathbf{G}_K^T\mathbf{y}\,, \tag{3.4.42}$$

which is identical to the particular solution (3.4.17). The residuals are

$$\tilde{\mathbf{n}} = \mathbf{y}-\tilde{\mathbf{y}} = \mathbf{y}-\mathbf{E}\tilde{\mathbf{x}} = \mathbf{y}-\mathbf{G}_K\mathbf{G}_K^T\mathbf{y} = (\mathbf{I}_L-\mathbf{G}_K\mathbf{G}_K^T)\mathbf{y} = \mathbf{Q}_G\mathbf{Q}_G^T\mathbf{y}\,. \tag{3.4.43}$$

One again has $\tilde{\mathbf{n}}^T\tilde{\mathbf{y}} = 0$.

Expression (3.4.43) shows that multiplication by $\mathbf{Q}_K\mathbf{Q}_K^T = \mathbf{I} - \mathbf{G}_K\mathbf{G}_K^T$ projects a vector onto the nullspace of $\mathbf{E}$, just as $\mathbf{G}_K\mathbf{G}_K^T$ projects onto its range. Such operators have an *idempotent* property,

$$(\mathbf{I} - \mathbf{G}_K\mathbf{G}_K^T)^n = (\mathbf{I} - \mathbf{G}_K\mathbf{G}_K^T),\ n = \text{integer}$$

–projection onto the nullspace is invariant. For future reference, notice that the reduced decomposition (3.4.38) permits writing,

$$\mathbf{E}^T(\mathbf{E}\mathbf{E}^T)^{-1}\mathbf{E} = \mathbf{E}(\mathbf{E}^T\mathbf{E})^{-1}\mathbf{E}^T = \mathbf{G}_K\mathbf{G}_K^T\,; \tag{3.4.44}$$

hence, (3.4.41) is

$$\tilde{\mathbf{x}} = \mathbf{E}^T(\mathbf{E}\mathbf{E}^T)^{-1}\mathbf{E}\mathbf{x} = \mathbf{E}(\mathbf{E}^T\mathbf{E})^{-1}\mathbf{E}^T\mathbf{x}\,, \tag{3.4.45}$$

and thus

$$\mathbf{Q}_G\mathbf{Q}_G^T = (\mathbf{I} - \mathbf{E}^T(\mathbf{E}\mathbf{E}^T)^{-1}\mathbf{E}) = (\mathbf{I} - \mathbf{E}(\mathbf{E}^T\mathbf{E})^{-1}\mathbf{E}^T)\,, \tag{3.4.46}$$

and the latter is also idempotent (see (3.3.11)).

The bias of the solution (3.4.17) or (3.4.41) is

$$< \tilde{\mathbf{x}} - \mathbf{x} > = \mathbf{G}_K\boldsymbol{\Lambda}_K^{-1}\mathbf{G}_K^T < \mathbf{y} > \ -\sum_{i=1}^{N} \alpha_i\mathbf{g}_i = -\mathbf{Q}_G\boldsymbol{\alpha}_G\,, \tag{3.4.47}$$

and so the solution is biased unless $\boldsymbol{\alpha}_G = 0$.

The uncertainty is

$$\begin{aligned}\mathbf{P} = D^2(\tilde{\mathbf{x}} - \mathbf{x}) &=< \mathbf{G}_K\boldsymbol{\Lambda}_K^{-1}\mathbf{G}_K^T(\mathbf{y}_0 + \mathbf{n} - \mathbf{y}_0)(\mathbf{y}_0 + \mathbf{n} - \mathbf{y}_0)^T\mathbf{G}_K\boldsymbol{\Lambda}_K^{-1}\mathbf{G}_K^T > \\ &\quad + < \mathbf{Q}_G\boldsymbol{\alpha}_G\boldsymbol{\alpha}_G^T\mathbf{Q}_G^T > \\ &= \mathbf{G}_K\boldsymbol{\Lambda}_K^{-1}\mathbf{G}_K^T < \mathbf{n}\mathbf{n}^T > \mathbf{G}_K\boldsymbol{\Lambda}_K^{-1}\mathbf{G}_K^T + \mathbf{Q}_G < \boldsymbol{\alpha}_G\boldsymbol{\alpha}_G^T > \mathbf{Q}_G^T \\ &= \mathbf{G}_K\boldsymbol{\Lambda}_K^{-1}\mathbf{G}_K^T\mathbf{R}_{nn}\mathbf{G}_K\boldsymbol{\Lambda}_K^{-1}\mathbf{G}_K^T + \mathbf{Q}_G\mathbf{R}_{\alpha\alpha}\mathbf{Q}_G^T \\ &= \mathbf{C}_{\tilde{x}\tilde{x}} + \mathbf{Q}_G\mathbf{R}_{\alpha\alpha}\mathbf{Q}_G^T\,,\end{aligned} \tag{3.4.48}$$

and $\mathbf{R}_{\alpha\alpha}$ are the second moments of the coefficients of the nullspace vectors.

Under the special circumstances that the residuals, $\mathbf{n}$, are white noise, with $\mathbf{R}_{nn} = \sigma_n^2\mathbf{I}$, (3.4.48) reduces to

$$\mathbf{P} = \sigma_n^2 \mathbf{G}_K \boldsymbol{\Lambda}_K^{-2} \mathbf{G}_K^T + \mathbf{Q}_G \mathbf{R}_{\alpha\alpha} \mathbf{Q}_G^T . \tag{3.4.49}$$

Either case shows that the uncertainty of the minimal solution is made up of two distinct parts. The first part, the solution covariance, $\mathbf{C}_{\tilde{x}\tilde{x}}$, arises owing to the noise present in the observations and generates uncertainty in the coefficients of the range vectors; the second contribution arises from the missing nullspace vector contribution. Either term can dominate. The magnitude of the noise term depends largely upon the ratio of the noise variance, σ_n^2, to the smallest nonzero singular value, λ_K^2. $\mathbf{R}_{\alpha\alpha}$ may be entirely unknown, or an estimate of its value might be available from prior information (e.g., on the basis of the difference between the expected variance of $\mathbf{x}$ and the estimated variance of $\tilde{\mathbf{x}}$).

3.4.2 The Singular Vector Expansion and Singular Value Decomposition

Instead of using the least-squares method already described to find solutions to sets of linear simultaneous equations, consider the possibility, suggested by the eigenvector method, of expanding the solution $\mathbf{x}$ in a set of orthonormal vectors. Equation (3.3.2) involves one vector, $\mathbf{x}$, of dimension N, and two vectors, $\mathbf{y}$, $\mathbf{n}$, of dimension M. We would like to use spanning orthonormal vectors but cannot expect, with two different vector dimensions involved, to use just one set: $\mathbf{x}$ can be expanded exactly in N, N-dimensional orthonormal vectors; and similarly, $\mathbf{y}$ and $\mathbf{n}$ can be exactly represented in M, M-dimensional orthonormal vectors. There are an infinite number of ways to select two such sets. But one particularly useful pair can be found, based upon the structure of $\mathbf{E}$.

The simple development leading to the discussion of the above solutions was based upon the theorem about the eigenvectors of a matrix $\mathbf{E}$ which was symmetric, so that they were guaranteed to be an orthonormal spanning set. Let us construct such a matrix out of an arbitrary $\mathbf{E}$. Put

$$\mathbf{B} = \begin{Bmatrix} \mathbf{0} & \mathbf{E}^T \\ \mathbf{E} & \mathbf{0} \end{Bmatrix} , \tag{3.4.50}$$

which by definition is not only square (dimension $M+N$ by $M+N$) but symmetric. Thus, $\mathbf{B}$ satisfies the theorem just alluded to, and the eigenvalue problem

$$\mathbf{B}\mathbf{q}_i = \lambda_i \mathbf{q}_i \tag{3.4.51}$$

will give rise to $M+N$ orthonormal eigenvectors $\mathbf{q}_i$ (an orthonormal spanning set) whether or not the λ_i are distinct or nonzero. Writing out (3.4.51),

$$\begin{Bmatrix} \mathbf{0} & \mathbf{E}^T \\ \mathbf{E} & \mathbf{0} \end{Bmatrix} \begin{bmatrix} q_{1i} \\ \cdot \\ q_{Ni} \\ q_{N+1,i} \\ \cdot \\ q_{N+M,i} \end{bmatrix} = \lambda_i \begin{bmatrix} q_{1i} \\ \cdot \\ q_{Ni} \\ q_{N+1,i} \\ \cdot \\ q_{N+M,i} \end{bmatrix}, \tag{3.4.52}$$

where q_{pi} is the p-th element of $\mathbf{q}_i$. Taking note of the zero matrices, (3.4.52) may be rewritten

$$\mathbf{E}^T \begin{bmatrix} q_{N+1,i} \\ \cdot \\ q_{N+M,i} \end{bmatrix} = \lambda_i \begin{bmatrix} q_{1i} \\ \cdot \\ q_{Ni} \end{bmatrix}, \tag{3.4.53}$$

$$\mathbf{E} \begin{bmatrix} q_{1i} \\ \cdot \\ q_{Ni} \end{bmatrix} = \lambda_i \begin{bmatrix} q_{N+1,i} \\ \cdot \\ q_{N+M,i} \end{bmatrix}. \tag{3.4.54}$$

Let

$$\mathbf{u}_i = \begin{bmatrix} q_{N+1,i} \\ \cdot \\ q_{N+M,i} \end{bmatrix}, \quad \mathbf{v}_i = \begin{bmatrix} q_{1i} \\ \cdot \\ q_{Ni} \end{bmatrix}, \text{ or, } \mathbf{q}_i = \begin{bmatrix} \mathbf{v}_i \\ \mathbf{u}_i \end{bmatrix}, \tag{3.4.55}$$

that is, defining the first N elements of $\mathbf{q}_i$ to be $\mathbf{v}_i$ and the last M to be $\mathbf{u}_i$. Then (3.4.53)–(3.4.54) are

$$\mathbf{E}\mathbf{v}_i = \lambda_i \mathbf{u}_i\,, \tag{3.4.56}$$

$$\mathbf{E}^T\mathbf{u}_i = \lambda_i \mathbf{v}_i\,. \tag{3.4.57}$$

If (3.4.56) is left multiplied by $\mathbf{E}^T$, and using (3.4.57), one has

$$\mathbf{E}^T\mathbf{E}\mathbf{v}_i = \lambda_i^2 \mathbf{v}_i\,. \tag{3.4.58}$$

Similarly, left multiplying (3.4.57) by $\mathbf{E}$ and using (3.4.56) produces

$$\mathbf{E}\mathbf{E}^T\mathbf{u}_i = \lambda_i^2 \mathbf{u}_i\,. \tag{3.4.59}$$

These last two equations show, surprisingly, that the $\mathbf{u}_i$, $\mathbf{v}_i$ each separately satisfy two independent eigenvector/eigenvalue problems of the square symmetric matrices $\mathbf{E}\mathbf{E}^T$, $\mathbf{E}^T\mathbf{E}$. If one of M, N is much smaller than the other, one need only solve the smaller of the two eigenvalues for either of $\mathbf{u}_i$, $\mathbf{v}_i$, with the other set calculated from (3.4.56) or (3.4.57).

The $\mathbf{u}_i$, $\mathbf{v}_i$ are called *singular vectors*, and the λ_i are the *singular values*. By convention, the λ_i are ordered in decreasing numerical value. Also by convention, they are all nonnegative (taking the negative values of λ_i produces singular vectors differing only by a sign from those corresponding to the positive roots, and thus they are not independent vectors). Equations (3.4.56)–(3.4.57) provide a relationship between each $\mathbf{u}_i$ and each $\mathbf{v}_i$. But because in general, $M \neq N$, there will be more of one set than another. The only way Equations (3.4.56)–(3.4.57) can be consistent is if $\lambda_i = 0$, $i > \min(M, N)$ [where $\min(M, N)$ is read as "the minimum of M and N"]. Suppose $M < N$. Then (3.4.59) is solved for $\mathbf{u}_i$, $1 \leq i \leq M$, and (3.4.57) is used to find the corresponding $\mathbf{v}_i$. There are $N - M$ $\mathbf{v}_i$ that are not generated this way but which can be found using the Gram-Schmidt method.

Let there be K nonzero λ_i; then

$$\mathbf{E}\mathbf{v}_i \neq 0\,, \quad 1 \leq i \leq K\,. \tag{3.4.60}$$

These $\mathbf{v}_i$ are known as the *range vectors of* **E** or the *solution range vectors.*" For the remaining N-K vectors $\mathbf{v}_i$,

$$\mathbf{E}\mathbf{v}_i = 0\,, \quad K + 1 \leq i \leq N\,, \tag{3.4.61}$$

known as the *nullspace vectors of* **E** or the *nullspace of the solution.* If $K < M$, there will be K of the $\mathbf{u}_i$ such that

$$\mathbf{E}^T\mathbf{u}_i = 0\,, \text{ or } \mathbf{u}_i^T\mathbf{E} \neq 0\,, \quad 1 \leq i \leq K\,, \tag{3.4.62}$$

which are the *range vectors of* $\mathbf{E}^T$ and M-K of the $\mathbf{u}_i$ such that

$$\mathbf{E}^T\mathbf{u}_i = 0\,, \text{ or } \mathbf{u}_i^T\mathbf{E} = 0\,, \quad K + 1 \leq i \leq M\,, \tag{3.4.63}$$

the *nullspace vectors of* $\mathbf{E}^T$ or the *data, or observation, nullspace vectors.* The nullspace of **E** is spanned by its nullspace vectors, the range of **E** is spanned by the range vectors, etc., in the sense, for example, that an arbitrary vector lying in the range is perfectly described by a sum of the range vectors.

Because the $\mathbf{u}_i$, $\mathbf{v}_i$ are complete orthonormal sets in their corresponding spaces, we can expand **x**, **y**, **n** without error:

$$\mathbf{x} = \sum_{i=1}^{N} \alpha_i \mathbf{v}_i\,, \quad \mathbf{y} = \sum_{j=1}^{M} \beta_i \mathbf{u}_i\,, \quad \mathbf{n} = \sum_{i=1}^{M} \gamma_i \mathbf{u}_i\,, \tag{3.4.64}$$

where **y** has been measured, so that we know $\beta_j = \mathbf{u}_j^T\mathbf{y}$. To find the solution, we need α_i, and to find the noise, we need the γ_i. Substitute (3.4.64) into the equations (3.3.2), and using (3.4.56)–(3.4.57),

$$\sum_{i=1}^{N} \alpha_i \mathbf{E}\mathbf{v}_i + \sum_{i=1}^{M} \gamma_i \mathbf{u}_i = \sum_{i=1}^{K} \alpha_i \lambda_i \mathbf{u}_i + \sum_{i=1}^{M} \gamma_i \mathbf{u}_i = \sum_{i=1}^{M} (\mathbf{u}_i^T\mathbf{y})\mathbf{u}_i\,. \tag{3.4.65}$$

Notice the differing upper limits on the summations. By the orthonormality of the singular vectors, (3.4.65) can be solved as

$$\alpha_i \lambda_i + \gamma_i = \mathbf{u}_i^T \mathbf{y}\,, \quad i = 1 \text{ to } M\,, \tag{3.4.66}$$

$$\alpha_i = (\mathbf{u}_i^T \mathbf{y} - \gamma_i)/\lambda_i\,, \quad \lambda_i \neq 0\,, \quad 1 \leq i \leq K\,. \tag{3.4.67}$$

In these equations, if $\lambda_i \neq 0$, nothing prevents setting $\gamma_i = 0$–that is,

$$\mathbf{u}_i^T \mathbf{n} = 0\,, \quad 1 \leq i \leq K \tag{3.4.68}$$

should we wish, which would have the effect of making the noise norm as small as possible. Then (3.4.67) produces

$$\alpha_i = \frac{\mathbf{u}_i^T \mathbf{y}}{\lambda_i}\,, \quad 1 \leq i \leq K\,. \tag{3.4.69}$$

But, because $\lambda_i = 0$, $i > K$, the only solution for these values of i to (3.4.66) is $\gamma_i = \mathbf{u}_i^T \mathbf{y}$, and α_i is indeterminate. These γ_i are nonzero, meaning that there is always a residual, except in the event (unlikely with real data) that

$$\mathbf{u}_i^T \mathbf{y} = 0\,, \quad K+1 \leq i \leq N\,. \tag{3.4.70}$$

This last equation is called a *solvability condition* in direct analogy to (3.4.16).

The solution obtained in this manner now has the following form:

$$\tilde{\mathbf{x}} = \sum_{i=1}^{K} \frac{\mathbf{u}_i^T \mathbf{y}}{\lambda_i} \mathbf{v}_i + \sum_{i=K+1}^{N} \alpha_i \mathbf{v}_i\,, \tag{3.4.71}$$

$$\tilde{\mathbf{y}} = \mathbf{E}\tilde{\mathbf{x}} = \sum_{i=1}^{K} (\mathbf{u}_i^T \mathbf{y}) \mathbf{u}_i\,, \tag{3.4.72}$$

$$\tilde{\mathbf{n}} = \sum_{i=K+1}^{M} (\mathbf{u}_i^T \mathbf{y}) \mathbf{u}_i\,. \tag{3.4.73}$$

The coefficients of the last N-K of the $\mathbf{v}_i$ in Equation (3.4.71), the solution nullspace vectors, are arbitrary, representing structures in the solution about which the equations provide no information. A nullspace is always present unless $K = N$. The solution residuals are directly proportional to the nullspace vectors of $\mathbf{E}^T$ and will vanish only if $K = M$, or the solvability conditions are met.

Just as in the square symmetric case, no choice of the coefficients of the solution nullspace vectors can have any effect on the size of the residuals. If we choose once again to exercise Occam's Razor, and regard the simplest solution as best, then setting the nullspace coefficients to zero,

$$\tilde{\mathbf{x}} = \sum_{i=1}^{K} \frac{\mathbf{u}_i^T \mathbf{y}}{\lambda_i} \mathbf{v}_i \,. \tag{3.4.74}$$

Along with (3.4.73), this is the *particular-SVD solution* (a terminology explained in the next subsection). It simultaneously minimizes the residuals and the solution norm. With $< \mathbf{n} > = 0$, the bias of (3.4.74) is

$$< \tilde{\mathbf{x}} - \mathbf{x} > = \; - \sum_{i=K+1}^{N} \alpha_i \mathbf{v}_i \,. \tag{3.4.75}$$

The solution uncertainty is

$$\mathbf{P} = \sum_{i=1}^{K} \sum_{j=1}^{K} \mathbf{v}_i \frac{\mathbf{u}_i^T < \mathbf{n}\mathbf{n}^T > \mathbf{u}_j}{\lambda_i \lambda_j} \mathbf{v}_j^T + \sum_{i=K+1}^{N} \sum_{j=K+1}^{N} \mathbf{v}_i < \alpha_i \alpha_j > \mathbf{v}_j^T \,. \tag{3.4.76}$$

If the noise is white with variance σ_n^2, or if a row-scaling matrix $\mathbf{W}^{-T/2}$ has been applied to make it so, then (3.4.76) becomes

$$\mathbf{P} = \sum_{i=1}^{K} \sigma_n^2 \frac{\mathbf{v}_i \mathbf{v}_i^T}{\lambda_i^2} + \sum_{i=K+1}^{N} < \alpha_i^2 > \mathbf{v}_i \mathbf{v}_i^T \tag{3.4.77}$$

where it was also assumed that $< \alpha_i \alpha_j > \; = \; < \alpha_i^2 > \delta_{ij}$ in the nullspace. The influence of very small singular values on the uncertainty is clear: In the solution (3.4.71) or (3.4.74), there are error terms $\mathbf{u}_i^T \mathbf{n} / \lambda_i$ that are greatly magnified by small or nearly vanishing singular values, introducing large terms proportional to σ_n^2 / λ_i^2 into (3.4.77).

The decision to set to zero the projection of the noise onto the range of $\mathbf{E}^T$ as we did in Equations (3.4.68), (3.4.73) needs to be examined. Should we make some other choice, the solution norm would decrease, but the residual norm would increase. Determining the desirability of such a tradeoff requires understanding of the noise structure–in particular, (3.4.68) imposes rigid structures onto the residuals.

3.4.2.1 The Singular Value Decomposition

The singular vectors and values have been used to provide a convenient pair of orthonormal spanning sets to solve an arbitrary set of simultaneous equations. The vectors and values have another use, however, in providing a decomposition of $\mathbf{E}$.

Define $\boldsymbol{\Lambda}$ as the $M \times N$ matrix whose diagonal elements are the λ_i, in order of descending values in the same order, $\mathbf{U}$ as the $M \times M$ matrix whose columns are the $\mathbf{u}_i$, $\mathbf{V}$ as the $N \times N$ matrix whose columns are the $\mathbf{v}_i$ and whose other elements are 0. As an example, suppose $M = 3$, $N = 4$; then

$$\boldsymbol{\Lambda} = \begin{Bmatrix} \lambda_i & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & \lambda_3 & 0 \end{Bmatrix}.$$

Alternatively, if $M = 4$, $N = 3$

$$\begin{Bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \\ 0 & 0 & 0 \end{Bmatrix},$$

therefore extending the definition of a diagonal matrix to nonsquare ones.

Precisely as with matrix $\mathbf{G}$ considered above, column orthonormality of $\mathbf{U}$, $\mathbf{V}$ implies that these matrices are orthogonal,

$$\mathbf{U}\mathbf{U}^T = \mathbf{I}_M\,, \tag{3.4.78}$$

$$\mathbf{U}^T\mathbf{U} = \mathbf{I}_M\,, \tag{3.4.79}$$

$$\mathbf{V}\mathbf{V}^T = \mathbf{I}_N\,, \tag{3.4.80}$$

$$\mathbf{V}^T\mathbf{V} = \mathbf{I}_N\,. \tag{3.4.81}$$

(It follows that $\mathbf{U}^{-1} = \mathbf{U}^T$, etc.) As with $\mathbf{G}$ in Section 3.4.1, should one or more columns of $\mathbf{U}$, $\mathbf{V}$ be deleted, the matrices will become semi-orthogonal.

The relations (3.4.56), (3.4.57) to (3.4.58), (3.4.59) can be written compactly as:

$$\mathbf{E}\mathbf{V} = \mathbf{U}\boldsymbol{\Lambda}\,, \tag{3.4.82}$$

$$\mathbf{E}^T\mathbf{U} = \mathbf{V}\boldsymbol{\Lambda}^T\,, \tag{3.4.83}$$

$$\mathbf{E}^T\mathbf{E}\mathbf{V} = \mathbf{V}\boldsymbol{\Lambda}^T\boldsymbol{\Lambda}\,, \tag{3.4.84}$$

$$\mathbf{E}\mathbf{E}^T\mathbf{U} = \mathbf{U}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^T\,. \tag{3.4.85}$$

If we left multiply (3.4.82) by $\mathbf{U}^T$ and invoke (3.4.79), then

$$\mathbf{U}^T\mathbf{E}\mathbf{V} = \mathbf{U}^T\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}\mathbf{V}^T = \boldsymbol{\Lambda}\,. \tag{3.4.86}$$

So $\mathbf{U}$, $\mathbf{V}$ diagonalize $\mathbf{E}$ (with *diagonal* having the extended meaning for a rectangular matrix as defined above.)

Right multiplying (3.4.82) by $\mathbf{V}^T$ produces

$$\mathbf{E} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^T\,. \tag{3.4.87}$$

This last equation represents a decomposition, called the *singular value* *decomposition* (SVD) of an arbitrary matrix, into two orthogonal matrices, $\mathbf{U}$, $\mathbf{V}$, and a usually nonsquare diagonal matrix $\boldsymbol{\Lambda}$.

There is one further step to take. Notice that for a rectangular $\boldsymbol{\Lambda}$, as in the examples above, one or more rows or columns must be all zero, depending upon the shape of the matrix. If any of the $\lambda_i = 0$, $i < \min(M, N)$, the corresponding rows or columns also will be all zeros. Let K be the number of nonvanishing singular values (the rank of $\mathbf{E}$). By inspection (multiplying it out), one finds that the last N-K columns of $\mathbf{V}$ and the last M-K columns of $\mathbf{U}$ are multiplied by zeros only. If these columns are dropped entirely from $\mathbf{U}$, $\mathbf{V}$ so that $\mathbf{U}$ becomes $M \times K$ and $\mathbf{V}$ becomes $N \times K$, and reducing $\boldsymbol{\Lambda}$ to a $K \times K$ square matrix, then the representation (3.4.87) remains exact, in the form

$$\mathbf{E} = \mathbf{U}_K \boldsymbol{\Lambda}_K \mathbf{V}_K^T\,, \tag{3.4.88}$$

the subscript indicating the number of columns, where $\mathbf{U}_K$, $\mathbf{V}_K$ are then only semi-orthogonal, and $\boldsymbol{\Lambda}_K$ is now square. Equation (3.4.88) should be compared to (3.4.38).

The singular value decomposition for arbitrary nonsquare matrices is apparently due to Carl Eckart (Eckart & Young, 1939; see the historical discussions in Haykin, 1986; Klema & Laub, 1980; or Stewart, 1993).[7] Derivations are given by Lanczos (1961), Noble and Daniel (1977), Strang (1986), and many other recent books on applied linear algebra. The crucial role it plays in inverse methods appears to have been first noticed by Wiggins (1972).

The SVD solution can be obtained by direct matrix manipulation rather than vector by vector. Consider once again finding the solution to the simultaneous equations (3.3.2), but first write $\mathbf{E}$ in its reduced SVD,

$$\mathbf{U}_K \boldsymbol{\Lambda}_K \mathbf{V}_K^T \mathbf{x} + \mathbf{n} = \mathbf{y}\,. \tag{3.4.89}$$

Left multiplying by $\mathbf{U}_K^T$ and invoking the semi-orthogonality of $\mathbf{U}_K$ produces

$$\boldsymbol{\Lambda}_K \mathbf{V}_K^T \mathbf{x} + \mathbf{U}_K^T \mathbf{n} = \mathbf{U}_K^T \mathbf{y}\,. \tag{3.4.90}$$

The inverse of $\boldsymbol{\Lambda}_K$ (square with all nonzero diagonal elements) is easily computed, and

$$\mathbf{V}_K^T \mathbf{x} + \boldsymbol{\Lambda}_K^{-1} \mathbf{U}_K^T \mathbf{n} = \boldsymbol{\Lambda}_K^{-1} \mathbf{U}_K^T \mathbf{y}\,. \tag{3.4.91}$$

But $\mathbf{V}_K^T \mathbf{x}$ is the dot product of the first K of the $\mathbf{v}_i$ with the unknown $\mathbf{x}$. Equation (3.4.91) thus represents statements about the relationship between

[7] Eckart, a physicist turned oceanographer, had a somewhat controversial career. The SVD may turn out to have been his most important, if least credited, contribution.

dot products of the unknown vector, $\mathbf{x}$, with a set of orthonormal vectors, and therefore must represent the expansion coefficients of the solution in those vectors. If we set

$$\mathbf{U}_K^T\mathbf{n} = 0\,, \tag{3.4.92}$$

then

$$\mathbf{V}_K^T\mathbf{x} = \mathit{\Lambda}_K^{-1}\mathbf{U}_K^T\mathbf{y}\,, \tag{3.4.93}$$

and hence

$$\tilde{\mathbf{x}} = \mathbf{V}_K\mathit{\Lambda}_K^{-1}\mathbf{U}_K^T\mathbf{y}\,, \tag{3.4.94}$$

identical to the solution (3.4.74), which the reader is urged to confirm by writing it out explicitly. Substituting this solution into (3.4.89),

$$\mathbf{U}_K\mathit{\Lambda}_K\mathbf{V}_K^T\mathbf{V}_K\mathit{\Lambda}_K^{-1}\mathbf{U}_K^T\mathbf{y} + \mathbf{n} = \mathbf{U}_K\mathbf{U}_K^T\mathbf{y} + \mathbf{n} = \mathbf{y}$$

or

$$\tilde{\mathbf{n}} = (\mathbf{I} - \mathbf{U}_K\mathbf{U}_K^T)\mathbf{y}\,. \tag{3.4.95}$$

Let the full $\mathbf{U}$ and $\mathbf{V}$ matrices be rewritten as

$$\mathbf{U} = \{\mathbf{U}_K \quad \mathbf{Q}_u\}\,, \tag{3.4.96}$$

$$\mathbf{V} = \{\mathbf{V}_K \quad \mathbf{Q}_v\} \tag{3.4.97}$$

where $\mathbf{Q}_u$, $\mathbf{Q}_v$ contain the nullspace vectors. Then

$$\mathbf{E}\tilde{\mathbf{x}} + \tilde{\mathbf{n}} = \mathbf{y}\,, \quad \mathbf{E}\tilde{\mathbf{x}} = \tilde{\mathbf{y}}\,,$$

$$\tilde{\mathbf{y}} = \mathbf{U}_K\mathbf{U}_K^T\mathbf{y}\,, \quad \tilde{\mathbf{n}} = \mathbf{Q}_u\mathbf{Q}_u^T\mathbf{y} = \sum_{j=K+1}^{M}(\mathbf{u}_i^T\mathbf{y})\mathbf{u}_i\,, \tag{3.4.98}$$

which is identical to (3.4.72). Note that $\mathbf{Q}_u\mathbf{Q}_u^T = (\mathbf{I} - \mathbf{U}_K\mathbf{U}_K^T)$, $\mathbf{Q}_v\mathbf{Q}_v^T = (\mathbf{I} - \mathbf{V}_K\mathbf{V}_K^T)$, which are idempotent. The general solution is

$$\tilde{\mathbf{x}} = \mathbf{V}_K\mathit{\Lambda}_K^{-1}\mathbf{U}_K^T\mathbf{y} + \mathbf{Q}_v\boldsymbol{\alpha}_v \tag{3.4.99}$$

where $\boldsymbol{\alpha}_v$ is now restricted to being the vector of coefficients of the nullspace vectors.

The solution uncertainty of (3.4.99) is

$$\begin{aligned}\mathbf{P} &= \mathbf{V}_K\mathit{\Lambda}_K^{-1}\mathbf{U}_K^T < \mathbf{n}\mathbf{n}^T > \mathbf{U}_K\mathit{\Lambda}_K^{-1}\mathbf{V}_K^T + \mathbf{Q}_v < \boldsymbol{\alpha}_v\boldsymbol{\alpha}_v^T > \mathbf{Q}_v^T \\ &= \mathbf{C}_{\tilde{x}\tilde{x}} + \mathbf{Q}_v < \boldsymbol{\alpha}_v\boldsymbol{\alpha}_v^T > \mathbf{Q}_v^T\end{aligned} \tag{3.4.100}$$

or

$$\mathbf{P} = \sigma_n^2\mathbf{V}_K\mathit{\Lambda}_K^{-2}\mathbf{V}_K^T + \mathbf{Q}_v\,\mathbf{R}_{\alpha\alpha}\mathbf{Q}_v^T \tag{3.4.101}$$

for white noise. The uncertainty of the residuals for white noise is

$$\mathbf{P}_{nn} = \sigma_n^2(\mathbf{I} - \mathbf{U}_K\mathbf{U}_K^T)\,. \tag{3.4.102}$$

Solution of simultaneous equations by SVD has several important advantages. Among other features, we can write down within one algebraic formulation the solution to systems of equations that can be under-, over-, or just-determined.[8] Unlike the eigenvalue/eigenvector solution for the square system, the singular values (eigenvalues) are always nonnegative and real, and the singular vectors (eigenvectors) can always be made a complete orthonormal set. Neither of these statements is true for the conventional eigenvector problem. Most important, however, the relations (3.4.56), (3.4.57) are a specific, quantitative statement of the connection between a set of orthonormal structures in the data and the corresponding presence of orthonormal structures in the solution. These relations provide a very powerful diagnostic method for understanding precisely why the solution takes on the form it does.

3.4.3 Some Simple Examples

The simplest underdetermined system is 1 × 2. Suppose $x_1 - 2x_2 = 3$ so that

$$\mathbf{E} = \{1 \quad -2\},\ \mathbf{U} = \{1\},\ \mathbf{V} = \left\{\begin{matrix} .447 & -.894 \\ -.894 & -.447 \end{matrix}\right\},\ \lambda_1 = 2.24$$

where the second column of $\mathbf{V}$ is in the nullspace of $\mathbf{E}$. The general solution is $\tilde{\mathbf{x}} = [.6 \quad -1.2]^T + \alpha_2\mathbf{v}_2$. Because $K = 1$ is the only possible choice here, it is readily confirmed that this solution satisfies the equation exactly, and a data nullspace is not possible.

The most elementary overdetermined problem is 2 × 1. Suppose

$$x_1 = 1\,,$$
$$x_1 = 3\,.$$

The appearance of two such equations is possible if there is noise in the observations, and they are written more properly as

$$x_1 + n_1 = 1\,,$$
$$x_1 + n_2 = 3\,.$$

[8] True, too, of the generalized least-squares formulation (3.3.66)–(3.3.68) or (3.3.38)–(3.3.40).

$\mathbf{E} = \{1 \quad 1\}^T$, $\mathbf{E}^T\mathbf{E}$ represents the eigenvalue problem of the smaller dimension, and

$$\mathbf{U} = \begin{Bmatrix} .707 & -.707 \\ .707 & .707 \end{Bmatrix}, \; \mathbf{V} = \{1\}, \; \lambda_1 = \sqrt{2}$$

where the second column of $\mathbf{U}$ lies in the data nullspace, there being no solution nullspace. The general solution is $\mathbf{x} = x_1 = 2$, which if substituted back into the original equations produces

$$\mathbf{E}\tilde{\mathbf{x}} = \begin{bmatrix} 2 \\ 2 \end{bmatrix} = \tilde{\mathbf{y}},$$

and hence there are residuals $\tilde{\mathbf{n}} = \tilde{\mathbf{y}} - \mathbf{y} = [1 \quad -1]^T$, which are necessarily proportional to $\mathbf{u}_2$. Evidently no other solution could produce a smaller l_2 norm residual than this one. The SVD produced a solution that compromised the contradiction between the two equations and is physically sensible.

The possibility of $K < M$, $K < N$ simultaneously is also easily seen. Consider the system:

$$\begin{Bmatrix} 1 & -2 & 1 \\ 3 & 2 & 1 \\ 4 & 0 & 2 \end{Bmatrix} \mathbf{x} = \begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix},$$

which appears superficially just-determined. But the singular values are $\lambda_1 = 5.67$, $\lambda_2 = 2.80$, $\lambda_3 = 0$. The vanishing of the third singular values means that the row and column vectors are not linearly independent sets (not spanning sets)–indeed, the third row vector is just the sum of the first two (but the third element of $\mathbf{y}$ is not the sum of the first two, making the equations inconsistent). Thus, there are both solution and data nullspaces, which the reader might wish to find. With a vanishing singular value, $\mathbf{E}$ can be written exactly using only two columns of $\mathbf{U}$, $\mathbf{V}$ and the linear dependence is given explicitly as $\mathbf{u}_3^T\mathbf{E} = 0$.

Consider now the underdetermined system

$$\begin{aligned} x_1 + x_2 - 2x_3 &= 1, \\ x_1 + x_2 - 2x_3 &= 2, \end{aligned}$$

which has no conventional solution at all, being a contradiction, and is thus simultaneously underdetermined and incompatible. If one of the coefficients is modified by a very small quantity, ϵ, to produce

$$\begin{aligned} x_1 + x_2 - (2 + \epsilon)x_3 &= 1, \\ x_1 + x_2 - 2x_3 &= 2, \end{aligned} \qquad (3.4.103)$$

not only is there a solution, there are an infinite number of them, which the reader should confirm by computing the basic SVD solution and the nullspace. Thus, the slightest perturbation in the coefficients has made the system jump from having no solution to having an infinite number, an obviously disconcerting situation. Such a system is *ill-conditioned.* How would we know the system is ill-conditioned? There are several indicators. First, the ratio of the two singular values is determined by ϵ. In (3.4.103), if we take $\epsilon = 10^{-10}$, the two singular values are $\lambda_1 = 3.46$, $\lambda_2 = 4.1 \times 10^{-11}$, an immediate warning that the two equations are nearly linearly dependent. (In a mathematical problem, the nonvanishing of the second singular value is enough to assure a solution. As will be discussed later, the inevitable slight errors in $\mathbf{y}$ suggest small singular values are best regarded as actually being zero.)

A similar problem exists with the system:

$$\begin{aligned} x_1 + x_2 - 2x_3 &= 1\,, \\ x_1 + x_2 - 2x_3 &= 1\,, \end{aligned}$$

which has an infinite number of solutions. But the change to

$$\begin{aligned} x_1 + x_2 - 2x_3 &= 1\,, \\ x_1 + x_2 - 2x_3 &= 1 + \epsilon \end{aligned}$$

for arbitrarily small ϵ produces a system with no solutions in the conventional mathematical sense, although the SVD will handle the system without any difficulty.

Problems like these are simple examples of the practical issues that arise once one recognizes that unlike mathematical textbook problems, observational ones always contain inaccuracies; any discussion of how to handle data in the presence of mathematical relations must account for these inaccuracies as intrinsic, not as something to be regarded as an afterthought. But the SVD itself is sufficiently powerful that it always contains the information to warn of ill-conditioning, and by determination of K to cope with it, producing useful solutions.

3.4.3.1 The Neumann Problem

Consider the classical Neumann problem described in Chapter 1. The problem is to be solved on a 10×10 grid as stated in Equation (1.2.7), $\mathbf{A}_3\phi = \mathbf{d}_3$. The singular values of $\mathbf{A}_3$ are plotted in Figure 3–10a; the largest one is $\lambda_1 = 7.8$, and the smallest nonzero one is $\lambda_{99} = 0.08$. As expected, $\lambda_{100} = 0$. The singular vector $\mathbf{v}_{100}$ corresponding to the zero singular value

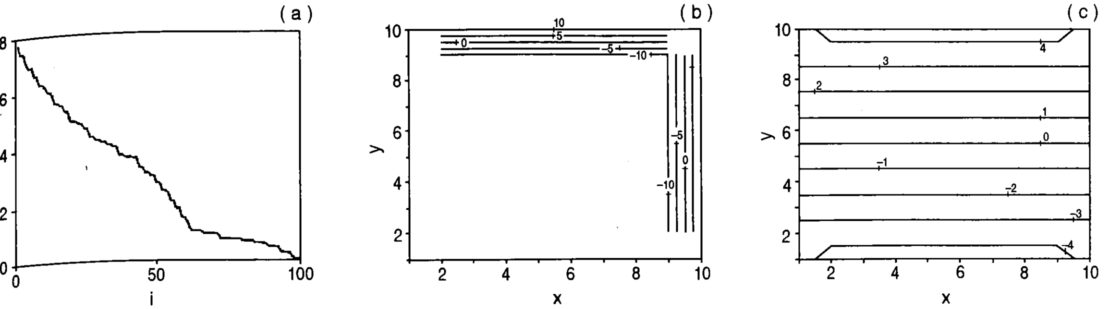

Figure 3–10. (a) Singular values of the coefficient matrix of the numerical Neumann problem. All λ_i are nonzero except the last one. (b) The 100th nullspace vector $\mathbf{u}_{100}$ of $\mathbf{A}_3$ such that $\mathbf{u}_{100}^T\mathbf{A}_3 = \mathbf{u}_{100}^T\mathbf{y} = 0$ defines the consistency or solvability conditions for the Neumann problem–that there be no net influx of source material from interior sources or across the boundaries. The form shown can be understood from the physical requirement that what enters across the left and bottom boundaries plus interior sources must balance what leaves across the top and right boundaries. The corresponding $\mathbf{v}_{100}$ is a constant. (c) Particular-SVD solution to the Neumann problem, with equal and opposite fluxes across the two horizontal walls, and vanishing flux across both vertical walls and in the interior. This specification satisfies the solvability conditions and leaves no residuals.

is not shown, because also as expected, it is a constant; $\mathbf{u}_{100}$, shown in Figure 3–10b, is not a constant and has considerable structure, which provides the solvability condition for the Neumann problem, $\mathbf{u}_{100}^T\mathbf{y} = 0$. The physical origin of the solvability condition is readily understood: The Neumann boundary conditions prescribe boundary flux rates, and the sum of the interior source strengths plus the boundary flux rates must equal zero; otherwise, no steady state is possible. If the boundary conditions are homogeneous, then no flow takes place through the boundary, and the interior sources must sum to zero. In particular, the value of $\mathbf{u}_{100}$ on the interior grid points is constant. The Neumann problem is thus a forward problem requiring one to deal with both a solution nullspace and a "data" solvability condition.

As an example of solution by the SVD, let there be unit positive flux into the box on the bottom boundary, unit positive flux out on the top and no interior sources. The resulting particular SVD solution is shown in Figure 3–10c. No residuals are left because the system was constructed as fully consistent, and there is an arbitrary constant that can be added ($\mathbf{v}_{100}$). The reader may wish to experiment with incompatible specifications for this problem. This is an example of a forward problem solved using an inverse method.

Related inverse problems are also easily formulated. The simplest of all would assert that ϕ is known and $\mathbf{d}_3$ is to be determined. In the present situation, one confirms that multiplication of the solution in Figure 3–10c by $\mathbf{A}_3$ produces $\mathbf{d}_3$. One can do a number of interesting experiments with the SVD. For example, if the equations imposing the boundary values are dropped, the resulting range vectors, $\mathbf{v}_i$, describe the particular solution of the partial differential equation, and the nullspace vectors describe the homogeneous one. In this way, one can pick apart the structure of the solution. A more interesting possibility is to withhold knowledge of the boundary conditions and ask for their determination, given the interior solution.

3.4.3.2 Relation of Least Squares to the SVD

What is the relationship of the SVD solution to the least-squares solutions? To some extent, the answer is already obvious from the orthonormality of the two sets of singular vectors. We begin by first asking when the simple least-squares solution will exist? Consider first the formally overdetermined problem, $M > N$. The solution (3.3.6) exists if and only if the matrix inverse exists. Substituting the SVD for $\mathbf{E}$, one finds

$$(\mathbf{E}^T\mathbf{E})^{-1} = (\mathbf{V}_N\mathbf{\Lambda}_N^T\mathbf{U}_N^T\mathbf{U}_N\mathbf{\Lambda}_N\mathbf{V}_N^T)^{-1} = (\mathbf{V}_N\mathbf{\Lambda}_N^2\mathbf{V}_N^T)^{-1} \tag{3.4.104}$$

where the semi-orthogonality of $\mathbf{U}_N$ has been used. Suppose that $K = N$, its maximum possible value; then $\mathbf{\Lambda}_N^2$ is $N \times N$ with *all nonzero diagonal elements* λ_i^2. The inverse in (3.4.104) may be found by inspection, using $\mathbf{V}_N\mathbf{V}_N^T = \mathbf{I}_N$,

$$(\mathbf{E}^T\mathbf{E})^{-1} = \mathbf{V}_N\mathbf{\Lambda}_N^{-2}\mathbf{V}_N^T\,. \tag{3.4.105}$$

Then the solution (3.3.6) becomes

$$\tilde{\mathbf{x}} = (\mathbf{V}_N\mathbf{\Lambda}_N^{-2}\mathbf{V}_N^T)\mathbf{V}_N\mathbf{\Lambda}_N\mathbf{U}_N^T = \mathbf{V}_N\mathbf{\Lambda}_N^{-1}\mathbf{U}_N^T\mathbf{y}\,, \tag{3.4.106}$$

which is identical to the SVD solution (3.4.94). If $K < N$, $\mathbf{\Lambda}_N^2$ has at least one zero on the diagonal, no matrix inverse exists, and the conventional least-squares solution is not defined. The condition for its existence is thus $K = N$, the so-called *full rank overdetermined* case. The condition $K < N$ is called *rank deficient.* The dependence of the least-squares solution magnitude upon the possible presence of very small, but nonvanishing, singular values is obvious.

That the full-rank overdetermined case is unbiased, as previously asserted, can now be seen from

$$< \tilde{\mathbf{x}} - \mathbf{x} > = \sum_{i=1}^{N} \frac{(\mathbf{u}_i^T < \mathbf{y} >)}{\lambda_i}\mathbf{v}_i - \mathbf{x} = \sum_{i=1}^{N} \frac{\mathbf{u}_i^T\mathbf{y}_0}{\lambda_i}\mathbf{v}_i - \mathbf{x} = \mathbf{0}\,,$$

if $< \mathbf{n} > = \mathbf{0}$, assuming that the correct $\mathbf{E}$ is being used.

The identity of the SVD solution and the overdetermined full-rank solution (3.3.6) is also readily shown by directly substituting

$$\mathbf{x} = \sum_{i=1}^{N} \alpha_i \mathbf{v}_i \tag{3.4.107}$$

into the objective function (3.3.2), using the relation (3.4.56) and the orthogonality of $\mathbf{u}_i$. One finds the minimum at $\alpha_i = \mathbf{u}_i^T \mathbf{y}/\lambda_i$, $\lambda_i \neq 0$. If any singular value vanishes, the vector orthogonality proves that no other choice of α_i, $i \leq K$, can reduce J further, and so the particular-SVD solution produces the best possible minimum even when the system is rank deficient.

Now consider another least-squares problem, the one with the conventional purely underdetermined least-squares solution (3.3.53). When does that exist? Substituting the SVD into (3.3.53),

$$\begin{aligned}\tilde{\mathbf{x}} &= \mathbf{V}_M \boldsymbol{\Lambda}_M \mathbf{U}_M^T (\mathbf{U}_M \boldsymbol{\Lambda}_M \mathbf{V}_M^T \mathbf{V}_M \boldsymbol{\Lambda}_M^T \mathbf{U}_M^T)^{-1} \mathbf{y} \\ &= \mathbf{V}_M \boldsymbol{\Lambda}_M \mathbf{U}_M^T (\mathbf{U}_M \boldsymbol{\Lambda}_M^2 \mathbf{U}_M^T)^{-1} \mathbf{y}\,.\end{aligned} \tag{3.4.108}$$

Again, the matrix inverse exists if and only if $\boldsymbol{\Lambda}_M^2$ has all nonzero diagonal elements, which occurs only when $K = M$. Under that specific condition, the inverse is obtained by inspection, and

$$\tilde{\mathbf{x}} = \mathbf{V}_M \boldsymbol{\Lambda}_M \mathbf{U}_M^T (\mathbf{U}_M \boldsymbol{\Lambda}_M^{-2} \mathbf{U}_M^T) \mathbf{y} = \mathbf{V}_M \boldsymbol{\Lambda}_M^{-1} \mathbf{U}_M^T \mathbf{y} \tag{3.4.109}$$

$$\tilde{\mathbf{n}} = 0\,, \tag{3.4.110}$$

which is once again the particular-SVD solution (3.4.94)–with $K = M$ and the nullspace coefficients set to zero. This situation is usually referred to as the *full-rank underdetermined case.* Again, the possible influence of small singular values is apparent, and an arbitrary sum of nullspace vectors can be added to (3.4.109).

The bias of (3.4.108) is given by the nullspace elements, and its formal uncertainty is from the nullspace contribution. With $\tilde{\mathbf{n}} = \mathbf{0}$, the formal sample noise variance vanishes, and the particular-SVD solution covariance $\mathbf{C}_{\tilde{x}\tilde{x}}$ would be zero, if the sample variance is used. If the formally overdetermined problem is converted to an exact underdetermined one as in Equation (3.3.42), then the uncertainty is calculated solely from the nullspace contribution.

The particular-SVD solution thus coincides with the two simplest forms of least-squares solution and generalizes both of them to the case where the matrix inverses do not exist. *All of the structure imposed by the SVD,* *in particular the restriction on the residuals in* (3.4.68), *is present in the least-squares solutions.* If the system is not of full rank, then the simple least-squares solutions do not exist. The SVD generalizes these results by determining the elements of the solution lying in the range of $\mathbf{E}$ and giving an explicit structure for the resulting nullspace vectors.

The SVD has much flexibility. For example, it permits one to modify the simplest underdetermined solution to remove its greatest shortcoming, the necessity that $\tilde{\mathbf{n}} = \mathbf{0}$. One simply truncates the solution (3.4.74) at $K' < K < M$, thus assigning all vectors $\mathbf{v}_i$, $K' + 1 \leq i \leq K$, to an *effective nullspace* (or substitutes K' for K everywhere). The resulting residual is then

$$\tilde{\mathbf{n}} = \sum_{i=K'+1}^{K} (\mathbf{u}_i^T \mathbf{y})\mathbf{u}_i \,, \tag{3.4.111}$$

with an uncertainty for $\tilde{\mathbf{x}}$, $\tilde{\mathbf{n}}$ given by (3.4.100)–(3.4.102), but with the upper limit being K' rather than K. Such truncation has the effect of reducing the solution covariance contribution to the uncertainty [recall (3.4.77)] but increasing the contribution owing to the nullspace (and increasing the potential bias). In the presence of singular values that are small compared to σ_n, the resulting overall reduction in uncertainty may be very great.

The solution now consists of three parts,

$$\tilde{\mathbf{x}} = \sum_{i=1}^{K'} \frac{\mathbf{u}_i^T \mathbf{y}}{\lambda_i} \mathbf{v}_i + \sum_{i=K'+1}^{K} \alpha_i \mathbf{v}_i + \sum_{i=K+1}^{N} \alpha_i \mathbf{v}_i \,, \tag{3.4.112}$$

where the middle sum contains the terms appearing with singular values too small to be employed for the given noise, and the third sum is the strict nullspace. Usually, one lumps the two nullspace sums together. The first sum, by itself, represents the particular-SVD solution in the presence of noise.

This consideration is extremely important: It says that despite the mathematical condition $\lambda_i \neq 0$, some structures in the solution cannot be estimated with sufficient reliability to be useful. The *effective rank* is then not the same as the mathematical rank.

Evidently, truncation of the SVD offers a simple method for controlling the ratio of solution and residual norms: As the nullspace grows by reducing K', it follows that the solution norm necessarily is reduced and that the residuals must grow, along with the size of the solution nullspace. The issue of how to choose K'–that is, *rank determination* in practice is an interesting one to which we will return.

The full-rank overdetermined least-squares solution leaves no solution nullspace but does produce a data nullspace (unless the special solvability conditions are met). In this case, we have the identity,

$$(\mathbf{I} - \mathbf{E}(\mathbf{E}^T\mathbf{E})^{-1}\mathbf{E}^T) = (\mathbf{I} - \mathbf{U}_M\mathbf{U}_M^T) = \mathbf{Q}_u\mathbf{Q}_u^T\,, \tag{3.4.113}$$

the idempotent projector of the data onto the nullspace of $\mathbf{E}^T$ (the matrix inverse is guaranteed to exist by the full-rank assumption). In the full-rank underdetermined case, there is no data nullspace, but there is a solution nullspace. In that situation, the relevant identity is

$$(\mathbf{I} - \mathbf{E}^T(\mathbf{E}\mathbf{E}^T)^{-1}\mathbf{E}) = (\mathbf{I} - \mathbf{V}_N\mathbf{V}_N^T) = \mathbf{Q}_v\mathbf{Q}_v^T\,, \tag{3.4.114}$$

the idempotent projector of $\mathbf{x}$ onto the solution nullspace. These identities follow immediately from introduction of the SVD and the definitions (3.4.96)–(3.4.97) and should be compared to the analogous result (3.4.46) for a square symmetric $\mathbf{E}$. The identities remain valid with both N, M replaced by the actual rank, K, for any $K \leq \min(N, M)$. Both identities prove useful in Chapter 6 for interpreting the Kalman filter and associated smoothers.

3.4.3.3 Row and Column Scaling

The effects on the least-squares solutions of the row and column scaling can now be understood. Suppose we have two equations

$$\begin{Bmatrix} 1 & 1 & 1 \\ 1 & 1.01 & 1 \end{Bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} n_1 \\ n_2 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix},$$

and there is no information about the noise covariance and no row scaling is reasonable, so $\mathbf{W} = \mathbf{I}$. The SVD of $\mathbf{E}$ is

$$\mathbf{U} = \begin{Bmatrix} 0.7059 & -0.7083 \\ 0.7083 & 0.7059 \end{Bmatrix}, \quad \mathbf{V} = \begin{Bmatrix} 0.5764 & -0.4096 & 0.7071 \\ 0.5793 & 0.8151 & 0.0000 \\ 0.5764 & -0.4096 & -0.7071 \end{Bmatrix},$$

$$\lambda_1 = 2.4536,\ \lambda_2 = .0058\,.$$

The SVD solutions, choosing ranks $K' = 1, 2$ in succession, are very nearly

$$\tilde{\mathbf{x}} \sim \frac{0.71(y_1 + y_2)}{2.45} \begin{bmatrix} .58 \\ .58 \\ .58 \end{bmatrix},$$

$$\sim \frac{0.71(y_1+y_2)}{2.45}\begin{bmatrix}.58\\.58\\.58\end{bmatrix}+\frac{0.71(y_1-y_2)}{.0058}\begin{bmatrix}-.41\\.82\\.41\end{bmatrix}, \qquad (3.4.115)$$

respectively, so that the first term simply averages the two measurements, y_i, and the difference between them contributes with great uncertainty in the second term of the rank 2 solution owing to the very small singular value.

Now suppose that the covariance matrix of the noise is known to be

$$\mathbf{R}_{nn}=\left\{\begin{matrix}1 & .999999\\ .999999 & 1\end{matrix}\right\}$$

(an extreme case, chosen for illustrative purposes). Then put $\mathbf{W}=\mathbf{R}_{nn}$,

$$\mathbf{W}^{1/2}=\left\{\begin{matrix}1.0000 & 1.0000\\ 0 & 0.0014\end{matrix}\right\},\ \mathbf{W}^{-T/2}=\left\{\begin{matrix}1.0000 & 0\\ -707.1063 & 707.1070\end{matrix}\right\}.$$

The new system to be solved is

$$\left\{\begin{matrix}1.0000 & 1.0000 & 1.0000\\ 0.0007 & 7.0718 & 0.0007\end{matrix}\right\}\begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix}=\begin{bmatrix}y_1\\707.1(-y_1+y_2)\end{bmatrix}$$

whose SVD is

$$\mathbf{U}=\left\{\begin{matrix}0.1456 & 0.9893\\ 0.9893 & -0.1456\end{matrix}\right\},\ \mathbf{V}=\left\{\begin{matrix}0.0205 & 0.7068 & 0.7071\\ 0.9996 & -0.0290 & 0.0000\\ 0.0205 & 0.7068 & -0.7071\end{matrix}\right\}$$

$$\lambda_1=7.1450,\ \lambda_2=1.3996.$$

The second singular value is now much larger relative to the first one, so that the two solutions are

$$\tilde{\mathbf{x}}\sim\frac{707(y_2-y_1)}{7.1}\begin{bmatrix}0\\1\\0\end{bmatrix},$$

$$\sim\frac{707(y_2-y_1)}{7.1}\begin{bmatrix}0\\1\\0\end{bmatrix}+\frac{y_1}{1.4}\begin{bmatrix}.71\\0\\.71\end{bmatrix}, \qquad (3.4.116)$$

and the rank 1 solution is obtained from the difference of the observations, in contrast to the unscaled solution. The result is quite sensible; given the information that the noise in the two equations is nearly perfectly correlated, it can be removed by subtraction.

At full rank, that is, $K=2$, it can be confirmed that the solutions

(3.4.115) and (3.4.116) are identical, as they must be. It was previously asserted that in a full-rank formally underdetermined system, row scaling is irrelevant to $\tilde{\mathbf{x}}$, $\tilde{\mathbf{n}}$, as may be seen as follows,

$$\begin{aligned}\tilde{\mathbf{x}} &= \mathbf{E}'^T\left(\mathbf{E}'\mathbf{E}'^T\right)^{-1}\mathbf{y}' = \mathbf{E}^T\mathbf{W}^{-1/2}(\mathbf{W}^{-T/2}\mathbf{E}\mathbf{E}^T\mathbf{W}^{-1/2})^{-1}\mathbf{W}^{-T/2}\mathbf{y}\\ &= \mathbf{E}\mathbf{W}^{-1/2}\mathbf{W}^{1/2}(\mathbf{E}\mathbf{E}^T)^{-1}\mathbf{W}^{T/2}\mathbf{W}^{-T/2}\mathbf{y} = \mathbf{E}^T\left(\mathbf{E}\mathbf{E}^T\right)^{-1}\mathbf{y} \quad (3.4.117)\end{aligned}$$

where we used the result $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$ and both inverses must exist, which is possible only in the full-rank situation. But the error covariance is quite different in the two cases:

$$\begin{aligned}\left(\mathbf{E}\mathbf{E}^T\right)^{-1} &= \left\{\begin{matrix} 1.510\times 10^4 & -1.505\times 10^4 \\ -1.505\times 10^4 & 1.505\times 10^4\end{matrix}\right\}\\ \left(\mathbf{E}'\mathbf{E}'^T\right)^{-1} &= \left\{\begin{matrix} 0.500 & -0.707 \\ -0.707 & 0.300\end{matrix}\right\}, \end{aligned} \qquad (3.4.118)$$

which would give rise to very different error estimates (using a prior estimate σ_n^2 of the noise variance, because here the noise residuals vanish, a degenerate limit). In effect, the information provided in the row scaling with $\mathbf{R}_{nn}$ permits the SVD to nearly eliminate the noise at rank 1 by an effective subtraction, whereas without that information, the noise is reduced in the solution (3.4.115) at rank 1 only by direct averaging.

There is a subtlety in row weighting. Suppose we have two equations of form

$$\begin{aligned}10x_1 + 5x_2 + x_3 &= 1\,,\\ 100x_1 + 50x_2 + 10x_3 &= 2\,,\end{aligned} \qquad (3.4.119)$$

after row scaling to make the expected noise variance in each the same. A rank 1 solution to these equations by SVD is $\tilde{\mathbf{x}} = [.0165 \quad .0083 \quad .0017]^T$, which produces residuals $\tilde{\mathbf{y}} - \mathbf{y} = [-0.79 \quad 0.079]^T$–much smaller in the second equation than in the first one.

Consider that the left side of the second equation is 10 times the first one; in effect we are saying that a measurement of 10 times the values of $10x_1 + 5x_2 + x_3$ has the same noise in it as a measurement of one times this same linear combination. The second equation clearly represents a much more accurate determination of this linear combination, and the equation should be given much more weight in determining the unknowns–and the SVD (and ordinary least squares) does precisely that. To the extent that one finds this result undesirable (one should be careful about why it is so found), there is an easy remedy–divide the equations by their row norms $\sum_j (E_{ij})^{1/2}$. But there may then be a contradiction if it was believed that the noise in all equations was the same to begin with.

An example of this situation is readily apparent in the box balances discussed in Chapter 2. Equations such as (2.4.2) for salt balance have row norms about 35 times larger than those (2.4.1) for the corresponding mass balance, simply because salinity is measured by convention on the Practical Salinity Scale, which produces ocean salinities near 35. Because there is nothing fundamental about the choice of salinity scale, it seems unreasonable to infer that the requirement of salt balance has an expected error 35 times smaller than for density. One usually proceeds in the obvious way by dividing the salt equations by their row norms as the first step. The second step is to ask whether anything further can be said about the relative errors of mass and salt balance, which would introduce a second, purely statistical row weight.

Consider two independent equations in two unknowns, for example,

$$\begin{aligned} x_1 + x_2 &= 1\,, \\ 2x_1 + x_2 &= 2 \end{aligned}$$

with unique solution $x_1 = 1$, $x_2 = 0$. Now suppose that the right-hand side of the second equation is totally unknown. We have several possibilities for handling the situation. (1) Drop the second equation, and solve the first one as an underdetermined system, giving as the minimum norm solution $\tilde{x}_1 = 1/2$, $\tilde{x}_2 = 1/2$. (2) Downweight the second equation, multiplying it by some very small number. The minimum norm solution is the same as in (1). The advantage over (1) is that most software will compute the right-hand side of the original equations after the solution has been estimated, and the original set-up is unaltered. We thus find out that an estimate of the right-hand side from this solution is 1.5. The disadvantage is that we work with a 2 × 2 system rather than the 1 × 2 of (1). (3) Regard the right-hand side of the second equation as a new formal unknown, and rewrite the system as

$$\begin{aligned} x_1 + x_2 &= 1\,, \\ 2x_1 + x_2 - q &= 0\,. \end{aligned} \tag{3.4.120}$$

Solving for the minimum norm underdetermined solution now, we obtain $\tilde{x}_1 = 0$, $\tilde{x}_2 = 1$, $\tilde{q} = 1$, and the estimate of the right-hand side of equation two is 1. Why is this answer different from that in (1) and (2)? The reason is that the presence of the third unknown in the second equation in (3.4.120) provides the information that the unknown right-hand side of equation (2) is of the same magnitude as that of the unknowns x_1, x_2–information that is removed by downweighting or eliminating the equation altogether. The investigator must make his own choice of solution, dependent upon particular circumstances. But see the next section.

3.4.3.4 Column Scaling

In the least-squares problem, we formally introduced a column scaling matrix **S**. Column scaling operates on the SVD solution exactly as it does in the least-squares solution, to which it reduces in the two special cases already described. That is, we should apply the SVD to sets of equations only where any knowledge of the solution element size has been removed first. If the SVD has been computed for such a column-scaled (and row-scaled) system, the solution is for the scaled unknown $\mathbf{x}'$, and the physical solution is

$$\tilde{\mathbf{x}} = \mathbf{S}^{T/2}\tilde{\mathbf{x}}' . \tag{3.4.121}$$

But there are occasions, with underdetermined systems, where a nonstatistical scaling may also be called for–the analogue to the situation considered above where a row scaling was introduced on the basis of possible nonstatistical considerations.

Example: Suppose we have one equation in two unknowns, the smallest example of an underdetermined system:

$$10x_1 + 1x_2 = 3 . \tag{3.4.122}$$

The particular-SVD solution produces $\tilde{\mathbf{x}} = [0.2970 \quad 0.0297]^T$ in which the magnitude of x_1 is much larger than that of x_2, and the result is readily understood: As we have seen, the SVD finds the exact solution, subject to making the solution norm as small as possible. Because the coefficient of x_1 in (3.4.122) is 10 times that of x_2, it is obviously more efficient in minimizing the norm to give x_1 a larger value than x_2.

Although we have demonstrated this dependence for a trivial example, similar behavior occurs for underdetermined systems in general. In many cases, this distribution of the elements of the solution vector $\mathbf{x}$ is desirable, the numerical value 10 appearing for good physical reasons. In other problems–and the geostrophic inversion problem is an example–the numerical values appearing in the coefficient matrix **E** are an accident (in the geostrophic problem, they are proportional to the distance steamed between hydrographic stations and the water depth). Unless one believed that velocities should be larger where the ship steamed further, or the water was deeper, then the solutions may behave unphysically. Indeed, in some situations the velocities are expected to be inverse to the water depth, and such a prior statistical hypothesis is best imposed after one has removed the structural accidents from the system. (The tendency for the solutions to be proportional to the column norms is not absolute. In particular, the equations themselves may actually preclude the proportionality.)

Take a positive-definite, diagonal matrix $\mathbf{S}$, and rewrite (3.3.2) as

$$\mathbf{E}\mathbf{S}^{T/2}\mathbf{S}^{-T/2}\mathbf{x}+\mathbf{n}=\mathbf{y}$$

Then,

$$\mathbf{E}'\mathbf{x}'+\mathbf{n}=\mathbf{y}\,.$$

Solving

$$\tilde{\mathbf{x}}'=\mathbf{E}'^{T}(\mathbf{E}'\mathbf{E}'^{T})^{-1}\mathbf{y}\,,\ \tilde{\mathbf{x}}=\mathbf{S}^{-T/2}\tilde{\mathbf{x}}'\,. \tag{3.4.123}$$

How should $\mathbf{S}$ be chosen? Apply the recipe (3.4.123) for the simple one-equation example of (3.4.122):

$$\begin{aligned}
\mathbf{E}' &= \{10/S_{11}^{1/2}\quad 1/S_{22}^{1/2}\},\ \mathbf{E}'\mathbf{E}'^{T}=\frac{100}{S_{11}}+\frac{1}{S_{22}},\\
(\mathbf{E}'\mathbf{E}'^{T})^{-1} &= \left(\frac{S_{11}S_{22}}{100S_{22}+S_{11}}\right),\\
\tilde{\mathbf{x}}' &= \begin{Bmatrix}10/S_{11}^{1/2}\\ 1/S_{22}^{1/2}\end{Bmatrix}\left[\frac{S_{11}S_{22}}{100S_{22}+S_{11}}\right][3]\,,\\
\tilde{\mathbf{x}} = \mathbf{S}^{-T/2}\mathbf{x}' &= \begin{Bmatrix}10/S_{11}\\ 1/S_{22}\end{Bmatrix}\left[\frac{S_{11}S_{22}}{100S_{22}+S_{11}}\right][3]\,.
\end{aligned} \tag{3.4.124}$$

The relative magnitudes of the elements of $\tilde{\mathbf{x}}$ are proportional to $10/S_{11}$, $1/S_{22}$. To make the numerical values of the elements of $\tilde{\mathbf{x}}$ the same, we should clearly choose $S_{11}=10$, $S_{22}=1$; that is, we should divide the elements of the first column of $\mathbf{E}$ by $\sqrt{10}$ and the second column by $\sqrt{1}$. The apparent rule (which is correct and general) is to divide each column of $\mathbf{E}$ by the square root of its length. The square root of the length may be surprising but arises because of the second multiplication by the elements of $\mathbf{S}^{-T/2}$ in (3.4.123). This form of column scaling should be regarded as nonstatistical in that it is based upon inferences from the numerical magnitudes of the columns of $\mathbf{E}$ and does not employ information about the statistics of the solution. Indeed, its purpose is to prevent the imposition of structure on the solution for which no statistical basis has been anticipated.

If the system is full-rank overdetermined, the column weights drop out, just as we claimed for least squares above. To see this, consider that in the full-rank case,

$$\begin{aligned}\tilde{\mathbf{x}}' &= \left(\mathbf{E}'^T\mathbf{E}'\right)^{-1}\mathbf{E}'^T\mathbf{y}\\ \tilde{\mathbf{x}} &= \mathbf{S}^{T/2}\left(\mathbf{S}^{1/2}\mathbf{E}^T\mathbf{E}\mathbf{S}^{T/2}\right)^{-1}\mathbf{S}^{1/2}\mathbf{E}^T\mathbf{y}\\ &= \mathbf{S}^{T/2}\mathbf{S}^{-T/2}\left(\mathbf{E}^T\mathbf{E}\right)^{-1}\mathbf{S}^{-1/2}\mathbf{S}^{1/2}\mathbf{E}^T\mathbf{y} = \left(\mathbf{E}^T\mathbf{E}\right)^{-1}\mathbf{E}^T\mathbf{y}\,. \quad (3.4.125)\end{aligned}$$

Note the importance of doing column scaling following the row scaling; otherwise, interpretation of the row norms becomes very difficult.

3.4.3.5 Solution and Observation Resolution

Typically, either or both of the set of vectors $\mathbf{v}_i$, $\mathbf{u}_i$ used to present $\mathbf{x}$, $\mathbf{y}$ will be deficient in the sense of the expansions in (3.4.2). Deficiency of one or the other or both is guaranteed if the effective system rank differs from one of M or N.

It follows immediately from Equations (3.4.3) that the particular-SVD solution is

$$\tilde{\mathbf{x}} = \mathbf{V}_K\mathbf{V}_K^T\mathbf{x} \qquad (3.4.126)$$

and the data vector with which both it and the general solution are consistent is

$$\tilde{\mathbf{y}} = \mathbf{U}_K\mathbf{U}_K^T\mathbf{y}\,. \qquad (3.4.127)$$

Define

$$\mathbf{T}_v = \mathbf{V}_K\mathbf{V}_K^T\,, \qquad (3.4.128)$$

$$\mathbf{T}_u = \mathbf{U}_K\mathbf{U}_K^T\,, \qquad (3.4.129)$$

the solution and observation (data) resolution matrices, respectively.

Interpretation of the data resolution matrix is slightly subtle. Suppose an element of $\mathbf{y}$ was fully resolved–that is, some row, j_0, of $\mathbf{U}_K\mathbf{U}_K^T$ were all zeros except for diagonal element j_0, which is one. Then a change of unity in y_{j_0} would produce a change in $\tilde{\mathbf{x}}$ that would leave unchanged all other elements of $\tilde{\mathbf{y}}$. If element j_0 is *not* fully resolved, then a change of unity in observation y_{j_0} produces a solution that leads to changes in other elements of $\tilde{\mathbf{y}}$. Stated slightly differently, if y_i is not fully resolved, the system lacks adequate information to distinguish equation i from a linear dependence on one or more other equations.

One can use these ideas to construct quantitative statements of which observations are the most important (data ranking). From Equation (3.4.5), trace$(\mathbf{T}_u) = K$, and the relative contribution to the solution of any particular constraint is given by the corresponding diagonal element of $\mathbf{T}_u$.

Consider the example (3.4.119) without row weighting. At rank 1,

$$\mathbf{T}_u = \left\{ \begin{matrix} 0.0099 & 0.099 \\ 0.099 & 0.9901 \end{matrix} \right\},$$

showing that the second equation has played a much more important role in the solution than the first one, despite the fact that we asserted the expected noise in both to be the same. The reason is that described above; the second equation in effect asserts that the measurement is 10 times more accurate than in the first equation–and the data resolution matrix informs us of that explicitly. All of the statements made previously about resolution matrices now apply to $\mathbf{T}_u$, $\mathbf{T}_v$.

If row and column scaling have been applied to the equations prior to application of the SVD, the covariance, uncertainty, and resolution expressions apply in those new, scaled spaces. The resolution in the original spaces is

$$\mathbf{T}_v = \mathbf{S}^{T/2}\mathbf{T}_{v'}\mathbf{S}^{-T/2}, \tag{3.4.130}$$

$$\mathbf{T}_u = \mathbf{W}^{T/2}\mathbf{T}_{u'}\mathbf{W}^{-T/2}, \tag{3.4.131}$$

so that

$$\tilde{\mathbf{x}} = \mathbf{T}_v\mathbf{x}, \quad \tilde{\mathbf{y}} = \mathbf{T}_u\mathbf{y} \tag{3.4.132}$$

where $\mathbf{T}_{v'}$, $\mathbf{T}_{u'}$ are the expressions (3.4.128), (3.4.129) in the scaled space. The uncertainty in the new space is $\mathbf{P} = \mathbf{S}^{-T/2}\mathbf{P}'\mathbf{S}^{-1/2}$ where $\mathbf{P}'$ is the expression (3.4.100) or (3.4.101) in the scaled space.

We have seen an interpretation of three matrices obtained from the SVD: $\mathbf{V}\mathbf{V}^T$, $\mathbf{U}\mathbf{U}^T$, $\mathbf{V}\Lambda^{-2}\mathbf{V}^T$. The reader may well wonder, on the basis of the symmetries between solution and data spaces, whether there is an interpretation of the remaining matrix $\mathbf{U}\Lambda^{-2}\mathbf{U}^T$? Such an interpretation exists, but it will emerge most simply when we discuss constrained least squares and Lagrange multipliers.

3.4.3.6 Relation to Tapered and Weighted Least-Squares

In using least squares, a shift was made from the simple objective functions (3.3.4) and (3.3.47) to the more complicated (3.3.22) or (3.3.29). The change was made to permit a degree of control of the relative norms of $\tilde{\mathbf{x}}$, $\tilde{\mathbf{n}}$, and through the use of $\mathbf{W}$, $\mathbf{S}$ of the individual elements and the resulting uncertainties and covariances. Application of the weight matrices $\mathbf{W}$, $\mathbf{S}$ through their Cholesky decompositions to the equations prior to the use of the SVD is equally valid, thus providing the same amount of influence over the solution elements. The SVD provides its control over the solution norms, uncertainties, and covariances through choice of the effective rank

K'. This approach is different from the use of the extended objective functions (3.3.22), but the SVD is actually useful in understanding the effect of such functions.

Assume any necessary **W**, **S** have been applied, but retain α^2 (= 1) as a marker. Then, the full SVD, including zero singular values and corresponding singular vectors, is substituted into (3.3.23),

$$\tilde{\mathbf{x}} = (\alpha^2\mathbf{I} + \mathbf{V}\Lambda^T\Lambda\mathbf{V}^T)^{-1}\mathbf{V}\Lambda^T\mathbf{U}^T\mathbf{y}\,,$$

and using the orthogonality of **V**, we have

$$\tilde{\mathbf{x}} = \mathbf{V}(\Lambda^T\Lambda + \alpha^2\mathbf{I})^{-1}\Lambda^T\mathbf{U}^T\mathbf{y}\,. \tag{3.4.133}$$

The matrix to be inverted is diagonal and so

$$\tilde{\mathbf{x}} = \sum_{i=1}^{N} \frac{\lambda_i(\mathbf{u}_i^T\mathbf{y})}{\lambda_i^2 + \alpha^2}\mathbf{v}_i\,. \tag{3.4.134}$$

It is now apparent what the effect of tapering has done in least squares. The word refers to the tapering down of the coefficients of the $\mathbf{v}_i$ from the values they would have in the pure SVD. In particular, the guarantee that matrices like $(\mathbf{E}^T\mathbf{E} + \alpha^2\mathbf{I})$ would always have an inverse despite vanishing singular values is seen to follow because the inverse of the sum always exists, irrespective of the rank of **E**. The simple addition of a positive constant to the diagonal of a singular matrix is a well-known method for making it have an inverse. Such methods are a form of what is usually known as *regularization* and are procedures for suppressing nullspaces.

The residuals of the tapered least-squares solution can be written in various forms. Equation (3.3.24) can be written (using the orthogonality of **U**, **V**),

$$\tilde{\mathbf{n}} = \alpha^2\mathbf{U}(\alpha^2\mathbf{I} + \Lambda\Lambda^T)^{-1}\mathbf{U}^T\mathbf{y} = \sum_{i=1}^{M} \frac{(\mathbf{u}_i^T\mathbf{y})\alpha^2}{\lambda_i^2 + \alpha^2}\mathbf{u}_i\,, \tag{3.4.135}$$

that is, the projection of the noise onto the range vectors $\mathbf{u}_i$ no longer vanishes. Some of the structure of the range of $\mathbf{E}^T$ is being attributed to noise, and it is no longer true that the residuals are subject to the rigid requirement (3.4.68) of having zero contribution from the range vectors. An increased noise norm is also deemed acceptable, as the price of keeping the solution norm small, by assuring that none of the coefficients in the sum (3.4.134) becomes overly large–values we can control by varying α^2; Wiggins (1972) discusses this form of solution. The covariance of this solution about its mean [Equation (3.3.25)] is readily rewritten as

$$\begin{aligned}\mathbf{C}_{\tilde{x}\tilde{x}} &= \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{\lambda_i\lambda_j \mathbf{u}_i \mathbf{R}_{nn} \mathbf{u}_j^T}{(\lambda_i^2+\alpha^2)(\lambda_j^2+\alpha^2)} \mathbf{v}_i \mathbf{v}_j^T \\ &= \sigma_n^2 \sum_{i=1}^{N} \frac{\lambda_i^2}{(\lambda_i^2+\alpha^2)^2} \mathbf{v}_i \mathbf{v}_i^T \\ &= \sigma_n^2 \mathbf{V}(\mathbf{\Lambda}^T\mathbf{\Lambda}+\alpha^2\mathbf{I}_N)^{-1}\mathbf{\Lambda}^T\mathbf{\Lambda}(\mathbf{\Lambda}^T\mathbf{\Lambda}+\alpha^2\mathbf{I}_N)^{-1}\mathbf{V}^T \qquad (3.4.136)\end{aligned}$$

where the second line is again the special case of white noise. The role of α^2 in controlling the solution variance, as well as the solution size, should be plain. The tapered least-squares solution is biased, but the presence of the bias can greatly reduce the solution variance. In agnostic situations where one has no real knowledge of any expected variation in the relative sizes of the individual elements of $\mathbf{x}$, $\mathbf{n}$, nor of any correlations amongst them, both $\mathbf{W}$, $\mathbf{S}$ are proportional to the identity. In this situation α^2 is often retained as a simple measure of the ratios of the diagonal elements of $\mathbf{W}$, $\mathbf{S}$ and used to control the relative norms of $\tilde{\mathbf{x}}$, $\tilde{\mathbf{n}}$. Study of the solution as a function of α^2 is known as *ridge regression* (Hoerl & Kennard, 1970a,b), but the interpretation of the results is clearer in the statistical methodology of Section 3.5. Elaborate techniques have been developed for determining the right value of α^2 (see Lawson & Hanson, 1974, or Hansen, 1992, for reviews).[9]

The uncertainty, $\mathbf{P}$, is readily found as

$$\begin{aligned}\mathbf{P} &= \alpha^2 \sum_{i=1}^{N} \frac{\mathbf{v}_i\mathbf{v}_i^T}{(\alpha^2+\lambda_i^2)^2} + \sigma_n^2 \sum_{i=1}^{N} \frac{\lambda_i^2 \mathbf{v}_i\mathbf{v}_i^T}{(\lambda_i^2+\alpha^2)^2} \\ &= \alpha^2\mathbf{V}(\mathbf{\Lambda}^T\mathbf{\Lambda}+\alpha^2\mathbf{I})^{-2}\mathbf{V}^T \\ &\quad + \sigma_n^2\mathbf{V}(\mathbf{\Lambda}^T\mathbf{\Lambda}+\alpha^2\mathbf{I})^{-1}\mathbf{\Lambda}^T\mathbf{\Lambda}(\mathbf{\Lambda}^T\mathbf{\Lambda}+\alpha^2\mathbf{I})^{-1}\mathbf{V}^T \qquad (3.4.137)\end{aligned}$$

where one uses formally, $\mathbf{x} = \mathbf{V}\mathbf{V}^T\mathbf{x}$, $< \mathbf{x}\mathbf{x}^T > = \alpha^2\mathbf{I}$, and the contribution from the noise components is clearly separated.

The truncated SVD and the tapered SVD–tapered least-squares solutions produce the same qualitative effect: It is possible to increase the noise norm while decreasing the solution norm. Although the solutions differ somewhat, they both achieve a purpose stated above–to extend ordinary least squares in such a way that one can control the relative norms. The quantitative difference between them is readily stated: The truncated form makes a clear separation between range and nullspace in both solution and residual spaces;

[9] Hansen's (1992) discussion is particularly interesting because he exploits the generalized SVD, which is used to simultaneously diagonalize two matrices.

the particular SVD solution contains only range vectors and no nullspace vectors. The residual contains only nullspace vectors and no range vectors. The tapered form permits a merger of the two different sets of vectors with both solution and residuals containing some contribution from both formal range and nullspaces.

One advantage of the tapered form over the truncated SVD or simple least squares is worth noticing. A common empirical measure of a good least-squares fit is through the requirement that the residuals should be unstructured–that is, as nearly white noise as possible: $< \tilde{\mathbf{n}} > = 0$, $< \tilde{\mathbf{n}}\tilde{\mathbf{n}}^T > = \mathbf{I}$ as estimated by the sample averages. But if ordinary least squares (3.3.6) or the equivalent truncated SVD are used, the residuals cannot actually conform to this requirement because they lack the range vectors. That is,

$$< \left(\sum_{K+1}^{M} \beta_i \mathbf{u}_i \right) \left(\sum_{K+1}^{M} \beta_j \mathbf{u}_j \right)^T > \neq \mathbf{I}, \quad K > 0\,, \tag{3.4.138}$$

because any white-noise process must include contributions from the entire spanning set. If $K \ll M$, this problem may be undetectable. But if K approaches M, the possible structure in the residuals is so restricted by the few nullspace vectors available that it may produce highly non-random values. These considerations become paramount in Section 3.6.

3.4.3.7 Resolution of Tapered Solutions to Simultaneous Equations

The tapered least-squares solutions have an implicit nullspace, arising from the terms corresponding to zero singular values, or values small compared to α^2. Such solutions are often computed directly in the form (3.3.23)–(3.3.25) without ever bothering with the SVD–to save computing. But that solution form does a good job of hiding the existence of what should still be regarded as an effective nullspace.

To obtain a measure of solution resolution in the absence of the explicit $\mathbf{v}_i$ vectors, consider a situation in which the true solution were $\mathbf{x}_{j_0} \equiv \delta_{j,j_0}$–that is, unity in the j_0 element and zero elsewhere. Then, in the absence of noise (the resolution analysis applies to the noise-free situation), the correct value of $\mathbf{y}$ would be

$$\mathbf{E}\mathbf{x}_{j_0} = \mathbf{y}_{j_0}\,, \tag{3.4.139}$$

defining $\mathbf{y}_{j_0}$. If one actually knew (had measured) $\mathbf{y}_{j_0}$, what solution $\mathbf{x}_{j_0}$ would be obtained?

Tapered least squares produces [in the form (3.3.61)]

$$\tilde{\mathbf{x}}_{j_0} = \mathbf{E}^T(\mathbf{E}\mathbf{E}^T + \alpha^2\mathbf{I})^{-1}\mathbf{y}_{j_0} = \mathbf{E}^T(\mathbf{E}\mathbf{E}^T + \alpha^2\mathbf{I})^{-1}\mathbf{E}\mathbf{x}_{j_0}\,, \tag{3.4.140}$$

which is row (or column) j_0 of

$$\mathbf{T}_v = \mathbf{E}^T(\mathbf{E}\mathbf{E}^T + \alpha^2\mathbf{I})^{-1}\mathbf{E}\,. \tag{3.4.141}$$

Thus, we can interpret any row of $\mathbf{T}_v$ as the solution resolution for a Kronecker delta, correct solution, in that element. It is an easy matter, using the SVD of $\mathbf{E}$ and letting $\alpha^2 \to 0$ to show that (3.4.141) reduces to $\mathbf{V}\mathbf{V}^T$. These expressions apply in the row- and column-scaled space; Equations (3.4.130)–(3.4.131) are used to scale and rotate them into the original spaces.

An obvious variant of (3.4.141) follows from the alternative least-squares solution (3.3.23) and is

$$\mathbf{T}_v = (\mathbf{E}^T\mathbf{E} + \alpha^2\mathbf{I})^{-1}\mathbf{E}^T\mathbf{E}\,. \tag{3.4.142}$$

A solution resolution matrix is obtained similarly: Let $\mathbf{y}_{j_1}$ be zero, except for one in element j_1. Then (3.3.61) produces

$$\tilde{\mathbf{x}}_{j_1} = \mathbf{E}^T(\mathbf{E}\mathbf{E}^T + \alpha^2\mathbf{I})^{-1}\mathbf{y}_{j_1}\,,$$

which if substituted into the original equations is

$$\mathbf{E}\tilde{\mathbf{x}}_{j_1} = \mathbf{E}\mathbf{E}^T(\mathbf{E}\mathbf{E}^T + \alpha^2\mathbf{I})^{-1}\mathbf{y}_{j_1} = \tilde{\mathbf{y}}_{j_1}\,,$$

and thus

$$\mathbf{T}_u = \mathbf{E}\mathbf{E}^T(\mathbf{E}\mathbf{E}^T + \alpha^2\mathbf{I})^{-1}\,. \tag{3.4.143}$$

The alternate form from (3.3.23) is

$$\mathbf{T}_u = \mathbf{E}(\mathbf{E}^T\mathbf{E} + \alpha^2\mathbf{I})^{-1}\mathbf{E}^T\,, \tag{3.4.144}$$

which reduces to $\mathbf{U}\mathbf{U}^T$ as $\alpha^2 \to 0$. If row- and column-scaling matrices have been applied, the resolution matrices are modified in analogy to (3.4.130)–(3.4.132).

3.5 Using a Steady Model–Combined Least Squares and Adjoints

Consider now a modest generalization of the constrained problem Equation (3.3.2) in which the unknowns $\mathbf{x}$ are also meant to satisfy some constraints exactly, or nearly exactly, for example

$$\mathbf{A}\mathbf{x} = \mathbf{q}\,, \tag{3.5.1}$$

but to satisfy the observations (3.3.2) only approximately, in a least-squares sense. Equations like (3.5.1) will be referred to as the *model.* An example of a model occurs in acoustic tomography where we may have measurements of both density and velocity, and they are connected by the thermal wind equations (this case is written out by Munk & Wunsch, 1982). The distinction between the model (3.5.1) and the observations is usually an arbitrary one; $\mathbf{A}$ may well be some subset of the rows of $\mathbf{E}$, for which the corresponding error is believed negligible. What follows can in fact be obtained by imposing the zero-noise limit for some of the rows of $\mathbf{E}$ in the solutions already described. Furthermore, whether the model should be satisfied exactly, or should contain a noise element, too, is situation dependent. The thermal wind relationship is an approximation, and model error would normally be included in its enforcement, in which case the distinction between model and observations is purely conceptual. One should be wary of introducing exact equalities into estimation problems, because they carry the strong possibility of introducing small eigenvalues, or near singular relationships, into the solution, which may dominate the results.

Several approaches are now available to us. Consider for example, the objective function,

$$J = (\mathbf{Ex} - \mathbf{y})^T(\mathbf{Ex} - \mathbf{y}) + \alpha^2(\mathbf{Ax} - \mathbf{q})^T(\mathbf{Ax} - \mathbf{q}) \tag{3.5.2}$$

where $\mathbf{W}$, $\mathbf{S}$ have been applied if necessary and α^2 is retained as a tradeoff parameter. (This objective function corresponds to the requirement of a solution of the concatenated equation sets,

$$\begin{Bmatrix} \mathbf{E} \\ \mathbf{A} \end{Bmatrix} \mathbf{x} + \begin{bmatrix} \mathbf{n} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{y} \\ \mathbf{q} \end{bmatrix} \tag{3.5.3}$$

in which $\mathbf{u}$ is the model noise, and the weight given to the model is $\alpha^2\mathbf{I}$.) By letting $\alpha^2 \to \infty$, the model can be forced to apply with arbitrary accuracy. For any finite α^2, the model is formally a soft constraint here because it is being applied only in a minimized sum of squares. The solution follows immediately from (3.3.6) with

$$\mathbf{E} \to \begin{Bmatrix} \mathbf{E} \\ \alpha\mathbf{A} \end{Bmatrix}, \quad \mathbf{y} \to \begin{Bmatrix} \mathbf{y} \\ \alpha\mathbf{q} \end{Bmatrix},$$

assuming the matrix inverse exists.

Alternatively, the model can be imposed as a perfect hard constraint with $\mathbf{u} = 0$. All prior covariances and scalings having been been applied, and Lagrange multipliers introduced, reduces the problem to one with an objective function

$$J = \mathbf{n}^T\mathbf{n} - 2\boldsymbol{\mu}^T(\mathbf{Ax} - \mathbf{q}) = (\mathbf{Ex} - \mathbf{y})^T(\mathbf{Ex} - \mathbf{y}) - 2\boldsymbol{\mu}^T(\mathbf{Ax} - \mathbf{q}), \quad (3.5.4)$$

which is just a variant of (3.3.47). To avoid confusion, it is important to realize that we have essentially interchanged the roles of the two terms in (3.3.47)–with the expression (3.5.1) to be exactly satisfied but the observations only approximately so.

Setting the derivatives of J with respect to $\mathbf{x}$, $\boldsymbol{\mu}$ to zero gives the normal equations

$$-\mathbf{E}^T\mathbf{y} + \mathbf{E}^T\mathbf{Ex} - \mathbf{A}^T\boldsymbol{\mu} = 0, \quad (3.5.5)$$

$$\mathbf{Ax} - \mathbf{q} = 0. \quad (3.5.6)$$

Equation (3.5.5) represents the adjoint, or *dual* model, for the adjoint or dual solution $\boldsymbol{\mu}$. We can distinguish two extreme cases, one in which $\mathbf{A}$ is square, $N \times N$, and of full rank, and one in which $\mathbf{E}$ has this property. In the first case,

$$\tilde{\mathbf{x}} = \mathbf{A}^{-1}\mathbf{q} \quad (3.5.7)$$

and from (3.5.5),

$$\mathbf{E}^T\mathbf{EA}^{-1}\mathbf{q} - \mathbf{E}^T\mathbf{y} = \mathbf{A}^T\boldsymbol{\mu} \quad (3.5.8)$$

or

$$\tilde{\boldsymbol{\mu}} = \mathbf{A}^{-T}(\mathbf{E}^T\mathbf{EA}^{-1}\mathbf{q} - \mathbf{E}^T\mathbf{y}). \quad (3.5.9)$$

Here, the values of $\mathbf{x}$ are completely determined by the full-rank, noiseless model, and the minimization of the deviation from the observations is passive. The Lagrange multipliers or adjoint solution, however, are useful, providing the sensitivity information, $\partial J/\partial\mathbf{q} = 2\boldsymbol{\mu}$. The uncertainty of this solution is zero because of the perfect model (3.5.6).

In the second case, from (3.5.5),

$$\tilde{\mathbf{x}} = (\mathbf{E}^T\mathbf{E})^{-1}[\mathbf{E}^T\mathbf{y} + \mathbf{A}^T\tilde{\boldsymbol{\mu}}] \equiv \tilde{\mathbf{x}}_u + (\mathbf{E}^T\mathbf{E})^{-1}\mathbf{A}^T\tilde{\boldsymbol{\mu}}$$

where $\tilde{\mathbf{x}}_u = (\mathbf{E}^T\mathbf{E})^{-1}\mathbf{E}^T\mathbf{y}$ is the ordinary, unconstrained least-squares solution. Substituting into (3.5.6) produces

$$\tilde{\boldsymbol{\mu}} = [\mathbf{A}(\mathbf{E}^T\mathbf{E})^{-1}\mathbf{A}^T]^{-1}(\mathbf{q} - \mathbf{A}\tilde{\mathbf{x}}_u) \quad (3.5.10)$$

and

$$\tilde{\mathbf{x}} = \tilde{\mathbf{x}}_u + (\mathbf{E}^T\mathbf{E})^{-1}\mathbf{A}^T[\mathbf{A}(\mathbf{E}^T\mathbf{E})^{-1}\mathbf{A}^T]^{-1}(\mathbf{q} - \mathbf{A}\tilde{\mathbf{x}}_u), \quad (3.5.11)$$

assuming $\mathbf{A}$ is full-rank underdetermined. The perfect model is underdetermined; its range is being fit perfectly, with its nullspace being employed to reduce the misfit to the data as far as possible. The uncertainty of this solution may be written (Seber, 1977)

$$\begin{aligned}\mathbf{P} &= D^2(\tilde{\mathbf{x}} - \mathbf{x}) \qquad (3.5.12)\\ &= \sigma^2 \left\{(\mathbf{E}^T\mathbf{E})^{-1} - (\mathbf{E}^T\mathbf{E})^{-1}\mathbf{A}^T \left[\mathbf{A}(\mathbf{E}^T\mathbf{E})^{-1}\mathbf{A}^T\right]^{-1} \mathbf{A}(\mathbf{E}^T\mathbf{E})^{-1}\right\},\end{aligned}$$

which represents a reduction in the uncertainty of the ordinary least-squares solution (first term on the right) by the information in the perfectly known constraints. The presence in the inverse of terms involving **A** in these solutions is a manifestation of the warning about the possible introduction of components dependent upon small eigenvalues of **A**.

Example: Consider the least-squares problem of solving

$$\begin{aligned} x_1 + n_1 &= 1\\ x_2 + n_2 &= 1\\ x_1 + x_2 + n_3 &= 3\end{aligned}$$

with uniform, uncorrelated noise of variance 1 in each of the equations. The solution is then

$$\tilde{\mathbf{x}} = [1.3333 \quad 1.3333]^T$$

with uncertainty

$$\mathbf{P} = \begin{Bmatrix} 0.6667 & -0.333 \\ -0.333 & 0.6667 \end{Bmatrix}.$$

But suppose that it is known or desired that $x_1 - x_2 = 1$. Then (3.5.11) produces $\tilde{\mathbf{x}} = [1.8333 \quad 0.8333]^T$, $\boldsymbol{\mu} = 0.5$, $J = 0.8333$, with reduced uncertainty

$$\mathbf{P} = \begin{Bmatrix} 0.1667 & 0.1667 \\ 0.1667 & 0.1667 \end{Bmatrix}.$$

If the constraint is shifted to $x_1 - x_2 = 1.1$, the new solution is

$$\tilde{\mathbf{x}} = [1.8833 \quad 0.7833]^T$$

and the new objective function is $J = 0.9383$, a shift consistent with $\boldsymbol{\mu}$ and (3.3.56).

If neither **A** nor **E** is full rank, then the inverses appearing in equations (3.5.7), (3.5.11) will not exist. Either form can be used then, by replacing the inverses by, say, the particular SVD inverse. But in that case, the solution will not be unique if the combined observations and model leave a

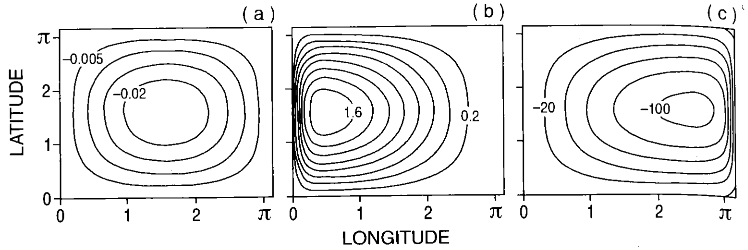

Figure 3–11. The Stommel Gulf Stream model solved using the adjoint: (a) depicts the windstress curl imposed, and (b) is the resulting transport streamfunction showing the expected westward intensification. The adjoint solution is shown in (c). Because it satisfies the adjoint equation, it shows an eastward intensification, consistent with a reversal of the sign of β.

nullspace in $\mathbf{x}$. The objective function (3.5.4) can be modified to have an extra term in $\mathbf{x}^T\mathbf{S}^{-1}\mathbf{x}$ if desired.

If the model has error terms, too, either in the forcing, $\mathbf{q}$, or in missing physics, it is modified to

$$\mathbf{A}\mathbf{x} + \mathbf{u} = \mathbf{q}\,. \tag{3.5.13}$$

A hard-constraint formulation can still be used, in which (3.5.13) is still to be exactly satisfied, imposed through an objective function of form,

$$J = (\mathbf{E}\mathbf{x} - \mathbf{y})^T(\mathbf{E}\mathbf{x} - \mathbf{y}) + \alpha^2\mathbf{u}^T\mathbf{u} - 2\boldsymbol{\mu}^T(\mathbf{A}\mathbf{x} + \mathbf{u} - \mathbf{q})\,. \tag{3.5.14}$$

It is again readily confirmed that the solutions using (3.5.2) or (3.5.14) are identical, and the hard/soft distinction is seen again to be artificial unless one truly has model equations with $\mathbf{u} = 0$. Equation (3.5.13) represents a model that is to be exactly satisfied; but it has an unknown *control* contribution, $\mathbf{u}$. Objective functions like (3.5.14) will be used extensively in Chapter 6. The most general form of objective function would be

$$J = \mathbf{n}^T\mathbf{R}^{-1}\mathbf{n} + \mathbf{x}^T\mathbf{S}^{-1}\mathbf{x} + \mathbf{u}^T\mathbf{Q}^{-1}\mathbf{u} - 2\boldsymbol{\mu}^T(\mathbf{A}\mathbf{x} + \mathbf{u} - \mathbf{q})\,. \tag{3.5.15}$$

If $\mathbf{A}$ is square and full rank, and $\mathbf{u} = 0$, one can readily confirm that $\mathbf{R}$ and $\mathbf{S}$ drop out of the solution.

Example: Let us apply these ideas to the Stommel Gulf Stream model. A code was written to solve by finite differences the nondimensional equation

$$\epsilon\nabla^2\phi + \frac{\partial\phi}{\partial x} = \hat{\mathbf{k}}\cdot\nabla\times\boldsymbol{\tau} \tag{3.5.16}$$

and is depicted in Figure 3–11 for the case $\epsilon = 0.05$ and $\phi = 0$ on the boundaries. The nondimensionalization and the basin dimension $0 \le x \le \pi$, $0 \le \phi \le \pi$ are those of Schröter and Wunsch (1986). The windstress curl was $\hat{\mathbf{k}}\cdot\nabla\times\boldsymbol{\tau} = -\sin x \sin y$.

The discretized form of the model is then the perfect $N \times N$ system

$$\mathbf{A}\mathbf{x} = \mathbf{q}\,, \quad \mathbf{x} = \{\phi_{ij}\}\,, \tag{3.5.17}$$

and $\mathbf{q}$ is the equivalently discretized windstress curl. The theory of partial differential equations shows that this system is full rank and generally well behaved. But let us ignore that knowledge and seek the values $\mathbf{x}$ that make the objective function (3.3.47)

$$J = \mathbf{x}^T\mathbf{x} - 2\boldsymbol{\mu}^T(\mathbf{A}\mathbf{x} - \mathbf{q}) \tag{3.5.18}$$

stationary with respect to $\mathbf{x}$, $\boldsymbol{\mu}$:

$$\mathbf{A}^T\boldsymbol{\mu} = \mathbf{x} \tag{3.5.19}$$

$$\mathbf{A}\mathbf{x} = \mathbf{q}\,. \tag{3.5.20}$$

$\mathbf{x}^T\mathbf{x}$ is readily identified with the solution potential energy. The solution $\boldsymbol{\mu}$, corresponding to the circulation of Figure 3–11b, is shown in Figure 3–11c. What is the interpretation? The Lagrange multipliers represent the sensitivity of the Stommel solution potential energy to perturbations in the windstress curl. We see that the sensitivity is greatest in the eastern half of the basin and indeed displays a boundary layer character. Schröter and Wunsch (1986) discuss this result in the context of the behavior of the Sverdrup interior of the Stommel model. A physical interpretation of the Lagrange multipliers can be inferred, given the simple structure of the governing equation (3.5.16) and the Dirichlet boundary conditions. This equation is not self-adjoint in the sense discussed by Morse and Feshbach (1953) or Lanczos (1961); the adjoint partial differential equation is of form

$$\epsilon\nabla^2\boldsymbol{\mu} - \frac{\partial\boldsymbol{\mu}}{\partial x} = \text{forcing}\ , \tag{3.5.21}$$

subject to mixed boundary conditions, and whose discrete form is (3.5.19), obtained by taking the transpose of the $\mathbf{A}$ matrix of the discretization. It is obvious from both (3.5.21) and Figure 3–11c that the adjoint solution represents flow streamlines in an ocean in which the sign of β has been reversed, resulting in an eastern boundary current, and the forcing is provided by the Stommel solution stream function. We can thus usefully think about the physics of an adjoint, or dual, ocean (or one might prefer the term *anti-ocean*) that governs the sensitivity of the real or "direct" ocean to parameter specifications. The structure of the $\boldsymbol{\mu}$ would change if J were changed.

The original objective function J is very closely analogous to the Lagrangian (not to be confused with the Lagrange multiplier) in classical mechanics. In mechanics, the gradients of the Lagrangian commonly are forces. The modified Lagrangian, J', is used in mechanics to impose various physical constraints, and the virtual force required to impose the constraints–for example, the demand that a particle follow a particular path–is the Lagrange multiplier. Lanczos (1970) has a good discussion. In an economics/management context, the multipliers are usually called *shadow prices* because they are intimately related to the question of how much profit (the objective function) will change with a change in the availability or cost of a product ingredient.

More generally, there is a close connection between the stationarity requirements imposed upon various objective functions throughout this book and the mathematics of classical mechanics. In Chapter 6, this analogy will be exploited to introduce the Hamiltonian form of governing equations.

A conventional dynamical model is one that is properly posed, which can be interpreted here as meaning that $\mathbf{A}$ is full rank of dimension $M \times M$ but it need not actually be so. Let us examine the state vector $\mathbf{x}$, Lagrange multiplier $\boldsymbol{\mu}$, pair in a little more detail. As already noted if (3.5.19) is underdetermined, then (3.5.20) is overdetermined (and vice versa). Some insight is obtained if the pair is rewritten in the SVD form

$$\mathbf{V}\boldsymbol{\Lambda}\mathbf{U}^T\boldsymbol{\mu} = \mathbf{x}\,, \tag{3.5.22}$$

$$\mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^T\mathbf{x} = \mathbf{q}\,. \tag{3.5.23}$$

Because of the structure of (3.5.22)–(3.5.23), the overdetermined system (whichever of the pair it is) will automatically satisfy the necessary solvability conditions, and an ill-posed model is readily handled.

Using the SVD inverse, (3.5.22)–(3.5.23) produce

$$\boldsymbol{\mu} = \mathbf{U}_K\boldsymbol{\Lambda}_K^{-2}\mathbf{U}_K^T\mathbf{q}\,. \tag{3.5.24}$$

We know that $\boldsymbol{\mu}$ are the sensitivity of J to perturbations in $\mathbf{q}$. Thus,

$$\frac{\partial J}{\partial \mathbf{q}} = 2\boldsymbol{\mu} = 2\mathbf{U}_K\boldsymbol{\Lambda}_K^{-2}\mathbf{U}_K^T\mathbf{q}\,, \tag{3.5.25}$$

where $\boldsymbol{\mu}$ contains the missing fourth matrix expression noticed in Section 3.4 and is the sensitivity of the objective function $\mathbf{x}^T\mathbf{x}$ to perturbations in the model elements $\mathbf{q}$. Taking the second derivative,

$$\frac{\partial^2 J}{\partial \mathbf{q}^2} = 2\mathbf{U}_K\boldsymbol{\Lambda}_K^{-2}\mathbf{U}_K^T\,, \tag{3.5.26}$$

is the Hessian of J. Evidently, if any of the λ_i become very small, the objective function will be extremely sensitive to small perturbations in the specification of $\mathbf{q}$, producing an effective nullspace of the problem. Equation (3.5.26) suggests that assertions that models are perfect can lead to difficulties. If the objective function (3.3.47) is used, Equation (3.5.26) represents the sensitivity to the data, $\mathbf{y}$.

3.5.1 Relation to Green's Functions

There is a close relationship between adjoint models and Green's functions. Consider any linear model, for example (1.2.4), the discrete Laplace equation with Dirichlet boundary conditions, which can be written as

$$\mathbf{Ax} = \rho \,. \tag{3.5.27}$$

To solve it, consider the collection of N adjoint problems

$$\mathbf{A}^T\mathbf{G}^T = \mathbf{I} \tag{3.5.28}$$

or

$$\mathbf{GA} = \mathbf{I}\,, \tag{3.5.29}$$

left multiplying (3.5.27) by $\mathbf{G}$, right multiplying (3.5.29) by $\mathbf{x}$, and subtracting,

$$\mathbf{G}\ \mathbf{Ax} - \mathbf{G}\ \mathbf{Ax} = \mathbf{G}\rho - \mathbf{x} \tag{3.5.30}$$

or

$$\mathbf{x} = \mathbf{G}\rho \,. \tag{3.5.31}$$

$\mathbf{G}$ is usually called the *Green's function*, which is seen to here satisfy a set of problems adjoint to the forward problem. The connection to the role of Green's functions in partial differential systems is laid out clearly in Morse and Feshbach (1953) and Lanczos (1961).

As in the general theory of Green's functions, there is an intimate relationship between the solution to (3.5.29) and the solution to the original problem, (3.5.27), for point disturbances. Consider the N separate problems

$$\mathbf{AX}_G = \mathbf{I} \tag{3.5.32}$$

where each column of $\mathbf{X}_G$ is the solution to a different right-hand side column of $\mathbf{I}_c$. As written, (3.5.32) is actually two different types of problem in a combined notation: In one type, a boundary value at one grid point is being set to 1, with zero everywhere else, and the interior sources, ρ, are all zero. In the other type, the boundary conditions are all zero, but one of the interior sources is being set to 1, all others being zero. Each separate problem in (3.5.32) gives rise to a solution vector $\mathbf{x}_G(j_0)$, where j_0 is the index of the nonzero boundary condition, or interior position source.

Now from (3.5.29), $\mathbf{X}_G = \mathbf{G} = \mathbf{A}^{-1}$–that is, each column of $\mathbf{G}$ corresponds to the solution to the forward problem for a unit disturbance at a particular boundary or grid point. Using the SVD for $\mathbf{A}$, it follows that

$$\mathbf{G} = \mathbf{U}\Lambda^{-1}\mathbf{V}^T \tag{3.5.33}$$

(if $\mathbf{A}$ is not of full rank, reduced SVDs are used).

Now consider a different problem. We wish to solve (3.5.27) but *in addition* have the independent knowledge that at an interior point, i_0,

$$x_{j_0} = \bar{\phi}_{j_0}\,, \tag{3.5.34}$$

which is not the same as specifying a disturbance at this point–rather it is a piece of information (compare Lanczos, 1961, p. 207). With the addition of the original equations and boundary conditions, the problem is now formally overspecified. Unless the value of $\bar{\phi}_{j_0}$ is chosen to be the specific value consistent with the solution to the original problem, or unless we are prepared to admit noise unknowns into the problem, the combination of (3.5.27) and (3.5.34) is a contradiction. Consider the consistency relationship that determines the unique value of $\bar{\phi}_{j_0}$, which would permit a solution by forming the overdetermined system

$$\mathbf{A}_1\mathbf{x} = \boldsymbol{\rho}_1, \mathbf{A}_1 = \begin{Bmatrix} \mathbf{A} \\ \delta_{jj_0} \end{Bmatrix}, \ \boldsymbol{\rho}_1 = \begin{bmatrix} \boldsymbol{\rho} \\ \bar{\phi}_{j_0} \end{bmatrix}. \tag{3.5.35}$$

The SVD of $\mathbf{A}_1$ will produce $N - \mathbf{v}_i$ and $N - \mathbf{u}_i$, corresponding to nonzero singular vectors. There will be one extra $\mathbf{u}_{N+1}$ in the nullspace. The solvability condition (3.4.70) is then

$$\mathbf{u}_{N+1}^T\boldsymbol{\rho}_1 = 0\,, \tag{3.5.36}$$

which is

$$\bar{\phi}_{j_0} = -\frac{\mathbf{u}'^T_{N+1}(j_0)\boldsymbol{\rho}}{u_{N+1,N+1}(j_0)} \tag{3.5.37}$$

where $\mathbf{u}'_{N+1}$ is defined as the vector containing only the first N elements of $\mathbf{u}_{N+1}$ and j_0 is written as an argument in $\mathbf{u}'_{N+1}$ to show its dependence upon the particular location of the information; $u_{N+1,N+1}(j_0)$ is the $N+1$st element of the nullspace vector. The calculation (3.5.37) can be done for each interior point j_0 and can be done for all interior points simultaneously by appending a separate equation of form (3.5.34) to (3.5.27) for each j_0 as

$$\mathbf{A}\mathbf{x} = \boldsymbol{\rho}\,,$$

$$\mathbf{I}\mathbf{x} = \bar{\phi} \tag{3.5.38}$$

where $\bar{\phi} = [\bar{\phi}_{j0}]$ is the vector of data points, or

$$\mathbf{A}_1\mathbf{x} = \rho_1, \mathbf{A}_1 = \begin{Bmatrix} \mathbf{A} \\ \mathbf{I} \end{Bmatrix}, \ \rho_1 = \begin{bmatrix} \rho \\ \bar{\phi} \end{bmatrix}. \tag{3.5.39}$$

Let the nullspace of $\mathbf{A}_1^T$, $\mathbf{u}_i$, $N+1 \leq i \leq 2N$, form a matrix, $\mathbf{Q}_u$. The solvability conditions are

$$\mathbf{Q}_u^T \rho_1 = 0 \tag{3.5.40}$$

or

$$\{\mathbf{u}_{iJ_2}\}^T \bar{\phi} = -\{\mathbf{u}_{iJ_1}\}^T \rho$$

where $\{\mathbf{u}_{iJ_2}\}$ is the matrix composed of the elements of the nullspace $\mathbf{u}_i$ in positions $N+1 \leq j \leq 2N$, and $\{\mathbf{u}_{iJ_1}\}$ are the first N elements of these vectors. Thus,

$$\bar{\phi} = -\{\mathbf{u}_{iJ_2}\}^{-T} \{\mathbf{u}_{iJ_1}\}^T \rho. \tag{3.5.41}$$

But ϕ is then a solution to the original problem, and it must follow [Equation (3.5.31)] that

$$\mathbf{G} = -\{\mathbf{u}_{iJ_2}\}^{-T} \{\mathbf{u}_{iJ_1}\}^T. \tag{3.5.42}$$

If ϕ are regarded as "data," then one can write instead,

$$\hat{\mathbf{x}} = \mathbf{G}\bar{\phi}. \tag{3.5.43}$$

Thus, although $\mathbf{G}$ is the solution to the physical problem of a point disturbance at a boundary point, or an interior point (keeping in mind the distinction between the physics of these two cases), it also serves as a way of determining the consistency of a set of "data," (3.5.39), with the solution to the original model system. If $\bar{\phi}$ are noisy, with white-noise contaminant, (3.5.43) will be a consistent solution to the equations agreeing as best as is possible with the observations.

3.6 Gauss-Markov Estimation, Mapmaking, and More Simultaneous Equations

The fundamental objective for least squares is minimization of the noise norm (3.3.4), although we complicated the discussion somewhat by introducing trade-offs against $\| \tilde{\mathbf{x}} \|$, various weights in the norms, and even the restriction that $\tilde{\mathbf{x}}$ should satisfy certain equations exactly. Least-squares methods, whether used directly as in (3.3.6) or indirectly through the vector representations of the SVD, are fundamentally deterministic–$\mathbf{W}$, $\mathbf{S}$, α^2 need not be given any statistical interpretation whatever–although sometimes one uses covariances for them. Statistics were used only to understand the sensitivity of the solutions to noise, and to obtain measures of the expected deviation of the solution from some supposed truth.

But there is another, radically different, approach to obtaining estimates of the solution to equation sets like (3.3.2), directed more clearly toward the physical goal: to find an estimate $\tilde{\mathbf{x}}$ which deviates as little as possible in the *mean square* from the true solution. That is, we wish to minimize the statistical quantities $< (\tilde{\mathbf{x}} - \mathbf{x})_i^2 >$. The next section is devoted to understanding how to find such an $\tilde{\mathbf{x}}$ (and the corresponding $\tilde{\mathbf{n}}$) through an excursion into statistical estimation theory. It is far from obvious that this $\tilde{\mathbf{x}}$ should bear any resemblance to one of the least-squares estimates, but as will be seen, under some circumstances the two are identical. Their possible identity is extremely useful but has apparently led many investigators to confuse the methodologies and therefore the interpretation of the result.

3.6.1 The Fundamental Result

Suppose we are interested in making an estimate of a physical variable $\mathbf{x}$, which might be a vector or a scalar and might be constant with space and time, or vary with either or both. To be definite, let $\mathbf{x}$ be a function of an independent variable $\mathbf{r}$, written discretely as $\mathbf{r}_j$ (it might be a vector of space coordinates, or a scalar time, or an accountant's label). Let us make some suppositions about what is usually called *prior information.* In particular, suppose we have an estimate of the low-order statistics describing $\mathbf{x}$–that is, we specify its mean and second moments:

$$< \mathbf{x} > = \mathbf{x}_0\,, \quad < \mathbf{x}(\mathbf{r}_i)\mathbf{x}(\mathbf{r}_j)^T > = \mathbf{R}_{xx}(\mathbf{r}_i,\, \mathbf{r}_j)\,. \tag{3.6.1}$$

To have a concrete problem, one might think of $\mathbf{x}$ as being temperature at 700-m depth in the ocean (a scalar) and $\mathbf{r}_j$ as a vector of horizontal positions; $\mathbf{x}$ is the vector each of whose elements is the scalar value at a different position. Alternatively, in one dimension, the elements of $\mathbf{x}$ would be the salinity along a surface transect by a ship. Then r_j is the scalar of position, either time or distance, and $\mathbf{r}$ is the vector of all such positions. But if the field of interest is the velocity vector, then each element of $\mathbf{x}$ is itself a vector, and one can extend the notation in a straightforward fashion. To keep the notation a little cleaner, however, we will usually treat the elements of $\mathbf{x}$ as scalars.

Now suppose there exist observations, y_i, as a function of the same coordinate $\mathbf{r}_i$, with known second moments

$$\mathbf{R}_{yy} = < \mathbf{y}\mathbf{y}^T > , \quad \mathbf{R}_{xy}(\mathbf{r}_i, \mathbf{r}_j) = < \mathbf{x}(\mathbf{r}_i)\mathbf{y}(\mathbf{r}_j)^T > , \quad 1 \leq i, j \leq M \tag{3.6.2}$$

(the individual observation elements can also be vectors–for example, two or three components of velocity and a temperature at a point–but as with $\mathbf{x}$, the modifications required to treat this case are straightforward, and we will maintain the simplicity of assuming scalar observations). Could the measurements be used to make an estimate of $\mathbf{x}$ at a point $\tilde{\mathbf{r}}_\alpha$, which may not coincide with one of the places (labels) where an observation is available? The idea is to exploit the concept that finite covariances carry predictive capabilities from known variables to unknown ones. A specific example would be to suppose the measurements are of temperature $y(\mathbf{r}_j) = y_0(\mathbf{r}_j)+n(\mathbf{r}_j)$, where n is the noise, and we wish to estimate the temperature at different locations, perhaps on a regular grid $\tilde{\mathbf{r}}_\alpha$, $1 \leq \alpha \leq N$. This special problem is one of gridding or mapmaking (the tilde is placed on $\mathbf{r}_\alpha$ as a device to emphasize that this is a location where an estimate is sought; the numerical values of these places or labels are known). Alternatively, and somewhat more interesting, perhaps the measurements are more indirect, with $y(r_i)$ representing a velocity field component at depth in the ocean and believed connected through the thermal wind equation to the temperature field. We might want to estimate the temperature from measurements of the velocity.

Given the discussion immediately following Equation (3.2.30), we seek an estimate $\tilde{x}(\tilde{\mathbf{r}}_\alpha)$, whose dispersion about its true value, $x(\tilde{\mathbf{r}}_\alpha)$, is as small as possible–that is,

$$\mathbf{P}(\tilde{\mathbf{r}}_\alpha, \tilde{\mathbf{r}}_\alpha) = < (\tilde{x}(\tilde{\mathbf{r}}_\alpha) - x(\tilde{\mathbf{r}}_\alpha))(\tilde{x}(\tilde{\mathbf{r}}_\beta) - x(\tilde{\mathbf{r}}_\beta)) > |_{\tilde{\mathbf{r}}_\alpha = \tilde{\mathbf{r}}_\beta}$$

is to be minimized (a minimum variance estimate). If we would like to answer the question for more than one point, and if we would like to understand the covariance of the errors of our estimates at various points $\tilde{\mathbf{r}}_\alpha$, then we can form a vector of values to be estimated, $\{\tilde{x}(\mathbf{r}_\alpha)\} \equiv \tilde{\mathbf{x}}$, and the uncertainty among them,

$$\begin{aligned}\mathbf{P}(\tilde{\mathbf{r}}_\alpha, \tilde{\mathbf{r}}_\beta) &= < (\tilde{x}(\tilde{\mathbf{r}}_\alpha) - x(\tilde{\mathbf{r}}_\alpha))(\tilde{x}(\tilde{\mathbf{r}}_\beta) - x(\tilde{\mathbf{r}}_\beta)) \\ &= < (\tilde{\mathbf{x}} - \mathbf{x})(\tilde{\mathbf{x}} - \mathbf{x})^T > , \qquad 1 \leq \alpha \leq N \, , \, 1 \leq \beta \leq N \, ,\end{aligned} \tag{3.6.3}$$

where the *diagonal* elements are to be *individually* minimized.

What should the relationship be between data and estimate? At least initially, one might try a linear combination of data,

$$\tilde{x}(\tilde{\mathbf{r}}_\alpha) = \sum_{j=1}^{M} B(\tilde{\mathbf{r}}_\alpha, \mathbf{r}_j) y(\mathbf{r}_j), \tag{3.6.4}$$

for all α, which makes the diagonal elements of $\mathbf{P}$ in (3.6.3) as small as possible. All the points $\tilde{\mathbf{r}}_\alpha$ can be treated simultaneously by letting $\mathbf{B}$ be an $M \times N$ matrix, and

$$\tilde{\mathbf{x}} = \mathbf{B}(\tilde{\mathbf{r}}_\alpha, \mathbf{r}_j)\mathbf{y}. \tag{3.6.5}$$

(This notation is mixed. Equation (3.6.5) is a shorthand for (3.6.4), in which the argument has been put into $\mathbf{B}$ explicitly as a reminder that there is a summation over all the data locations $\mathbf{r}_j$ for all mapping locations $\tilde{\mathbf{r}}_\alpha$.)

An important theorem, usually called the *Gauss-Markov theorem*, produces the values of $\mathbf{B}$ so as to minimize the diagonal elements of $\mathbf{P}$. The following heuristic derivation is based on that in Liebelt (1967): Substituting (3.6.5) into (3.6.3) and expanding,

$$\begin{aligned}\mathbf{P}(\tilde{\mathbf{r}}_\alpha, \tilde{\mathbf{r}}_\beta) &= < (\mathbf{B}(\tilde{\mathbf{r}}_\alpha, \mathbf{r}_j)\mathbf{y} - x(\tilde{\mathbf{r}}_\alpha))(\mathbf{B}(\tilde{\mathbf{r}}_\beta, \mathbf{r}_l)\mathbf{y} - x(\tilde{\mathbf{r}}_\beta))^T > \\ &\equiv < (\mathbf{B}\mathbf{y} - \mathbf{x})(\mathbf{B}\mathbf{y} - \mathbf{x})^T > |_{\alpha\beta} \\ &= \mathbf{B} < \mathbf{y}\mathbf{y}^T > \mathbf{B}^T - < \mathbf{x}\mathbf{y}^T > \mathbf{B}^T - \mathbf{B} < \mathbf{y}\mathbf{x}^T > + < \mathbf{x}\mathbf{x}^T > |_{\alpha\beta}\end{aligned} \tag{3.6.6}$$

(keep in mind that $\mathbf{y}$ is a function of the data positions $\mathbf{r}_j$, $\tilde{\mathbf{x}}$ is a function of the estimation positions $\tilde{\mathbf{r}}_\beta$, and $\mathbf{B}$ is a function of both). Using $\mathbf{R}_{xy} = \mathbf{R}_{yx}^T$, Equation (3.6.6) is

$$\mathbf{P} = \mathbf{B}\mathbf{R}_{yy}\mathbf{B}^T - \mathbf{R}_{xy}\mathbf{B}^T - \mathbf{B}\mathbf{R}_{xy}^T + \mathbf{R}_{xx}. \tag{3.6.7}$$

Notice that because $\mathbf{R}_{xx}$ represents the moments of $\mathbf{x}$ evaluated at the estimation positions, it is a function of $\tilde{\mathbf{r}}_\alpha$, $\tilde{\mathbf{r}}_\beta$, whereas $\mathbf{R}_{xy}$ involves covariances of $\mathbf{y}$ at the data positions with $\mathbf{x}$ at the estimation positions, and is consequently a function $\mathbf{R}_{xy}(\tilde{\mathbf{r}}_\alpha, \mathbf{r}_j)$.

Now, by completing the square (3.1.26), Equation (3.6.7) becomes

$$\mathbf{P} = (\mathbf{B} - \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1})\mathbf{R}_{yy}(\mathbf{B} - \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1})^T - \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{R}_{xy}^T + \mathbf{R}_{xx}. \tag{3.6.8}$$

The diagonal elements of (3.6.8) are the variances of the estimate at points $\tilde{\mathbf{r}}_\alpha$ about their true values. Because $\mathbf{R}_{xx}$ and $\mathbf{R}_{yy}$ are positive-definite, they and their inverses have positive diagonal elements (if they are only positive semidefinite, $\mathbf{R}_{yy}^{-1}$ has to be redefined, but we will ignore this pathology). By the symmetries present, then, the diagonal elements of all three terms in (3.6.8) are positive. Thus, minimization of any diagonal element of $\mathbf{P}$ is obtained by choosing $\mathbf{B}$ so that the first term vanishes, or

$$\mathbf{B}(\tilde{\mathbf{r}}_\alpha, \mathbf{r}_j) = \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1} \,. \tag{3.6.9}$$

Then the minimum variance estimate is

$$\tilde{\mathbf{x}}(\tilde{\mathbf{r}}_\alpha) = \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{y} \,, \tag{3.6.10}$$

and the actual minimum value of the diagonal elements of $\mathbf{P}$ is found by substituting back into (3.6.7), producing

$$\mathbf{P}(\tilde{\mathbf{r}}_\alpha, \tilde{\mathbf{r}}_\beta) = \mathbf{R}_{xx}(\tilde{\mathbf{r}}_\alpha, \tilde{\mathbf{r}}_\beta) - \mathbf{R}_{xy}(\tilde{\mathbf{r}}_\alpha, \mathbf{r}_j)\mathbf{R}_{yy}^{-1}(\mathbf{r}_j, \mathbf{r}_k)\mathbf{R}_{xy}^T(\tilde{\mathbf{r}}_\beta, \mathbf{r}_k) \,. \tag{3.6.11}$$

The bias of (3.6.10) is

$$< \tilde{\mathbf{x}} - \mathbf{x} > = \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1} < \mathbf{y} > -\mathbf{x} \,. \tag{3.6.12}$$

If $< \mathbf{y} > = \mathbf{x} = 0$, the estimator is unbiased and called a *best linear unbiased estimator*, or *BLUE*; otherwise, it is biased. It is not difficult to show that $\tilde{\mathbf{x}}$ is also the maximum likelihood estimate if the solution is jointly normal.

3.6.2 Linear Algebraic Equations

The result (3.6.9)–(3.6.11) is the abstract general case and is deceptively simple. Understanding it is far from trivial, and for many applications, some simplifications are very useful. Suppose the observations are related to the unknown vector $\mathbf{x}$ as in our canonical problem–that is, through a set of linear equations: $\mathbf{Ex} + \mathbf{n} = \mathbf{y}$. The measurement moments, $\mathbf{R}_{yy}$, can be computed directly:

$$\mathbf{R}_{yy} = < (\mathbf{Ex} + \mathbf{n})(\mathbf{Ex} + \mathbf{n})^T > = \mathbf{E}\mathbf{R}_{xx}\mathbf{E}^T + \mathbf{R}_{nn} \tag{3.6.13}$$

where the unnecessary but simplifying and often excellent assumption was made that the cross-terms of form

$$\mathbf{R}_{xn} = \mathbf{R}_{nx}^T = \mathbf{0} \,, \tag{3.6.14}$$

so that

$$\mathbf{R}_{xy} = < \mathbf{x}(\mathbf{Ex} + \mathbf{n})^T > = \mathbf{R}_{xx}\mathbf{E}^T \,, \tag{3.6.15}$$

that is, there is no correlation between the measurement noise and the actual state vector (e.g., that the noise in a temperature measurement does not depend upon whether the true value is 10° or 25°).

Under these circumstances, Equations (3.6.10), (3.6.11) take on the form:

$$\tilde{\mathbf{x}} = \mathbf{R}_{xx}\mathbf{E}^T(\mathbf{E}\mathbf{R}_{xx}\mathbf{E}^T + \mathbf{R}_{nn})^{-1}\mathbf{y}\,, \tag{3.6.16}$$

$$\tilde{\mathbf{n}} = \left\{\mathbf{I} - \mathbf{E}\mathbf{R}_{xx}\mathbf{E}^T(\mathbf{E}\mathbf{R}_{xx}\mathbf{E}^T + \mathbf{R}_{nn})^{-1}\right\}\mathbf{y}\,, \tag{3.6.17}$$

$$\mathbf{P} = \mathbf{R}_{xx} - \mathbf{R}_{xx}\mathbf{E}^T(\mathbf{E}\mathbf{R}_{xx}\mathbf{E}^T + \mathbf{R}_{nn})^{-1}\mathbf{E}\mathbf{R}_{xx}\,, \tag{3.6.18}$$

$$\mathbf{P}_{nn} = \left\{\mathbf{I} - \mathbf{E}\mathbf{R}_{xx}\mathbf{E}^T(\mathbf{E}\mathbf{R}_{xx}\mathbf{E}^T + \mathbf{R}_{nn})^{-1}\right\} \times \mathbf{R}_{nn}\left\{\mathbf{I} - \mathbf{E}\mathbf{R}_{xx}\mathbf{E}^T(\mathbf{E}\mathbf{R}_{xx}\mathbf{E}^T + \mathbf{R}_{nn})^{-1}\right\}\,. \tag{3.6.19}$$

These latter expressions are extremely important; they permit discussion of the solution to a set of linear algebraic equations in the presence of noise using information concerning the statistics of the noise and of the solution. Notice that they are *identical to the least-squares expression* (3.3.66)–(3.3.70) *if* $\mathbf{S} = \mathbf{R}_{xx}$, $\mathbf{W} = \mathbf{R}_{nn}$.

From the matrix inversion lemma, Equations (3.6.16)–(3.6.18) can be rewritten

$$\tilde{\mathbf{x}} = (\mathbf{R}_{xx}^{-1} + \mathbf{E}^T\mathbf{R}_{nn}^{-1}\mathbf{E})^{-1}\mathbf{E}^T\mathbf{R}_{nn}^{-1}\mathbf{y}\,, \tag{3.6.20}$$

$$\tilde{\mathbf{n}} = \left\{\mathbf{I} - \mathbf{E}(\mathbf{R}_{xx}^{-1} + \mathbf{E}^T\mathbf{R}_{nn}^{-1}\mathbf{E})^{-1}\mathbf{E}^T\mathbf{R}_{nn}^{-1}\right\}\mathbf{y}\,, \tag{3.6.21}$$

$$\mathbf{P} = (\mathbf{R}_{xx}^{-1} + \mathbf{E}^T\mathbf{R}_{nn}^{-1}\mathbf{E})^{-1}\,, \tag{3.6.22}$$

$$\mathbf{P}_{nn} = \left\{\mathbf{I} - \mathbf{E}(\mathbf{R}_{xx}^{-1} + \mathbf{E}^T\mathbf{R}_{nn}^{-1}\mathbf{E})^{-1}\mathbf{E}^T\mathbf{R}_{nn}^{-1}\right\} \times \mathbf{R}_{nn}\left\{\mathbf{I} - \mathbf{E}(\mathbf{R}_{xx}^{-1} + \mathbf{E}^T\mathbf{R}_{nn}^{-1}\mathbf{E})^{-1}\mathbf{E}^T\mathbf{R}_{nn}^{-1}\right\}\,. \tag{3.6.23}$$

Although these alternate forms are algebraically and numerically identical to (3.6.16)–(3.6.19), the size of the matrices to be inverted changes from $M \times M$ matrices to $N \times N$, where $\mathbf{E}$ is $M \times N$ (but note that $\mathbf{R}_{nn}$ is $M \times M$; the efficacy of this alternate form may depend upon whether the *inverse* of $\mathbf{R}_{nn}$ is known). Depending upon the relative magnitudes of M, N, one form may be much preferable to the other; finally, (3.6.22) has an important interpretation that will be discussed when we come to recursive methods. Recall, too, the options we had with the SVD of solving $M \times M$ or $N \times N$ problems. Equations (3.6.20)–(3.6.23) are identical to the alternative form least-squares solution (3.3.38)–(3.3.41) if $\mathbf{S} = \mathbf{R}_{xx}$, $\mathbf{W} = \mathbf{R}_{nn}$.

The solution (3.6.16)–(3.6.18) or (3.6.20)–(3.6.22) is an estimator; it was found by demanding a solution with the minimum dispersion about the true solution and is seen, surprisingly, to be identical with the tapered, weighted least-squares solution when the least-squares objective function weights are chosen to be the corresponding second-moment matrices of $\mathbf{x}$, $\mathbf{n}$. This correspondence of the two solutions often leads them to be seriously confused. It is essential to recognize that the logic of the derivations are quite distinct: We were free in the least-squares derivation to use weight matrices that were anything we wished–as long as appropriate inverses existed.

The correspondence of least squares with minimum variance estimation can be understood by recognizing that the Gauss-Markov estimator was derived by minimizing a quadratic objective function. The least-squares estimate was obtained from minimizing a summation that was a sample *estimate* of the Gauss-Markov objective function with $\mathbf{S}$, $\mathbf{W}$ properly chosen. The coincidence of the answers can be exploited in a number of ways. For example, we infer immediately that a resolution discussion directly analogous to Equations (3.4.142)–(3.4.144) for the least-squares solution is possible for the Gauss-Markov solution.

As with any statistical estimator, one must make *posterior* checks that the behavior of $\tilde{\mathbf{x}}$, $\tilde{\mathbf{n}}$ is consistent with the assumed prior statistics reflected in $\mathbf{R}_{xx}$, $\mathbf{R}_{nn}$, and any assumptions about their means or other properties. Such posterior checks are both essential and very demanding. One sometimes hears it said that estimation using Gauss-Markov and related methods is "pulling solutions out of the air" because the prior moment matrices $\mathbf{R}_{xx}$, $\mathbf{R}_{nn}$ often are only poorly known. But producing solutions that pass the test of consistency with the prior covariances can be very difficult. Solutions tend to be somewhat insensitive to the details of the prior statistics, and it is easy to become overly concerned with the detailed structure of $\mathbf{R}_{xx}$, $\mathbf{R}_{nn}$. As stated previously, it is also rare to be faced with a situation in which one is truly ignorant of the moments–*true ignorance* meaning that arbitrarily large or small numerical values of x_i, n_i would be acceptable. In the box inversions of Chapter 2 (to be revisited in Chapter 4), deep ocean velocity fields of order 1000 cm/s would be absurd, and their absurdity is readily asserted by choosing $\mathbf{R}_{xx} = \mathrm{diag}((10\ \mathrm{cm/s})^2)$, which reflects a mild belief that velocities are 0(10 cm/s) with no known correlations with each other. Testing of statistical estimates against prior hypotheses is a highly developed field in applied statistics, and we leave it to the references (e.g., Seber, 1977) for their discussion. Should such tests be failed, one must reject the solutions $\tilde{\mathbf{x}}$, $\tilde{\mathbf{n}}$ and ask why they failed–as it usually implies an incorrect model, ($\mathbf{E}$), or misunderstanding of the observational noise structure.

If the intention is that least-squares solutions are to be equivalent to the

Gauss-Markov ones, they must pass the same statistical tests. The simplest of these is that the weight matrices, thought to represent the covariances of $\mathbf{x}$, $\mathbf{n}$, should be shown after the fact to have been reasonable. Objective functions such as (3.3.36) must also have values consistent with the hypothesis. For example, substitution of the (scaled and weighted) solutions that are appropriate to (3.3.36) produce

$$< \tilde{J} > = < \mathbf{x}^T\mathbf{x} > \; + \; < \mathbf{n}^T\mathbf{n} > = N + M - K\,, \tag{3.6.24}$$

with an actual value consistent with a χ^2_ν probability density, and the values of the individual terms of $\tilde{J}$ should, as previously discussed, prove consistent with a χ^2_1 distribution. Note that the number of degrees of freedom, ν, in $\tilde{J}$ would be approximately $N + M - K$, where K is the rank of $\mathbf{E}$. (Draper & Smith, 1982, Chapters 2 and 3 discuss such problems in detail.)

3.6.2.1 Use of Basis Functions

A superficially different way of dealing with prior statistical information is often commonly used. Suppose that the indices of x_i refer to a spatial or temporal position, call it r_i, so that $x_i = x(r_i)$. Then it is often sensible to consider expanding the unknown $\mathbf{x}$ in a set of basis functions, F_j–for example, sines and cosines, Chebyshev polynomials, ordinary polynomials, etc. One might write

$$x(r_i) = \sum_{j=1}^{L} \alpha_j F_j(r_i)$$

or

$$\mathbf{x} = \mathbf{F}\boldsymbol{\alpha}\,, \quad \mathbf{F} = \left\{ \begin{matrix} F_1(r_1) & F_2(r_1) & \cdots & F_L(r_1) \\ F_1(r_2) & F_2(r_2) & \cdots & F_L(r_2) \\ \cdot & \cdot & \cdot & \cdot \\ F_1(r_N) & F_2(r_N) & \cdots & F_L(r_N) \end{matrix} \right\}, \quad \boldsymbol{\alpha}^T = [\alpha_1 \cdots \alpha_L]^T\,,$$

which, when substituted into (3.3.2), produces

$$\mathbf{L}\boldsymbol{\alpha} + \mathbf{n} = \mathbf{y}\,, \quad \mathbf{L} = \mathbf{E}\mathbf{F}\,. \tag{3.6.25}$$

If $L < M < N$, one can convert an underdetermined system into one that is formally overdetermined and, of course, the reverse is possible as well. More generally, one transfers the discussion of solution covariance, etc., to the expansion coefficients $\boldsymbol{\alpha}$. If there are special conditions applying to $\mathbf{x}$, such as boundary conditions at certain positions, r_B, a choice of basis function satisfying those conditions could be more convenient than appending them as additional equations.

It should be apparent, however, that the solution to (3.6.25) will have a covariance structure dictated in large part by that contained in the basis functions chosen; thus, there is no fundamental gain in employing basis functions although they may be convenient, numerically or otherwise. If $\mathbf{P}_{\alpha\alpha}$ denotes the uncertainty of $\boldsymbol{\alpha}$, then

$$\mathbf{P} = \mathbf{F}\mathbf{P}_{\alpha\alpha}\mathbf{F}^T \tag{3.6.26}$$

is the uncertainty of $\mathbf{x}$.

Example: The underdetermined system

$$\left\{\begin{matrix} 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 1 \end{matrix}\right\} \mathbf{x} + \mathbf{n} = \begin{bmatrix} 1 \\ -1 \end{bmatrix},$$

with noise variance $< \mathbf{n}\mathbf{n}^T > = .01\mathbf{I}$, has a solution, if $\mathbf{R}_{xx} = \mathbf{I}$, of

$$\tilde{\mathbf{x}} = \mathbf{E}^T(\mathbf{E}\mathbf{E}^T + .01\mathbf{I})^{-1}\mathbf{y} = [0 \pm 0.7 \quad 0.5 \pm 0.7 \quad 0.5 \pm 0.7 \quad 0 \pm 0.7]^T,$$
$$\tilde{\mathbf{n}} = [.0025 \pm 0.002 \quad -.0025 \pm 0.002]^T.$$

If the solution was thought to be large scale and smooth, one might use the covariance

$$\mathbf{R}_{xx} = \left\{\begin{matrix} 1 & .999 & .998 & .997 \\ .999 & 1 & .999 & .998 \\ .998 & .999 & 1 & .999 \\ .997 & .998 & .999 & 1 \end{matrix}\right\},$$

which produces a solution

$$\tilde{\mathbf{x}} = [0.18 \pm 0.05 \quad 0.32 \pm 0.04 \quad 0.32 \pm 0.04 \quad 0.18 \pm 0.05]^T,$$

$$\tilde{\mathbf{n}} = [4.6 \times 10^{-4} \pm 6.5 \times 10^{-5}, \ -0.71 \pm 0.07]^T,$$

which has a larger-scale property as desired and a smaller standard error.

If one attempts a solution as a first-order polynomial,

$$x_i = a + br_i, \ r_1 = 0, \ r_2 = 1, \ r_3 = 2, \ldots,$$

the system will become two equations in the two unknowns a, b:

$$\mathbf{E}\mathbf{F}\begin{bmatrix} a \\ b \end{bmatrix} = \left\{\begin{matrix} 4 & 6 \\ -2 & -6 \end{matrix}\right\} \begin{bmatrix} a \\ b \end{bmatrix} + \mathbf{n} = \begin{bmatrix} 1 \\ -1 \end{bmatrix},$$

and if the covariance of a, b is the identity matrix,

$$[\tilde{a}, \tilde{b}] = [4 \times 10^{-4} \pm 0.07, \ 0.2 \pm 0.04],$$

$$\tilde{\mathbf{x}} = [4 \times 10^{-3} \pm 0.07 \quad 0.17 \pm 0.04 \quad 0.33 \pm 0.02 \quad 0.5 \pm 0.05]^T ,$$

$$\tilde{\mathbf{n}} = [0.0002 \pm 0.0004 , \; -1.00 \pm 0.0005]^T ,$$

which is also large scale and smooth but clearly different than that from the Gauss-Markov estimator. Although this latter solution has been obtained from a just-determined system, it is not clearly better. If a linear trend is expected in the solution, then the polynomial expansion is certainly convenient–although such a structure can be imposed through use of $\mathbf{R}_{xx}$ by specifying a growing variance with r_i.

3.6.3 Determining a Mean Value

Let the measurements of the physical quantity continue to be denoted y_i and suppose that each is made up of an unknown large-scale mean m, plus a deviation from that mean of θ_i. Then,

$$m + \theta_i = y_i \, , \quad 1 \leq i \leq M \tag{3.6.27}$$

or

$$\mathbf{D}m + \boldsymbol{\theta} = \mathbf{y} \, , \; \mathbf{D}^T = [1 \quad 1 \quad 1 \quad \cdots \quad 1]^T , \tag{3.6.28}$$

and we seek a best estimate, $\tilde{m}$, of m. In (3.6.27) or (3.6.28) the unknown $\mathbf{x}$ has become the scalar m, and the deviation of the field from its mean is the noise, that is, $\boldsymbol{\theta} \equiv \mathbf{n}$, whose true mean is zero. The problem is evidently a special case of the use of basis functions, in which only one function–a zeroth-order polynomial, m–is retained.

Set $\mathbf{R}_{nn} = \mathbf{C}_{nn} = < \boldsymbol{\theta}\boldsymbol{\theta}^T >$. If, for example, we were looking for a large-scale mean temperature in a field of oceanic mesoscale eddies, then $\mathbf{R}_{nn}$ is the sum of the covariance of the eddy field plus that of observational errors and any other fields contributing to the difference between y_i and the true mean m. To be general, suppose $\mathbf{R}_{xx} = < m^2 > = m_0^2$ and from (3.6.20),

$$\begin{aligned} \tilde{m} &= \left\{ \frac{1}{m_0^2} + \mathbf{D}^T \mathbf{R}_{nn}^{-1} \mathbf{D} \right\}^{-1} \mathbf{D}^T \mathbf{R}_{nn}^{-1} \mathbf{y} \\ &= \frac{1}{1/m_0^2 + \mathbf{D}^T \mathbf{R}_{nn}^{-1} \mathbf{D}} \mathbf{D}^T \mathbf{R}_{nn}^{-1} \mathbf{y} \end{aligned} \tag{3.6.29}$$

$$\tilde{\mathbf{n}} = \tilde{\boldsymbol{\theta}} = \mathbf{y} - \mathbf{D}\tilde{m} \tag{3.6.30}$$

($\mathbf{D}^T \mathbf{R}_{nn}^{-1} \mathbf{D}$ is a scalar). The expected uncertainty of this estimate is (3.6.22),

$$\mathbf{P} = \left\{ \frac{1}{m_0^2} + \mathbf{D}^T \mathbf{R}_{nn}^{-1} \mathbf{D} \right\}^{-1} = \frac{1}{1/m_0^2 + \mathbf{D}^T \mathbf{R}_{nn}^{-1} \mathbf{D}} \, . \tag{3.6.31}$$

As $m_0^2 \to \infty$, Equations (3.6.29)–(3.6.31) become the same expressions given by Bretherton, Davis, & Fandry (1976) for the mean of a field.

The estimates may appear somewhat unfamiliar; they reduce to more common expressions in certain limits. Let the θ_i be uncorrelated, with uniform variance σ^2; $\mathbf{R}_{nn}$ is then diagonal and (3.6.29) reduces to

$$\tilde{m} = \frac{1}{(1/m_0^2 + M/\sigma^2)\sigma^2} \sum_{i=1}^{M} y_i = \frac{m_0^2}{\sigma^2 + M m_0^2} \sum_{i=1}^{M} y_i \,, \tag{3.6.32}$$

where the relations $\mathbf{D}^T\mathbf{D} = M$, $\mathbf{D}^T\mathbf{y} = \sum_{i=1}^{M} y_i$ were used. The expected value of the estimate is

$$< \tilde{m} > = \frac{m_0^2}{\sigma^2 + M m_0^2} \sum_{i}^{M} < y_i > = \frac{m_0^2}{\sigma^2 + M m_0^2} M m \neq m \,, \tag{3.6.33}$$

that is, it is biased, as inferred above, unless $< y_i > = 0$, implying $m = 0$. $\mathbf{P}$ becomes

$$P = \frac{1}{1/m_0^2 + M/\sigma^2} = \frac{\sigma^2 m_0^2}{\sigma^2 + M m_0^2} \,. \tag{3.6.34}$$

Under the further assumption that $m_0^2 \to \infty$,

$$\tilde{m} = \frac{1}{M} \sum_{i=1}^{M} y_i \,, \tag{3.6.35}$$

$$P = \sigma^2 / M \,, \tag{3.6.36}$$

which are the ordinary average and its variance [the latter expression is the well-known square root of M rule for the standard deviation of an average–see Equation (3.5.27)]; $< \tilde{m} >$ in (3.6.35) is readily seen to be the true mean, but (3.6.29) is biased. However, the magnitude of (3.6.36) always exceeds that of (3.6.34)–acceptance of bias in the estimate (3.6.32) reduces the uncertainty of the result—a common trade-off in estimation problems.

Equations (3.6.29)–(3.6.31) are the general estimation rule–accounting through $\mathbf{R}_{nn}$ for correlations in the observations and their irregular distribution. Because many samples are not independent, (3.6.34) or (3.6.36) may be extremely optimistic. Equations (3.6.29)–(3.6.31) give the appropriate expression for the variance when the data are correlated (that is, when there are fewer degrees of freedom than the number of sample points). On the other hand, knowledge of the covariance structure of the noise can be exploited to reduce the uncertainty of the mean: Recall the reduced errors [Equation (3.4.118)] when the noise was known to be strongly positively correlated, thus permitting its reduction by subtraction.

The use of the prior estimate, m_0^2, is interesting. Letting m_0^2 go to infinity does not mean that an infinite mean is expected [(3.6.35) is finite]. This limit is merely a statement that there is no information whatever, before we start, as to the size of the true average–it could be arbitrarily large. Such a situation is, of course, unlikely, and even though we might choose not to use information concerning the probable size of the solution, we should remain aware that we could do so (the importance of the prior estimate diminishes as M grows–so that with an infinite amount of data it has no effect at all on the estimate).

It is very important not to be tempted into making a first estimate of m_0^2 by using (3.6.35), substituting into (3.6.32), thinking to reduce the error variance. For the Gauss-Markov theorem to be valid, the prior information must be truly independent of the data being used. If a prior estimate of m itself is available rather than just its mean square, the problem should be reformulated as one for the estimate of the perturbation about this value.

It is quite common in mapping and interpolation problems (taken up immediately below), to first estimate the mean of the field, to remove it, and then to map the residuals–here called θ. Such a procedure is a special case of a methodology often called (particularly in the geological literature) *kriging*,[10] which is discussed in Chapter 5.

3.6.4 Making a Map; Sampling, Interpolation, and Objective Mapping

A familiar oceanographic problem is to draw a set of contours from data that may have been observed irregularly in space. Such maps are usually the first step in understanding what one is measuring. A somewhat more sophisticated use of a map is to produce a regular grid of values to use in a numerical ocean model. An example with an irregular data distribution is shown in Figure 2–14. Ships observe the windfield wherever they happen to be, and the data are interpolated by investigators onto regular grids that are then used to drive model oceans.

Obtaining numbers on a regular grid from irregularly distributed observations is basically an interpolation or mapping problem–and as such may seem somewhat trivial. But it is far from trivial when one adds the requirement that the map should be accompanied by a useful estimate of the error of the values calculated at a grid point. For regions of a map surrounded by densely spaced data, the gridded value can be expected to be more accurate

[10] Pronounced with a soft "g."

than at a location effectively extrapolated from distant data points. (But what does one mean by "distant?") When using a map with a complex model, such inhomogeneities in the accuracy may be extremely important–as calculations could be in error because of large mapping errors in distant parts of the domain. Bretherton et al. (1976) introduced the subject of quantitative mapmaking into oceanography, and the book by Thièbaux and Pedder (1987) is devoted to the problem.

3.6.4.1 Sampling

The first question that must be addressed is whether the sampling of the field is adequate to make a useful map. This subject is a large and interesting one in its own right, and there are a number of useful references, including Bracewell (1978), Freeman (1965), Jerri (1977), or Butzer and Stens (1992), 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 (see Figure 3–12a) to produce the infinite set of sample values $\{f(n\Delta q)\}$, $-\infty \leq n \leq \infty$. The sampling theorem, or sometimes the Shannon-Whittaker Sampling Theorem[11] is a statement of necessary and sufficient conditions so that $f(q)$ can be perfectly reconstructable from the sample values. Let the Fourier transform of $f(q)$ be defined as

$$\hat{f}(r) = \int_{-\infty}^{\infty} f(q) e^{2i\pi rq} dq\,. \tag{3.6.37}$$

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

$$|\hat{f}(r)| = 0\,, \quad |r| \geq 1/(2\Delta q)\,. \tag{3.6.38}$$

The theorem produces an explicit formula for the reconstruction, the Shannon-Whittaker formula, which is

$$f(q) = \sum_{n=-\infty}^{\infty} f(n\Delta q) \frac{\sin[(2\pi/2\Delta q)(q - n\Delta q)]}{(2\pi/2\Delta q)(q - n\Delta q)}\,. \tag{3.6.39}$$

Mathematically, the Shannon-Whittaker theorem is surprising because it provides a condition under which a function at an uncountable infinity of points can be perfectly reconstructed from information only at a countable infinity of them. For present purposes, an intuitive interpretation is all

[11] In the Russian literature, Kotel'nikov's theorem.

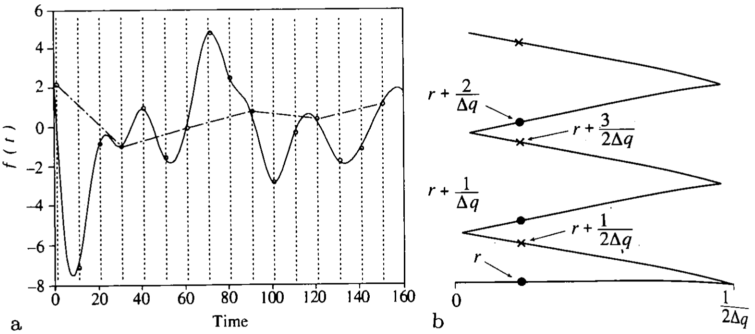

Figure 3–12. (a) Classical aliasing of a curve by undersampling. The original function is shown by the solid line. Samples every 10 time units (open circles) do an adequate job of capturing the variability in the function. But if the sampling interval is increased to 30 time units, the original function is grossly misrepresented (dashed line), and an attempt to calculate the derivative of the original curve from the coarse sampling interval would prove disastrously wrong. (b) Uniformly undersampled high frequencies or wavenumbers masquerade (alias) as lower frequencies or wavenumbers according to the Equation (3.6.40). The net result is a *folding* of the frequencies outside the baseband [Equation (3.6.41)] into apparent frequencies within the baseband. Here, a function $f(q)$ with Fourier transform $\hat{f}(r)$ is sampled at intervals Δq. The Fourier transform of the sampled function, $\hat{f}_s(r)$, sums contributions from the original Fourier transform as $\hat{f}_s(r) = \hat{f}(r) + \hat{f}(r \pm 1/\Delta q) + \hat{f}(r \pm 2/\Delta q) + \cdots$ (dotted points), which is potentially radically different from $\hat{f}(r)$. Points denoted "x" are aliased into a negative value $0 \geq r > -1/2\Delta q$–for example, $(r + 1/2\Delta q) - 1/\Delta q = r - 1/2\Delta q$.

we seek, and this is perhaps best done by considering a special case in which the conditions of the theorem are violated. Figure 3–12a displays an undersample curve. It is quite clear that there is at least one other curve, the one depicted with the broken line, which is completely consistent with all the sample points and which cannot be distinguished from it. If a pure sinusoid of frequency r_0 is sampled at intervals Δq, $\Delta q > 1/2r_0$, a little thought shows that the apparent frequency of this new sinusoid is

$$r_a = r_0 \pm \frac{n}{\Delta q} \tag{3.6.40}$$

such that

$$|r_a| \leq \frac{1}{2\Delta q}. \tag{3.6.41}$$

The samples cannot distinguish the true high-frequency sinusoid from a low frequency one, and the high frequency can be said to masquerade or *alias* as the lower-frequency one.[12] The Fourier transform of a sampled function is

[12] 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.

easily seen to be periodic with period $1/\Delta q$ in the transform domain–that is, in the r space (Bracewell, 1978, and Hamming, 1973, have particularly clear discussions). 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 Figure 3–12b). Frequencies of absolute value larger than $1/2\Delta q$, the so-called Nyquist frequency, cannot be distinguished from those in the baseband, and they alias into it.

The consequences of aliasing range from the negligible to the disastrous. A simple, possibly trivial, 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 (3.6.40), the apparent frequency of the tide will be .0676 cycles/day for a period of about 14.8 days. To the extent that the observer understands what is going on, he will not conclude that the principal lunar tide has a period of 14.8 days but will realize that he can compute the true period through (3.6.40) from the apparent one. But if he did not understand what was happening, he might produce some bizarre theory.[13]

Few situations are this simple. Consider Figure 2–2j, which shows a section of silicate across the North Atlantic Ocean. The observed variation includes all wavenumbers; an estimate of the wavenumber spectrum is displayed in Figure 3–13. (Spectra are discussed in many textbooks, e.g., Priestley, 1981. For present purposes it can be regarded as simply a numerical estimate of the wavenumber content of Figure 2–2i.) As one thins the sampling, the higher wavenumbers will be aliased into lower ones. Because the distribution is dominated by the low wavenumbers, the effects of this aliasing are perhaps not readily apparent to the eye. But for use in numerical models, one almost inevitably must differentiate observed fields one or more times. Consider what would happen if one tried to estimate the derivative of the curve in Figure 3–12a from the samples: The numerical values of the aliased data would be radically different from the correct values. It is this type of concern that leads one to single out the aliasing problem for special attention.

The reader may object that the Shannon-Whittaker theorem applies only to an infinite number of perfect samples and that one never has either per-

[13] There is a story, perhaps apocryphal, that one investigator was measuring the mass flux of the Gulf Stream at a fixed time each day. He was preparing to publish the exciting discovery that there was a strong 14-day periodicity to the flow, before someone pointed out to him that he was aliasing the tidal currents with a 12.42-hour period.

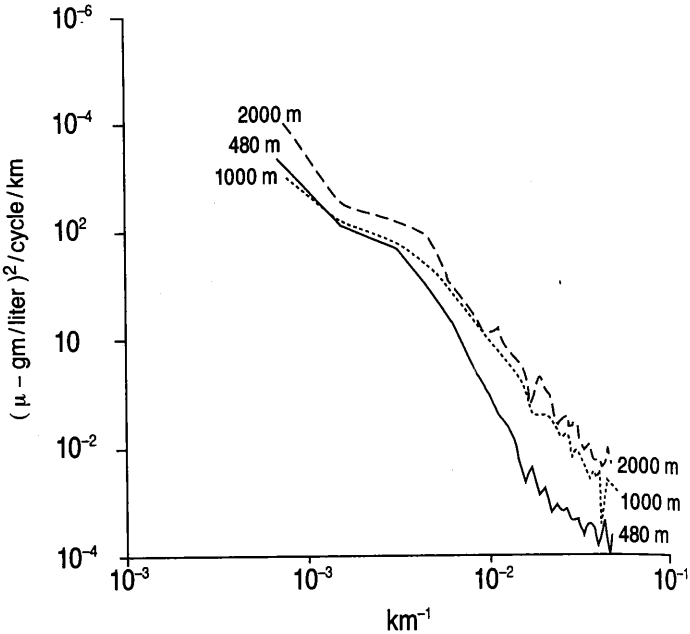

Figure 3–13. Estimated wavenumber spectrum of the silicate distribution along 25°N in the Atlantic, at three different depths. Such spectra are said to be *red*, because the energy increases as the wavelength increases (wavenumber decreases). It is the large-scale structure that is most visible in sections and maps. But if the field is differentiated, as it must be to use in conservation equations, the large-scale structures are suppressed, and the high wavenumbers–short scales–remain and are subject to aliases from inadequate sampling.

fect 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 (it follows from the so-called Paley-Wiener criterion and is usually stated in the form that "timelimited signals cannot be bandlimited"). Nonetheless, the rule of thumb that results from (3.6.39) 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

$$\Delta q \ll 1/2r_0 \,. \tag{3.6.42}$$

Many extensions and variations of the sampling theorem exist–taking account of the finite time duration (e.g., see Landau & Pollak, 1962), the use of burst-sampling and known-function derivatives, etc. (see Freeman, 1965; Jerri, 1977). Most of these variations are sensitive to noise. There are also extensions to multiple dimensions (e.g., Petersen & Middleton, 1962), which are required for mapmaking purposes. (An application, with discussion of the noise sensitivity, can be found in Wunsch, 1989.)

The subject will be left here for present purposes, with the comment that sampling theorems can be unforgiving–that is, once a function is undersampled, and unless the aliased signal is as simple as a known tidal contribution, it will be mappable, and differentiable, etc., but perhaps in a way that can be disastrously misleading. Consideration of sampling is critical to any discussion of field data.

3.6.5 One-Dimensional Interpolation

Supposing that the field has been adequately sampled, consider using only two observations $[y_1 \quad y_2]^T = [x_1 + n_1 \quad x_2 + n_2]^T$ located at positions $[r_1 \quad r_2]^T$ where n_i are the observation noise. We require an estimate of $x(\tilde{r})$, where $r_1 < \tilde{r} < r_2$. The formula (3.6.39) is unusable; there are only two noisy observations, not an infinite number of perfect ones. We could try instead using linear interpolation:

$$\tilde{x}(\tilde{r}) = \frac{|r_2 - \tilde{r}|}{|r_2 - r_1|} y(r_1) + \frac{|r_1 - \tilde{r}|}{|r_2 - r_1|} y(r_2) \,. \tag{3.6.43}$$

If there are data points, r_i, $1 \le i \le M$, then another possibility is Aitken-Lagrange interpolation (Davis & Polonsky, 1965):

$$\tilde{x}(\tilde{r}) = \sum_{j=1}^{M} l_j(\tilde{r}) y_j \,, \tag{3.6.44}$$

$$l_j(\tilde{r}) = \frac{(\tilde{r} - r_1) \cdots (\tilde{r} - r_{j-1})(\tilde{r} - r_{j+1}) \cdots (\tilde{r} - r_M)}{(r_j - r_1) \cdots (r_j - r_{j-1})(r_j - r_{j+1}) \cdots (r_j - r_M)} \,. \tag{3.6.45}$$

Figure 3–14 shows these two examples.

Equation (3.6.43), (3.6.44)–(3.6.45) 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 $< x > = < n > = 0$ or that a known value has been removed from both (this just keeps our notation a bit simpler). Then,

$$\begin{aligned} \mathbf{R}_{xy}(\tilde{r}, r_j) &\equiv < x(\tilde{r}) y(r_j) > = < x(\tilde{r})(x(r_j) + n(r_j)) > \\ &= \mathbf{R}_{xx}(\tilde{r}, r_j) \end{aligned} \tag{3.6.46}$$

$$\begin{aligned} \mathbf{R}_{yy}(r_i, r_j) &\equiv < (x(r_i) + n(r_i))(x(r_j) + n(r_j)) > \\ &= \mathbf{R}_{xx}(r_i, r_j) + \mathbf{R}_{nn}(r_i, r_j) \,, \end{aligned} \tag{3.6.47}$$

where it has been assumed that $< x(r) n(q) > = 0$ for all r, q.

From (3.6.9), the best linear interpolator is

$$\tilde{\mathbf{x}} = \mathbf{B}\mathbf{y}, \quad \mathbf{B}(\tilde{r}, \tilde{r}_i) = \sum_{j=1}^{M} \mathbf{R}_{xx}(\tilde{r}, r_j) \left\{ \mathbf{R}_{xx} + \mathbf{R}_{nn} \right\}_{ji}^{-1} \tag{3.6.48}$$

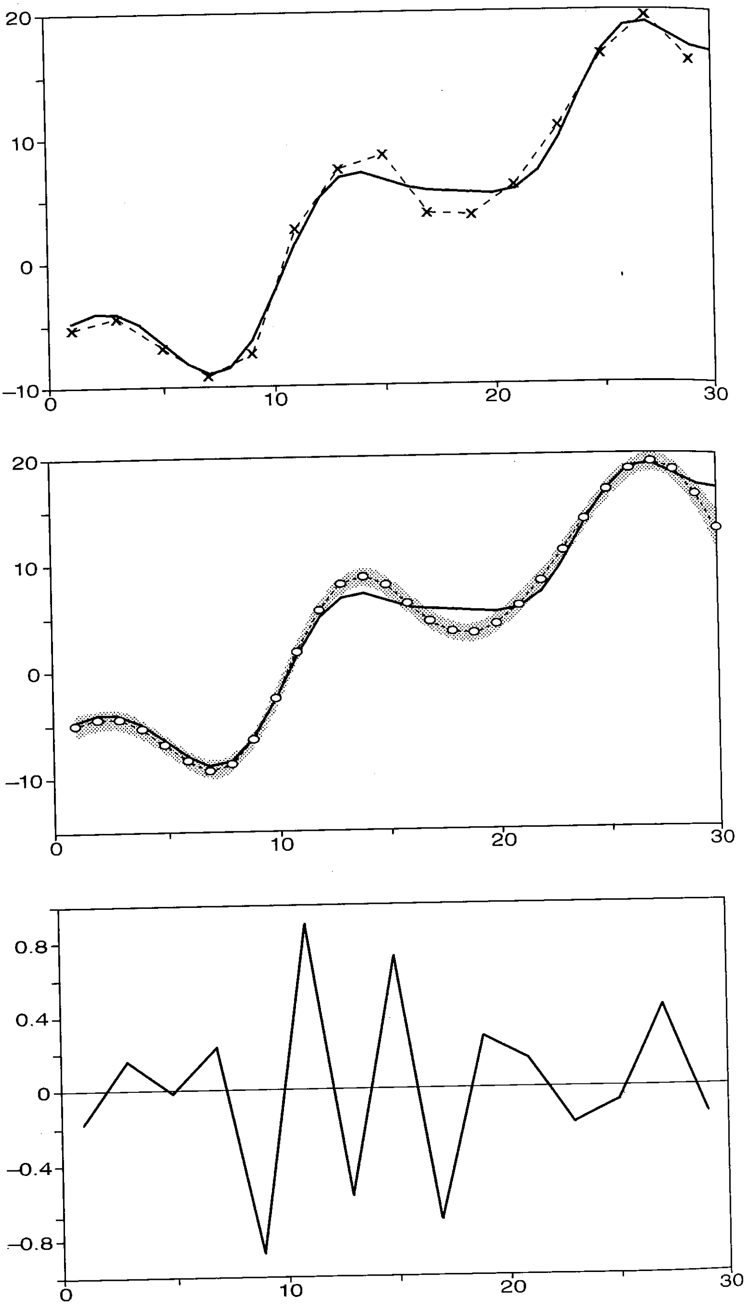

Figure 3–14. In the upper and middle panels, the solid curve represents the "true" values, generated as a function having covariance $S = 100\exp(-r^2/30)$. "Data" were then generated at every second point $(1, 3, \ldots)$ with pseudorandom white noise of variance 1. The upper panel shows linear interpolation of the data. Notice that the result interpolates literally, passing exactly through the point. The middle panel shows the result of using the Gauss-Markov estimate on the same pseudo-data. Estimated points do not agree exactly with the data, and an estimate of the expected one standard error is shown–perhaps the most important difference from the upper panel. Estimate and truth are generally consistent within two standard errors. The lowest panel displays the noise estimate at each data point from the Gauss-Markov estimate and appears, visually, suitably unstructured.

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

$$\mathbf{P}(\tilde{r}, \tilde{r}) = \mathbf{R}_{xx}(\tilde{r}, \tilde{r}) - \sum_{i}^{M} \sum_{j}^{M} \mathbf{R}_{xx}(\tilde{r}, r_j)\{\mathbf{R}_{xx}+\mathbf{R}_{nn}\}_{ji}^{-1}\mathbf{R}_{xx}(r_i, \tilde{r}) \quad (3.6.49)$$

[here $\mathbf{R}_{xx}$, $\mathbf{P}(\tilde{r}, \tilde{r})$ are both scalars], and $\tilde{\mathbf{n}} = \mathbf{y} - \tilde{\mathbf{x}}$.

Like the linear interpolation or the Aitken-Lagrange formula, or most other interpolation formulas, the optimal interpolator is simply a linear combination of the data. If any other set of weights $\mathbf{B}$ is chosen, then the interpolation is not as good in the mean-square error sense as it could be; the error of any such scheme can be obtained by substituting it into (3.6.7) and evaluating the result (the true covariances still need to be known.)

Looking back now at the two familiar formulas (3.6.43)–(3.6.45), 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 either of (3.6.43), (3.6.44)–(3.6.45) was 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 use uniformly spaced grid values 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 $\tilde{r}_\alpha = r_i$–that is, on the data points themselves. Such a result is characteristic of *true interpolation.* 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. The Gauss-Markov estimate is thus not a true interpolator; it is instead a *smoother* (smoothers will be encountered again in Chapter 6). One can recover true interpolation from the Gauss-Markov estimate if $||\mathbf{R}_{nn}|| \to 0$, but being conscious that the matrix being inverted in (3.6.48) and (3.6.49) can become singular. If no noise is present, then the observed value is the correct one to use at a data point. The weights $\mathbf{B}$ can be complicated if there is any structure at all in either of $\mathbf{R}_{xx}$, $\mathbf{R}_{nn}$. The estimator takes explicit account of the expected spatial structure of both $\mathbf{x}$ and $\mathbf{n}$ 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 $\|\mathbf{R}_{nn}\| \gg \|\mathbf{R}_{xx}\|$, $\tilde{\mathbf{x}} \to 0$, manifesting the bias in the estimator–a bias introduced in the Gauss-Markov estimators so as to minimize the uncertainty (minimum variance about the true value). Thus, interpolated values tend toward zero, particularly far from the data points. For this reason, it is common to use expressions such as (3.6.29), (3.6.30) to first remove the mean, prior to mapping the residual, adding the estimated mean back in afterward. The interpolated values of the residuals are unbiased, 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 (e.g., see Ripley, 1981, Section 5.2), but the corrections are commonly ignored.

The noise-free case would not normally be mapped with the Gauss-Markov estimator, and the presence of a realistic $\mathbf{R}_{nn}$ usually prevents singularity in $\mathbf{R}_{xx} + \mathbf{R}_{nn}$. Nonetheless, the general possibility of singularity should be examined and interpreted. This sum matrix is symmetric, and its SVD reduces to the symmetric form, (3.4.37),

$$\mathbf{R}_{xx} + \mathbf{R}_{nn} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T \,. \tag{3.6.50}$$

If the sum covariance is positive-definite, $\boldsymbol{\Lambda}$ will be square with $K = M$, and the inverse will exist. If the sum covariance is not positive-definite but is only semidefinite, one or more of its 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 if one is truly confident that $\mathbf{y}$ does not contain such structures. In that case, the solution

$$\tilde{\mathbf{x}} = \mathbf{R}_{xx}(\mathbf{R}_{xx} + \mathbf{R}_{nn})^{-1}\mathbf{y} = \mathbf{R}_{xx}(\mathbf{U}\boldsymbol{\Lambda}^{-1}\mathbf{U}^T)\mathbf{y} \tag{3.6.51}$$

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 $\mathbf{R}_{xx}+\mathbf{R}_{nn} \to \mathbf{R}_{xx}+\mathbf{R}_{nn}+\epsilon^2\mathbf{I}$, one is assured by the discussion of tapering that $\mathbf{R}_{xx} + \mathbf{R}_{nn}$ is no longer singular–all structures in the field being assigned to either the noise or the solution (or in part to both).

Anyone using a Gauss-Markov estimator to make maps must do checks that the result is consistent with the prior estimates of $\mathbf{R}_{xx}$, $\mathbf{R}_{nn}$. Such checks include determining whether the difference 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.

A variant mapping problem is the construction of a streamfunction, $\Psi(\tilde{\mathbf{r}}_i)$, on a uniform grid, $\tilde{\mathbf{r}}_i$, from noisy measurements of a velocity field $[u(\mathbf{r}_j)$, $v(\mathbf{r}_j)]$ at a collection of data points, $\mathbf{r}_j$. One has then a set of relations of the form

$$\begin{aligned}\Psi(\tilde{\mathbf{r}}_q) - \Psi(\tilde{\mathbf{r}}_{q'}) &= \Delta y u(\mathbf{r}_j) + n(\mathbf{r}_j)\\ \Psi(\tilde{\mathbf{r}}_s) - \Psi(\tilde{\mathbf{r}}_{s'}) &= -\Delta x v(\mathbf{r}_k) + n(\mathbf{r}_k)\end{aligned}$$

where $\tilde{\mathbf{r}}_q$, $\tilde{\mathbf{r}}_{q'}$ are the grid points bracketing observation point $\mathbf{r}_j$ in the y–direction, over a distance Δy, and $\tilde{\mathbf{r}}_s$, $\tilde{\mathbf{r}}_{s'}$ bracket point $\mathbf{r}_k$ in the x–direction over a distance Δx, and are just another version of the problem of estimating the solution to a set of simultaneous equations.

3.6.6 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 $\mathbf{n}$. The mean value of the field is first estimated and subtracted from the measurements, and we proceed as though the true mean were zero. This problem was discussed by Bretherton et al. (1976); in meteorology, the method is associated with Gandin (1965). Fundamentally, it is nothing more than an application of the Gauss-Markov theorem in two (most commonly) dimensions. Fuller discussions may be found in Thièbaux and Pedder (1987) and Daley (1991).

One proceeds exactly as in the case where the positions are scalars, minimizing the expected mean-square difference between the estimated and the true field $\mathbf{x}(\tilde{\mathbf{r}}_\alpha)$. The result is (3.6.48), (3.6.49) 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 $\mathbf{B}$ become matrices. The observations could also be vectors at each point.

An example of a two-dimensional map is shown in Figure 3–15a: The data points are the dots, while estimates of y on the uniform grid were wanted. The prior noise was described as $<\mathbf{n}> = 0$, $\mathbf{R}_{nn} = < n_i n_j >= \sigma_n^2 \delta_{ij}$, $\sigma_n^2 = 1$, and the true field covariance was $<\mathbf{x}> = 0$, $\mathbf{R}_{xx} = < \mathbf{x}(\mathbf{r}_i)\mathbf{x}(\mathbf{r}_j) >$ $= P_0 \exp -|\mathbf{r}_i - \mathbf{r}_j|^2/L_2$, $P_0 = 25$, $L_2 = 9$. Figure 3–15b shows the estimated values and 3–15c the error variance estimate of the mapped values. Far from the data points, the estimated values are 0–that is, the mapped field goes asymptotically to the estimated true mean, and the error variance goes to the full value of 25, which cannot be exceeded. That is to say, when mapping far from any data point, the only real information available is provided by the prior statistics–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, with the best estimate being the mean itself.

The mapped field has a complex error structure 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 for accounting for the spatial change in the expected errors of this forcing.

In practice, most published objective mapping (often called *OI* for *objective 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 on $|\mathbf{r}_i - \mathbf{r}_j|$ and not upon the two positions separately nor upon their orientation). The assumption is often qualitatively reasonable, but much of the ocean circulation is neither spatially stationary nor isotropic. 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 (see the discussion in Fukumori, Martel, & Wunsch, 1991).

3.6.7 Linear Combinations of Estimates and Mapping Derivatives

A common problem in setting up a general circulation model is to specify the fields of quantities like temperature, salinity, etc., on a regular model grid. One also often must specify the derivatives of these fields for use in equations like that of the advection-diffusion equation,

$$\frac{\partial C}{\partial t} + \mathbf{v} \cdot \nabla C = K \nabla^2 C \tag{3.6.52}$$

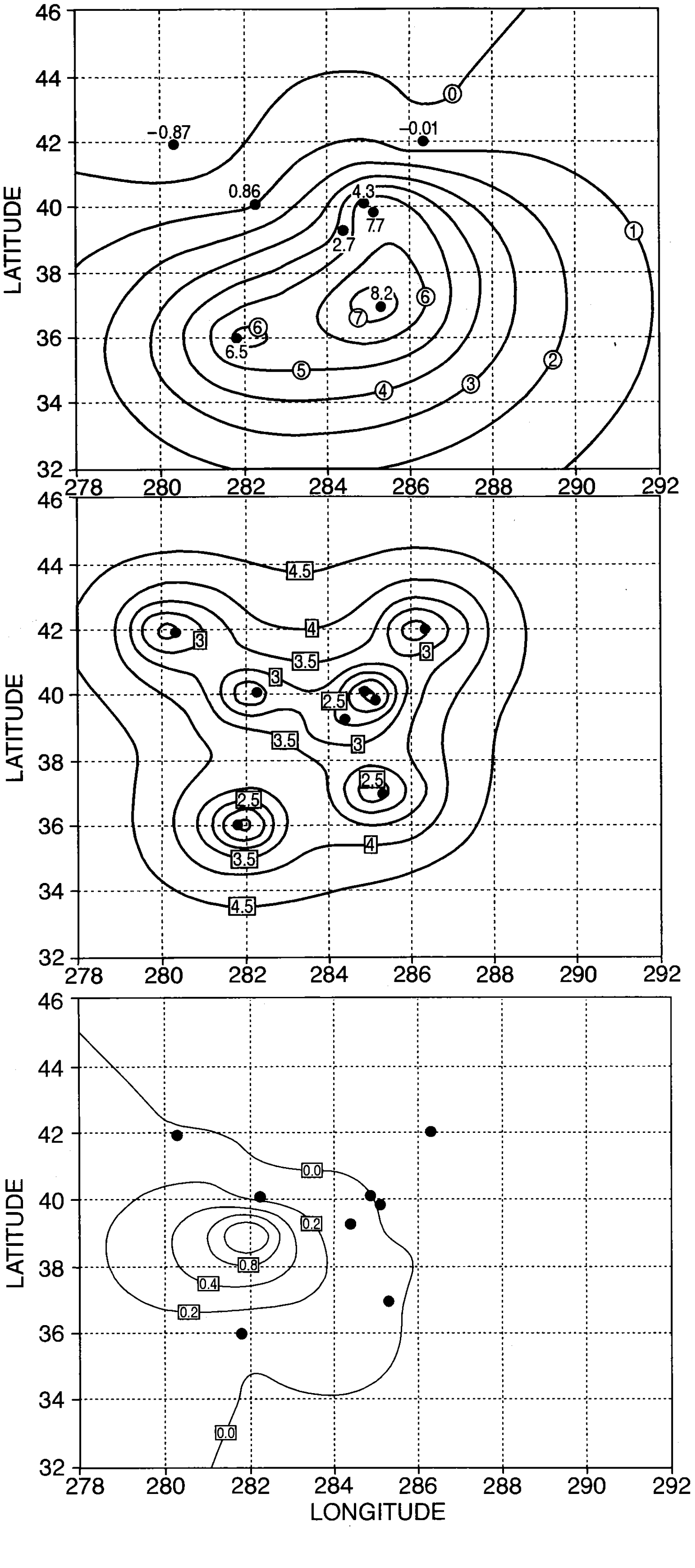

Figure 3–15. (a) Data points are assumed to be available at the locations marked with a solid dot and the values shown. The estimated field values on the regular grid (integer values of latitude and longitude) emulate what happens when data are interpolated for purposes of driving a model. The solution covariance was $R(|\mathbf{r}_i - \mathbf{r}_j|) = 25\exp(-|\mathbf{r}_i - \mathbf{r}_j|/9)$. The observation noise was assumed to be white of variance unity. Far from the data, the mapping tends toward the expected value (here, zero) because no other information about the correct value is available. (b) Standard error $\sqrt{P_{ii}}$ for the mapped field shown in (a). The values tend to 5–that is, $\sqrt{25}$–far from the data points, as the largest possible square error can never exceed the prior estimate of 25. In general, the expected errors are smallest near the data points. (c) One of the rows of $\mathbf{P}$, corresponding to the grid point on which the contours are centered (39°N, 282°E), displaying the correlations that occur in the expected errors of the mapped field at neighboring grid points. The variance was normalized to 1 for plotting.

where C is any scalar field of interest. Suppose one wished to estimate a derivative as a one-sided difference,

$$\frac{\partial C(\tilde{r}_1)}{\partial r} \sim \frac{C(\tilde{r}_1) - C(\tilde{r}_2)}{\tilde{r}_1 - \tilde{r}_2}. \tag{3.6.53}$$

Then one might think simply to subtract the two estimates made from Equation (3.6.48), producing

$$\Delta r \frac{\partial C(\tilde{r}_1)}{\partial r} \sim (\mathbf{R}_{xx}(\tilde{r}_1, r_j) - \mathbf{R}_{xx}(\tilde{r}_2, r_j))(\mathbf{R}_{xx} + \mathbf{R}_{nn})^{-1}_{jk}\mathbf{y}(r_k) \tag{3.6.54}$$

(a sum on j and k is implied).

Alternatively, suppose we tried to estimate $\partial C/\partial r$ directly from (3.6.5), 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:

$$\begin{aligned}\mathbf{R}_{xy} &= < (C(\tilde{r}_1) - C(\tilde{r}_2))(C(r_j) + n(r_j)) > \\ &= \mathbf{R}_{xx}(\tilde{r}_1, r_j) - \mathbf{R}_{xx}(\tilde{r}_2, r_j),\end{aligned} \tag{3.6.55}$$

which when substituted into (3.6.9) produces (3.6.54). *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 (see Luenberger, 1969). Of particular importance is the estimate of an arbitrary linear combination of elements of $\tilde{\mathbf{x}}$, such as the finite difference derivative just considered, and the essential computation of their uncertainty. Consider any estimate $\tilde{\mathbf{x}}$, and a weighted sum

$$\tilde{H} = \mathbf{a}^T\tilde{\mathbf{x}} \tag{3.6.56}$$

where the constant vector $\mathbf{a}$ may be mostly zeros. The expected value of the sum is

$$< \tilde{H} > = \mathbf{a}^T < \tilde{\mathbf{x}} >, \tag{3.6.57}$$

whose bias depends directly on that of $\tilde{\mathbf{x}}$. If the uncertainty of $\tilde{\mathbf{x}}$ is $\mathbf{P}$, then one has immediately

$$< (\tilde{H} - H)^2 > = \mathbf{a}^T < (\tilde{\mathbf{x}} - \mathbf{x})(\tilde{\mathbf{x}} - \mathbf{x})^T > \mathbf{a} = \mathbf{a}^T\mathbf{P}\mathbf{a}. \tag{3.6.58}$$

3.7 Improving Solutions Recursively

An important idea in both least-squares approximation and estimation theory derives from the need to improve the result of an earlier computation with the arrival of some new data. In what follows, we initially will use the language of least squares, but because of the coincidence of the results for least squares with appropriate weight matrices, and minimum variance estimation, we will obtain the correct result for estimation problems, too.

Suppose we have solved the system (3.3.2), using any one of the procedures discussed above. Because we will add data, some extra notation is needed. Rewrite (3.3.2) as

$$\mathbf{E}(1)\mathbf{x}(1) + \mathbf{n}(1) = \mathbf{y}(1) \tag{3.7.1}$$

where the noise $\mathbf{n}(1)$ has zero mean and covariance matrix $\mathbf{R}_{nn}(1)$. Let the estimate of the solution to (3.7.1) from one of the estimators be written as $\tilde{\mathbf{x}}(1)$, with uncertainty $\mathbf{P}(1)$. As a specific example, suppose (3.7.1) is full-rank overdetermined and was solved using row-weighted least-squares solution as

$$\tilde{\mathbf{x}}(1) = \left(\mathbf{E}(1)^T\mathbf{R}_{nn}(1)^{-1}\mathbf{E}(1)\right)^{-1}\mathbf{E}(1)^T\mathbf{R}_{nn}(1)^{-1}\mathbf{y}(1) \tag{3.7.2}$$

with corresponding $\mathbf{P}(1)$ (no column weights are used because we know they are irrelevant for a full-rank overdetermined problem).

Some new observations, $\mathbf{y}(2)$, are obtained, with the error covariance of the new observations given by $\mathbf{R}_{nn}(2)$ so that the problem is now

$$\left\{\begin{matrix}\mathbf{E}(1)\\ \mathbf{E}(2)\end{matrix}\right\}\mathbf{x} + \begin{bmatrix}\mathbf{n}(1)\\ \mathbf{n}(2)\end{bmatrix} = \begin{bmatrix}\mathbf{y}(1)\\ \mathbf{y}(2)\end{bmatrix} \tag{3.7.3}$$

where $\mathbf{x}$ is the same unknown. We assume $< \mathbf{n}(2) > = \mathbf{0}$ and

$$< \mathbf{n}(1)\mathbf{n}(2)^T > = \mathbf{0}\,, \tag{3.7.4}$$

that is, no correlation of the old and new measurement error (this assumption is very important, and particular attention is called to it). A solution to (3.7.3) should give a better estimate of $\mathbf{x}$ than (3.7.1) alone because more observations are available. It is sensible to row weight the concatenated set to

$$\left\{\begin{matrix}\mathbf{R}_{nn}(1)^{-T/2}\mathbf{E}(1)\\ \mathbf{R}_{nn}(2)^{-T/2}\mathbf{E}(2)\end{matrix}\right\}\mathbf{x} + \begin{bmatrix}\mathbf{R}_{nn}(1)^{-T/2}\mathbf{n}(1)\\ \mathbf{R}_{nn}(2)^{-T/2}\mathbf{n}(2)\end{bmatrix} = \begin{bmatrix}\mathbf{R}_{nn}(1)^{-T/2}\mathbf{y}(1)\\ \mathbf{R}_{nn}(2)^{-T/2}\mathbf{y}(2)\end{bmatrix}\,. \tag{3.7.5}$$

Recursive weighted least squares seeks the solution to (3.7.5) without inverting the new, larger matrix by taking advantage of the existing knowledge of $\mathbf{x}$ already in hand from (3.7.2). Because of (3.7.4), the objective function corresponding to finding the minimum weighted error norm is

$$\begin{aligned} J = {} & [\mathbf{y}(1) - \mathbf{E}(1)\mathbf{x}]^T\, \mathbf{R}_{nn}(1)^{-1}\, [\mathbf{y}(1) - \mathbf{E}(1)\mathbf{x}] \\ & + [\mathbf{y}(2) - \mathbf{E}(2)\mathbf{x}]\, \mathbf{R}_{nn}(2)^{-1}\, [\mathbf{y}(2) - \mathbf{E}(2)\mathbf{x}]\,. \end{aligned} \tag{3.7.6}$$

Taking the derivatives with respect to $\mathbf{x}$, the least-squares solution is

$$\begin{aligned}\tilde{\mathbf{x}}(2) = &\left\{\mathbf{E}(1)^T\mathbf{R}_{nn}(1)^{-1}\mathbf{E}(1) + \mathbf{E}(2)^T\mathbf{R}_{nn}(2)^{-1}\mathbf{E}(2)\right\}^{-1} \times \\ &\left\{\mathbf{E}(1)^T\mathbf{R}_{nn}(1)^{-1}\mathbf{y}(1) + \mathbf{E}(2)^T\mathbf{R}_{nn}(2)^{-1}\mathbf{y}(2)\right\}. \qquad (3.7.7)\end{aligned}$$

But one can manipulate (3.7.7) into (e.g., see Brogan, 1985, or Stengel, 1986)

$$\begin{aligned}\tilde{\mathbf{x}}(2) &= \tilde{\mathbf{x}}(1) + \mathbf{P}(1)\mathbf{E}(2)^T\left[\mathbf{E}(2)\mathbf{P}(1)\mathbf{E}(2)^T + \mathbf{R}_{nn}(2)\right]^{-1}\left[\mathbf{y}(2) - \mathbf{E}(2)\tilde{\mathbf{x}}(1)\right] \\ &= \tilde{\mathbf{x}}(1) + \mathbf{K}(2)\left[\mathbf{y}(2) - \mathbf{E}(2)\tilde{\mathbf{x}}(1)\right] \qquad (3.7.8)\end{aligned}$$

where

$$\mathbf{K}(2) = \mathbf{P}(1)\mathbf{E}(2)^T\left[\mathbf{E}(2)\mathbf{P}(1)\mathbf{E}(2)^T + \mathbf{R}_{nn}(2)\right]^{-1} \qquad (3.7.9)$$

$$\tilde{\mathbf{n}}(2) = \mathbf{y}(2) - \mathbf{E}(2)\tilde{\mathbf{x}}(2), \qquad (3.7.10)$$

and this improved estimate has uncertainty

$$\mathbf{P}(2) = \mathbf{P}(1) - \mathbf{K}(2)\mathbf{E}(2)\mathbf{P}(1). \qquad (3.7.11)$$

These last equations are algebraically identical to the alternate forms found from the matrix inversion lemma (3.1.25):

$$\begin{aligned}\tilde{\mathbf{x}}(2) = &\left(\mathbf{E}(2)^T\mathbf{R}_{nn}(2)^{-1}\mathbf{E}(2)\right)^{-1}\left\{\mathbf{P}(1) + \left[\mathbf{E}(2)^T\mathbf{R}_{nn}(2)^{-1}\mathbf{E}(2)\right]^{-1}\right\}^{-1}\tilde{\mathbf{x}}(1) \\ &+ \mathbf{P}(1)\left\{\mathbf{P}(1) + \left[\mathbf{E}(2)^T\mathbf{R}_{nn}(2)^{-1}\mathbf{E}(2)\right]^{-1}\right\}^{-1}\left[\mathbf{E}(2)^T\mathbf{R}_{nn}(2)^{-1}\mathbf{E}(2)\right]^{-1} \times \\ &\mathbf{E}(2)\mathbf{R}_{nn}(2)^{-1}\mathbf{y}(2), \qquad (3.7.12)\end{aligned}$$

$$\mathbf{P}(2) = \left\{\mathbf{P}(1)^{-1} + \mathbf{E}(2)^T\mathbf{R}_{nn}(2)^{-1}\mathbf{E}(2)\right\}^{-1}. \qquad (3.7.13)$$

The two different boxed sets differ only in the matrix sizes to be inverted, and a choice between them is typically based upon computational loads (in some large problems, matrix inversion may prove less onerous than matrix multiplication).

The solution is just the least-squares solution to the full set but rearranged after a bit of algebra. The original data, $\mathbf{y}(1)$, and coefficient matrix, $\mathbf{E}(1)$, have disappeared, to be replaced by the first solution $\tilde{\mathbf{x}}(1)$ and its uncertainty $\mathbf{P}(1)$. That is to say, one need not retain the original data and $\mathbf{E}(1)$ for the new solution to be computed. Furthermore, because the new solution depends only upon $\tilde{\mathbf{x}}(1)$, $\mathbf{P}(1)$, the particular methodology originally employed for obtaining them is irrelevant (i.e., they might have actually been obtained from an educated guess or through some arbitrarily complex model computation). Finally, the structure of the improved solution (3.7.8) is interesting and suggestive. It is made up of two terms: the previous estimate plus a term proportional to the difference between the new observations $\mathbf{y}(2)$, and a prediction of what those observations should have been were the first estimate the wholly correct one and the new observations perfect. It thus has the form of a *predictor-corrector.*

The difference between the prediction and the forecast can be called the *prediction error*, but recall that there is observational noise in $\mathbf{y}(2)$. The new estimate is a weighted average of this difference and the prior estimate, with the weighting depending upon the details of the uncertainty of prior estimate and new data. The behavior of the updated estimate is worth understanding in various limits. For example, suppose the initial uncertainty estimate is diagonal, $\mathbf{P}(1) = \Delta^2\mathbf{I}$, or that one rotates $\mathbf{x}$ into a new space of uncorrelated uncertainty. Then

$$\mathbf{K}(2) = \mathbf{E}(2)^T(\mathbf{E}(2)\mathbf{E}(2)^T + \mathbf{R}_{nn}(2)/\Delta^2)^{-1}\,. \tag{3.7.14}$$

If the norm of $\mathbf{R}_{nn}(2)$ is small compared to that of $\Delta^2\mathbf{I}$ and if (to be specific only) the second set of observations is by itself full-rank underdetermined, then

$$\mathbf{K}(2) \to \mathbf{E}(2)^T(\mathbf{E}(2)\mathbf{E}(2)^T)^{-1}$$

and

$$\begin{aligned}\tilde{\mathbf{x}}(2) &= \tilde{\mathbf{x}}(1) + \mathbf{E}(2)^T(\mathbf{E}(2)\mathbf{E}(2)^T)^{-1}[\mathbf{y}(2) - \mathbf{E}(2)\tilde{\mathbf{x}}(1)]\\ &= \left[\mathbf{I} - \mathbf{E}(2)^T(\mathbf{E}(2)\mathbf{E}(2)^T)^{-1}\mathbf{E}(2)\right]\tilde{\mathbf{x}}(1) + \mathbf{E}(2)^T(\mathbf{E}(2)\mathbf{E}(2)^T)^{-1}\mathbf{y}(2)\end{aligned} \tag{3.7.15}$$

where $[\mathbf{I} - \mathbf{E}(2)^T(\mathbf{E}(2)\mathbf{E}(2)^T)\mathbf{E}(2)^{-1}]$ will be recognized as the nullspace projector (3.4.114) of $\mathbf{E}(2)$. The update is replacing the first estimate by the estimate from the second set of observations, which were deemed perfect, but keeping unchanged any components of $\tilde{\mathbf{x}}(1)$ in the nullspace of $\mathbf{E}(2)$ because no new information is available about them. Should the new observations be fully determined and perfect, then the previous estimate is wholly replaced by the estimate made from the new, low-noise observations.

At the opposite extreme, when the new observations are very noisy compared to the previous ones, $\mathbf{K}(2)$ will be comparatively small, and the previous estimate is left largely unchanged. The general case represents a weighted average of the previous and new data, the weighting depending both upon the relative noise in each and upon the structure of the observations relative to the structure of $\mathbf{x}$.

The matrix being inverted in (3.7.8)–(3.7.11) is the sum of the measurement error covariance $\mathbf{R}_{nn}(2)$, and the error covariance of the "forecast" $\mathbf{E}(2)\tilde{\mathbf{x}}(1)$. To see this, let $\gamma(1)$ be the error component in $\tilde{\mathbf{x}}(1)$, which by definition has covariance $< \gamma(1)\gamma(1)^T > = \mathbf{P}(1)$. Then the expected covariance of the error of prediction is $< \mathbf{E}(2)\gamma(1)\gamma(1)^T\mathbf{E}(2)^T > = \mathbf{E}(2)\mathbf{P}(1)\mathbf{E}(2)^T$, which appears in $\mathbf{K}(2)$. Because of the assumptions (3.7.4) and $< \gamma(1)x(1)^T > = \mathbf{0}$, it follows that

$$< \mathbf{y}(1)[\mathbf{y}(2) - \mathbf{E}(2)\tilde{\mathbf{x}}(1)] > = \mathbf{0}\,. \tag{3.7.16}$$

That is, the *innovation*, $\mathbf{y}(2) - \mathbf{E}(2)\tilde{\mathbf{x}}(1)$, is uncorrelated with the previous measurement.

It is useful to notice that Equations (3.5.11)–(3.5.12), the solution to the least-squares problem subject to certain perfect constraints imposed by a Lagrange multiplier, can be recovered from (3.7.8)–(3.7.13) by putting $\mathbf{E}(2) = \mathbf{A}$, $\mathbf{y}(2) = \mathbf{q}$, $\mathbf{P}(1) = (\mathbf{E}^T\mathbf{E})^{-1}$, $\mathbf{R}_{nn}(2) \to 0$. That is, this earlier solution can be conceived of as having been obtained by first solving the conventional least-squares problem and then being modified by the later information that $\mathbf{A}\mathbf{x} = \mathbf{q}$ with very high accuracy.

Finally, suppose given $\tilde{\mathbf{x}}(1)$, $\tilde{\mathbf{y}}(1)$ that we regard $\Delta\mathbf{y} = \mathbf{y}(2) - \mathbf{E}(2)\tilde{\mathbf{x}}(1)$ as the discrepancy between an initial estimate of $\mathbf{x}$ and what the new data suggest is correct. Putting $\Delta\mathbf{x} = \tilde{\mathbf{x}}(2)-\tilde{\mathbf{x}}(1)$, we have an ordinary estimation problem with $< \Delta\mathbf{x}\,\Delta\mathbf{x}^T > = \mathbf{P}(1)$,

$$\mathbf{E}(2)\Delta\mathbf{x} + \mathbf{n}(2) = \Delta\mathbf{y}\,.$$

The solution by the Gauss-Markov estimate (3.6.16)–(3.6.18) (or least squares) is

$$\Delta\tilde{\mathbf{x}} = \mathbf{P}(1)\mathbf{E}(2)^T\left(\mathbf{E}(2)\mathbf{P}(1)\mathbf{E}(2)^T + \mathbf{R}_{nn}(2)\right)^{-1}\Delta\mathbf{y}\,,$$

which if added to $\tilde{\mathbf{x}}(1)$ produces (3.7.8).

The possibility of a recursion based on either of (3.7.8)–(3.7.11) or (3.7.12), (3.7.13) should now be obvious–all argument–1 variables being replaced by argument–2 variables, which in turn are replaced by argument–3 variables, etc. A practical example, as applied to altimetric data, may be seen in Wunsch (1991).

The computational load of the recursive solution needs to be addressed. If all of the constraints are available at once and used, the solution can be found as in any least-squares problem without ever computing the solution uncertainty (although its utility without the uncertainty may be doubted).

But if the constraints are divided and used in two or more groups, then the uncertainty must be computed one or more times to carry out the improvement. In general, owing to the need to compute the uncertainties, it is more efficient to use all of the constraints at once (if available, and if the computer can handle them) than it is to divide them into groups–unless special structures are present in the $\mathbf{E}(t)$. Oceanography has a particular need for recursive methods, however. The global-scale data flow is not well organized, and data tend to drift in to scientists over lengthy periods of time. It is a considerable advantage to be able to improve estimates when previously unavailable data finally appear.

The comparatively simple interpretation of the recursive, weighted least-squares problem will be used in Chapter 6 to derive the Kalman filter and suboptimal filters in a very simple form. It also becomes the key to understanding *assimilation* schemes such as *nudging*, *forcing to climatology*, and *robust diagnostic* methods.

If the initial set of equations (3.7.1) is actually underdetermined and should it have been solved using the SVD, one must be careful that $\mathbf{P}(1)$ includes the estimated error owing to the missing nullspace. Otherwise, these elements would be assigned zero error variance, and the new data could never affect them.

Consider another special case. Let there be a prior best estimate of the solution, which we will call $\tilde{\mathbf{x}}(0)$, and which is assumed to be zero. Specify an initial uncertainty (most often diagonal),

$$\mathbf{P}(0) = < \tilde{\mathbf{x}}(0)\tilde{\mathbf{x}}(0)^T > ,$$

and assuming a true mean of zero, treat this estimate and its uncertainty as the first set of data, replacing all the "1" estimates with those now relabeled "0":

$$\tilde{\mathbf{x}}(1) = \mathbf{0} + \mathbf{K}(1)\left[\mathbf{y}(1) - \mathbf{E}(1)\mathbf{0}\right] , \tag{3.7.17}$$

$$\mathbf{K}(1) = \mathbf{P}(0)\mathbf{E}(1)^T\left[\mathbf{E}(1)\mathbf{P}(0)\mathbf{E}(1)^T + \mathbf{R}_{nn}(1)\right]^{-1} ,$$

$$\mathbf{P}(1) = \mathbf{P}(0) - \mathbf{K}(1)\mathbf{E}(1)\mathbf{P}(0) . \tag{3.7.18}$$

The objective mapping estimate discussed in (3.6.16)–(3.6.18) is identical with (3.7.17)–(3.7.18). There, the prior estimate of the field at an interpolation point is 0; the prior uncertainty of the field corresponds to its estimated second moments, $\mathbf{R}_{xx}$, and the observation noise covariance is $\mathbf{R}_{nn}$. $\mathbf{K}$ interpolates from the data points $\mathbf{r}_i$ to the grid points $\tilde{\mathbf{r}}_\alpha$. Thus, as asserted, the Gauss-Markov mapping estimate coincides with the least-squares one, and we can regard objective mapping as a special case of recursive least squares.

Let us confirm more generally that the recursive least-squares result is identical to a recursive estimation procedure. Suppose there exist two estimates of an unknown vector $\mathbf{x}$, denoted $\tilde{\mathbf{x}}_a$, $\tilde{\mathbf{x}}_b$ with estimated uncertainties $\mathbf{P}_a$, $\mathbf{P}_b$, respectively. They are either unbiased or have the same bias–that is, $<\tilde{\mathbf{x}}_a> = <\tilde{\mathbf{x}}_b> = \mathbf{x}_B$. How should the two be combined to give a third estimate, $\tilde{\mathbf{x}}^+$, with minimum error? Try a linear combination

$$\tilde{\mathbf{x}}^+ = \mathbf{L}_a\tilde{\mathbf{x}}_a + \mathbf{L}_b\tilde{\mathbf{x}}_b\,. \tag{3.7.19}$$

If the new estimate is to be unbiased or is to retain the prior bias, it follows that

$$<\tilde{\mathbf{x}}^+> = \mathbf{L}_a <\tilde{\mathbf{x}}_a> + \mathbf{L}_b <\tilde{\mathbf{x}}_b> \tag{3.7.20}$$

or

$$\begin{aligned}\mathbf{x}_B &= \mathbf{L}_a\mathbf{x}_B + \mathbf{L}_b\mathbf{x}_B\\ \mathbf{L}_b &= \mathbf{I} - \mathbf{L}_a\,.\end{aligned} \tag{3.7.21}$$

Then the uncertainty is

$$\begin{aligned}<(\tilde{\mathbf{x}}^+ - \mathbf{x})(\tilde{\mathbf{x}}^+ - \mathbf{x})^T> &= <(\mathbf{L}_a\tilde{\mathbf{x}}_a + (\mathbf{I}-\mathbf{L}_a)\tilde{\mathbf{x}}_b - \mathbf{x})(\mathbf{L}_a\tilde{\mathbf{x}}_a + (\mathbf{I}-\mathbf{L}_a)\tilde{\mathbf{x}}_b - \mathbf{x})^T>\\ &= \mathbf{L}_a\mathbf{P}_a\mathbf{L}_a^T + (\mathbf{I}-\mathbf{L}_a)\mathbf{P}_b(\mathbf{I}-\mathbf{L}_a)^T\end{aligned} \tag{3.7.22}$$

where the assumption that the errors in $\mathbf{x}_a$, $\mathbf{x}_b$ are uncorrelated has been used. This expression is positive-definite; minimizing with respect to $\mathbf{L}_a$ yields immediately

$$\mathbf{L}_a = \mathbf{P}_b(\mathbf{P}_a + \mathbf{P}_b)^{-1}, \quad \mathbf{L}_b = \mathbf{P}_a(\mathbf{P}_a + \mathbf{P}_b)^{-1}\,.$$

The new combined estimate is then

$$\tilde{\mathbf{x}}^+ = \mathbf{P}_b(\mathbf{P}_a + \mathbf{P}_b)^{-1}\tilde{\mathbf{x}}_a + \mathbf{P}_a(\mathbf{P}_a + \mathbf{P}_b)^{-1}\tilde{\mathbf{x}}_b\,.$$

This last expression can be rewritten by adding and subtracting $\tilde{\mathbf{x}}_a$ as

$$\begin{aligned}\tilde{\mathbf{x}}^+ &= \tilde{\mathbf{x}}_a + \mathbf{P}_b(\mathbf{P}_a+\mathbf{P}_b)^{-1}\tilde{\mathbf{x}}_a + \mathbf{P}_a(\mathbf{P}_a+\mathbf{P}_b)^{-1}\tilde{\mathbf{x}}_b - (\mathbf{P}_a+\mathbf{P}_b)(\mathbf{P}_a+\mathbf{P}_b)^{-1}\tilde{\mathbf{x}}_a\\ &= \tilde{\mathbf{x}}_a + \mathbf{P}_a(\mathbf{P}_a + \mathbf{P}_b)^{-1}(\tilde{\mathbf{x}}_b - \tilde{\mathbf{x}}_a)\,.\end{aligned} \tag{3.7.23}$$

The uncertainty of the estimate (3.7.23) is easily evaluated as

$$\mathbf{P}^+ = (\mathbf{P}_a^{-1} + \mathbf{P}_b^{-1})^{-1}\,, \tag{3.7.24}$$

which, by straightforward application of the matrix inversion lemma (3.1.24), is

$$\mathbf{P}^{+}=\mathbf{P}_a-\mathbf{P}_a(\mathbf{P}_a+\mathbf{P}_b)^{-1}\mathbf{P}_a\,. \tag{3.7.25}$$

Equations (3.7.23)–(3.7.25) are the general rules for combining two estimates with uncorrelated errors.

Now suppose that $\tilde{\mathbf{x}}_a$ and its uncertainty are known and that there are measurements

$$\mathbf{E}(2)\mathbf{x}+\mathbf{n}(2)=\mathbf{y}(2) \tag{3.7.26}$$

with $<\mathbf{n}(2)>=0$, $<\mathbf{n}(2)\mathbf{n}(2)^T>=\mathbf{R}_{nn}(2)$. From this second set of observations, we *estimate* the solution, using the Gauss-Markov estimator (3.6.20)–(3.6.22) with no prior estimate of the solution variance–that is, $\|\mathbf{R}_{xx}^{-1}\|\rightarrow 0$–so that

$$\tilde{\mathbf{x}}_b=(\mathbf{E}(2)^T\mathbf{R}_{nn}(2)^{-1}\mathbf{E}(2))^{-1}\mathbf{E}(2)^T\mathbf{R}_{nn}(2)^{-1}\mathbf{y}(2) \tag{3.7.27}$$

$$\mathbf{P}=(\mathbf{E}(2)^T\mathbf{R}_{nn}(2)^{-1}\mathbf{E}(2))^{-1}\,. \tag{3.7.28}$$

Subsituting (3.7.27), (3.7.28) into (3.7.23)–(3.7.25) and again using the matrix inversion lemma gives

$$\tilde{\mathbf{x}}^{+}=\tilde{\mathbf{x}}_a+\mathbf{P}_a\mathbf{E}(2)^T[\mathbf{E}(2)\mathbf{P}_a\mathbf{E}(2)^T+\mathbf{R}_{nn}(2)]^{-1}[\mathbf{y}(2)-\mathbf{E}(2)\tilde{\mathbf{x}}_a]\,, \tag{3.7.29}$$

which is the same as (3.7.8); thus a recursive minimum variance estimate coincides with a corresponding weighted least-squares recursion. The covariance may also be confirmed to be (3.7.11). The alternate forms (3.7.12), (3.7.13) are also correct.

3.8 Estimation from Linear Constraints–A Summary

A number of different procedures for producing estimates of the solution to a set of noisy simultaneous equations of arbitrary dimension have been described here. The reader may wonder which of the variants makes the most sense to use in practice. There is no single best answer because in the presence of noise one is dealing with a statistical estimation problem, and one must be guided by model context and goals. A few general remarks might be helpful.

In any problem where data are to be used to make inferences about physical parameters, one typically needs some approximate idea of just how large the solution is likely to be and how large the residuals probably are. In this nearly agnostic case, where almost nothing else is known, and the problem is very large, the weighted, tapered least-squares solution (Section 3.3.2) is a good first choice—it is easily and efficiently computed and coincides with the Gauss-Markov and tapered SVD solutions for this situation if the weight matrices are the appropriate variances. Sparse matrix methods for its solution exist (e.g., Paige & Saunders, 1982) should that be necessary. Coincidence with the Gauss-Markov solution means that one can reinterpret it as a minimum-variance solution should one wish (and for Gaussian variables, it is also the maximum likelihood solution).

It is a comparatively easy matter to vary the trade-off parameter, α^2, to explore the consequences of any errors in specifying the noise and solution variances. Once a value for α^2 is known, the tapered SVD can be computed to understand the relationships between solution and data structures, their resolution, and their variance. For problems of small to moderate size (the meaning of *moderate* is constantly shifting, but it is difficult to examine and interpret matrices of more than order 1000 × 1000), the SVD, whether in the truncated or tapered forms is probably the method of choice–because it provides the fullest information about data and its relationship to the solution. Its only disadvantages are that one can be easily overwhelmed by the available information, particularly if a range of solutions must be examined, and it cannot take advantage of sparsity in large problems. The SVD has a flexibility beyond even what we have discussed should the investigator know enough to justify it. One could, for example, change the degree of tapering in each of the terms of (3.4.133)–(3.4.134) should there be reason to repartition the variance between solution and noise, or some terms could be dropped out of the truncated form at will.

The more general situation, in which structured solution and noise covariances are available, is then readily understood. These matrices are used to reduce the problem by coordinate transformation to ones in which the structure has been removed. At that point, the methods for unstructured problems are used, with the resulting solution, residuals, covariances, and resolution matrices being transformed back to the original physical spaces.

Both ordinary weighted least squares and the SVD applied to row- and column-weighted equations are best thought of as approximation, rather than estimation, methods and thus have a lot to recommend them. In particular, the truncated SVD does not produce a minimum variance estimate the way the tapered version can. On the other hand, the tapered SVD (along with the Gauss-Markov estimate, or the tapered least-squares solutions) produces the minimum variance property by tolerating a bias in the solution. Whether the bias is more desirable than a larger uncertainty is a decision that the user must make. But the reader is warned against the belief that there is any single best method whose determination should take precedence over understanding the problem physics.

A useful working definition now of an inverse method, distinguishing them from mere curve fitting, is that it quantifies the extent to which elements of a system have been determined by focusing on uncertainties in the solution. Different approaches have different desirable features, including (1) separation of nullspace uncertainties from those owing to observational noise, (2) ability to use prior statistical knowledge, (3) determination of orthogonal solution structures in terms of orthogonal data structures and of their relative importance (data ranking), and (4) ability to trade resolution against stability.

The statistical discussion here has been qualitative and intuitive with no claim to rigor. To some extent, the subject of inferring the ocean circulation from observations and dynamics has not yet evolved to the point where more than semiquantitative statistical tests seem warranted. One can expect that ultimately more refined tests leading to adoption or rejection of particular dynamical models will one day become necessary. The reader wishing a more careful account of the statistical underpinnings of the subject can make a beginning with Tarantola (1987) or Backus (1970a,b; 1988a) and the references there.