2

Basic machinery

2.1 Background

The purpose of this chapter is to record a number of results that are useful in finding and understanding the solutions to sets of usually noisy simultaneous linear equations and in which formally there may be too much or too little information. A lot of the material is elementary; good textbooks exist, to which the reader will be referred. Some of what follows is discussed primarily so as to produce a consistent notation for later use. But some topics are 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 linear 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. Minimum variance estimation is introduced through the Gauss–Markov theorem as an alternative method for obtaining solutions to simultaneous equations, and its relation to and distinction from least-squares is discussed. The chapter ends with a brief discussion of recursive least-squares and estimation; this part is essential background for the study of time-dependent problems in Chapter 4.

2.2 Matrix and vector algebra

This subject is very large and well-developed and it is not my intention to repeat material better found elsewhere.[1] Only a brief survey of essential results is provided.

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

$$\mathbf{A} = \{A_{ij}\}, \quad i = 1, 2, \ldots, M, \qquad j = 1, 2, \ldots, N.$$

Normally a matrix is denoted by a bold-faced capital letter. A vector is a special case of an $M \times 1$ matrix, written as a bold-face lower case letter, for example, $\mathbf{q}$. Corresponding capital or lower case letters for Greek symbols are also indicated in bold-face. Unless otherwise stipulated, vectors are understood to be columnar. The transpose of a matrix $\mathbf{A}$ is written $\mathbf{A}^{\mathrm{T}}$ and is defined as $\{A^{\mathrm{T}}\}_{ij} = A_{ji}$, an interchange of the rows and columns of $\mathbf{A}$. Thus $(\mathbf{A}^{\mathrm{T}})^{\mathrm{T}} = \mathbf{A}$. Transposition applied to vectors is sometimes used to save space in printing, for example, $\mathbf{q} = [q_1, q_2, \ldots, q_N]^{\mathrm{T}}$ is the same as

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

Matrices and vectors

A conventional measure of length of a vector is $\sqrt{\mathbf{a}^{\mathrm{T}}\mathbf{a}} = \sqrt{\sum_i^N a_i^2} = \|\mathbf{a}\|$. The inner, or dot, product between two $L \times 1$ vectors $\mathbf{a}$, $\mathbf{b}$ is written $\mathbf{a}^{\mathrm{T}}\mathbf{b} \equiv \mathbf{a} \cdot \mathbf{b} = \sum_{i=1}^{L} a_i b_i$ and is a scalar. Such an inner product is the "projection" of $\mathbf{a}$ onto $\mathbf{b}$ (or vice versa). It is readily shown that $|\mathbf{a}^{\mathrm{T}}\mathbf{b}| = \|\mathbf{a}\|\ \|\mathbf{b}\|\,|\cos\phi| \leq \|\mathbf{a}\|\ \|\mathbf{b}\|$, where the magnitude of $\cos\phi$ 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

$$\mathbf{f} = \sum_{i=1}^{N} \alpha_i \mathbf{e}_i, \tag{2.1}$$

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

$$\mathbf{e}_j - \sum_{i=1,\, i\neq j}^{N} \beta_i \mathbf{e}_i \neq 0, \qquad j = 1, 2, \ldots, N. \tag{2.2}$$

A subset of the $\mathbf{e}_j$ are said to span a subspace (all vectors perfectly representable by the subset). For example, $[1, -1, 0]^{\mathrm{T}}$, $[1, 1, 0]^{\mathrm{T}}$ span the subspace of all vectors $[v_1, v_2, 0]^{\mathrm{T}}$. A "spanning set" completely describes the subspace too, but might have additional, redundant vectors. Thus the vectors $[1, -1, 0]^{\mathrm{T}}$, $[1, 1, 0]^{\mathrm{T}}$, $[1, 1/2, 0]$ span the subspace but are not a basis for it.

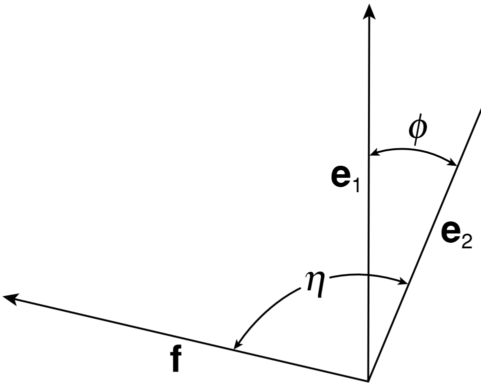

Figure 2.1 Schematic of expansion of an arbitrary vector $\mathbf{f}$ in two vectors $\mathbf{e}_1, \mathbf{e}_2$ which may nearly coincide in direction.

The expansion coefficients α_i in (2.1) are obtained by taking the dot product of (2.1) with each of the vectors in turn:

$$\sum_{i=1}^{N} \alpha_i \mathbf{e}_k^T \mathbf{e}_i = \mathbf{e}_k^T \mathbf{f}, \quad k = 1, 2, \ldots, N, \tag{2.3}$$

which is 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 a necessary one. With a basis, 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 $\mathbf{f}$ is contained in the dot products.

The concept of "nearly dependent" vectors is helpful and can be understood heuristically. Consider Fig. 2.1, 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 $\mathbf{f}$ in the plane. The simultaneous equations become

$$\begin{aligned} \alpha_1 \mathbf{e}_1^T \mathbf{e}_1 + \alpha_2 \mathbf{e}_1^T \mathbf{e}_2 &= \mathbf{e}_1^T \mathbf{f}, \\ \alpha_1 \mathbf{e}_2^T \mathbf{e}_1 + \alpha_2 \mathbf{e}_2^T \mathbf{e}_2 &= \mathbf{e}_2^T \mathbf{f}. \end{aligned} \tag{2.4}$$

The vectors become nearly parallel as the angle ϕ in Fig. 2.1 goes to zero; as long as they are not identically parallel, they can still be used mathematically to represent $\mathbf{f}$ perfectly. An important feature is that even if the lengths of $\mathbf{e}_{1,} \mathbf{e}_2$, $\mathbf{f}$ are all order-one, the expansion coefficients α_1, α_2 can have unbounded magnitudes when the angle ϕ becomes small and $\mathbf{f}$ is nearly orthogonal to both (measured by angle η).

That is to say, we find readily from (2.4) that

$$\alpha_1 = \frac{\left(\mathbf{e}_1^T\mathbf{f}\right)\left(\mathbf{e}_2^T\mathbf{e}_2\right) - \left(\mathbf{e}_2^T\mathbf{f}\right)\left(\mathbf{e}_1^T\mathbf{e}_2\right)}{\left(\mathbf{e}_1^T\mathbf{e}_1\right)\left(\mathbf{e}_2^T\mathbf{e}_2\right) - \left(\mathbf{e}_1^T\mathbf{e}_2\right)^2}, \tag{2.5}$$

$$\alpha_2 = \frac{\left(\mathbf{e}_2^T\mathbf{f}\right)\left(\mathbf{e}_1^T\mathbf{e}_1\right) - \left(\mathbf{e}_1^T\mathbf{f}\right)\left(\mathbf{e}_2^T\mathbf{e}_1\right)}{\left(\mathbf{e}_1^T\mathbf{e}_1\right)\left(\mathbf{e}_2^T\mathbf{e}_2\right) - \left(\mathbf{e}_1^T\mathbf{e}_2\right)^2}. \tag{2.6}$$

Suppose for simplicity that $\mathbf{f}$ has unit length, and that the $\mathbf{e}_i$ have also been normalized to unit length as shown in Fig. 2.1. Then,

$$\alpha_1 = \frac{\cos(\eta - \phi) - \cos\phi\cos\eta}{1 - \cos^2\phi} = \frac{\sin\eta}{\sin\phi}, \tag{2.7}$$

$$\alpha_2 = \cos\eta - \sin\eta\cot\phi \tag{2.8}$$

and whose magnitudes can become arbitrarily large as $\phi \to 0$. One can imagine a situation in which $\alpha_1\mathbf{e}_1$ and $\alpha_2\mathbf{e}_2$ were separately measured and found to be very large. One could then erroneously infer that the sum vector, $\mathbf{f}$, was equally large. This property of the expansion in non-orthogonal vectors potentially producing large coefficients becomes important later (Chapter 5) as a way of gaining insight into the behavior of so-called non-normal operators. The generalization to higher dimensions is left to the reader's intuition. One anticipates that as ϕ becomes very small, numerical problems can arise in using these "almost parallel" vectors.

Gram–Schmidt process

One often has a set of p independent, but non-orthonormal 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 the first k of the $\mathbf{h}_i$ have been orthonormalized to a new set, $\mathbf{g}_i$. To generate vector $k+1$, let

$$\mathbf{g}_{k+1} = \mathbf{h}_{k+1} - \sum_j^k \gamma_j \mathbf{g}_j. \tag{2.9}$$

Because $\mathbf{g}_{k+1}$ must be orthogonal to the preceding $\mathbf{g}_i$, $i = 1, \ldots, k$, take the dot products of (2.9) 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 dividing by its length.

Given the first k of N necessary vectors, an additional $N - k$ independent vectors, $\mathbf{h}_i$, are needed. There are several possibilities. The necessary extra vectors might be generated by filling their elements with random numbers. Or a very simple trial set like $\mathbf{h}_{k+1} = [1, 0, 0, \ldots, 0]^T$, $\mathbf{h}_{k+2} = [0, 1, 0, \ldots 0], \ldots$ might be adequate. If one is unlucky, the set chosen might prove not to be independent of the existing $\mathbf{g}_i$.

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" for purposes of numerical stability.[2]

2.2.1 Matrix multiplication and identities

It has been found convenient and fruitful to usually define multiplication of two matrices $\mathbf{A}$, $\mathbf{B}$, written as $\mathbf{C} = \mathbf{AB}$, such that

$$C_{ij} = \sum_{p=1}^{P} A_{ip} B_{pj}. \tag{2.10}$$

For the definition (2.10) to make sense, $\mathbf{A}$ must be an $M \times P$ matrix and $\mathbf{B}$ 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. Define $\mathbf{A}^2 = \mathbf{AA}$, etc. Other definitions of matrix multiplication exist, and are useful, but are not needed here.

The mathematical operation in (2.10) may appear arbitrary, but a physical interpretation is available: Matrix multiplication is the dot product of all of the rows of $\mathbf{A}$ with all of the columns of $\mathbf{B}$. Thus multiplication of a vector by a matrix represents the projections of the rows of the matrix onto the vector.

Define a matrix, $\mathbf{E}$, each of whose columns is the corresponding vector $\mathbf{e}_i$, and a vector, $\boldsymbol{\alpha} = \{\alpha_i\}$, in the same order. Then the expansion (2.1) can be written compactly as

$$\mathbf{f} = \mathbf{E}\boldsymbol{\alpha}. \tag{2.11}$$

A "symmetric matrix" is one for which $\mathbf{A}^{\mathrm{T}} = \mathbf{A}$. The product $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ represents the array of all the dot products of the columns of $\mathbf{A}$ with themselves, and similarly, $\mathbf{AA}^{\mathrm{T}}$ represents the set of all dot products of all the rows of $\mathbf{A}$ with themselves. It follows that $(\mathbf{AB})^{\mathrm{T}} = \mathbf{B}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}$. Because we have $(\mathbf{AA}^{\mathrm{T}})^{\mathrm{T}} = \mathbf{AA}^{\mathrm{T}}$, $(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{\mathrm{T}} = \mathbf{A}^{\mathrm{T}}\mathbf{A}$, both of these matrices are symmetric.

The "trace" of a square $M \times M$ matrix $\mathbf{A}$ is defined as trace($\mathbf{A}$) $= \sum_i^M A_{ii}$. A "diagonal matrix" is square and zero except for the terms along the main diagonal, although we will later generalize this definition. The operator diag($\mathbf{q}$) forms a square diagonal matrix with $\mathbf{q}$ 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. $\mathbf{IA} = \mathbf{A}$, $\mathbf{AI} = \mathbf{I}$, for any $\mathbf{A}$ for which the products make sense. If there is a matrix $\mathbf{B}$, such that $\mathbf{BE} = \mathbf{I}$, then $\mathbf{B}$ is the "left inverse" of $\mathbf{E}$. If $\mathbf{B}$ is the left inverse of $\mathbf{E}$ and $\mathbf{E}$ is square, a standard result is that it must also be a right inverse: $\mathbf{EB} = \mathbf{I}$, $\mathbf{B}$ is then called "the inverse of $\mathbf{E}$" and is usually written $\mathbf{E}^{-1}$. Square matrices with inverses are "non-singular." Analytical expressions exist for a few inverses; more generally, linear algebra books explain how to find them numerically when they exist. If $\mathbf{E}$ is not square, one must distinguish left and right inverses, 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. A notational shorthand is $(\mathbf{A}^{-1})^{\mathrm{T}} = (\mathbf{A}^{\mathrm{T}})^{-1} \equiv \mathbf{A}^{-\mathrm{T}}$.

The "length," or norm, of a vector has already been introduced. But several choices are possible; for present purposes, the conventional l_2 norm already defined,

$$\|\mathbf{f}\|_2 \equiv (\mathbf{f}^{\mathrm{T}}\mathbf{f})^{1/2} = \left(\sum_{i=1}^{N} f_i^2\right)^{1/2}, \tag{2.12}$$

is most useful; often the subscript is omitted. Equation (2.12) leads in turn to the measure of distance between two vectors, $\mathbf{a}$, $\mathbf{b}$, as

$$\|\mathbf{a} - \mathbf{b}\|_2 = \sqrt{(\mathbf{a} - \mathbf{b})^{\mathrm{T}}(\mathbf{a} - \mathbf{b})}, \tag{2.13}$$

which is 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,

$$\|\mathbf{c}\|_W = \sqrt{\textstyle\sum_i c_i W_{ii} c_i}, \tag{2.14}$$

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

$$\|\mathbf{c}\|_W = \sqrt{\mathbf{c}^{\mathrm{T}}\mathbf{W}\mathbf{c}}, \tag{2.15}$$

where $\mathbf{W}$ is an arbitrary matrix (but usually, for physical reasons, symmetric and positive definite,[3] implying that $\mathbf{c}^{\mathrm{T}}\mathbf{W}\mathbf{c} \geq 0$).

2.2.2 Linear simultaneous equations

Consider a set of M-linear equations in N-unknowns,

$$\mathbf{Ex} = \mathbf{y}. \tag{2.16}$$

Because of the appearance of simultaneous equations in situations in which the y_i are observed, and where $\mathbf{x}$ are parameters whose values are sought, it is often convenient to refer to (2.16) as a set of measurements of $\mathbf{x}$ that produced the observations or data, $\mathbf{y}$. If $M > N$, the system is said to be "formally overdetermined." If $M < N$, it is "underdetermined," and if $M = N$, it is "formally just-determined." The use of the word "formally" has a purpose we will come to. 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 (2.16) by $\mathbf{E}^{-1}$:

$$\mathbf{E}^{-1}\mathbf{E}\mathbf{x} = \mathbf{I}\mathbf{x} = \mathbf{x} = \mathbf{E}^{-1}\mathbf{y}.$$

The reader is cautioned that although matrix inverses are a very powerful theoretical tool, one is usually ill-advised to solve large sets of simultaneous equations by employing $\mathbf{E}^{-1}$; better numerical methods are available for the purpose.[4]

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 $\mathbf{e}_i$, then (2.16) is

$$x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + \cdots + x_N\mathbf{e}_N = \mathbf{y}. \tag{2.17}$$

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 coefficients. Depending upon the ratio of M to N, that is, the number of equations compared to the number of 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.

An alternative interpretation of simultaneous linear equations denotes the rows of $\mathbf{E}$ as $\mathbf{r}_i^{\mathrm{T}}$, $i = 1, 2, \ldots, M$. Then Eq. (2.16) is a set of M-inner products,

$$\mathbf{r}_i^{\mathrm{T}}\mathbf{x} = y_i, \quad i = 1, 2, \ldots, M. \tag{2.18}$$

That is, the set of simultaneous equations is also 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.

A special set of simultaneous equations for square matrices, **A**, is labelled the "eigenvalue/eigenvector problem,"

$$\mathbf{Ae} = \lambda\mathbf{e}. \tag{2.19}$$

In this set of linear simultaneous equations one seeks a special vector, **e**, such that for some as yet unknown scalar eigenvalue, λ, there is a solution. An $N \times N$ matrix will have up to N solutions $(\lambda_i, \mathbf{e}_i)$, but the nature of these elements and their relations require considerable effort to deduce. We will look at this problem more later; for the moment, it again suffices to say that numerical methods for solving Eq. (2.19) are well-known.

2.2.3 Matrix norms

A number of useful definitions of a matrix size, or norm, exist. The so-called "spectral norm" or "2-norm" defined as

$$\|\mathbf{A}\|_2 = \sqrt{\text{maximum eigenvalue of } (\mathbf{A}^T\mathbf{A})} \tag{2.20}$$

is usually adequate. Without difficulty, it may be seen that this definition is equivalent to

$$\|\mathbf{A}\|_2 = \max \frac{\mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x}}{\mathbf{x}^T\mathbf{x}} = \max \frac{\|\mathbf{Ax}\|_2}{\|\mathbf{x}\|_2} \tag{2.21}$$

where the maximum is defined over all vectors **x**.[5] Another useful measure is the "Frobenius norm,"

$$\|\mathbf{A}\|_F = \sqrt{\sum\nolimits_{i=1}^{M} \sum\nolimits_{j=1}^{N} A_{ij}^2} = \sqrt{\text{trace}(\mathbf{A}^T\mathbf{A})}. \tag{2.22}$$

Neither norm requires **A** to be square. These norms permit one to derive various useful results. Consider the following illustration. Suppose **Q** is square, and $\|\mathbf{Q}\| < 1$, then

$$(\mathbf{I} + \mathbf{Q})^{-1} = \mathbf{I} - \mathbf{Q} + \mathbf{Q}^2 - \cdots, \tag{2.23}$$

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.

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.[6] Software systems like MATLAB, Maple, IDL, and Mathematica 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.

2.2.4 Identities: differentiation

There 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,"

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

is positive for all possible vectors $\mathbf{x}$. (It suffices to consider only symmetric $\mathbf{A}$ because for a general matrix, $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} = \mathbf{x}^{\mathrm{T}}[(\mathbf{A}+\mathbf{A}^{\mathrm{T}})/2]\mathbf{x}$, which follows from the scalar property of the quadratic form.) If $J \geq 0$ for all $\mathbf{x}$, $\mathbf{A}$ is "positive semi-definite," or "non-negative definite." Linear algebra books show that a necessary and sufficient requirement for positive definiteness is that $\mathbf{A}$ has only positive eigenvalues (Eq. 2.19) and a semi-definite one must have all non-negative eigenvalues.

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. The derivative of a matrix by a scalar is just the matrix of element by element derivatives. Alternatively, if s is any scalar, its derivative by a vector,

$$\frac{\partial s}{\partial \mathbf{q}} = \left[\frac{\partial s}{\partial q_1} \cdots \frac{\partial s}{\partial q_N}\right]^{\mathrm{T}}, \tag{2.24}$$

is a column vector (the gradient; some authors define it to be a row vector). The derivative of one vector by another is defined as a matrix:

$$\frac{\partial \mathbf{r}}{\partial \mathbf{q}} = \left\{\frac{\partial r_i}{\partial q_j}\right\} = \left\{\begin{array}{cccc} \frac{\partial r_1}{\partial q_1} & \frac{\partial r_2}{\partial q_1} & \cdot & \frac{\partial r_M}{\partial q_1} \\ \frac{\partial r_1}{\partial q_2} & \cdot & \cdot & \frac{\partial r_M}{\partial q_2} \\ \cdot & \cdot & \cdot & \cdot \\ \frac{\partial r_1}{\partial q_N} & \cdot & \cdot & \frac{\partial r_M}{\partial q_N} \end{array}\right\} \equiv \mathbf{B}. \tag{2.25}$$

If $\mathbf{r}$, $\mathbf{q}$ are of the same dimension, the determinant of $\mathbf{B} = \det(\mathbf{B})$ is the "Jacobian" of $\mathbf{r}$.[7]

The second derivative of a scalar,

$$\frac{\partial^2 s}{\partial \mathbf{q}^2} = \left\{\frac{\partial}{\partial \mathbf{q}_i}\frac{\partial s}{\partial \mathbf{q}_j}\right\} = \left\{\begin{array}{ccccc} \frac{\partial^2 s}{\partial q_1^2} & \frac{\partial^2 s}{\partial q_1 q_2} & \cdot & \cdot & \frac{\partial^2 s}{\partial q_1 q_N} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \frac{\partial^2 s}{\partial q_N \partial q_1} & \cdot & \cdot & \cdot & \frac{\partial^2 s}{\partial q_N^2} \end{array}\right\}, \tag{2.26}$$

is the "Hessian" of s and is the derivative of the gradient of s.

Assuming conformability, the inner product, $J = \mathbf{r}^{\mathrm{T}}\mathbf{q} = \mathbf{q}^{\mathrm{T}}\mathbf{r}$, is a scalar. The differential of J is

$$\mathrm{d}J = \mathrm{d}\mathbf{r}^{\mathrm{T}}\mathbf{q} + \mathbf{r}^{\mathrm{T}}\mathrm{d}\mathbf{q} = \mathrm{d}\mathbf{q}^{\mathrm{T}}\mathbf{r} + \mathbf{q}^{\mathrm{T}}\mathrm{d}\mathbf{r}, \tag{2.27}$$

and hence the partial derivatives are

$$\frac{\partial(\mathbf{q}^{\mathrm{T}}\mathbf{r})}{\partial\mathbf{q}} = \frac{\partial(\mathbf{r}^{\mathrm{T}}\mathbf{q})}{\partial\mathbf{q}} = \mathbf{r}, \tag{2.28}$$

$$\frac{\partial(\mathbf{q}^{\mathrm{T}}\mathbf{q})}{\partial\mathbf{q}} = 2\mathbf{q}. \tag{2.29}$$

It follows immediately that, for matrix/vector products,

$$\frac{\partial}{\partial\mathbf{q}}(\mathbf{B}\mathbf{q}) = \mathbf{B}^{\mathrm{T}}, \qquad \frac{\partial}{\partial\mathbf{q}}(\mathbf{q}^{\mathrm{T}}\mathbf{B}) = \mathbf{B}. \tag{2.30}$$

The first of these is used repeatedly, and attention is called to the apparently trivial fact that differentiation of $\mathbf{Bq}$ with respect to $\mathbf{q}$ produces the transpose of $\mathbf{B}$ – the origin, as seen later, of so-called adjoint models. For a quadratic form,

$$\begin{aligned} J &= \mathbf{q}^{\mathrm{T}}\mathbf{A}\mathbf{q} \\ \frac{\partial J}{\partial\mathbf{q}} &= (\mathbf{A} + \mathbf{A}^{\mathrm{T}})\mathbf{q}, \end{aligned} \tag{2.31}$$

and the Hessian of the quadratic form is $2\mathbf{A}$ if $\mathbf{A} = \mathbf{A}^{\mathrm{T}}$.

Differentiation of a scalar function (e.g., J in Eq. 2.31) or a vector by a matrix, $\mathbf{A}$, is readily defined.[8] Differentiation of a matrix by another matrix results in a third, very large, matrix. One special case of the *differential* of a matrix function proves useful later on. It can be shown[9] that

$$\mathrm{d}\mathbf{A}^n = (\mathrm{d}\mathbf{A})\,\mathbf{A}^{n-1} + \mathbf{A}\,(\mathrm{d}\mathbf{A})\,\mathbf{A}^{n-2} + \cdots + \mathbf{A}^{n-1}(\mathrm{d}\mathbf{A}), \tag{2.32}$$

where $\mathbf{A}$ is square. Thus the derivative with respect to some scalar, k, is

$$\frac{\mathrm{d}\mathbf{A}^n}{\mathrm{d}k} = \frac{(\mathrm{d}\mathbf{A})}{\mathrm{d}k}\mathbf{A}^{n-1} + \mathbf{A}^{n-2}\frac{(\mathrm{d}\mathbf{A})}{\mathrm{d}k}\mathbf{A} + \cdots + \mathbf{A}^{n-1}\left(\frac{\mathrm{d}\mathbf{A}}{\mathrm{d}k}\right). \tag{2.33}$$

There are a few, unfortunately unintuitive, matrix inversion identities that are essential later. They are derived by considering the square, partitioned matrix,

$$\begin{Bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^{\mathrm{T}} & \mathbf{C} \end{Bmatrix}, \tag{2.34}$$

where $\mathbf{A}^{\mathrm{T}} = \mathbf{A}$, $\mathbf{C}^{\mathrm{T}} = \mathbf{C}$, but $\mathbf{B}$ can be rectangular of conformable dimensions in (2.34).[10] The most important of the identities, sometimes called the “matrix inversion lemma" is, in one form,

$$\begin{aligned}\{\mathbf{C}-\mathbf{B}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{B}\}^{-1} &= \{\mathbf{I}-\mathbf{C}^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{B}\}^{-1}\mathbf{C}^{-1}\\ &= \mathbf{C}^{-1}-\mathbf{C}^{-1}\mathbf{B}^{\mathrm{T}}(\mathbf{B}\mathbf{C}^{-1}\mathbf{B}^{\mathrm{T}}-\mathbf{A})^{-1}\mathbf{B}\mathbf{C}^{-1},\end{aligned} \tag{2.35}$$

where it is assumed that the inverses exist.[11] A variant is

$$\mathbf{A}\mathbf{B}^{\mathrm{T}}(\mathbf{C}+\mathbf{B}\mathbf{A}\mathbf{B}^{\mathrm{T}})^{-1} = (\mathbf{A}^{-1}+\mathbf{B}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{B})^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{C}^{-1}. \tag{2.36}$$

Equation (2.36) is readily confirmed by left-multiplying both sides by $(\mathbf{A}^{-1}+\mathbf{B}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{B})$, and right-multiplying by $(\mathbf{C}+\mathbf{B}\mathbf{A}\mathbf{B}^{\mathrm{T}})$ and showing that the two sides of the resulting equation are equal.

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

$$\mathbf{A}\mathbf{C}\mathbf{A}^{\mathrm{T}}-\mathbf{B}\mathbf{A}^{\mathrm{T}}-\mathbf{A}\mathbf{B}^{\mathrm{T}} = (\mathbf{A}-\mathbf{B}\mathbf{C}^{-1})\mathbf{C}(\mathbf{A}-\mathbf{B}\mathbf{C}^{-1})^{\mathrm{T}}-\mathbf{B}\mathbf{C}^{-1}\mathbf{B}^{\mathrm{T}}. \tag{2.37}$$

2.3 Simple statistics: regression

2.3.1 Probability densities, moments

Some statistical ideas are required, but the discussion is confined to stating some basic notions and to developing a notation.[12] We require the idea of a probability density for a random variable x. This subject is a very deep one, but our approach is heuristic.[13] Suppose that an arbitrarily large number of experiments can be conducted for the determination of the values of x, denoted X_i, $i=1,2,\ldots,M$, 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 of an arbitrarily large number of experiments, $M\to\infty$, divided into bins of arbitrarily small value ranges, and normalized by M, to produce the fraction of the total appearing in the ranges. Let the corresponding limiting frequency function be denoted $p_x(X)\mathrm{d}X$, interpreted as the fraction (probability) of values of x lying in the range, $X\le x\le X+\mathrm{d}X$. As a consequence of the definition, $p_x(X)\ge 0$ and

$$\int_{\text{all }X} p_x(X)\,\mathrm{d}X = \int_{-\infty}^{\infty} p_x(X)\,\mathrm{d}X = 1. \tag{2.38}$$

The infinite integral is a convenient way of representing an integral over "all X," as p_x simply vanishes for impossible values of X. (It should be noted that this so-called frequentist approach has fallen out of favor, with Bayesian assumptions being regarded as ultimately more rigorous and fruitful. For introductory purposes, however, empirical frequency functions appear to provide an adequate intuitive basis for proceeding.)

The "average," or "mean," or "expected value" is denoted $\langle x \rangle$ and defined as

$$\langle x \rangle \equiv \int_{\mathrm{all}\ X} X p_x(X) \mathrm{d}X \ = m_1. \tag{2.39}$$

The mean is the center of mass of the probability density. 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 $\langle x \rangle$. If the deviation from the true mean is denoted x' so that $x = \langle x \rangle + x'$, such a forecast has the virtue that we are assured the average forecast error, $\langle x' \rangle$, 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 non-random quantity, $\langle ax \rangle = a\langle x \rangle$ and $\langle ax + y \rangle = a \langle x \rangle + \langle y \rangle$.

Quantity $\langle x \rangle$ is the "first-moment" of the probability density. Higher order moments are defined as

$$m_n = \langle x^n \rangle = \int_{-\infty}^{\infty} X^n p_x(X) \mathrm{d}X,$$

where n are the non-negative integers. A useful theoretical result is that a knowledge of all the moments is usually enough to completely define the probability density themselves. (There are troublesome situations with, e.g., non-existent moments, as with the so-called Cauchy distribution, $p_x(X) = (2/\pi)(1/(1 + X^2)) \ X \geq 0$, whose mean is infinite.) For many important probability densities, including the Gaussian, a knowledge of the first two moments $n = 1, 2$ is sufficient to define all the others, and hence the full probability density. It is common to define the moments for $n > 1$ about the mean, so that one has

$$\mu_n = \langle (x - \langle x \rangle)^n \rangle = \int_{-\infty}^{\infty} (X - \langle X \rangle)^n \, p_x(X) \mathrm{d}X.$$

μ_2 is the variance and often written $\mu_2 = \sigma^2$, where σ is the "standard deviation."

2.3.2 Sample estimates: bias

In observational sciences, one normally must estimate the values defining the probability density from the data itself. Thus the first moment, the mean, is often computed as the "sample average,"

$$\tilde{m}_1 = \langle x \rangle_M \equiv \frac{1}{M} \sum_{i=1}^{M} X_i. \tag{2.40}$$

The notation $\tilde{m}_1$ is used to distinguish the sample estimate from the true value, m_1. On the other hand, if the experiment of computing $\tilde{m}_1$ from M samples could be repeated many times, the mean of the sample estimates would be the true mean. This conclusion is readily seen by considering the expected value of the difference from the true mean:

$$\begin{aligned}\langle \langle x\rangle_M - \langle x\rangle\rangle &= \left\langle \frac{1}{M}\sum_{i=1}^{M} X_i - \langle x\rangle \right\rangle \\ &= \frac{1}{M}\sum_{i=1}^{M}\langle X_i\rangle - \langle x\rangle = \frac{M}{M}\langle x\rangle - \langle x\rangle = 0.\end{aligned}$$

Such an estimate is said to be "unbiassed": its expected value is the quantity one seeks.

The interpretation is that, for finite M, 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 will fluctuate about the true mean, with equal probability of being higher or lower. 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 "biassed." A simple example of a biassed estimator is the "sample variance," defined as

$$s^2 \equiv \frac{1}{M}\sum_{i}^{M}(X_i - \langle x\rangle_M)^2. \tag{2.41}$$

For reasons explained later in this chapter (p. 42), one finds that

$$\langle s^2\rangle = \frac{M-1}{M}\sigma^2 \neq \sigma^2,$$

and thus the expected value is not the true variance. (This particular estimate is "asymptotically unbiassed," as the bias vanishes as $M \to \infty$.)

We are assured that the sample mean is unbiassed. But the probability that $\langle x\rangle_M = \langle x\rangle$, that is that we obtain exactly the true value, is very small. It helps to have a measure of the extent to which $\langle x\rangle_M$ is likely to be very far from $\langle x\rangle$. To do so, 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:

$$\sigma^2 = \langle (x - \langle x\rangle)^2\rangle.$$

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

$$D^2(z) = \langle z^2 \rangle.$$

Thus, the variance of x is $D^2(x - \langle x \rangle)$.

The variance of $\langle x \rangle_M$ about the correct value is obtained by a little algebra using the bracket notation,

$$D^2((\langle x \rangle_M - x)^2) = \frac{\sigma^2}{M}. \tag{2.42}$$

This expression shows the well-known result that as M becomes large, any tendency of the sample mean to lie far from the true value will diminish. It does not prove that some particular value will not, by accident, be far away, merely that it becomes increasingly unlikely as M grows. (In statistics textbooks, the Chebyschev inequality is used to formalize this statement.)

An estimate that is unbiassed and whose expected dispersion about the true value goes to zero with M is evidently desirable. In more interesting estimators, a bias is often present. Then for a fixed number of samples, M, there would be two distinct sources of deviation (error) from the true value: (1) the bias – how far, on average, it is expected to be from the true value, and (2) the tendency – from purely random events – for the value to differ from the true value (the random error). In numerous cases, one discovers that tolerating a small bias error can greatly reduce the random error – and thus the bias may well be worth accepting for that reason. In some cases therefore, a bias is deliberately introduced.

2.3.3 Functions and sums 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

$$\langle g(x) \rangle = \int_{-\infty}^{\infty} g(X) p_x(X) \mathrm{d}X, \tag{2.43}$$

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 obtained from

$$p_g(G) = p_x(X(G)) \frac{\mathrm{d}x}{\mathrm{d}g} \mathrm{d}G, \tag{2.44}$$

where $\mathrm{d}x/\mathrm{d}g$ is the ratio of the differential intervals occupied by x and g and can be understood by reverting to the original definition of probability densities from histograms.

The Gaussian, or normal, probability density is one that is mathematically handy (but is potentially dangerous as a general model of the behavior of natural processes – many geophysical and fluid processes are demonstrably non-Gaussian). For a single random 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 $\langle x \rangle = m_x$, $\langle (x - \langle x \rangle)^2 \rangle = \sigma_x^2$.

One important special case is the transformation of the Gaussian to another Gaussian of zero-mean and unit standard deviation, $G(0, 1)$,

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

which can always be done, and thus,

$$p_z(Z) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{Z^2}{2}\right].$$

A second important special case of a change of variable is $g = z^2$, where z is Gaussian of zero mean and unit variance. Then the probability density of g is

$$p_g(G) = \frac{1}{G^{1/2}\sqrt{2\pi}} \exp(-G/2), \tag{2.45}$$

a special probability density usually denoted χ_1^2 ("chi-square-sub-1"), the probability density for the square of a Gaussian. One finds $\langle g \rangle = 1$, $\langle (g - \langle g \rangle)^2 \rangle = 2$.

2.3.4 Multivariable probability densities: correlation

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 normalized by the total number of samples, taking the limit as the number of samples goes to infinity, and the bin sizes go to zero, to produce a joint probability density $p_{xy}(X, Y)$. $p_{xy}(X, Y)\,\mathrm{d}X\,\mathrm{d}Y$ is then the fraction of occurrences such that $X \le x \le X + \mathrm{d}X, Y \le y \le Y + \mathrm{d}Y$. A simple example would be the probability density for the simultaneous measurement of the two components of horizontal velocity at a point in a fluid. Again, from the definition, $p_{xy}(X, Y) \ge 0$ and

$$\int_{-\infty}^{\infty} p_{xy}(X, Y)\,\mathrm{d}Y = p_x(X), \tag{2.46}$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_{xy}(X, Y)\mathrm{d}X\,\mathrm{d}Y = 1. \tag{2.47}$$

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 read as "the probability of x, given that y has occurred," or, "the probability of x conditioned on y." It follows immediately from the definition of the probability density that

$$p_{x|y}(X|Y) = \frac{p_{xy}(X,\, Y)}{p_y(Y)} \tag{2.48}$$

(This equation can be interpreted by going back to the original experimental concept, and understanding the restriction on x, given that y is known to lie within a strip paralleling the X axis).

Using the joint frequency function, define the average product as

$$\langle xy \rangle = \int\int_{\mathrm{all}X,Y} XY p_{xy}(X,\, Y)\mathrm{d}X\, \mathrm{d}Y. \tag{2.49}$$

Suppose that upon examining the joint frequency function, one finds that $p_{xy}(X,\, Y) = p_x(X)p_y(Y)$, that is, it factors into two distinct functions. In that case, x, y are said to be "independent." Many important results follow, including

$$\langle xy \rangle = \langle x \rangle \langle y \rangle.$$

Non-zero mean values are often primarily a nuisance. One can always define modified variables, e.g., $x' = x - \langle x \rangle$, such that the new variables have zero mean. Alternatively, one computes statistics centered on the mean. Should the centered product $\langle (x - \langle x \rangle)(y - \langle y \rangle) \rangle$ be non-zero, x, y are said to "co-vary" or to be "correlated." If $\langle (x - \langle x \rangle)(y - \langle y \rangle) \rangle = 0$, then the two variables are "uncorrelated." If x, y are independent, they are uncorrelated. Independence thus implies lack of correlation, but the reverse is not necessarily true. (These are theoretical relationships, and if $\langle x \rangle$, $\langle y \rangle$ are determined from observation, as described below, one must carefully distinguish estimated behavior from that expected theoretically.)

If the two variables are independent, then (2.48) is

$$p_{x|y}(X|Y) = p_x(X), \tag{2.50}$$

that is, the probability of x given y does not depend upon Y, and thus

$$p_{xy}\,(X, Y) = p_x\,(X)\,p_y\,(Y)$$

– and *there is then no predictive power for one variable given knowledge of the other.*

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

$$x = ay + n, \tag{2.51}$$

where n represents any contributions to x that remain unknown despite knowledge of y, and a is a constant. Then,

$$\langle x \rangle = a\langle y \rangle + \langle n \rangle, \tag{2.52}$$

and (2.51) can be re-written as

$$x - \langle x \rangle = a(y - \langle y \rangle) + (n - \langle n \rangle),$$

or

$$x' = ay' + n', \quad \text{where } x' = x - \langle x \rangle, \quad \text{etc.} \tag{2.53}$$

From this last equation,

$$a = \frac{\langle x'y' \rangle}{\langle y'^2 \rangle} = \frac{\langle x'y' \rangle}{(\langle y'^2 \rangle \langle x'^2 \rangle)^{1/2}} \frac{\langle x'^2 \rangle^{1/2}}{\langle y'^2 \rangle^{1/2}} = \rho \frac{\langle x'^2 \rangle^{1/2}}{\langle y'^2 \rangle^{1/2}}, \tag{2.54}$$

where it was supposed that $\langle y'n' \rangle = 0$, thus defining n'. The quantity

$$\rho \equiv \frac{\langle x'y' \rangle}{\langle y'^2 \rangle^{1/2} \langle x'^2 \rangle^{1/2}} \tag{2.55}$$

is the "correlation coefficient" and has the property,[14] $|\rho| \leq 1$. 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

$$\langle x'^2 \rangle = a^2 \langle y'^2 \rangle + \langle n'^2 \rangle = \rho^2 \langle x'^2 \rangle + \langle n'^2 \rangle,$$

or

$$(1 - \rho^2)\langle x'^2 \rangle = \langle n'^2 \rangle. \tag{2.56}$$

That is, $(1 - \rho^2)\langle x'^2 \rangle$ is the fraction of the variance in x' that is *unpredictable* from knowledge of y' and is the "unpredictable power." Conversely, $\rho^2 \langle x'^2 \rangle$ is the "predictable" power in x' given knowledge of y'. The limits as $\rho \to 0, \ 1$ are readily apparent.

Thus we interpret the statement that two variables x', y' "are correlated" or "co-vary" to mean that knowledge of one permits at least a partial prediction of the other, the expected success of the prediction depending upon the magnitude of ρ. If ρ is not zero, the variables cannot be independent, and the conditional probability $p_{x|y}(X|Y) \neq p_x(X)$. 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 co-vary, but are known not to be independent, a linear model like (2.51) would not be useful – a non-linear one would be required. Such non-linear 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.

In most cases, quantities like ρ, $\langle x'^2 \rangle$ are determined from the available measurements, e.g., of the form

$$a y_i + n_i = x_i, \tag{2.57}$$

and are not known exactly. They are thus sample values, are not equal to the true values, and must be interpreted carefully in terms of their inevitable biasses and variances. This large subject of regression analysis is left to the references.[15]

2.3.5 Change of variables

Suppose we have two random variables x, y with joint probability density $p_{xy}(X, Y)$. They are known as functions of two new variables $x = x(\xi_1, \xi_2)$, $y = y(\xi_1, \xi_2)$ and an inverse mapping $\xi_1 = \xi_1(x, y)$, $\xi_2 = \xi_2(x, y)$. 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 the x, y space to the ξ_1, ξ_2 space, that is,

$$p_{\xi_1 \xi_2}(\Xi_1, \Xi_2) = p_{xy}(X(\Xi_1, \Xi_2),\ Y(\Xi_1, \Xi_2)) \frac{\partial(X, Y)}{\partial(\Xi_1, \Xi_2)}, \tag{2.58}$$

where $\partial(X, Y)/\partial(\Xi_1, \Xi_2)$ is the Jacobian of the transformation between the two variable sets. As in any such transformation, one must be alert for zeros or infinities in the Jacobian, indicative of multiple valuedness in the mapping. Texts on multivariable calculus discuss such issues in detail.

Example *Suppose x_1, x_2 are independent Gaussian random variables of zero mean and variance σ^2. Then*

$$p_{\mathbf{x}}(\mathbf{X}) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{\left(X_1^2 + X_2^2\right)}{2\sigma^2} \right).$$

Define new random variables

$$r = (x_1^2 + x_2^2)^{1/2}, \qquad \phi = \tan^{-1}(x_2/x_1), \tag{2.59}$$

whose mapping in the inverse direction is

$$x_1 = r\cos\phi, \qquad y_1 = r\sin\phi, \tag{2.60}$$

that is, the relationship between polar and Cartesian coordinates. The Jacobian of the transformation is $J_a = r$. Thus

$$p_{r,\phi}(R, \Phi) = \frac{R}{2\pi}\frac{2}{\sigma^2}\exp(-R^2/\sigma^2), \quad 0 \le r,\ -\pi \le \phi \le \pi \tag{2.61}$$

The probability density for r alone is obtained by integrating

$$p_r(R) = \int_{-\pi}^{\pi} p_{r,\phi} d\phi = \frac{R}{\sigma^2}\exp[-R^2/(2\sigma^2)], \tag{2.62}$$

which is known as a Rayleigh distribution. By inspection then,

$$p_\phi(\Phi) = \frac{1}{2\pi},$$

which is the uniform distribution, independent of Φ. (These results are very important in signal processing.)

To generalize to n dimensions, let there be N variables, x_i, $i = 1, 2, \ldots, N$, with known joint probability density $p_{x_1 \cdots x_N}$. Let there be N new variables, ξ_i, that are known functions of the x_i, and an inverse mapping between them. Then the joint probability density for the new variables is just

$$\begin{aligned} &p_{\xi_1 \cdots \xi_N}(\Xi_1, \ldots, \Xi_N) \\ &\quad = p_{x_1 \cdots x_N}(\Xi_1(X_1, \ldots, X_N) \ldots, \Xi_N(X_1, \ldots, X_N)) \frac{\partial(X_1, \ldots, X_N)}{\partial(\Xi_1, \ldots, \Xi_N)}. \end{aligned} \tag{2.63}$$

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,

$$p_{x,y}(X, Y) = \frac{1}{2\pi\sigma_x\sigma_y}\exp\left(-\frac{(X - m_x)^2}{2\sigma_x^2} - \frac{(Y - m_y)^2}{2\sigma_y^2}\right). \tag{2.64}$$

Let two new random variables, ξ_1, ξ_2, be defined as a linear combination of x, y,

$$\begin{aligned} \xi_1 &= a_{11}(x - m_x) + a_{12}(y - m_y) + m_{\xi_1} \\ \xi_2 &= a_{21}(x - m_x) + a_{22}(y - m_y) + m_{\xi_2}, \end{aligned} \tag{2.65}$$

or, in vector form,

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

where $\mathbf{x} = [x, y]^{\mathrm{T}}$, $\mathbf{m}_x = [m_x, m_y]^{\mathrm{T}}$, $\mathbf{m}_y = [m_{\xi_1}, m_{\xi_2}]^{\mathrm{T}}$, and the numerical values satisfy the corresponding functional relations,

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

etc. Suppose that the relationship (2.65) 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

$$\mathbf{x} = \mathbf{B}(\boldsymbol{\xi} - \mathbf{m}_\xi) + \mathbf{m}_x. \tag{2.66}$$

Then the Jacobian of the transformation is

$$\frac{\partial(X, Y)}{\partial(\Xi_1, \Xi_2)} = b_{11}b_{22} - b_{12}b_{21} = \det(\mathbf{B}) \tag{2.67}$$

(det(**B**) is the determinant). Equation (2.65) produces

$$\begin{aligned} \langle \xi_1 \rangle &= m_{\xi_1} \\ \langle \xi_2 \rangle &= m_{\xi_2} \\ \langle (\xi_1 - \langle \xi_1 \rangle)^2 \rangle &= a_{11}^2\sigma_x^2 + a_{12}^2\sigma_y^2, \quad \langle (\xi_2 - \langle \xi_2 \rangle)^2 \rangle = a_{21}^2\sigma_x^2 + a_{22}^2\sigma_y^2 \\ \langle (\xi_1 - \langle \xi_1 \rangle)(\xi_2 - \langle \xi_2 \rangle) \rangle &= a_{11}a_{21}\sigma_x^2 + a_{12}a_{22}\sigma_y^2 \neq 0. \end{aligned} \tag{2.68}$$

In the special case,

$$\mathbf{A} = \begin{Bmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{Bmatrix}, \quad \mathbf{B} = \begin{Bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{Bmatrix}, \tag{2.69}$$

the transformation (2.69) is a simple coordinate rotation through angle ϕ, and the Jacobian is 1. The new second-order moments are

$$\langle (\xi_1 - \langle \xi_1 \rangle)^2 \rangle = \sigma_{\xi_1}^2 = \cos^2\phi\ \sigma_x^2 + \sin^2\phi\ \sigma_y^2, \tag{2.70}$$

$$\langle (\xi_2 - \langle \xi_2 \rangle)^2 \rangle = \sigma_{\xi_2}^2 = \sin^2\phi\ \sigma_x^2 + \cos^2\phi\ \sigma_y^2, \tag{2.71}$$

$$\langle (\xi_1 - \langle \xi_1 \rangle)(\xi_2 - \langle \xi_2 \rangle) \rangle \equiv \mu_{\xi_1\xi_2} = \left(\sigma_y^2 - \sigma_x^2\right)\cos\phi\ \sin\phi. \tag{2.72}$$

The new probability density is

$$p_{\xi_1\xi_2}(\Xi_1, \Xi_2) = \frac{1}{2\pi\sigma_{\xi_1}\sigma_{\xi_2}\sqrt{1-\rho_\xi^2}} \tag{2.73}$$

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

where $\rho_\xi = (\sigma_y^2 - \sigma_x^2)\sin\phi\cos\phi/(\sigma_x^2 + \sigma_y^2)^{1/2} = \mu_{\xi_1\xi_2}/\sigma_{\xi_1}\sigma_{\xi_2}$ 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 (2.73) 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 (2.69), which produces two other variables x, y that are uncorrelated, and therefore *independent*. Two such uncorrelated variables x, y will necessarily have different variances, otherwise ξ_1, ξ_2 would have zero correlation, too, by Eq. (2.72) .

As an important by-product, it is concluded that two jointly Gaussian random variables that are uncorrelated, are also independent. This property is one of the reasons Gaussians are so nice to work with; but it is not generally true of uncorrelated variables.

Vector random processes

Simultaneous discussion of two random processes, x, y, can regarded as discussion of a vector random process $[x,\ y]^{\mathrm{T}}$, and suggests a generalization to N dimensions. Label N random processes as x_i and define them as the elements of a vector $\mathbf{x} = [x_1, x_2, \ldots, x_N]^{\mathrm{T}}$. Then the mean is a vector: $\langle \mathbf{x} \rangle = \mathbf{m}_x$, and the covariance is a matrix:

$$\mathbf{C}_{xx} = D^2(\mathbf{x} - \langle \mathbf{x} \rangle) = \langle (\mathbf{x} - \langle \mathbf{x} \rangle)(\mathbf{x} - \langle \mathbf{x} \rangle)^{\mathrm{T}} \rangle, \tag{2.74}$$

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

$$\mathbf{C}_{xy} = \langle (\mathbf{x} - \langle \mathbf{x} \rangle)(\mathbf{y} - \langle \mathbf{y} \rangle)^{\mathrm{T}} \rangle, \tag{2.75}$$

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

It proves convenient to introduce two further moment matrices in addition to the covariance matrices. The "second moment" matrices will be defined as

$$\mathbf{R}_{xx} \equiv D^2(\mathbf{x}) = \langle \mathbf{x}\mathbf{x}^{\mathrm{T}} \rangle, \qquad \mathbf{R}_{xy} = \langle \mathbf{x}\mathbf{y}^{\mathrm{T}} \rangle,$$

that is, not taken about the means. Note $\mathbf{R}_{xy} = \mathbf{R}_{yx}^{\mathrm{T}}$, 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" (it is sometimes called the "error covariance") and is

$$\mathbf{P} \equiv D^2(\tilde{\mathbf{x}} - \mathbf{x}) = \langle(\tilde{\mathbf{x}} - \mathbf{x})(\tilde{\mathbf{x}} - \mathbf{x})^{\mathrm{T}}\rangle.$$

$\mathbf{P}$ is similar to $\mathbf{C}$, but differs in being taken about the true value, rather than about the mean value; the distinction can be very important.

If there are N variables, ξ_i, $i = 1, 2, \ldots, N$, they will be said to have an "N-dimensional jointly normal probability density" if it is of the form

$$p_{\xi_1,\ldots,\xi_N}(\Xi_1, \ldots, \Xi_N) = \frac{\exp\left[-\frac{1}{2}(\Xi - \mathbf{m})^{\mathrm{T}}\mathbf{C}_{\xi\xi}^{-1}(\Xi - \mathbf{m})\right]}{(2\pi)^{N/2}\sqrt{\det(\mathbf{C}_{\xi\xi})}}. \tag{2.76}$$

One finds $\langle\boldsymbol{\xi}\rangle = \mathbf{m}$, $\langle(\boldsymbol{\xi} - \mathbf{m})(\boldsymbol{\xi} - \mathbf{m})^{\mathrm{T}}\rangle = \mathbf{C}_{\xi\xi}$. Equation 2.73 is a special case for $N = 2$, and so the earlier forms are consistent with this general definition.

Positive definite symmetric matrices can be factored as

$$\mathbf{C}_{\xi\xi} = \mathbf{C}_{\xi\xi}^{\mathrm{T}/2}\mathbf{C}_{\xi\xi}^{1/2}, \tag{2.77}$$

which is called the "Cholesky decomposition," where $\mathbf{C}_{\xi\xi}^{1/2}$ is an upper triangular matrix (all zeros below the main diagonal) and non-singular.[16] It follows that the transformation (a rotation and stretching)

$$\mathbf{x} = \mathbf{C}_{\xi\xi}^{-\mathrm{T}/2}(\boldsymbol{\xi} - \mathbf{m}) \tag{2.78}$$

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

$$\begin{aligned} p_{x_1,\ldots,x_N}(X_1, \ldots, X_N) &= \frac{\exp -\frac{1}{2}\left(X_1^2 + \cdots + X_N^2\right)}{(2\pi)^{N/2}} \\ &= \frac{\exp\left(-\frac{1}{2}X_1^2\right)}{(2\pi)^{1/2}} \cdots \frac{\exp\left(-\frac{1}{2}X_N^2\right)}{(2\pi)^{1/2}}, \end{aligned} \tag{2.79}$$

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 called Gaussian "white noise," and has the property $\langle x_i x_j\rangle = 0,\ i \neq j$.[17]

2.3.6 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 (2.43). Let x be a random variable and consider the expected value of the function e^{ixt}:

$$\langle \mathrm{e}^{ixt} \rangle = \int_{-\infty}^{\infty} p_x(X)\, \mathrm{e}^{iXt} \mathrm{d}X \equiv \phi_x(t), \tag{2.80}$$

which is the Fourier transform of $p_x(X)$; $\phi_x(t)$ is 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? One starts by 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,

$$\phi_z(t) = \langle \mathrm{e}^{izt} \rangle = \left\langle \mathrm{e}^{i(x+y)t} \right\rangle = \langle \mathrm{e}^{ixt} \rangle \langle \mathrm{e}^{iyt} \rangle,$$

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

$$\phi_z(t) = \phi_x(t)\phi_y(t). \tag{2.81}$$

That is, the characteristic function for a sum of two independent variables is the product of the characteristic functions. The "convolution theorem"[18] 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. That is,

$$p_z(Z) = \int_{-\infty}^{\infty} p_x(r)\, p_y(Z - r)\, \mathrm{d}r. \tag{2.82}$$

We will not explore this relation in any detail, leaving the reader to pursue the subject in the references.[19] 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 Eq. (2.81) that if a variable ξ is defined as

$$\xi = x_1^2 + x_2^2 + \cdots + x_\nu^2, \tag{2.83}$$

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

$$p_\xi(\Xi) = \frac{\Xi^{\nu/2-1}}{2^{\nu/2}\Gamma\left(\frac{\nu}{2}\right)} \exp(-\Xi/2), \tag{2.84}$$

known as χ_ν^2 – "chi-square sub-ν." The chi-square probability density is central to the discussion of the expected sizes of vectors, such as $\tilde{\mathbf{n}}$, measured as $\tilde{\mathbf{n}}^{\mathrm{T}}\tilde{\mathbf{n}} = \|\tilde{\mathbf{n}}\|^2 = \sum_i \tilde{n}_i^2$ if the elements of $\tilde{\mathbf{n}}$ can be assumed to be independent and Gaussian. One has $\langle \xi \rangle = \nu$, $\langle (\xi - \langle \xi \rangle)^2 \rangle = 2\nu$. Equation (2.45) is the special case $\nu = 1$.

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 (2.76) and (2.79) have N degrees-of-freedom and (2.84) has ν of them. If a sample average (2.40) 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 non-diagonal? This question of how to interpret averages of correlated variables will be explicitly discussed later (p. 133).

Consider the special case of the sample variance Eq. (2.41) – which we claimed was biassed. The reason is that even if the sample values, x_i, are independent, 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, $\langle x\rangle_2 = (x_1+x_2)/2$. The two-term sample variance is

$$s^2 = \tfrac{1}{2}[(x_1-\langle x\rangle_2)^2 + (x_2-\langle x\rangle_2)^2],$$

but knowledge of x_1, and of the sample average, permits perfect prediction of $x_2 = 2\,\langle x\rangle_2 - x_1$. The second term in the sample variance as written is not independent of the first term, and thus there is just one independent piece of information in the two-term sample variance. To show it in general, assume without loss of generality that $\langle x\rangle = 0$, so that $\sigma^2 = \langle x^2\rangle$. The sample variance about the sample mean (which will not vanish) of independent samples is given by Eq. (2.41), and so

$$\begin{aligned}
\langle s^2\rangle &= \frac{1}{M}\sum_{i=1}^{M}\left\langle\left(x_i - \frac{1}{M}\sum_{j=1}^{M}x_j\right)\left(x_i - \frac{1}{M}\sum_{p=1}^{M}x_p\right)\right\rangle \\
&= \frac{1}{M}\sum_{i=1}^{M}\left\{\langle x_i^2\rangle - \frac{1}{M}\sum_{j=1}^{M}\langle x_ix_j\rangle - \frac{1}{M}\sum_{p=1}^{M}\langle x_ix_p\rangle + \frac{1}{M^2}\sum_{j=1}^{M}\sum_{p=1}^{M}\langle x_jx_p\rangle\right\} \\
&= \frac{1}{M}\sum_{i=1}^{M}\left\{\sigma^2 - \frac{\sigma^2}{M}\sum_{j}\delta_{ij} - \frac{\sigma^2}{M}\sum_{p}\delta_{ip} + \frac{\sigma^2}{M^2}\sum_{j}\sum_{p}\delta_{jp}\right\} \\
&= \frac{(M-1)}{M}\sigma^2 \neq \sigma^2.
\end{aligned}$$

Stationarity

Consider a vector random variable, with element x_i where the subscript i denotes a position in time or space. Then x_i, x_j are two different random variables – for example, the temperature at two different positions in a moving fluid, 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 all the underlying statistics are independent of i.[20] Specifically, $\langle x_i \rangle = \langle x_j \rangle \equiv \langle x \rangle$, $D^2(x_i) = D^2(x_j) = D^2(x)$, etc. Furthermore, x_i, x_j have a covariance

$$C_{xx}(i, j) = \langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \rangle, \tag{2.85}$$

that is, independent of i, j, and might as well be written $C_{xx}(|i - j|)$, depending only upon the difference $|i - j|$. The distance, $|i - j|$, is often called the "lag." $C_{xx}(|i - j|)$ is called the "autocovariance" of $\mathbf{x}$ or just the covariance, because x_i, x_j are now regarded as intrinsically the same process.[21] 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 for map-making and objective analysis. For stationary processes, all moments having the same $|i - j|$ are identical; it is seen that all diagonals of such a matrix $\{C_{xx}(i, j)\}$, are constant, for example, $\mathbf{C}_{\xi\xi}$ in Eq. (2.76). Matrices with constant diagonals are thus defined by the vector $C_{xx}(|i - j|)$, and are said to have a "Toeplitz form."

2.4 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 the discussion to discrete models and finite numbers of measurements (all that ever goes into a digital computer), almost everything can also be viewed as a form of ordinary least-squares, providing a much more intuitive approach than one through functional analysis. It is thus useful to go back and review what "everyone knows" about this most-familiar of all approximation methods.

2.4.1 Basic formulation

Consider the elementary problem motivated by the "data" shown in Fig. 2.2. t is supposed to be an independent variable, which could be time, a spatial coordinate, or just an index. Some physical variable, call it $\theta(t)$, perhaps temperature at a point in a laboratory tank, has been measured at coordinates $t = t_i$, $i = 1, 2, \ldots, M$, as depicted in the figure.

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

$$y(t_i) = \theta(t_i) + n(t_i) = a + bt_i + n(t_i), \tag{2.86}$$

where $n(t)$ is the inevitable measurement noise. The straight-line relationship might as well be referred to as a "model," as it represents the present conception of the data structure. We want to determine a, b.

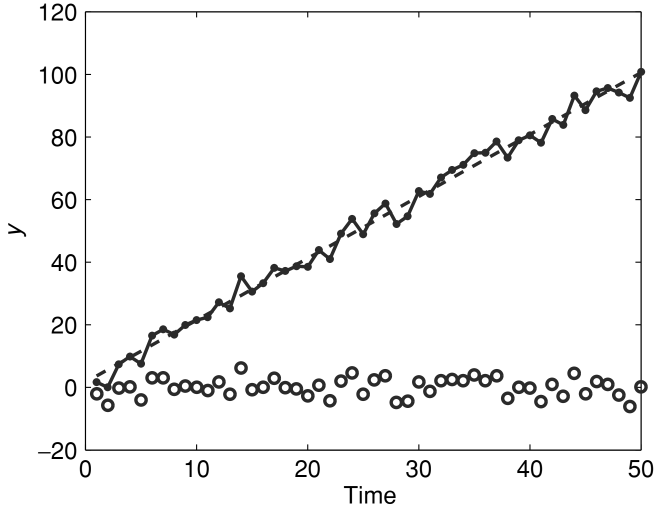

Figure 2.2 "Data" generated through the rule $y = 1 + 2t + n_t$, where $\langle n_t \rangle = 0$, $\langle n_i n_j \rangle = 9\delta_{ij}$, shown as solid dots connected by the solid line. The dashed line is the simple least-squares fit, $\tilde{y} = 1.69 \pm 0.83 + (1.98 \pm 0.03)t$. Residuals are plotted as open circles, and at least visually, show no obvious structure. Note that the fit is correct within its estimated standard errors. The sample variance of the estimated noise, not the theoretical value, was used for calculating the uncertainty.

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

$$\mathbf{E}\mathbf{x} + \mathbf{n} = \mathbf{y}, \tag{2.87}$$

where

$$\mathbf{E} = \left\{\begin{matrix} 1 & t_1 \\ 1 & t_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & t_M \end{matrix}\right\}, \quad \mathbf{x} = \begin{bmatrix} a \\ b \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y(t_1) \\ y(t_2) \\ \cdot \\ \cdot \\ y(t_M) \end{bmatrix}, \quad \mathbf{n} = \begin{bmatrix} n(t_1) \\ n(t_2) \\ \cdot \\ \cdot \\ n(t_M) \end{bmatrix}. \tag{2.88}$$

Equation sets like (2.87) are seen in many practical situations, including the ones described in Chapter 1. The matrix **E** in general represents arbitrarily complicated linear relations between the parameters **x**, and the observations **y**. In some real cases, it has many thousands of rows and columns. Its construction involves specifying what those relations are, and, in a very general sense, it requires a "model" of the data set. Unfortunately, the term "model" is used in a variety of other ways in this context, including statistical assumptions, and often for auxiliary relationships among the elements of **x** that are independent of those contained in **E**. To separate these difference usages, we will sometimes append various adjectives to the use ("statistical model," "exact relationships," etc.).

One sometimes sees (2.87) written as

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

or even

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

But Eq. (2.87) is preferable, because it explicitly recognizes that $\mathbf{n} = \mathbf{0}$ is exceptional. Sometimes, by happenstance or arrangement, one finds $M = N$ and that $\mathbf{E}$ has an inverse. But the obvious solution, $\mathbf{x} = \mathbf{E}^{-1}\mathbf{y}$, leads to the conclusion that $\mathbf{n} = \mathbf{0}$, which should be unacceptable if the $\mathbf{y}$ are the result of measurements. We will need to return to this case, but, for now, consider the commonplace problem where $M > N$.

Then, one often sees a "best possible" solution – defined as producing the smallest possible value of $\mathbf{n}^{\mathrm{T}}\mathbf{n}$, that is, the minimum of

$$J = \sum_{i=1}^{M} n_i^2 = \mathbf{n}^{\mathrm{T}}\mathbf{n} = (\mathbf{y} - \mathbf{Ex})^{\mathrm{T}}(\mathbf{y} - \mathbf{Ex}). \tag{2.89}$$

(Whether the smallest noise solution really is the best one is considered later.) In the special case of the straight-line model,

$$J = \sum_{i=1}^{M} (y_i - a - bt_i)^2. \tag{2.90}$$

J is an example of what is called an "objective," "cost" or "misfit" function.[22]

Taking the differential of (2.90) with respect to a, b or $\mathbf{x}$ (using (2.27)), and setting it to zero produces

$$\begin{aligned} \mathrm{d}J &= \sum_i \frac{\partial J}{\partial x_i}\mathrm{d}x_i = \left(\frac{\partial J}{\partial \mathbf{x}}\right)^{\mathrm{T}} \mathrm{d}\mathbf{x} \\ &= 2\mathrm{d}\mathbf{x}^{\mathrm{T}}(\mathbf{E}^{\mathrm{T}}\mathbf{y} - \mathbf{E}^{\mathrm{T}}\mathbf{Ex}) = 0. \end{aligned} \tag{2.91}$$

This equation is of the form

$$\mathrm{d}J = \sum a_i \mathrm{d}x_i = 0. \tag{2.92}$$

It is an elementary result of multivariable calculus that an extreme value (here a minimum) of J is found where $\mathrm{d}J = 0$. Because the x_i are free to vary independently, $\mathrm{d}J$ will vanish only if the coefficients of the $\mathrm{d}x_i$ are separately zero, or

$$\mathbf{E}^{\mathrm{T}}\mathbf{y} - \mathbf{E}^{\mathrm{T}}\mathbf{Ex} = \mathbf{0}. \tag{2.93}$$

That is,

$$\mathbf{E}^{\mathrm{T}}\mathbf{Ex} = \mathbf{E}^{\mathrm{T}}\mathbf{y}, \tag{2.94}$$

which are called the "normal equations." Note that Eq. (2.93) asserts that the columns of $\mathbf{E}$ are orthogonal (that is "normal") to $\mathbf{n} = \mathbf{y} - \mathbf{E}\mathbf{x}$. Making the sometimes-valid assumption that $(\mathbf{E}^{\mathrm{T}}\mathbf{E})^{-1}$ exists,

$$\tilde{\mathbf{x}} = (\mathbf{E}^{\mathrm{T}}\mathbf{E})^{-1}\mathbf{E}^{\mathrm{T}}\mathbf{y}. \tag{2.95}$$

Note on notation: Solutions to equations involving data will be denoted $\tilde{\mathbf{x}}$, to show that they are an estimate of the solution and not necessarily identical to the "true" one in a mathematical sense.

Second derivatives of J with respect to $\mathbf{x}$, make clear that we have a minimum and not a maximum. The relationship between 2.95 and the "correct" value is obscure. $\tilde{\mathbf{x}}$ can be substituted everywhere for $\mathbf{x}$ in Eq. (2.89), but usually the context makes clear the distinction between the calculated and true values. Figure 2.2 displays the fit along with the residuals,

$$\tilde{\mathbf{n}} = \mathbf{y} - \mathbf{E}\tilde{\mathbf{x}} = [\mathbf{I} - \mathbf{E}(\mathbf{E}^{\mathrm{T}}\mathbf{E})^{-1}\mathbf{E}^{\mathrm{T}}]\mathbf{y}. \tag{2.96}$$

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

$$\mathbf{H} = \mathbf{E}(\mathbf{E}^{\mathrm{T}}\mathbf{E})^{-1}\mathbf{E}^{\mathrm{T}} \tag{2.97}$$

occurs sufficiently often that it is worth a special symbol. Note the "idempotent" property $\mathbf{H}^2 = \mathbf{H}$. If the solution, $\tilde{\mathbf{x}}$, is substituted into the original equations, the result is

$$\mathbf{E}\tilde{\mathbf{x}} = \mathbf{H}\mathbf{y} = \tilde{\mathbf{y}}, \tag{2.98}$$

and

$$\tilde{\mathbf{n}}^{\mathrm{T}}\tilde{\mathbf{y}} = [(\mathbf{I} - \mathbf{H})\,\mathbf{y}]^{\mathrm{T}}\,\mathbf{H}\mathbf{y} = \mathbf{0}. \tag{2.99}$$

The residuals are orthogonal (normal) to the inferred noise-free "data" $\tilde{\mathbf{y}}$.

All of this is easy and familiar and applies to any set of simultaneous linear equations, not just the straight-line example. Before proceeding, let us apply some of the statistical machinery to understanding (2.95). Notice that no statistics were used in obtaining (2.95), 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 $\mathbf{n} = \mathbf{0}$. Assume further that $\langle\mathbf{n}\rangle = \mathbf{0}$ and that $\mathbf{R}_{nn} = \mathbf{C}_{nn} = \langle\mathbf{n}\mathbf{n}^{\mathrm{T}}\rangle$ is known. Then the expected value of $\tilde{\mathbf{x}}$ is

$$\langle\tilde{\mathbf{x}}\rangle = (\mathbf{E}^{\mathrm{T}}\mathbf{E})^{-1}\mathbf{E}^{\mathrm{T}}\mathbf{y}_0. \tag{2.100}$$

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,

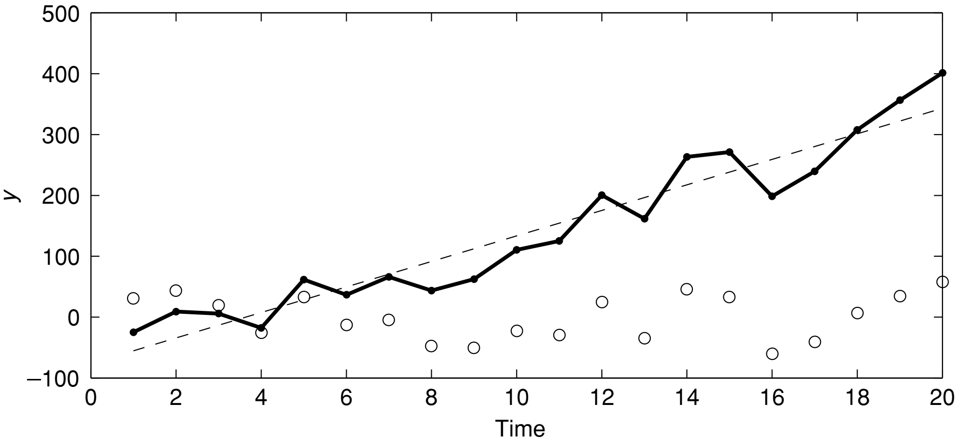

Figure 2.3 Here the "data" were generated from a quadratic rule, $y = 1 + t^2 + n(t)$, $\langle n^2 \rangle = 900$. Note that only the first 20 data points are used. An incorrect straight line fit was used resulting in $\tilde{y} = (-76.3 \pm 17.3) + (20.98 \pm 1.4)t$, which is incorrect, but the residuals, at least visually, do not appear unacceptable. At this point some might be inclined to claim the model has been "verified," or "validated."

and Eq. (2.95) provides an unbiassed estimate of the true solution, $\langle \tilde{\mathbf{x}} \rangle = \mathbf{x}$. A more transparent demonstration of this result will be given later in this chapter (see p. 103).

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 biassed. An example of such a fit is shown in Figs. 2.3 and 2.4. Such errors are conceptually distinguishable from the noise of observation, and are properly labeled "model errors."

Assume, however, that the correct model is being used, and therefore that $\langle \tilde{\mathbf{x}} \rangle = \mathbf{x}$. Then the uncertainty of the solution is

$$\begin{aligned} \mathbf{P} = \mathbf{C}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}} &= \langle (\tilde{\mathbf{x}} - \mathbf{x})(\tilde{\mathbf{x}} - \mathbf{x})^{\mathrm{T}} \rangle \\ &= (\mathbf{E}^{\mathrm{T}}\mathbf{E})^{-1}\mathbf{E}^{\mathrm{T}} \langle \mathbf{n}\mathbf{n}^{\mathrm{T}} \rangle \, \mathbf{E}(\mathbf{E}^{\mathrm{T}}\mathbf{E})^{-1} \\ &= (\mathbf{E}^{\mathrm{T}}\mathbf{E})^{-1} \, \mathbf{E}^{\mathrm{T}} \, \mathbf{R}_{nn} \, \mathbf{E}(\mathbf{E}^{\mathrm{T}}\mathbf{E})^{-1}. \end{aligned} \tag{2.101}$$

In the special case, $\mathbf{R}_{nn} = \sigma_n^2 \mathbf{I}$, that is, no correlation between the noise in different equations (white noise), Eq. (2.101) simplifies to

$$\mathbf{P} = \sigma_n^2 (\mathbf{E}^{\mathrm{T}}\mathbf{E})^{-1}. \tag{2.102}$$

If we are not confident that $\langle \tilde{\mathbf{x}} \rangle = \mathbf{x}$, perhaps because of doubts about the straight-line model, Eqs. (2.101) and (2.102) are still interpretable, but as $C_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}} =$

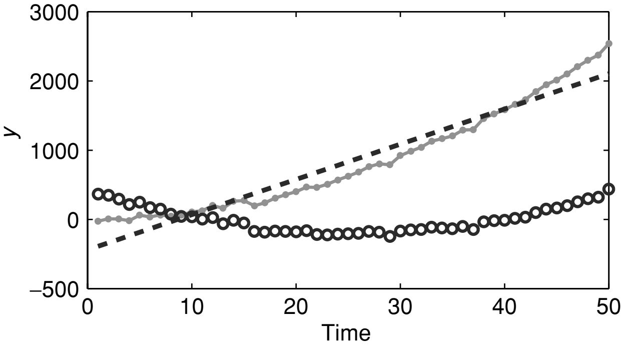

Figure 2.4 The same situation as in Fig. 2.3, except the series was extended to 50 points. Now the residuals (°) are visually structured, and one would have a powerful suggestion that some hypothesis (something about the model or data) is not correct. This straight-line fit should be rejected as being inconsistent with the assumption that the residuals are unstructured: the model has been "invalidated."

$D^2(\tilde{\mathbf{x}} - \langle\tilde{\mathbf{x}}\rangle)$, the covariance of $\tilde{\mathbf{x}}$. The "standard error" of $\tilde{x}_i$ is usually defined to be $\pm\sqrt{C_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}_{ii}}}$ 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 Fig. 2.2, $\tilde{\mathbf{x}}^{\mathrm{T}} = [\tilde{a}, \tilde{b}] = [1.69 \pm 0.83, 1.98 \pm 0.03]$, which are within one standard deviation of the true values, $[a, b] = [1, 2]$. If the noise in $\mathbf{y}$ is Gaussian, it follows that the probability density of $\tilde{\mathbf{x}}$ is also Gaussian, with mean $\langle\tilde{\mathbf{x}}\rangle$ and covariance $\mathbf{C}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}$. Of course, if $\mathbf{n}$ is not Gaussian, then the estimate won't be either, and one must be wary of the utility of the standard errors. A Gaussian, or other, assumption should be regarded as part of the model definition. The uncertainty of the residuals is

$$\begin{aligned}\mathbf{C}_{nn} &= \langle(\tilde{\mathbf{n}} - \langle\tilde{\mathbf{n}}\rangle)(\tilde{\mathbf{n}} - \langle\tilde{\mathbf{n}}\rangle)^{\mathrm{T}}\rangle = (\mathbf{I} - \mathbf{H})\,\mathbf{R}_{nn}\,(\mathbf{I} - \mathbf{H})^{\mathrm{T}} \\ &= \sigma_n^2\,(\mathbf{I} - \mathbf{H})^2 = \sigma_n^2\,(\mathbf{I} - \mathbf{H}),\end{aligned} \tag{2.103}$$

where zero-mean white noise was assumed, and $\mathbf{H}$ was defined in Eq. (2.97). The true noise, $\mathbf{n}$, was assumed to be white, but the estimated noise, $\tilde{\mathbf{n}}$, has a non-diagonal covariance and so in a formal sense does not have the expected structure. We return to this point below.

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 Fig. 2.5, the correct model used to generate the data was the same as in Fig. 2.2, but the noise level is very high. The parameters $[\tilde{a}, \tilde{b}]$ are numerically inexact, but consistent within one standard error with the correct values, which is all one can hope for.

In Fig. 2.3, a quadratic model $y = 1 + t^2 + n(t)$ was used to generate the numbers, with $\langle n^2\rangle = 900$. Using only the first 20 points, and fitting an incorrect model,

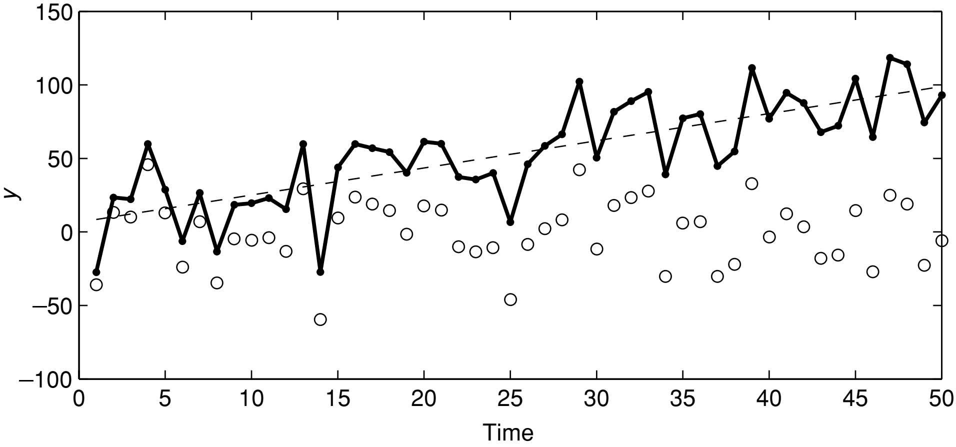

Figure 2.5 The same situation as in Fig. 2.2, $y = 1 + 2t$, except $\langle n^2 \rangle = 900$ to give very noisy data. Now the best-fitting straight line is $y = (6.62 \pm 6.50) + (1.85 \pm 0.22)\,t$, which includes the correct answer within one standard error. Note that the intercept value is indistinguishable from zero.

produces a reasonable straight-line fit to the data as shown. Modeling a quadratic field with a linear model produces a systematic or "model" error, which is not easy to detect here. One sometimes hears it said that "least-squares failed" in situations such as this one. But this conclusion shows a fundamental misunderstanding: least-squares did exactly what it was asked to do – to produce the best-fitting straight line to the data. Here, one might conclude that "the straight-line fit *is* consistent with the data." Such a conclusion is completely different from asserting that one has proven a straight-line fit correctly "explains" the data or, in modeler's jargon, that the model has been "verified" or "validated." If the outcome of the fit were sufficiently important, one might try more powerful tests on the $\tilde{n}_i$ than a mere visual comparison. Such tests might lead to rejection of the straight-line hypothesis; but even if the tests are passed, the model has *never* been verified: it has only been shown to be consistent with the available data.

If the situation remains unsatisfactory (perhaps one suspects the model is inadequate, but there are not enough data to produce sufficiently powerful tests), it can be very frustrating. But sometimes the only remedy is to obtain more data. So, in Fig. 2.4, the number of observations was extended to 50 points. Now, even visually, the $\tilde{n}_i$ are obviously structured, and one would almost surely reject any hypothesis that a straight line was an adequate representation of the data. *The model has been invalidated.* A quadratic rule, $y = a + bt + ct^2$, produces an acceptable solution (see Fig. 2.6).

One must always confirm, after the fact, that J, which is a direct function of the residuals, behaves as expected when the solution is substituted. In particular, its

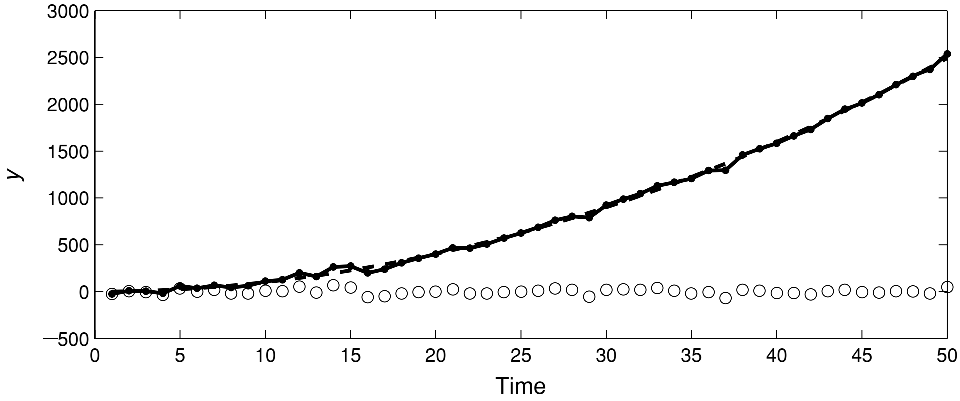

Figure 2.6 Same as Fig. 2.4, except a more complete model, $y = a + bt + ct^2$, was used, and which gives acceptable residuals.

expected value,

$$\langle J \rangle = \sum_{i}^{M} \left\langle n_i^2 \right\rangle = M - N, \tag{2.104}$$

assuming that the n_i have been scaled so that each has an expected value $\langle n_i^2 \rangle = 1$. That there are only $M - N$ independent terms in (2.104) follows from the N supposed-independent constraints linking the variables. For any particular solution, $\tilde{\mathbf{x}}$, $\tilde{\mathbf{n}}$, J will be a random variable, whose expectation is (2.104). Assuming the n_i are at least approximately Gaussian, J itself is the sum of $M - N$ independent χ_1^2 variables, and is therefore distributed in χ_{M-N}^2. One can and should make histograms of the individual n_i^2 to check them against the expected χ_1^2 probability density. This type of argument leads to the large literature on hypothesis testing.

As an illustration of the random behavior of residuals, 30 equations, $\mathbf{Ex} + \mathbf{n} = \mathbf{y}$, in 15 unknowns were constructed, such that $\mathbf{E}^{\mathrm{T}}\mathbf{E}$ was non-singular. Fifty different values of $\mathbf{y}$ were then constructed by generating 50 separate $\mathbf{n}$ using a pseudo-random number generator. An ensemble of 50 different solutions were calculated using (2.95), producing $50 \times 30 = 1500$ separate values of $\tilde{n}_i^2$. These are plotted in Fig. 2.7 and compared to χ_1^2. The corresponding value, $\tilde{J}^{(p)} = \sum_1^{30} \tilde{n}_i^2$, was found for each set of equations, and also plotted. A corresponding frequency function for $\tilde{J}^{(p)}$ is compared in Fig. 2.7 to χ_{15}^2, with reasonably good results. The empirical mean value of all $\tilde{J}_i$ is 14.3. Any particular solution may, completely correctly, produce individual residuals $\tilde{n}_i^2$ differing considerably from the mean of $\langle \chi_1^2 \rangle = 1$, and, similarly, their sums, $J^{(p)}$, may differ greatly from $\langle \chi_{15}^2 \rangle = 15$. But one can readily calculate the probability of finding a much larger or smaller value, and employ it to help evaluate the possibility that one has used an incorrect model.

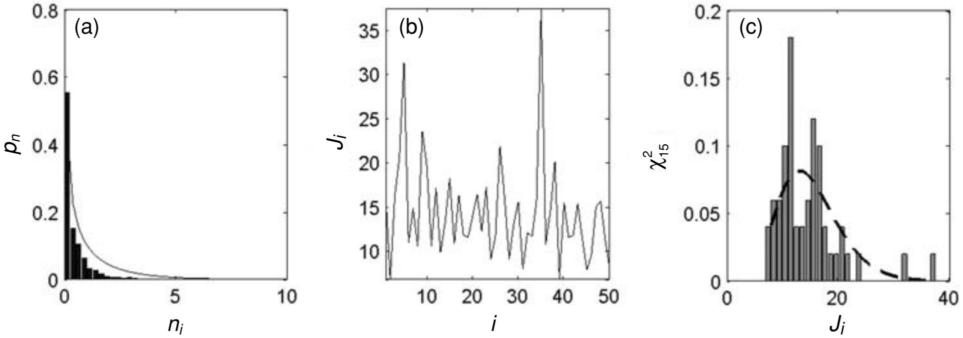

Figure 2.7 (a) χ^2_1 probability density and the empirical frequency function of *all* residuals, $\tilde{n}_i^2$, from 50 separate experiments for the simple least-squares solution of $\mathbf{Ex} + \mathbf{n} = \mathbf{y}$. There is at least rough agreement between the theoretical and calculated frequency functions. (b) The 50 values of J_i computed from the same experiments in (a). (c) The empirical frequency function for J_i as compared to the theoretical value of χ^2_{15} (dashed line). Tests exist (not discussed here) of the hypothesis that the calculated J_i are consistent with the theoretical distribution.

Visual tests for randomness of residuals have obvious limitations, and elaborate statistical tests in addition to the comparison with χ^2 exist to help determine objectively whether one should accept or reject the hypothesis that no significant structure remains in a sequence of numbers. Books on regression analysis[23] should be consulted for general methodologies. As an indication of what can be done, Fig. 2.8 shows the "sample autocorrelation"

$$\tilde{\phi}_{nn}(\tau) = \frac{1/M \sum_{i=1}^{M-|\tau|} \tilde{n}_i \tilde{n}_{i+\tau}}{1/M \sum_{i=1}^{M} \tilde{n}_i^2} \tag{2.105}$$

for the residuals of the fits shown in Figs. 2.4 and 2.6. For white noise,

$$\langle \tilde{\phi}(\tau) \rangle = \delta_{0\tau}, \tag{2.106}$$

and deviations of the estimated $\tilde{\phi}(t)$ from Eq. (2.106) can be used in simple tests. The adequate fit (Fig. 2.6) produces an autocorrelation of the residuals indistinguishable from a delta function at the origin, while the inadequate fit shows a great deal of structure that would lead to the conclusion that the residuals are too different from white noise to be acceptable. (Not all cases are this obvious.)

As already pointed out, the residuals of the least-squares fit cannot be expected to be precisely white noise. Because there are M relationships among the parameters of the problem (M equations), and the number of $\tilde{\mathbf{x}}$ elements determined is N, there are $M - N$ degrees of freedom in the determination of $\tilde{\mathbf{n}}$ and structures are imposed upon them. The failure, for this reason, of $\tilde{\mathbf{n}}$ strictly to be white noise, is generally only an issue in practice when $M - N$ becomes small compared to M.[24]

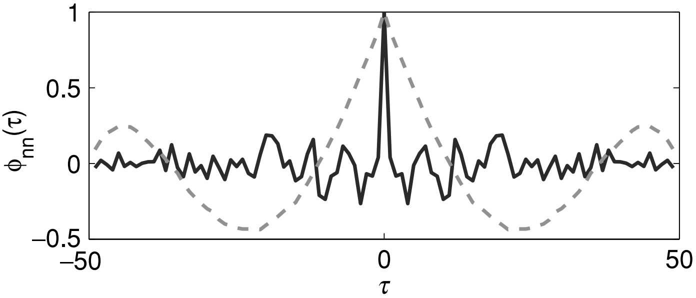

Figure 2.8 Autocorrelations of the estimated residuals in Figs. 2.4 (dashed line) and 2.6 (solid). The latter is indistinguishable, by statistical test, from a delta function at the origin, and so, with this test, the residuals are not distinguishable from white noise.

2.4.2 Weighted and tapered least-squares

The least-squares solution (Eqs. (2.95) and (2.96)) was derived by minimizing the objective function (2.89), in which each residual element is given equal 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,

$$J = \sum_i W_{ii}^{-1} n_i^2, \tag{2.107}$$

where W_{ii} are any numbers desired. The choice $W_{ii} = 1$, as used above, might be reasonable, but it is an arbitrary one that without further justification does not produce a solution with any special claim to significance. In the least-squares context, we are free to make any other reasonable choice, including demanding that some residuals should be much smaller than others – perhaps just to see if it is possible.

A general formalism is obtained by defining a diagonal weight matrix, $W = \text{diag}(W_{ii})$. Dividing each equation by $\sqrt{W_{ii}}$,

$$W_{ii}^{-\mathrm{T}/2} \sum_i E_{ij} x_j + W_{ii}^{-\mathrm{T}/2} n_i = W_{ii}^{-\mathrm{T}/2} y_i, \tag{2.108}$$

or

$$\begin{gathered} \mathbf{E}'\mathbf{x} + \mathbf{n}' = \mathbf{y}' \\ \mathbf{E}' = \mathbf{W}^{-\mathrm{T}/2}\mathbf{E}, \quad \mathbf{n}' = \mathbf{W}^{-\mathrm{T}/2}\mathbf{n}, \quad \mathbf{y}' = \mathbf{W}^{-\mathrm{T}/2}\mathbf{y} \end{gathered} \tag{2.109}$$

where we used the fact that the square root of a diagonal matrix is the diagonal matrix of element-by-element square roots. The operation in (2.108) or (2.109) is usually called "row scaling" because it operates on the rows of **E** (as well as on **n**, **y**).

For the new equations (2.109), the objective function,

$$\begin{aligned} J &= \mathbf{n}'^{\mathrm{T}}\mathbf{n}' = (\mathbf{y}' - \mathbf{E}'\mathbf{x})^{\mathrm{T}}(\mathbf{y}' - \mathbf{E}'\mathbf{x}) \\ &= \mathbf{n}^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{n} = (\mathbf{y} - \mathbf{E}\mathbf{x})^{\mathrm{T}}\mathbf{W}^{-1}(\mathbf{y} - \mathbf{E}\mathbf{x}), \end{aligned} \tag{2.110}$$

weights the residuals as desired. If, for some reason, **W** is non-diagonal, but symmetric and positive-definite, then it has a Cholesky decomposition (see p. 40) and

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

and (2.109) remains valid more generally.

The values $\tilde{\mathbf{x}}$, $\tilde{\mathbf{n}}$, minimizing (2.110), are

$$\boxed{\begin{aligned} \tilde{\mathbf{x}} &= (\mathbf{E}'^{\mathrm{T}}\mathbf{E}')^{-1}\mathbf{E}'^{\mathrm{T}}\mathbf{y}' = (\mathbf{E}^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{E})^{-1}\mathbf{E}^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{y}, \\ \tilde{\mathbf{n}} &= \mathbf{W}^{\mathrm{T}/2}\mathbf{n}' = [\mathbf{I} - \mathbf{E}(\mathbf{E}^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{E})^{-1}\mathbf{E}^{\mathrm{T}}\mathbf{W}^{-1}]\mathbf{y}, \qquad (2.111) \\ \mathbf{C}_{xx} &= (\mathbf{E}^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{E})^{-1}\mathbf{E}^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{R}_{nn}\mathbf{W}^{-1}\mathbf{E}(\mathbf{E}^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{E})^{-1}. \qquad (2.112) \end{aligned}}$$

Uniform diagonal weights are a special case. The rationale for choosing differing diagonal weights, or a non-diagonal **W**, is probably not very obvious to the reader. Often one chooses $\mathbf{W} = \mathbf{R}_{nn} = \{\langle n_i n_j \rangle\}$, that is, the weight matrix is chosen to be the expected second moment matrix of the residuals. Then

$$\langle \mathbf{n}'\mathbf{n}'^{\mathrm{T}} \rangle = \mathbf{I},$$

and Eq. (2.112) simplifies to

$$\mathbf{C}_{xx} = \left(\mathbf{E}^{\mathrm{T}}\mathbf{R}_{nn}^{-1}\mathbf{E}\right)^{-1}. \tag{2.113}$$

In this special case, the weighting (2.109) has a ready interpretation: The equations (and hence the residuals) are rotated and stretched so that in the new coordinate system of n'_i, the covariances are all diagonal and the variances are all unity. Under these circumstances, an objective function

$$J = \sum_i n_i'^2,$$

as used in the original form of least-squares (Eq. (2.89)), is a reasonable choice.

Example *Consider the system*

$$\begin{aligned} x_1 + x_2 + n_1 &= 1 \\ x_1 - x_2 + n_2 &= 2 \\ x_1 - 2x_2 + n_3 &= 4. \end{aligned}$$

Then if $\langle n_i \rangle = 0$, $\langle n_i^2 \rangle = \sigma^2$, *the least-squares solution is* $\tilde{\mathbf{x}} = [2.0, 0.5]^{\mathrm{T}}$. *Now suppose that*

$$\langle n_i n_j \rangle = \begin{Bmatrix} 1 & 0.99 & 0.98 \\ 0.99 & 1 & 0.99 \\ 0.98 & 0.99 & 4 \end{Bmatrix}.$$

Then from Eq. (2.112), $\tilde{\mathbf{x}} = [1.51, -0.48]^{\mathrm{T}}$. *Calculation of the two different solution uncertainties is left to the reader.*

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 an approximation procedure 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 whose significance is best understood after we examine a different, statistical, estimation procedure.

Whether the equations are scaled or not, the previous limitations of the simple least-squares solutions remain. In particular, we still have the problem that the solution may produce elements in $\tilde{\mathbf{x}}$, $\tilde{\mathbf{n}}$ whose relative values are not in accord with expected or reasonable behavior, and the solution uncertainty or variances could be unusably large, as the solution is determined, mechanically, and automatically, from combinations such as $(\mathbf{E}^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{E})^{-1}$. Operators like these are neither controllable nor very easy to understand; if any of the matrices are singular, they will not even exist.

It was long ago recognized that some control over the magnitudes of $\tilde{\mathbf{x}}$, $\tilde{\mathbf{n}}$, $\mathbf{C}_{xx}$ could be obtained in the simple least-squares context by modifying the objective function (2.107) to have an additional term:

$$J' = \mathbf{n}^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{n} + \gamma^2\mathbf{x}^{\mathrm{T}}\mathbf{x} \tag{2.114}$$

$$= (\mathbf{y} - \mathbf{E}\mathbf{x})^{\mathrm{T}}\mathbf{W}^{-1}(\mathbf{y} - \mathbf{E}\mathbf{x}) + \gamma^2\mathbf{x}^{\mathrm{T}}\mathbf{x}, \tag{2.115}$$

in which γ^2 is a positive constant.

If the minimum of (2.114) is sought by setting the derivatives with respect to $\mathbf{x}$ to zero, then we arrive at the following:

$$\tilde{\mathbf{x}} = (\mathbf{E}^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{E} + \gamma^2\mathbf{I})^{-1}\mathbf{E}^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{y} \tag{2.116}$$

$$\tilde{\mathbf{n}} = \mathbf{y} - \mathbf{E}\tilde{\mathbf{x}} \tag{2.117}$$

$$\mathbf{C}_{xx} = (\mathbf{E}^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{E} + \gamma^2\mathbf{I})^{-1}\mathbf{E}^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{R}_{nn}\mathbf{W}^{-1}\mathbf{E}(\mathbf{E}^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{E} + \gamma^2\mathbf{I})^{-1}. \tag{2.118}$$

By letting $\gamma^2 \to 0$, the solution to (2.111) and (2.112) is recovered, and if $\gamma^2 \to \infty$, $\|\tilde{\mathbf{x}}\|_2 \to 0$, $\tilde{\mathbf{n}} \to \mathbf{y}$; γ^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 γ^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,

$$\langle\tilde{\mathbf{x}}\rangle = [\mathbf{E}^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{E} + \gamma^2\mathbf{I}]^{-1}\mathbf{E}^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{y}_0. \tag{2.119}$$

If the true solution is believed to be (2.100), then this new solution is biassed. But the variance of $\tilde{\mathbf{x}}$ has been reduced, (2.118), by introduction of $\gamma^2 > 0$ – that is, the acceptance of a bias reduces the variance, possibly very greatly. Equations (2.116) and (2.117) are sometimes known as the "tapered least-squares" solution, a label whose implication becomes clear later. $\mathbf{C}_{nn}$, which is not displayed, is readily found by direct computation as in Eq. (2.103).

The most basic, and commonly seen, form of this solution assumes that $\mathbf{W} = \mathbf{R}_{nn} = \mathbf{I}$, so that

$$\boxed{\begin{aligned}\tilde{\mathbf{x}} &= (\mathbf{E}^{\mathrm{T}}\mathbf{E} + \gamma^2\mathbf{I})^{-1}\mathbf{E}^{\mathrm{T}}\mathbf{y}, && (2.120)\\ \mathbf{C}_{xx} &= (\mathbf{E}^{\mathrm{T}}\mathbf{E} + \gamma^2\mathbf{I})^{-1}\mathbf{E}^{\mathrm{T}}\mathbf{E}(\mathbf{E}^{\mathrm{T}}\mathbf{E} + \gamma^2\mathbf{I})^{-1}. && (2.121)\end{aligned}}$$

A physical motivation for the modified objective function (2.114) is obtained by noticing that a preference for a bounded $\|\mathbf{x}\|$ is easily produced by adding an equation set, $\mathbf{x} + \mathbf{n}_1 = \mathbf{0}$, so that the combined set is

$$\mathbf{E}\mathbf{x} + \mathbf{n} = \mathbf{y}, \tag{2.122}$$

$$\mathbf{x} + \mathbf{n}_1 = \mathbf{0}, \tag{2.123}$$

or

$$\begin{gathered}\mathbf{E}_1\mathbf{x} + \mathbf{n}_2 = \mathbf{y}_2,\\ \mathbf{E}_1 = \begin{Bmatrix}\mathbf{E}\\ \gamma\mathbf{I}\end{Bmatrix}, \quad \mathbf{n}_2^{\mathrm{T}} = \left[\mathbf{n}^{\mathrm{T}}\ \gamma\mathbf{n}_1^{\mathrm{T}}\right], \quad \mathbf{y}_2^{\mathrm{T}} = [\mathbf{y}^{\mathrm{T}}\ \mathbf{0}^{\mathrm{T}}],\end{gathered} \tag{2.124}$$

in which $\gamma > 0$ expresses a preference for fitting the first or second sets more closely. Then J in Eq. (2.114) becomes the natural objective function to use. A preference that $\mathbf{x} \approx \mathbf{x}_0$ is readily imposed instead, with an obvious change in (2.114) or (2.123).

Note the important points, to be shown later, that the matrix inverses in Eqs. (2.116) and (2.117) will *always* exist, as long as $\gamma^2 > 0$, and that the expressions remain valid even if $M < N$. Tapered least-squares produces 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 $\tilde{x}_i$.

To gain some of that control, we can further generalize the objective function by introducing another non-singular $N \times N$ weight matrix, $\mathbf{S}$ (which is usually symmetric), and

$$J = \mathbf{n}^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{n} + \mathbf{x}^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{x} \tag{2.125}$$

$$= (\mathbf{y} - \mathbf{E}\mathbf{x})^{\mathrm{T}}\mathbf{W}^{-1}(\mathbf{y} - \mathbf{E}\mathbf{x}) + \mathbf{x}^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{x}, \tag{2.126}$$

for which Eq. (2.114) is a special case. Setting the derivatives with respect to **x** to zero results in the following:

$$\tilde{\mathbf{x}} = (\mathbf{E}^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{E} + \mathbf{S}^{-1})^{-1}\mathbf{E}^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{y}, \tag{2.127}$$

$$\tilde{\mathbf{n}} = \mathbf{y} - \mathbf{E}\tilde{\mathbf{x}}, \tag{2.128}$$

$$\mathbf{C}_{xx} = (\mathbf{E}^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{E} + \mathbf{S}^{-1})^{-1}\mathbf{E}^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{R}_{nn}\mathbf{W}^{-1}\mathbf{E}(\mathbf{E}^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{E} + \mathbf{S}^{-1})^{-1}, \tag{2.129}$$

which are identical to Eqs. (2.116)–(2.118) with $\mathbf{S}^{-1} = \gamma^2\mathbf{I}$. $\mathbf{C}_{xx}$ simplifies if $\mathbf{R}_{nn} = \mathbf{W}$.

Suppose that **S**, **W** are positive definite and symmetric and thus have Cholesky decompositions. Then employing both matrices directly on the equations, $\mathbf{E}\mathbf{x} + \mathbf{n} = \mathbf{y}$,

$$\mathbf{W}^{-\mathrm{T}/2}\mathbf{E}\mathbf{S}^{-\mathrm{T}/2}\mathbf{S}^{\mathrm{T}/2}\mathbf{x} + \mathbf{W}^{-\mathrm{T}/2}\mathbf{n} = \mathbf{W}^{-\mathrm{T}/2}\mathbf{y} \tag{2.130}$$

$$\mathbf{E}'\mathbf{x}' + \mathbf{n}' = \mathbf{y}' \tag{2.131}$$

$$\mathbf{E}' = \mathbf{W}^{-\mathrm{T}/2}\mathbf{E}\mathbf{S}^{\mathrm{T}/2},\ \mathbf{x}' = \mathbf{S}^{-\mathrm{T}/2}\mathbf{x},\ \mathbf{n}' = \mathbf{W}^{-\mathrm{T}/2}\mathbf{n},\ \mathbf{y}' = \mathbf{W}^{-\mathrm{T}/2}\mathbf{y} \tag{2.132}$$

The use of **S** in this way is called “column scaling” because it weights the columns of **E**. With Eqs. (2.131) the obvious objective function is

$$J = \mathbf{n}'^{\mathrm{T}}\mathbf{n}' + \mathbf{x}'^{\mathrm{T}}\mathbf{x}', \tag{2.133}$$

which is identical to Eq. (2.125) in the original variables, and the solution must be that in Eqs. (2.127)–(2.129).

Like **W**, one is completely free to choose **S** as one pleases. A common example is to write, where **F** is $N \times N$, that

$$\mathbf{S} = \mathbf{F}^{\mathrm{T}}\mathbf{F}$$

$$\mathbf{F} = \gamma^2 \left\{\begin{matrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & \cdots & 0 & 1 \end{matrix}\right\}, \tag{2.134}$$

The effect of (2.134) is to minimize a term $\gamma^2 \sum_i (x_i - x_{i+1})^2$, which can be regarded as a “smoothest” solution, and, using γ^2 to trade smoothness against the size of $\|\tilde{\mathbf{n}}\|_2$, **F** is obtained from the Cholesky decomposition of **S**.

By invoking the matrix inversion lemma, an alternate form for Eqs. (2.127)–(2.129) is found:

$$\boxed{\begin{aligned}\tilde{\mathbf{x}} &= \mathbf{S}\mathbf{E}^{\mathrm{T}}(\mathbf{E}\mathbf{S}\mathbf{E}^{\mathrm{T}}+\mathbf{W})^{-1}\mathbf{y}, && (2.135)\\ \tilde{\mathbf{n}} &= \mathbf{y}-\mathbf{E}\tilde{\mathbf{x}}, && (2.136)\\ \mathbf{C}_{xx} &= \mathbf{S}\mathbf{E}^{\mathrm{T}}(\mathbf{E}\mathbf{S}\mathbf{E}^{\mathrm{T}}+\mathbf{W})^{-1}\mathbf{R}_{nn}(\mathbf{E}\mathbf{S}\mathbf{E}^{\mathrm{T}}+\mathbf{W})^{-1}\mathbf{E}\mathbf{S}. && (2.137)\end{aligned}}$$

A choice of which form to use is sometimes made on the basis of the dimensions of the matrices being inverted. Note again that $\mathbf{W}=\mathbf{R}_{nn}$ is a special case.

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 over determined 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 equations. Another way to make the point is to re-write Eq. (2.87) without distinction between $\mathbf{x}$, $\mathbf{n}$ as

$$\mathbf{E}_1\boldsymbol{\xi} = \mathbf{y}, \tag{2.138}$$

$$\mathbf{E}_1 = \{\mathbf{E}, \mathbf{I}_M\}, \quad \boldsymbol{\xi}^{\mathrm{T}} = [\mathbf{x}, \mathbf{n}]^{\mathrm{T}}. \tag{2.139}$$

A combined weight matrix,

$$\mathbf{S}_1 = \begin{Bmatrix}\mathbf{S} & \mathbf{0}\\ \mathbf{0} & \mathbf{W}\end{Bmatrix}, \tag{2.140}$$

would be used, and any distinction between the $\mathbf{x}$, $\mathbf{n}$ solution elements is suppressed. Equation (2.138) describes a formally underdetermined system, derived from the formally over determined observed one. This identity leads us to the problem of formal underdetermination in the next section.

In general with least-squares problems, the solution sought can be regarded as any of the following equivalents:

1. The $\tilde{\mathbf{x}}$, $\tilde{\mathbf{n}}$ satisfying

$$\mathbf{E}\mathbf{x}+\mathbf{n} = \mathbf{y}. \tag{2.141}$$

2. $\tilde{\mathbf{x}}$, $\tilde{\mathbf{n}}$ satisfying the normal equations arising from J (2.125).
3. $\tilde{\mathbf{x}}$, $\tilde{\mathbf{n}}$ producing the minimum of J in (2.125).

The point of this list lies with item (3): algorithms exist to find minima of functions by deterministic methods ("go downhill" from an initial guess),[25] or stochastic search methods (Monte Carlo) or even, conceivably, through a shrewd guess by the investigator. If an acceptable minimum of J is found, by whatever means, it is an acceptable solution (subject to further testing, and the possibility that there is more than one such solution). Search methods become essential for the non-linear problems taken up later.

2.4.3 Underdetermined systems and Lagrange multipliers

What does one do when the number, M, of equations is less than the number, N, of unknowns and no more observations are possible? We have seen that the claim that a problem involving observations is ever over determined is misleading – because each equation or observation always has a noise unknown, but to motivate some of what follows, it is helpful to first pursue the conventional approach.

One often attempts when $M < N$ 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. For example, let $p_i(t)$, $i = 0, 1, \ldots,$ be a set of polynomials, e.g., Chebyschev or Laguerre, etc. Consider data produced from the formula

$$y(t) = 1 + a_M p_M(t) + n(t), \tag{2.142}$$

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

$$y = \sum_{j=0}^{Q} a_j p_j(t), \quad Q < M \tag{2.143}$$

may give a good, even perfect fit; but it would be incorrect. The reduction in model parameters in such a case biasses the result, perhaps hopelessly so. One is better off retaining the underdetermined system and making inferences concerning the possible values of a_i rather than using the form (2.143), in which any possibility of learning something about a_M has been eliminated.

Example *Consider a tracer problem, not unlike those encountered in medicine, hydrology, oceanography, etc. A box (Fig. 1.2) is observed to contain a steady tracer concentration C_0, and is believed to be fed at the rates J_1, J_2 from two reservoirs each with tracer concentration of C_1, C_2 respectively. One seeks to determine J_1, J_2. Tracer balance is*

$$J_1C_1 + J_2C_2 - J_0C_0 = 0, \tag{2.144}$$

where J_0 is the rate at which fluid is removed. Mass balance requires that

$$J_1 + J_2 - J_0 = 0. \tag{2.145}$$

Evidently, there are but two equations in three unknowns (and a perfectly good solution would be $J_1 = J_2 = J_3 = 0$); but, as many have noticed, we can nonetheless, determine the relative fraction of the fluid coming from each reservoir. Divide both equations through by J_0,

$$\frac{J_1}{J_0}C_1 + \frac{J_2}{J_0}C_2 = C_0$$
$$\frac{J_1}{J_0} + \frac{J_2}{J_0} = 1,$$

producing two equations in two unknowns, $J_1/J_0,\ J_2/J_0$, which has a unique stable solution (noise is being ignored). Many examples can be given of such calculations in the literature – determining the flux ratios – apparently definitively. But suppose the investigator is suspicious that there might be a third reservoir with tracer concentration C_3. Then the equations become

$$\frac{J_1}{J_0}C_1 + \frac{J_2}{J_0}C_2 + \frac{J_3}{J_0}C_3 = C_0$$
$$\frac{J_1}{J_0} + \frac{J_2}{J_0} + \frac{J_3}{J_0} = 1,$$

which are now underdetermined with two equations in three unknowns. If it is obvious that no such third reservoir exists, then the reduction to two equations in two unknowns is the right thing to do. But if there is even a suspicion of a third reservoir (or more), one should solve these equations with one of the methods we will develop – permitting construction and understanding of all possible solutions.

In general terms, parameter reduction can lead to model errors, that is, bias errors, which can produce wholly illusory results.[26] A common situation, particularly in problems involving tracer movements in groundwater, ocean, or atmosphere, is fitting a one or two-dimensional model to data that represent a fully three-dimensional field. The result may be apparently pleasing, but possibly completely erroneous.

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

$$\mathrm{d}J = \mathrm{d}\mathbf{x}^{\mathrm{T}} \frac{\partial J}{\partial \mathbf{x}} = 2\mathbf{x}^{\mathrm{T}}\mathrm{d}\mathbf{x} = 0, \tag{2.146}$$

but, unlike the case in Eq. (2.91), the coefficients of the individual $\mathrm{d}x_i$ can no longer be separately set to zero (i.e., $\mathbf{x} = 0$ is an incorrect solution) because the $\mathrm{d}x_i$ no longer vary independently, but are restricted to values satisfying $\mathbf{Ax} = \mathbf{b}$. 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

$$\begin{bmatrix} x_1 \\ \vdots \\ x_M \end{bmatrix} = \begin{bmatrix} \xi_1(x_{M+1}, \ldots, x_N) \\ \vdots \\ \xi_M(x_{M+1}, \ldots, x_N) \end{bmatrix}.$$

Then the first M elements of x_i may be eliminated, and the objective function becomes

$$J = [\xi_1(x_{M+1}, \ldots, x_N)^2 + \cdots + \xi_M(x_{M+1}, \ldots, x_N)^2] + \left[x_{M+1}^2 + \cdots + x_N^2\right],$$

in which the remaining x_i, $M + i = 1, 2, \ldots, N$ are independently varying. In the present case, one can choose (arbitrarily) the first M unknowns, $\mathbf{q} = [x_i]$, and define the last $N - M$ unknowns $\mathbf{r} = [x_i]$, $i = N - M + 1, \ldots, N$, and rewrite the equations as

$$\{\mathbf{A}_1\ \mathbf{A}_2\} \begin{bmatrix} \mathbf{q} \\ \mathbf{r} \end{bmatrix} = \mathbf{b} \tag{2.147}$$

where $\mathbf{A}_1$ is $M \times M$, $\mathbf{A}_2$ is $M \times (N - M)$. Then solving the first set for $\mathbf{q}$,

$$\mathbf{q} = \mathbf{b} - \mathbf{A}_2\mathbf{r}. \tag{2.148}$$

$\mathbf{q}$ can be eliminated from J leaving an unconstrained minimization problem in the independent variables, $\mathbf{r}$. If $\mathbf{A}_1^{-1}$ does not exist, one can try any other subset of the x_i to eliminate until a suitable group is found. This approach is completely correct, but finding an explicit solution for L elements of $\mathbf{x}$ in terms of the remaining ones may be difficult or inconvenient.

Example *Solve*

$$x_1 - x_2 + x_3 = 1,$$

for the solution of minimum norm. The objective function is $J = x_1^2 + x_2^2 + x_3^2$. *With one equation, one variable can be eliminated. Arbitrarily choosing* $x_1 = 1 +$

$x_2 - x_3, J = (1 + x_2 - x_3)^2 + x_2^2 + x_3^2$. *$x_2, x_3$ are now independent variables, and the corresponding derivatives of J can be independently set to zero.*

Example *A somewhat more interesting example involves two equations in three unknowns:*

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

and we choose to find a solution minimizing,

$$J = x_1^2 + x_2^2 + x_3^2.$$

Solving for two unknowns x_1, x_2 from

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

produces $x_2 = -1/2, x_1 = 3/2 - x_3$, and then

$$J = (3/2 - x_3)^2 + 1/4 + x_3^2,$$

whose minimum with respect to x_3 (the only remaining variable) is $x_3 = 3/4$, and the full solution is

$$x_1 = \frac{3}{4}, \quad x_2 = -\frac{1}{2}, \quad x_3 = \frac{3}{4}.$$

Lagrange multipliers and adjoints

When it is inconvenient to find such an explicit representation by 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

$$
\begin{aligned}
J' &= J - 2\boldsymbol{\mu}^{\mathrm{T}}(\mathbf{A}\mathbf{x} - \mathbf{b}) \\
&= \mathbf{x}^{\mathrm{T}}\mathbf{x} - 2\boldsymbol{\mu}^{\mathrm{T}}(\mathbf{A}\mathbf{x} - \mathbf{b}),
\end{aligned}
\tag{2.149}
$$

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.[27] Consider first a very simple example of one equation in two unknowns,

$$x_1 - x_2 = 1. \tag{2.150}$$

We seek the minimum norm solution,

$$J = x_1^2 + x_2^2, \tag{2.151}$$

subject to Eq. (2.150). The differential

$$\mathrm{d}J = 2x_1\mathrm{d}x_1 + 2x_2\mathrm{d}x_2 = 0 \tag{2.152}$$

leads to the unacceptable solution $x_1 = x_2 = 0$ if we incorrectly set the coefficients of $\mathrm{d}x_1$, $\mathrm{d}x_2$ to zero. Consider instead a modified objective function

$$J' = J - 2\mu\,(x_1 - x_2 - 1)\,, \tag{2.153}$$

where μ is unknown. The differential of J' is

$$\mathrm{d}J' = 2x_1\mathrm{d}x_1 + 2x_2\mathrm{d}x_2 - 2\mu\,(\mathrm{d}x_1 - \mathrm{d}x_2) - 2\,(x_1 - x_2 - 1)\,\mathrm{d}\mu = 0, \tag{2.154}$$

or

$$\mathrm{d}J'/2 = \mathrm{d}x_1\,(x_1 - \mu) + \mathrm{d}x_2\,(x_2 + \mu) - \mathrm{d}\mu\,(x_1 - x_2 - 1) = 0. \tag{2.155}$$

We are free to choose $x_1 = \mu$, which kills off the differential involving $\mathrm{d}x_1$. But then only the differentials $\mathrm{d}x_2, \mathrm{d}\mu$ remain; as they can vary independently, their coefficients must vanish separately, and we have

$$x_2 = -\mu \tag{2.156}$$

$$x_1 - x_2 = 1. \tag{2.157}$$

Note that the second of these recovers the original equation. Substituting $x_1 = \mu$, we have $2\mu = 1$, or $\mu = 1/2$, and $x_1 = 1/2, x_2 = -1/2,\ J = 0.5$, and one can confirm that this is indeed the "constrained" minimum. (A "stationary" value of J' was found, not an absolute minimum value, because J' is no longer necessarily positive; it has a saddle point, which we have found.)

Before writing out the general case, note the following question: Suppose the constraint equation was changed to

$$x_1 - x_2 = \Delta. \tag{2.158}$$

How much would J change as Δ is varied? With variable Δ, (2.154) becomes

$$\mathrm{d}J' = 2\mathrm{d}x_1\,(x_1 - \mu) + 2\mathrm{d}x_2\,(x_2 + \mu) - 2\mathrm{d}\mu\,(x_1 - x_2 - \Delta) + 2\mu\mathrm{d}\Delta. \tag{2.159}$$

But the first three terms on the right vanish, and hence

$$\frac{\partial J'}{\partial \Delta} = 2\mu = \frac{\partial J}{\partial \Delta}, \tag{2.160}$$

because $J = J'$ at the stationary point (from (2.158)). *Thus 2μ is the sensitivity of the objective function J to perturbations in the right-hand side of the* *constraint equation*. If Δ is changed from 1 to 1.2, it can be confirmed that the approximate change in the value of J is 0.2, as one deduces immediately from Eq. (2.160). Keep in mind, however, that this sensitivity corresponds to infinitesimal perturbations.

We now develop this method generally. Reverting to Eq. (2.149),

$$\begin{aligned} \mathrm{d}J' &= \mathrm{d}J - 2\boldsymbol{\mu}^{\mathrm{T}}\mathbf{A}\mathrm{d}\mathbf{x} - 2\left(\mathbf{A}\mathbf{x}-\mathbf{b}\right)^{\mathrm{T}}\mathrm{d}\boldsymbol{\mu} \\ &= \left(\frac{\partial J}{\partial x_1} - 2\boldsymbol{\mu}^{\mathrm{T}}\mathbf{a}_1\right)\mathrm{d}x_1 + \left(\frac{\partial J}{\partial x_2} - 2\boldsymbol{\mu}^{\mathrm{T}}\mathbf{a}_2\right)\mathrm{d}x_2 + \cdots \qquad (2.161) \\ &\quad + \left(\frac{\partial J}{\partial x_N} - 2\boldsymbol{\mu}^{\mathrm{T}}\mathbf{a}_N\right)\mathrm{d}x_N - 2(\mathbf{A}\mathbf{x}-\mathbf{b})^{\mathrm{T}}\mathrm{d}\boldsymbol{\mu} \\ &= (2x_1 - 2\boldsymbol{\mu}^{\mathrm{T}}\mathbf{a}_1)\mathrm{d}x_1 + (2x_2 - 2\boldsymbol{\mu}^{\mathrm{T}}\mathbf{a}_2)\mathrm{d}x_2 + \cdots \qquad (2.162) \\ &\quad + (2x_N - 2\boldsymbol{\mu}^{\mathrm{T}}\mathbf{a}_N)\mathrm{d}x_N - 2(\mathbf{A}\mathbf{x}-\mathbf{b})^{\mathrm{T}}\mathrm{d}\boldsymbol{\mu} = 0 \end{aligned}$$

Here the $\mathbf{a}_i$ are the corresponding columns of $\mathbf{A}$. The coefficients of the first M differentials $\mathrm{d}x_i$ can be set to zero by assigning $x_i = \boldsymbol{\mu}^{\mathrm{T}}\mathbf{a}_i$, leaving $N - M$ differentials $\mathrm{d}x_i$ whose coefficients must separately vanish (hence they *all* vanish, but for two separate reasons), plus the coefficient of the $M - \mathrm{d}\mu_i$, which must also vanish separately. This recipe produces, from Eq. (2.162),

$$\frac{1}{2}\frac{\partial J'}{\partial \mathbf{x}} = \mathbf{x} - \mathbf{A}^{\mathrm{T}}\boldsymbol{\mu} = 0, \qquad (2.163)$$

$$\frac{1}{2}\frac{\partial J'}{\partial \boldsymbol{\mu}} = \mathbf{A}\mathbf{x} - \mathbf{b} = \mathbf{0}, \qquad (2.164)$$

where the first equation set is the result of the vanishing of the coefficients of $\mathrm{d}x_i$ and the second, which is the original set of equations, arises from the vanishing of the coefficients of the $\mathrm{d}\mu_i$. 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 M unknown μ_i. The first set is N equations for $\boldsymbol{\mu}$ in terms of $\mathbf{x}$, and the second set is M equations in $\mathbf{x}$ in terms of $\mathbf{y}$. Taken together, these are $M + N$ equations in $M + N$ unknowns, and hence just-determined no matter what the ratio of M to N.

Equation (2.163) is

$$\mathbf{A}^{\mathrm{T}}\boldsymbol{\mu} = \mathbf{x}, \qquad (2.165)$$

and, substituting for $\mathbf{x}$ into (2.164),

$$\begin{aligned} \mathbf{A}\mathbf{A}^{\mathrm{T}}\boldsymbol{\mu} &= \mathbf{b}, \\ \tilde{\mu} &= (\mathbf{A}\mathbf{A}^{\mathrm{T}})^{-1}\mathbf{b}, \qquad (2.166) \end{aligned}$$

assuming the inverse exists, and that

$$\tilde{\mathbf{x}} = \mathbf{A}^{\mathrm{T}}(\mathbf{A}\mathbf{A}^{\mathrm{T}})^{-1}\mathbf{b} \tag{2.167}$$

$$\tilde{\mathbf{n}} = \mathbf{0} \tag{2.168}$$

$$\mathbf{C}_{xx} = 0. \tag{2.169}$$

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

Equations (2.167)–(2.169) are the classical solution, 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[28] and where the term proportional to $\boldsymbol{\mu}$ necessarily vanishes. The operator $\mathbf{A}^{\mathrm{T}}(\mathbf{A}\mathbf{A}^{\mathrm{T}})^{-1}$ is sometimes called a "Moore–Penrose inverse."

Equation (2.165) for $\boldsymbol{\mu}$ in terms of $\mathbf{x}$ involves the coefficient matrix $\mathbf{A}^{\mathrm{T}}$. An intimate connection exists between matrix transposes and adjoints of differential equations (see the appendix to this chapter), and thus $\boldsymbol{\mu}$ is sometimes called the "adjoint solution," with $\mathbf{A}^{\mathrm{T}}$ defining the "adjoint model"[29] in Eq. (2.165), and $\mathbf{x}$ acting as a forcing term. The original $\mathbf{Ax} = \mathbf{b}$ were assumed formally underdetermined, and thus the adjoint model equations in (2.165) are necessarily formally over determined.

Example *The last example now using matrix vector notation is*

$$\mathbf{A} = \left\{\begin{matrix} 1 & 1 & 1 \\ 1 & -1 & 1 \end{matrix}\right\}, \mathbf{b} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

$$J = \mathbf{x}^{\mathrm{T}}\mathbf{x} - 2\boldsymbol{\mu}^{\mathrm{T}}\left(\mathbf{Ax} - \mathbf{b}\right)$$

$$\frac{\mathrm{d}}{\mathrm{d}\mathbf{x}}(\mathbf{x}^{\mathrm{T}}\mathbf{x} - 2\boldsymbol{\mu}^{\mathrm{T}}\left(\mathbf{Ax} - \mathbf{b}\right)) = 2\mathbf{x} - 2\mathbf{A}^{\mathrm{T}}\boldsymbol{\mu} = \mathbf{0}$$

$$\mathbf{Ax} = \mathbf{b}$$

$$\mathbf{x} = \mathbf{A}^{\mathrm{T}}(\mathbf{A}\mathbf{A}^{\mathrm{T}})^{-1}\mathbf{b}$$

$$\mathbf{x} = [3/4, -1/2, 3/4]^{\mathrm{T}}.$$

Example *Write out* J':

$$J' = x_1^2 + x_2^2 + x_3^2 - 2\mu_1\left(x_1 + x_2 + x_3 - 1\right) - 2\mu_2\left(x_1 - x_2 + x_3 - 2\right)$$

$$\begin{aligned} \mathrm{d}J' = {} & (2x_1 - 2\mu_1 - 2\mu_2)\mathrm{d}x_1 + (2x_2 - 2\mu_1 + 2\mu_2)\mathrm{d}x_2 + (2x_3 - 2\mu_1 - 2\mu_2)\mathrm{d}x_3 \\ & + \left(-2x_1 - 2x_2 + 2 - 2x_3\right)\mathrm{d}\mu_1 + \left(-2x_1 + 2x_2 - 2x_3 + 4\right)\mathrm{d}\mu_2 \\ = {} & 0. \end{aligned}$$

Set $x_1 = \mu_1 + \mu_{2,}\, x_2 = \mu_1 - \mu_2$ so that the first two terms vanish, and set the coefficients of the differentials of the remaining, independent terms to zero:

$$\begin{aligned}
\frac{dJ'}{dx_1} &= 2x_1 - 2\mu_1 - 2\mu_2 = 0,\\
\frac{dJ'}{dx_2} &= 2x_2 - 2\mu_1 + 2\mu_2 = 0,\\
\frac{dJ'}{dx_3} &= 2x_3 - 2\mu_1 - 2\mu_2 = 0,\\
\frac{dJ'}{d\mu_1} &= -2x_1 - 2x_2 + 2 - 2x_3 = 0,\\
\frac{dJ'}{d\mu_2} &= -2x_1 + 2x_2 - 2x_3 + 4 = 0.
\end{aligned}$$

Then,

$$\begin{aligned}
dJ' &= (2x_3 - 2\mu_1 - 2\mu_2)dx_3 + (-2x_1 - 2x_2 + 2 - 2x_3)d\mu_1\\
&\quad + (-2x_1 + 2x_2 - 2x_3 + 4)d\mu_2\\
&= 0,
\end{aligned}$$

or

$$\begin{aligned}
x_1 &= \mu_1 + \mu_2\\
x_2 &= \mu_1 - \mu_2\\
x_3 - \mu_1 - \mu_2 &= 0\\
-x_1 - x_2 + 1 - x_3 &= 0\\
-x_1 + x_2 - x_3 + 2 &= 0.
\end{aligned}$$

That is,

$$\begin{aligned}
\mathbf{x} &= \mathbf{A}^{\mathrm{T}}\boldsymbol{\mu}\\
\mathbf{A}\mathbf{x} &= \mathbf{b},
\end{aligned}$$

or

$$\left\{\begin{matrix} \mathbf{I} & -\mathbf{A}^{\mathrm{T}} \\ \mathbf{A} & \mathbf{0} \end{matrix}\right\}\begin{bmatrix} \mathbf{x} \\ \boldsymbol{\mu} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{b} \end{bmatrix}.$$

In this particular case, the first set can be solved for $\mathbf{x} = \mathbf{A}^{\mathrm{T}}\boldsymbol{\mu}$:

$$\begin{aligned}
\boldsymbol{\mu} &= (\mathbf{A}\mathbf{A}^{\mathrm{T}})^{-1}\mathbf{b} = [\,1/8 \quad 5/8\,]^{\mathrm{T}},\\
\mathbf{x} &= \mathbf{A}^{\mathrm{T}}\begin{bmatrix} \frac{1}{8} \\ \frac{5}{8} \end{bmatrix} = [\,3/4 \quad -1/2 \quad 3/4\,]^{\mathrm{T}}.
\end{aligned}$$

Example *Suppose, instead, we wanted to minimize*

$$J = (x_1 - x_2)^2 + (x_2 - x_3)^2 = \mathbf{x}^{\mathrm{T}}\mathbf{F}^{\mathrm{T}}\mathbf{F}\mathbf{x},$$

where

$$\mathbf{F} = \begin{Bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{Bmatrix}$$

$$\mathbf{F}^{\mathrm{T}}\mathbf{F} = \begin{Bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{Bmatrix}^{\mathrm{T}} \begin{Bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{Bmatrix} = \begin{Bmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{Bmatrix}.$$

Such an objective function might be used to find a "smooth" solution. One confirms that

$$\begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{Bmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{Bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = x_1^2 - 2x_1x_2 + 2x_2^2 - 2x_2x_3 + x_3^2$$
$$= (x_1 - x_2)^2 + (x_2 - x_3)^2 .$$

The stationary point of

$$J' = \mathbf{x}^{\mathrm{T}}\mathbf{F}^{\mathrm{T}}\mathbf{F}\mathbf{x} - 2\boldsymbol{\mu}^{\mathrm{T}}(\mathbf{A}\mathbf{x} - \mathbf{b})$$

leads to

$$\mathbf{F}^{\mathrm{T}}\mathbf{F}\mathbf{x} = \mathbf{A}^{\mathrm{T}}\boldsymbol{\mu}$$
$$\mathbf{A}\mathbf{x} = \mathbf{b}.$$

But,

$$\mathbf{x} \neq (\mathbf{F}^{\mathrm{T}}\mathbf{F})^{-1}\mathbf{A}^{\mathrm{T}}\boldsymbol{\mu},$$

because there is no inverse (guaranteed). But the coupled set

$$\begin{Bmatrix} \mathbf{F}^{\mathrm{T}}\mathbf{F} & -\mathbf{A}^{\mathrm{T}} \\ \mathbf{A} & \mathbf{0} \end{Bmatrix} \begin{bmatrix} \mathbf{x} \\ \boldsymbol{\mu} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{b} \end{bmatrix}$$

does have a solution.

The physical interpretation of $\boldsymbol{\mu}$ can be obtained as above by considering the way in which J would vary with infinitesimal changes in $\mathbf{b}$. As in the special case above, $J = J'$ at the stationary point. Hence,

$$\mathrm{d}J' = \mathrm{d}J - 2\boldsymbol{\mu}^{\mathrm{T}}\mathbf{A}\mathrm{d}\mathbf{x} - 2(\mathbf{A}\mathbf{x} - \mathbf{b})^{\mathrm{T}}\mathrm{d}\boldsymbol{\mu} + 2\boldsymbol{\mu}^{\mathrm{T}}\mathrm{d}\mathbf{b} = 0, \qquad (2.170)$$

or, since the first three terms on the right vanish at the stationary point,

$$\frac{\partial J'}{\partial \mathbf{b}} = \frac{\partial J}{\partial \mathbf{b}} = 2\boldsymbol{\mu}. \tag{2.171}$$

Thus, as inferred previously, the Lagrange multipliers are the sensitivity of J, at the stationary point, to perturbations in the parameters **b**. This conclusion leads, in Chapter 4, to the scrutiny of the Lagrange multipliers as a means of understanding the sensitivity of models and the flow of information within them.

Now revert to $\mathbf{Ex} + \mathbf{n} = \mathbf{y}$, that is, equations containing noise. If these are first column scaled using $\mathbf{S}^{-\mathrm{T}/2}$, Eqs. (2.167)–(2.169) are in the primed variables, and the solution in the original variables is

$$\tilde{\mathbf{x}} = \mathbf{SE}^{\mathrm{T}}(\mathbf{ESE}^{\mathrm{T}})^{-1}\mathbf{y}, \tag{2.172}$$

$$\tilde{\mathbf{n}} = \mathbf{0}, \tag{2.173}$$

$$\mathbf{C}_{xx} = \mathbf{0}, \tag{2.174}$$

and the result depends directly upon **S**. If a row scaling with $\mathbf{W}^{-\mathrm{T}/2}$ is used, it is readily shown that **W** disappears from the solution and has no effect on it (see p. 107).

Equations (2.172)–(2.174) are a solution, but there is the same fatal defect as in Eq. (2.173) – $\tilde{\mathbf{n}} = \mathbf{0}$ is usually unacceptable when **y** are observations. Furthermore, $\|\tilde{\mathbf{x}}\|$ is again uncontrolled, and $\mathbf{ESE}^{\mathrm{T}}$ may not have an inverse.

Noise vector **n** must be regarded as fully an element of the solution, as much as **x**. Equations representing observations can always be written as in (2.138), and can be solved exactly. Therefore, we now use a modified objective function, allowing for general **S**, **W**,

$$J = \mathbf{x}^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{x} + \mathbf{n}^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{n} - 2\boldsymbol{\mu}^{\mathrm{T}}(\mathbf{Ex} + \mathbf{n} - \mathbf{y}), \tag{2.175}$$

with both **x**, **n** appearing in the objective function. Setting the derivatives of (2.175) with respect to **x**, **n**, $\boldsymbol{\mu}$ to zero, and solving the resulting normal equations produces the following:

$$\tilde{\mathbf{x}} = \mathbf{SE}^{\mathrm{T}}(\mathbf{ESE}^{\mathrm{T}} + \mathbf{W})^{-1}\mathbf{y}, \tag{2.176}$$

$$\tilde{\mathbf{n}} = \mathbf{y} - \mathbf{E}\tilde{\mathbf{x}}, \tag{2.177}$$

$$\mathbf{C}_{xx} = \mathbf{SE}^{\mathrm{T}}(\mathbf{ESE}^{\mathrm{T}} + \mathbf{I})^{-1}\mathbf{R}_{nn}(\mathbf{ESE}^{\mathrm{T}} + \mathbf{I})^{-1}\mathbf{ES}, \tag{2.178}$$

$$\tilde{\boldsymbol{\mu}} = W^{-1}\tilde{\mathbf{n}}, \tag{2.179}$$

Equations (2.176)–(2.179) are identical to Eqs. (2.135)–(2.137) or to the alternate form Eqs. (2.127) – (2.129) derived from an objective function without Lagrange multipliers.

Equations (2.135)–(2.137) and (2.176)–(2.178) result from two very different appearing objective functions – one in which the equations are imposed in the mean-square, and one in which they are imposed exactly, using Lagrange multipliers. Constraints in the mean-square will be termed "soft," and those imposed exactly are "hard."[30] The distinction is, however, largely illusory: although (2.87) are being imposed exactly, it is only the presence of the error term, **n**, which 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 (rare) circumstances, which we will discuss briefly below, one may wish to impose exact constraints upon the elements of $\tilde{x}_i$. The solution in (2.167)–(2.169) was derived from the noise-free hard constraint, $\mathbf{Ax} = \mathbf{b}$, but we ended by rejecting it as generally inapplicable.

Once again, **n** is only by convention discussed separately from **x**, and is fully a part of the solution. The combined form (2.138), which literally treats **x**, **n** as the solution, are imposed through a hard constraint on the objective function,

$$J = \boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\xi} - 2\boldsymbol{\mu}^{\mathrm{T}}(\mathbf{E}_1\boldsymbol{\xi} - \mathbf{y}), \tag{2.180}$$

where $\boldsymbol{\xi} = [(\mathbf{S}^{-\mathrm{T}/2}\mathbf{x})^{\mathrm{T}}, (\mathbf{W}^{-\mathrm{T}/2}\mathbf{n})^{\mathrm{T}}]^{\mathrm{T}}$, which is Eq. (2.175). (There are numerical advantages, however, in working with objects in two spaces of dimensions M and N, rather than a single space of dimension $M + N$.)

2.4.4 Interpretation of discrete adjoints

When the operators are matrices, as they are in discrete formulations, then the adjoint is just the transposed matrix. Sometimes the adjoint has a simple physical interpretation. Suppose, e.g., that scalar y was calculated from a sum,

$$y = \mathbf{Ax}, \;\; \mathbf{A} = \left\{ 1\;1\;.\;1\;1 \right\}. \tag{2.181}$$

Then the adjoint operator applied to y is evidently

$$\mathbf{r} = \mathbf{A}^{\mathrm{T}}y = \left\{ 1\;1\;1\;.\;1 \right\}^{\mathrm{T}} y = \mathbf{x}. \tag{2.182}$$

Thus the adjoint operator "sprays" the average back out onto the originating vector, and might be thought of as an inverse operator.

A more interesting case is a first-difference forward operator,

$$\mathbf{A} = \left\{ \begin{matrix} -1 & 1 & & & & \\ & -1 & 1 & & & \\ & & -1 & 1 & & \\ & & & . & . & . \\ & & & & -1 & 1 \\ & & & & & -1 \end{matrix} \right\}, \tag{2.183}$$

that is,

$$y_i = x_{i+1} - x_i, \tag{2.184}$$

(with the exception of the last element, $y_N = -x_N$).

Then its adjoint is

$$\mathbf{A}^{\mathrm{T}} = \left\{ \begin{matrix} -1 & & & & & \\ 1 & -1 & & & & \\ & 1 & -1 & & & \\ & & \cdot & \cdot & & \\ & & & 1 & -1 & \\ & & & & 1 & -1 \end{matrix} \right\}, \tag{2.185}$$

which is a first-difference *backward* operator with $\mathbf{z} = \mathbf{A}^{\mathrm{T}}\mathbf{y}$, producing $z_i = y_{i-1} - y_i$, again with the exception of the end point, which is now z_1.

In general, the transpose matrix, or adjoint operator is *not* simply interpretable as an inverse operation as the summation/spray-out case might have suggested.[31] A more general understanding of the relationship between adjoints and inverses will be obtained in the next section.

2.5 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 even in the simplest form of least-squares, the solution is dependent upon the inverses of $\mathbf{E}^{\mathrm{T}}\mathbf{E}$, or $\mathbf{E}\mathbf{E}^{\mathrm{T}}$. In practice, their existence cannot be guaranteed, and we need to understand 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 relatable to particular elements of the data y_i. For many purposes, particularly physical insight, understanding the structure of the solution is essential. We will return to examine the least-squares solutions using some extra machinery.

2.5.1 Simple vector expansions

Consider again the elementary problem (2.1) of representing an L-dimensional vector $\mathbf{f}$ as a sum of a basis of L-orthonormal vectors $\mathbf{g}_i, i = 1, 2, \ldots, L, \mathbf{g}_i^{\mathrm{T}}\mathbf{g}_j = \delta_{ij}$.

Without error,

$$\mathbf{f} = \sum_{j=1}^{L} a_j \mathbf{g}_j, \quad a_j = \mathbf{g}_j^{\mathrm{T}} \mathbf{f}. \tag{2.186}$$

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:

$$\begin{aligned}\tilde{\mathbf{f}} &= \sum_{j=1}^{K} b_j \mathbf{g}_j \\ &= \mathbf{f} + \delta\mathbf{f}_1,\end{aligned} \tag{2.187}$$

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 only if it is orthogonal to the K vectors retained in the approximation, and then only if $b_j = a_j$ as given by (2.186). The only way the error could be reduced further 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 $\mathbf{b} = \mathbf{a} = \mathbf{G}_K^{\mathrm{T}} \mathbf{f}$ is the vector of coefficients $a_j = \mathbf{g}_j^{\mathrm{T}} \mathbf{f}$, $j = 1, 2, \ldots, K$, and the finite representation (2.187) is (one should write it out)

$$\tilde{\mathbf{f}} = \mathbf{G}_K \mathbf{a} = \mathbf{G}_K \left(\mathbf{G}_K^{\mathrm{T}} \mathbf{f}\right) = \left(\mathbf{G}_K \mathbf{G}_K^{\mathrm{T}}\right) \mathbf{f}, \quad \mathbf{a} = \{a_i\}, \tag{2.188}$$

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). Column i of $\mathbf{G}_K \mathbf{G}_K^{\mathrm{T}}$ produces the weighted linear combination of the true elements of $\mathbf{f}$ that will appear as $\tilde{f}_i$.

Because the columns of $\mathbf{G}_K$ are orthonormal, $\mathbf{G}_K^{\mathrm{T}} \mathbf{G}_K = \mathbf{I}_K$, that is, the $K \times K$ identity matrix; but $\mathbf{G}_K \mathbf{G}_K^{\mathrm{T}} \neq \mathbf{I}_L$ unless $K = L$. (That $\mathbf{G}_L \mathbf{G}_L^{\mathrm{T}} = \mathbf{I}_L$ for $K = L$ follows from the theorem for *square* matrices that shows a left inverse is also a right inverse.) If $K < L$, $\mathbf{G}_K$ is "semi-orthogonal." If $K = L$, it is "orthogonal"; in this case, $\mathbf{G}_L^{-1} = \mathbf{G}_L^{\mathrm{T}}$. If it is only semi-orthogonal, $\mathbf{G}_K^{\mathrm{T}}$ is a left inverse, but not a right inverse. Any orthogonal matrix has the property that its transpose is identical to its inverse.

The matrix $\mathbf{G}_K \mathbf{G}_K^{\mathrm{T}}$ is known as a "resolution matrix," with a simple interpretation. Suppose the true value of $\mathbf{f}$ were $\mathbf{f}_{j_0} = [0\ 0\ 0\ \ldots\ 0\ 1\ 0\ .\ 0\ ..\ 0]^{\mathrm{T}}$, that is, a Kronecker delta δ_{jj_0}, with unity in element j_0 and zero otherwise. Then the incomplete expansion (2.187) or (2.188) would not reproduce the delta function, but

$$\tilde{\mathbf{f}}_{j_0} = \mathbf{G}_K \mathbf{G}_K^{\mathrm{T}} \mathbf{f}_{j_0}, \tag{2.189}$$

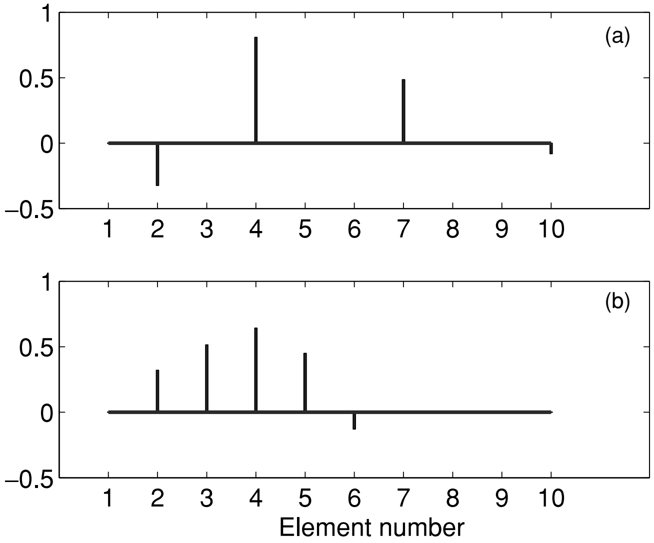

Figure 2.9 (a) Example of a row, j_0, of a 10×10 resolution matrix, perhaps the fourth one, showing widely distributed averaging in forming $\mathbf{f}_{j0}$. (b) The so-called compact resolution, in which the solution is readily interpreted as representing a local average of the true solution. Such situations are not common.

which is column j_0 of $\mathbf{G}_K \mathbf{G}_K^{\mathrm{T}}$. Each column (or row) of the resolution matrix tells one what the corresponding form of the approximating vector would be, if its true form were a Kronecker delta.

To form a Kronecker delta requires a spanning set of vectors. An analogous elementary result of Fourier analysis shows that a Dirac delta function demands contributions from all frequencies to represent a narrow, very high pulse. Removal of some of the requisite vectors (sinusoids) produces peak 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 Fig. 2.9) without distant sidelobes), leading to the tidy interpretation of the result as a local average of the true values, called "compact resolution."[32]

A resolution matrix has the property

$$\operatorname{trace}\left(\mathbf{G}_K \mathbf{G}_K^{\mathrm{T}}\right) = K, \tag{2.190}$$

which follows from noting that

$$\operatorname{trace}\left(\mathbf{G}_K \mathbf{G}_K^{\mathrm{T}}\right) = \operatorname{trace}\left(\mathbf{G}_K^{\mathrm{T}} \mathbf{G}_K\right) = \operatorname{trace}(\mathbf{I}_K) = K.$$

2.5.2 Square-symmetric problem: eigenvalues/eigenvectors

Orthogonal vector expansions are particularly simple to use and interpret, but might seem irrelevant when dealing with simultaneous equations where neither the row nor column vectors of the coefficient matrix are so simply related. What we will show, however, is that we can always find sets of orthonormal vectors to greatly simplify the job of solving simultaneous equations. To do so, we digress to recall the basic elements of the "eigenvector/eigenvalue problem" mentioned in passing on p. 26.

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

$$\mathbf{E}\mathbf{g}_i = \lambda_i \mathbf{g}_i, \quad i = 1, 2, \ldots, M, \tag{2.191}$$

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.* It is easy to see that if two λ_j, λ_k are distinct, then the corresponding eigenvectors are orthogonal:

$$\mathbf{E}\mathbf{g}_j = \lambda_j \mathbf{g}_j, \tag{2.192}$$

$$\mathbf{E}\mathbf{g}_k = \lambda_k \mathbf{g}_k. \tag{2.193}$$

Left multiply the first of these by $\mathbf{g}_k^T$, and the second by $\mathbf{g}_j^T$, and subtract:

$$\mathbf{g}_k^T \mathbf{E}\mathbf{g}_j - \mathbf{g}_j^T \mathbf{E}\mathbf{g}_k = (\lambda_j - \lambda_k)\mathbf{g}_k^T \mathbf{g}_j. \tag{2.194}$$

But because $\mathbf{E} = \mathbf{E}^T$, the left-hand side vanishes, and hence $\mathbf{g}_k^T \mathbf{g}_j$ by the assumption $\lambda_j \neq \lambda_k$. A similar construction proves that the λ_i are all real, and an elaboration shows that for coincident λ_i, the corresponding eigenvectors can always be made orthogonal.

Example *To contrast with the above result, consider the non-symmetric, square matrix*

$$\left\{\begin{matrix} 1 & 2 & 3 \\ 0 & 1 & 4 \\ 0 & 0 & 1 \end{matrix}\right\}.$$

Solution to the eigenvector/eigenvalue problem produces $\lambda_i = 1$, *and* $\mathbf{u}_i = [1, 0, 0]^T$, $i = 1, 2, 3$. *The eigenvectors are not orthogonal, and are certainly not a spanning set. On the other hand, the eigenvector/eigenvalues of*

$$\left\{\begin{matrix} 1 & -1 & -2 \\ -1 & 2 & -1 \\ 1.5 & 2 & -2.5 \end{matrix}\right\}$$

are

$$\mathbf{g}_1 = \begin{bmatrix} -0.29+0.47i \\ -0.17+0.25i \\ 0.19+0.61i \end{bmatrix}, \mathbf{g}_2 = \begin{bmatrix} -0.29-0.47i \\ -0.17-0.25i \\ 0.19-0.61i \end{bmatrix}, \mathbf{g}_3 = \begin{bmatrix} -0.72 \\ 0.90 \\ 0.14 \end{bmatrix},$$
$$\lambda_j = [-1.07+1.74i, -1.07-1.74\,i, 2.64]$$

and (rounded) are not orthogonal, but are a basis. The eigenvalues/eigenvectors appear in complex conjugate pairs and in some contexts are called "principal oscillation patterns" (POPs).

Suppose for the moment that we have the square, symmetric, special case, and recall how eigenvectors can be used to solve (2.16). By convention, the pairs $(\lambda_i, \mathbf{g}_i)$ are ordered in the sense of decreasing λ_i. If some λ_i are repeated, an arbitrary order choice is made.

With an orthonormal, spanning set, both the known $\mathbf{y}$ and the unknown $\mathbf{x}$ can be written as

$$\mathbf{x} = \sum_{i=1}^{M} \alpha_i \mathbf{g}_i, \quad \alpha_i = \mathbf{g}_i^{\mathrm{T}} \mathbf{x}, \tag{2.195}$$

$$\mathbf{y} = \sum_{i=1}^{M} \beta_i \mathbf{g}_i, \quad \beta_i = \mathbf{g}_i^{\mathrm{T}} \mathbf{y}. \tag{2.196}$$

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

Substitute (2.195) into (2.16), to give

$$\mathbf{E} \sum_{i=1}^{M} \alpha_i \mathbf{g}_i = \sum_{i=1}^{M} \left(\mathbf{g}_i^{\mathrm{T}} \mathbf{y}\right) \mathbf{g}_i, \tag{2.197}$$

or, using the eigenvector property,

$$\sum_{i=1}^{M} \alpha_i \lambda_i \mathbf{g}_i = \sum_{i}^{M} \left(\mathbf{g}_i^{\mathrm{T}} \mathbf{y}\right) \mathbf{g}_i. \tag{2.198}$$

But the expansion vectors are orthonormal and so

$$\lambda_i \alpha_i = \mathbf{g}_i^{\mathrm{T}} \mathbf{y}, \tag{2.199}$$

$$\alpha_i = \frac{\mathbf{g}_i^{\mathrm{T}} \mathbf{y}}{\lambda_i}, \tag{2.200}$$

$$\mathbf{x} = \sum_{i=1}^{M} \frac{\mathbf{g}_i^{\mathrm{T}} \mathbf{y}}{\lambda_i} \mathbf{g}_i. \tag{2.201}$$

Apart from the obvious difficulty if an eigenvalue vanishes, the problem is now completely solved. Define a diagonal matrix, $\mathbf{\Lambda}$, with elements, λ_i, in descending numerical value, and the matrix $\mathbf{G}$, whose columns are the corresponding $\mathbf{g}_i$ in the same order, the solution to (2.16) can be written, from (2.195), (2.199)–(2.201) as

$$\boldsymbol{\alpha} = \mathbf{\Lambda}^{-1}\mathbf{G}^{\mathrm{T}}\mathbf{y}, \tag{2.202}$$

$$\mathbf{x} = \mathbf{G}\mathbf{\Lambda}^{-1}\mathbf{G}^{\mathrm{T}}\mathbf{y}, \tag{2.203}$$

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

Vanishing eigenvalues, $i = i_0$, cause trouble and must be considered. Let the corresponding eigenvectors be $\mathbf{g}_{i_0}$. Then any part of the solution which 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 which 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 non-zero 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 $\mathbf{g}_2 = [1, -1]^{\mathrm{T}}/\sqrt{2}$. (The purpose of showing the observation twice is to produce an $\mathbf{E}$ that is square.)

Suppose there are $K < M$ non-zero λ_i. Then for $i > K$, Eq. (2.199) is

$$0\alpha_i = \mathbf{g}_i^{\mathrm{T}}\mathbf{y}, \quad K + i = 1, 2, \ldots, M, \tag{2.204}$$

and two cases must be distinguished.

Case (1)

$$\mathbf{g}_i^{\mathrm{T}}\mathbf{y} = 0, \quad K + i = 1, 2, \ldots, M. \tag{2.205}$$

We could then put $\alpha_i = 0$, $K + i = 1, 2, \ldots, M$, and the solution can be written

$$\tilde{\mathbf{x}} = \sum_{i=1}^{K} \frac{\mathbf{g}_i^{\mathrm{T}}\mathbf{y}}{\lambda_i}\mathbf{g}_i, \tag{2.206}$$

and $\mathbf{E}\tilde{\mathbf{x}} = \mathbf{y}$, *exactly*. Equation (2.205) is often known as a "solvability condition." A tilde has been placed over $\mathbf{x}$ because a solution of the form

$$\tilde{\mathbf{x}} = \sum_{i=1}^{K} \frac{\mathbf{g}_i^{\mathrm{T}}\mathbf{y}}{\lambda_i}\mathbf{g}_i + \sum_{i=K+1}^{M} \alpha_i \mathbf{g}_i, \tag{2.207}$$

with the remaining α_i taking on arbitrary values, also satisfies the equations exactly. That is, the true value of $\mathbf{x}$ *could* contain structures proportional to the nullspace vectors of $\mathbf{E}$, but the equations (2.16) neither require their presence, nor provide information necessary to determine their amplitudes. We thus have a situation with a "solution nullspace." Define the matrix $\mathbf{G}_K$ to be $M \times K$, carrying only the first K of the $\mathbf{g}_i$, that is, the range vectors, $\boldsymbol{\Lambda}_K$, to be $K \times K$ with only the first K, non-zero eigenvalues, and the columns of $\mathbf{Q}_G$ are the $M - K$ nullspace vectors (it is $M \times (M - K)$), then the solutions (2.206) and (2.207) are

$$\tilde{\mathbf{x}} = \mathbf{G}_K \boldsymbol{\Lambda}_K^{-1} \mathbf{G}_K^{\mathrm{T}} \mathbf{y}, \tag{2.208}$$

$$\tilde{\mathbf{x}} = \mathbf{G}_K \boldsymbol{\Lambda}_K^{-1} \mathbf{G}_K^{\mathrm{T}} \mathbf{y} + \mathbf{Q}_G \boldsymbol{\alpha}_G, \tag{2.209}$$

where $\boldsymbol{\alpha}_G$ is the vector of unknown nullspace coefficients. The solution in (2.208), with no nullspace contribution, will be called the "particular" solution. If $\mathbf{y} = \mathbf{0}$, however, Eq. (2.16) is a homogeneous set of equations, then the nullspace represents the only possible 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

$$\mathbf{G}\mathbf{G}^{\mathrm{T}} = \mathbf{I} = \mathbf{G}_K \mathbf{G}_K^{\mathrm{T}} + \mathbf{Q}_G \mathbf{Q}_G^{\mathrm{T}}, \tag{2.210}$$

or

$$\mathbf{Q}_G \mathbf{Q}_G^{\mathrm{T}} = \mathbf{I} - \mathbf{G}_K \mathbf{G}_K^{\mathrm{T}}. \tag{2.211}$$

Vectors $\mathbf{Q}_G$ span the nullspace of $\mathbf{G}$.

Case (2)

$$\mathbf{g}_i^{\mathrm{T}} \mathbf{y} \neq 0, \quad i > K + 1, \tag{2.212}$$

for one or more of the nullspace vectors. In this case, Eq. (2.199) is the contradiction

$$0\alpha_i \neq 0,$$

and Eq. (2.198) is actually

$$\sum_{i=1}^{K} \lambda_i \alpha_i \mathbf{g}_i = \sum_{i=1}^{M} \left(\mathbf{g}_i^T \mathbf{y}\right) \mathbf{g}_i, \quad K < M, \tag{2.213}$$

that is, with differing upper limits on the sums. Therefore, the solvability condition is not satisfied. Owing to the orthonormality of the $\mathbf{g}_i$, there is no choice of α_i, $i = 1, \ldots, K$ on the left that can match the last $M - K$ terms on the right. Evidently there is no solution in the conventional sense unless (2.205) 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 (2.200), $i = 1, 2, \ldots, K$. No choice of nullspace vector coefficients, nor any other value of the coefficients of the range vectors, can reduce the norm of $\tilde{\mathbf{n}}$. The best solution is then also (2.206) or (2.208).

In this situation, we are no longer solving the equations (2.16), but rather are dealing with a set that could be written

$$\mathbf{E}\mathbf{x} \sim \mathbf{y}, \tag{2.214}$$

where the demand is for a solution that is the "best possible," in the sense just defined. Such statements of approximation are awkward, so as before rewrite (2.214) as

$$\mathbf{E}\mathbf{x} + \mathbf{n} = \mathbf{y}, \tag{2.215}$$

where $\mathbf{n}$ is the residual. If $\tilde{\mathbf{x}}$ is given by (2.207), then by (2.213),

$$\tilde{\mathbf{n}} = \sum_{i=K+1}^{M} \left(\mathbf{g}_i^T \mathbf{y}\right) \mathbf{g}_i. \tag{2.216}$$

Notice that $\tilde{\mathbf{n}}^T \tilde{\mathbf{y}} = \mathbf{0}$: $\tilde{\mathbf{y}}$ is orthogonal to the residuals.

Example *Let*

$$x_1 + x_2 = 1,$$
$$x_1 + x_2 = 3.$$

Then using $\lambda_1 = 2,\ \mathbf{g}_1 = [1, 1]^T/\sqrt{2},\ \lambda_2 = 0,\ \mathbf{g}_2 = [1, -1]^T/\sqrt{2},$ *one has* $\tilde{\mathbf{x}} = [1/2, 1/2]^T \propto \mathbf{g}_1,\ \tilde{\mathbf{y}} = [2, 2]^T \propto \mathbf{g}_1,\ \tilde{\mathbf{n}} = [-1, 1]^T \propto \mathbf{g}_2.$

This outcome, where M equations in M unknowns were found in practice not to be able to determine some solution structures, is labeled "formally just-determined.” The expression “formally” alludes to the fact that the appearance of a just-determined system did not mean that the characterization was true in practice. One or more vanishing eigenvalues mean that neither the rows nor columns of **E** are spanning sets.

Some decision has to be made about the coefficients of the nullspace vectors in (2.209). The form could be used as it stands, regarding it as the “general solution.” The analogy with the solution of differential equations should be apparent – typically, there is a particular solution and a homogeneous solution – here 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 missing.

Despite the presence of indeterminate elements in the solution, a great deal is known about them: 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 observing these missing structures. The reader is also reminded of the discussion of the Neumann problem in Chapter 1.

Another approach is to define a “simplest” solution, appealing to what is usually known as “Ockham’s razor,” or the “principle 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 (2.208), which is without any nullspace contributions, is less structured than any other solution. (It is often, but not always, true that the nullspace vectors are more “wiggly” than those in the range. The nullspace of the Neumann problem is a counter example. 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 plausible choice. It follows from the orthogonality of the $\mathbf{g}_i$ that this particular solution is also the one of minimum solution norm. Later, other choices for the nullspace vectors will be made. If $\mathbf{y} = \mathbf{0}$, then the nullspace *is* the solution.

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

$$\tilde{\mathbf{x}} = \mathbf{G}_K \mathbf{G}_K^{\mathrm{T}} \mathbf{x} = \mathbf{x} - \mathbf{Q}_G \boldsymbol{\alpha}_G, \quad \tilde{\mathbf{y}} = \mathbf{G}_K \mathbf{G}_K^{\mathrm{T}} \mathbf{y}, \quad \tilde{\mathbf{n}} = \mathbf{Q}_G \mathbf{Q}_G^{\mathrm{T}} \mathbf{y}. \tag{2.217}$$

These results are very important, so we recapitulate them: (2.207) or (2.209) is the general solution. There are three vectors involved, one of them, **y**, known, and two of them, **x**, **n**, unknown. Because of the assumption that **E** has an orthonormal basis of eigenvectors, all three of these vectors can be expanded exactly as

$$\mathbf{x} = \sum_{i=1}^{M} \alpha_i \mathbf{g}_i, \quad \mathbf{n} = \sum_{i=1}^{M} \gamma_i \mathbf{g}_i, \quad \mathbf{y} = \sum_{i=1}^{M} (\mathbf{y}^{\mathrm{T}} \mathbf{g}_i) \mathbf{g}_i. \tag{2.218}$$

Substituting into (2.215),

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

From the orthogonality property,

$$\lambda_i \alpha_i + \gamma_i = \mathbf{y}^{\mathrm{T}} \mathbf{g}_i, \quad i = 1, 2, \ldots, K, \tag{2.219}$$

$$\gamma_i = \mathbf{y}^{\mathrm{T}} \mathbf{g}_i, \quad K + i = 1, 2, \ldots, M. \tag{2.220}$$

In dealing with the first relationship, a choice is required. If we set

$$\gamma_i = \mathbf{g}_i^{\mathrm{T}} \mathbf{n} = 0, \quad i = 1, 2, \ldots, K, \tag{2.221}$$

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 attention later.

The relative contributions of any structure in **y**, determined by the projection, $\mathbf{g}_i^{\mathrm{T}} \mathbf{y}$, will depend upon the ratio $\mathbf{g}_i^{\mathrm{T}} \mathbf{y} / \lambda_i$. Comparatively weak values of $\mathbf{g}_i^{\mathrm{T}} \mathbf{y}$ may well be amplified by small, but non-zero, elements of λ_i. One must keep track of both $\mathbf{g}_i^{\mathrm{T}} \mathbf{y}$, and $\mathbf{g}_i^{\mathrm{T}} \mathbf{y} / \lambda_i$.

Before leaving this special case, note one more useful property of the eigenvector/eigenvalues. For the moment, let **G** have all its columns, containing both the range and nullspace vectors, with the nullspace vectors being last in arbitrary order. It is thus an $M \times M$ matrix. Correspondingly, let $\mathbf{\Lambda}$ contain all the eigenvalues on its diagonal, including the zero ones; it too, is $M \times M$. Then the eigenvector definition (2.191) produces

$$\mathbf{E}\mathbf{G} = \mathbf{G}\mathbf{\Lambda}. \tag{2.222}$$

Multiply both sides of (2.222) by $\mathbf{G}^{\mathrm{T}}$:

$$\mathbf{G}^{\mathrm{T}} \mathbf{E} \mathbf{G} = \mathbf{G}^{\mathrm{T}} \mathbf{G} \mathbf{\Lambda} = \mathbf{\Lambda}. \tag{2.223}$$

G is said to “diagonalize” **E**. Now multiply both sides of (2.223) on the left by **G** and on the right by $\mathbf{G}^{\mathrm{T}}$:

$$\mathbf{G}\mathbf{G}^{\mathrm{T}} \mathbf{E} \mathbf{G} \mathbf{G}^{\mathrm{T}} = \mathbf{G} \mathbf{\Lambda} \mathbf{G}^{\mathrm{T}}. \tag{2.224}$$

Using the orthogonality of $\mathbf{G}$,

$$\mathbf{E} = \mathbf{G}\boldsymbol{\Lambda}\mathbf{G}^{\mathrm{T}}. \tag{2.225}$$

This is a useful representation of $\mathbf{E}$, consistent with its symmetry.

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 (2.225), 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, $\boldsymbol{\Lambda}$ can be reduced to its $K \times K$ form, and the decomposition (2.225) is still exact – in the form

$$\mathbf{E} = \mathbf{G}_K\boldsymbol{\Lambda}_K\mathbf{G}_K^{\mathrm{T}}. \tag{2.226}$$

The representation (decomposition) in either Eq. (2.225) or (2.226) is identical to

$$\mathbf{E} = \lambda_1\mathbf{g}_1\mathbf{g}_1^{\mathrm{T}} + \lambda_2\mathbf{g}_2\mathbf{g}_2^{\mathrm{T}} + \cdots + \lambda_K\mathbf{g}_K\mathbf{g}_K^{\mathrm{T}}. \tag{2.227}$$

That is, a square symmetric matrix can be exactly represented by a sum of products of orthonormal vectors $\mathbf{g}_i\mathbf{g}_i^{\mathrm{T}}$ multiplied by a scalar, λ_i.

Example *Consider the matrix from the last example,*

$$\mathbf{E} = \begin{Bmatrix} 1 & 1 \\ 1 & 1 \end{Bmatrix}.$$

We have

$$\mathbf{E} = \frac{2}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix}\begin{bmatrix} 1 & 1 \end{bmatrix}\frac{1}{\sqrt{2}} + \frac{0}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \end{bmatrix}\begin{bmatrix} 1 & -1 \end{bmatrix}\frac{1}{\sqrt{2}}.$$

The simultaneous equations (2.215) are

$$\mathbf{G}_K\boldsymbol{\Lambda}_K\mathbf{G}_K^{\mathrm{T}}\mathbf{x} + \mathbf{n} = \mathbf{y}. \tag{2.228}$$

Left multiply both sides by $\boldsymbol{\Lambda}_K^{-1}\mathbf{G}_K^{\mathrm{T}}$ to get

$$\mathbf{G}_K^{\mathrm{T}}\mathbf{x} + \boldsymbol{\Lambda}_K^{-1}\mathbf{G}_K^{\mathrm{T}}\mathbf{n} = \boldsymbol{\Lambda}_K^{-1}\mathbf{G}_K^{\mathrm{T}}\mathbf{y}. \tag{2.229}$$

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

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

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

$$\tilde{\mathbf{x}} \equiv \mathbf{G}_K\mathbf{G}_K^{\mathrm{T}}\mathbf{x} \equiv \mathbf{G}_K\boldsymbol{\Lambda}_K^{-1}\mathbf{G}_K^{\mathrm{T}}\mathbf{y}, \tag{2.230}$$

$$\tilde{\mathbf{y}} = \mathbf{E}\tilde{\mathbf{x}} = \mathbf{G}_K\mathbf{G}_K^{\mathrm{T}}\mathbf{y}, \tag{2.231}$$

which is identical to the particular solution (2.206). The residuals are

$$\tilde{\mathbf{n}} = \mathbf{y} - \tilde{\mathbf{y}} = \mathbf{y} - \mathbf{E}\tilde{\mathbf{x}} = \left(\mathbf{I}_M - \mathbf{G}_K\mathbf{G}_K^{\mathrm{T}}\right)\mathbf{y} = \mathbf{Q}_G\mathbf{Q}_G^{\mathrm{T}}\mathbf{y}, \tag{2.232}$$

with $\tilde{\mathbf{n}}^{\mathrm{T}}\tilde{\mathbf{y}} = 0$. Notice that matrix $\mathbf{H}$ of Eq. (2.97) is just $\mathbf{G}_K\mathbf{G}_K^{\mathrm{T}}$, and hence $(\mathbf{I} - \mathbf{H})$ is the projector of $\mathbf{y}$ onto the nullspace vectors.

The expected value of the solution (2.206) or (2.230) is

$$\langle\tilde{\mathbf{x}} - \mathbf{x}\rangle = \mathbf{G}_K\boldsymbol{\Lambda}_K^{-1}\mathbf{G}_K^{\mathrm{T}}\langle\mathbf{y}\rangle - \sum_{i=1}^{N}\alpha_i\mathbf{g}_i = -\mathbf{Q}_G\boldsymbol{\alpha}_G, \tag{2.233}$$

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

The uncertainty is given by

$$\begin{aligned}\mathbf{P} = D^2(\tilde{\mathbf{x}} - \mathbf{x}) &= \left\langle\mathbf{G}_K\boldsymbol{\Lambda}_K^{-1}\mathbf{G}_K^{\mathrm{T}}(\mathbf{y}_0 + \mathbf{n} - \mathbf{y}_0)(\mathbf{y}_0 + \mathbf{n} - \mathbf{y}_0)^{\mathrm{T}}\mathbf{G}_K\boldsymbol{\Lambda}_K^{-1}\mathbf{G}_K^{\mathrm{T}}\right\rangle \\ &\quad + \left\langle\mathbf{Q}_G\boldsymbol{\alpha}_G\boldsymbol{\alpha}_G^{\mathrm{T}}\mathbf{Q}_G^{\mathrm{T}}\right\rangle \\ &= \mathbf{G}_K\boldsymbol{\Lambda}_K^{-1}\mathbf{G}_K^{\mathrm{T}}\langle\mathbf{n}\mathbf{n}^{\mathrm{T}}\rangle\mathbf{G}_K\boldsymbol{\Lambda}_K^{-1}\mathbf{G}_K^{\mathrm{T}} + \mathbf{Q}_G\left\langle\boldsymbol{\alpha}_G\boldsymbol{\alpha}_G^{\mathrm{T}}\right\rangle\mathbf{Q}_G^{\mathrm{T}} \\ &= \mathbf{G}_K\boldsymbol{\Lambda}_K^{-1}\mathbf{G}_K^{\mathrm{T}}\mathbf{R}_{nn}\mathbf{G}_K\boldsymbol{\Lambda}_K^{-1}\mathbf{G}_K^{\mathrm{T}} + \mathbf{Q}_G\mathbf{R}_{\alpha\alpha}\mathbf{Q}_G^{\mathrm{T}} \\ &= \mathbf{C}_{xx} + \mathbf{Q}_G\mathbf{R}_{\alpha\alpha}\mathbf{Q}_G^{\mathrm{T}},\end{aligned} \tag{2.234}$$

defining the second moments, $\mathbf{R}_{\alpha\alpha}$, of the coefficients of the nullspace vectors. Under the special circumstances that the residuals, $\mathbf{n}$, are white noise, with $\mathbf{R}_{nn} = \sigma_n^2\mathbf{I}$, (2.234) reduces to

$$\mathbf{P} = \sigma_n^2\mathbf{G}_K\boldsymbol{\Lambda}_K^{-2}\mathbf{G}_K^{\mathrm{T}} + \mathbf{Q}_G\mathbf{R}_{\alpha\alpha}\mathbf{Q}_G^{\mathrm{T}}. \tag{2.235}$$

Either case shows that the uncertainty of the minimal solution is made up of two distinct parts. The first part, the solution covariance, $\mathbf{C}_{xx}$, 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 non-zero eigenvalue, λ_K^2.

Example *Suppose that*

$$\begin{gathered}\mathbf{E}\mathbf{x} = \mathbf{y}, \\ \left\{\begin{matrix}1 & 1 \\ 1 & 1\end{matrix}\right\}\begin{bmatrix}x_1 \\ x_2\end{bmatrix} = \mathbf{y} = \begin{bmatrix}1 \\ 3\end{bmatrix},\end{gathered} \tag{2.236}$$

which is inconsistent and has no solution in the conventional sense. Solving,

$$\mathbf{E}\mathbf{g}_i = \lambda_i\mathbf{g}_i, \tag{2.237}$$

or

$$\left\{\begin{matrix} 1-\lambda & 1 \\ 1 & 1-\lambda \end{matrix}\right\} \begin{bmatrix} g_{i1} \\ g_{i2} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{2.238}$$

This equation requires that

$$g_{i1} \begin{bmatrix} 1-\lambda \\ 1 \end{bmatrix} + g_{i2} \begin{bmatrix} 1 \\ 1-\lambda \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

or

$$\begin{bmatrix} 1-\lambda \\ 1 \end{bmatrix} + \frac{g_{i2}}{g_{i1}} \begin{bmatrix} 1 \\ 1-\lambda \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

which is

$$\frac{g_{i2}}{g_{i1}} = -(1-\lambda)$$
$$\frac{g_{i2}}{g_{i1}} = -\frac{1}{1-\lambda}.$$

Both equations are satisfied only if $\lambda = 2, 0$. *This method, which can be generalized, in effect derives the usual statement that for Eq. (2.238) to have a solution, the determinant,*

$$\begin{vmatrix} 1-\lambda & 1 \\ 1 & 1-\lambda \end{vmatrix},$$

must vanish. The first solution is labeled $\lambda_1 = 2$, *and substituting back in produces* $\mathbf{g}_1 = \frac{1}{\sqrt{2}}[1, 1]^{\mathrm{T}}$, *when given unit length. Also* $\mathbf{g}_2 = \frac{1}{\sqrt{2}}[-1, 1]^{\mathrm{T}},\ \lambda_2 = 0$. *Hence,*

$$\mathbf{E} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix} 2 \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}^{\mathrm{T}} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \end{bmatrix}. \tag{2.239}$$

The equations have no solution in the conventional sense. There is, however, a sensible "best" solution:

$$\tilde{\mathbf{x}} = \frac{\mathbf{g}_1^{\mathrm{T}} \mathbf{y}}{\lambda_1} \mathbf{g}_1 + \alpha_2 \mathbf{g}_2, \tag{2.240}$$
$$= \left(\frac{4}{2\sqrt{2}}\right) \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \alpha_2 \frac{1}{\sqrt{2}} \begin{bmatrix} -1 \\ 1 \end{bmatrix} \tag{2.241}$$
$$= \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \alpha_2 \frac{1}{\sqrt{2}} \begin{bmatrix} -1 \\ 1 \end{bmatrix}. \tag{2.242}$$

Notice that

$$\mathbf{E}\tilde{\mathbf{x}} = \begin{bmatrix} 2 \\ 2 \end{bmatrix} + 0 \neq \begin{bmatrix} 1 \\ 3 \end{bmatrix}. \tag{2.243}$$

The solution has compromised the inconsistency. No choice of α_2 can reduce the residual norm. The equations would more sensibly have been written

$$\mathbf{Ex} + \mathbf{n} = \mathbf{y},$$

and the difference, $\mathbf{n} = \mathbf{y} - \mathbf{E}\tilde{\mathbf{x}}$ *is proportional to* $\mathbf{g}_2$. *A system like (2.236) would most likely arise from measurements (if both equations are divided by 2, they represent two measurements of the average of* (x_1, x_2), *and* $\mathbf{n}$ *would be best regarded as the noise of observation).*

Example *Suppose the same problem as in the last example is solved using Lagrange multipliers, that is, minimizing*

$$J = \mathbf{n}^{\mathrm{T}}\mathbf{n} + \gamma^2\mathbf{x}^{\mathrm{T}}\mathbf{x} - \mathbf{2}\boldsymbol{\mu}^{\mathrm{T}}\left(\mathbf{y} - \mathbf{Ex} - \mathbf{n}\right).$$

Then, the normal equations are

$$\begin{aligned}
\frac{1}{2}\frac{\partial J}{\partial \mathbf{x}} &= \gamma^2\mathbf{x} + \mathbf{E}^{\mathrm{T}}\boldsymbol{\mu} = \mathbf{0},\\
\frac{1}{2}\frac{\partial J}{\partial \mathbf{n}} &= \mathbf{n} + \boldsymbol{\mu} = \mathbf{0},\\
\frac{1}{2}\frac{\partial J}{\partial \boldsymbol{\mu}} &= \mathbf{y} - \mathbf{Ex} - \mathbf{n} = \mathbf{0},
\end{aligned}$$

which produces

$$\begin{aligned}
\tilde{\mathbf{x}} &= \mathbf{E}^{\mathrm{T}}(\mathbf{EE}^{\mathrm{T}} + \gamma^2\mathbf{I})^{-1}\mathbf{y}\\
&= \begin{Bmatrix} 1 & 1\\ 1 & 1 \end{Bmatrix} \left\{\begin{Bmatrix} 2 & 2\\ 2 & 2 \end{Bmatrix} + \gamma^2 \begin{Bmatrix} 1 & 0\\ 0 & 1 \end{Bmatrix}\right\}^{-1} \begin{bmatrix} 1\\ 3 \end{bmatrix}.
\end{aligned}$$

The limit $\gamma^2 \to \infty$ *is readily evaluated. Letting* $\gamma^2 \to 0$ *involves inverting a singular matrix. To understand what is going on, use*

$$\mathbf{E} = \mathbf{G}\boldsymbol{\Lambda}\mathbf{G}^{\mathrm{T}} = \mathbf{g}_1\lambda_1\mathbf{g}_1^{\mathrm{T}} + 0 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1\\ 1 \end{bmatrix} 2\frac{1}{\sqrt{2}}\begin{bmatrix} 1\\ 1 \end{bmatrix}^{\mathrm{T}} + 0 \tag{2.244}$$

Hence,

$$\mathbf{EE}^{\mathrm{T}} = \mathbf{G}\boldsymbol{\Lambda}^2\mathbf{G}^{\mathrm{T}}.$$

Note that the full $\mathbf{G}$, $\boldsymbol{\Lambda}$ *are being used, and* $\mathbf{I} = \mathbf{GG}^{\mathrm{T}}$. *Thus,*

$$(\mathbf{EE}^{\mathrm{T}} + \gamma^2\mathbf{I}) = (\mathbf{G}\boldsymbol{\Lambda}^2\mathbf{G}^{\mathrm{T}} + \mathbf{G}(\gamma^2)\mathbf{G}^{\mathrm{T}}) = \mathbf{G}(\boldsymbol{\Lambda}^2 + \gamma^2\mathbf{I})\mathbf{G}^{\mathrm{T}}.$$

By inspection, the inverse of this last matrix is

$$(\mathbf{EE}^{\mathrm{T}} + \mathbf{I}/\gamma^2)^{-1} = \mathbf{G}(\boldsymbol{\Lambda}^2 + \gamma^2\mathbf{I})^{-1}\mathbf{G}^{\mathrm{T}}.$$

But $(\mathbf{\Lambda}^2+\gamma^2\mathbf{I})^{-1}$ *is the inverse of a diagonal matrix,*

$$(\mathbf{\Lambda}^2+\gamma^2\mathbf{I})^{-1}=\operatorname{diag}\left\{1/\left(\lambda_i^2+\gamma^2\right)\right\}.$$

Then

$$\begin{aligned}\tilde{\mathbf{x}}&=\mathbf{E}^{\mathrm{T}}(\mathbf{E}\mathbf{E}^{\mathrm{T}}+\gamma^2\mathbf{I})^{-1}\mathbf{y}=\mathbf{G}\mathbf{\Lambda}\mathbf{G}^{\mathrm{T}}\left(\mathbf{G}\operatorname{diag}\left\{1/\left(\lambda_i^2+\gamma^2\right)\right\}\mathbf{G}^{\mathrm{T}}\right)\mathbf{y}\\&=\mathbf{G}\operatorname{diag}\left\{\lambda_i/\left(\lambda_i^2+\gamma^2\right)\right\}\mathbf{G}^{\mathrm{T}}\mathbf{y}\\&=\sum_{i=1}^{K}\mathbf{g}_i\frac{\lambda_i}{\lambda_i^2+\gamma^2}\mathbf{g}_i^{\mathrm{T}}\mathbf{y}=\frac{1}{\sqrt{2}}\begin{bmatrix}1\\1\end{bmatrix}\frac{2}{2+\gamma^2}\frac{1}{\sqrt{2}}\begin{bmatrix}1\\1\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}1\\3\end{bmatrix}+0\\&=\frac{4}{2+\gamma^2}\begin{bmatrix}1\\1\end{bmatrix}.\end{aligned}$$

The solution always exists as long as $\gamma^2>0$. *It is a tapered-down form of the solution with* $\gamma^2=0$ *if all* $\lambda_i\neq 0$. Therefore,

$$\mathbf{n}=\begin{bmatrix}1\\3\end{bmatrix}-\frac{4}{2+\gamma^2}\mathbf{E}\begin{bmatrix}1\\1\end{bmatrix}=\begin{bmatrix}1\\3\end{bmatrix}-\frac{4}{2+\gamma^2}\begin{bmatrix}2\\2\end{bmatrix},$$

so that, as $\gamma^2\to\infty$, *the solution* $\tilde{\mathbf{x}}$ *is minimized, becoming 0, and the residual is equal to* $\mathbf{y}$.

2.5.3 *Arbitrary systems*

The singular vector expansion and singular value decomposition

It may be objected that this entire development is of little interest, because most problems, including those outlined in Chapter 1, produced $\mathbf{E}$ matrices that could not be guaranteed to have complete orthonormal sets of eigenvectors. Indeed, the problems considered produce matrices that are usually non-square, 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.[33] When a square matrix of dimension N 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 next find a way to avoid these intricacies, and yet deal with sets of simultaneous equations of arbitrary dimensions, not just square ones. The square, symmetric case nonetheless provides full analogues to all of the issues in the more general case, and the reader may find it helpful to refer back to this situation for insight.

Consider the possibility, suggested by the eigenvector method, of expanding the solution $\mathbf{x}$ in a set of orthonormal vectors. Equation (2.87) involves one vector, $\mathbf{x}$, of dimension N, and two vectors, $\mathbf{y}$, $\mathbf{n}$, of dimension M. We would like to use orthonormal basis 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 using the structure of $\mathbf{E}$, a particularly useful pair can be identified.

The simple development of the solutions in the square, symmetric case resulted from the theorem concerning the complete nature of the eigenvectors of such a matrix. So construct a new matrix,

$$\mathbf{B} = \begin{Bmatrix} \mathbf{0} & \mathbf{E}^{\mathrm{T}} \\ \mathbf{E} & \mathbf{0} \end{Bmatrix}, \tag{2.245}$$

which by definition is square (dimension $M + N$ by $M + N$) and symmetric. Thus, $\mathbf{B}$ satisfies the theorem just alluded to, and the eigenvalue problem,

$$\mathbf{B}\mathbf{q}_i = \lambda_i \mathbf{q}_i, \tag{2.246}$$

will give rise to $M + N$ orthonormal eigenvectors $\mathbf{q}_i$ (an orthonormal basis) whether or not the λ_i are distinct or non-zero. Writing out (2.246),

$$\begin{Bmatrix} \mathbf{0} & \mathbf{E}^{\mathrm{T}} \\ \mathbf{E} & \mathbf{0} \end{Bmatrix} \begin{bmatrix} q_{1i} \\ \cdot \\ q_{Ni} \\ q_{N+1,i} \\ \cdot \\ q_{N+M,i} \end{bmatrix} = \lambda_i \begin{bmatrix} q_{1i} \\ \cdot \\ q_{Ni} \\ q_{N+1,i} \\ \cdot \\ q_{N+M,i} \end{bmatrix}, \quad i = 1, 2, \ldots, M + N, \tag{2.247}$$

where q_{pi} is the pth element of $\mathbf{q}_i$. Taking note of the zero matrices, (2.247) may be rewritten as

$$\mathbf{E}^{\mathrm{T}} \begin{bmatrix} q_{N+1,i} \\ \cdot \\ q_{N+M,i} \end{bmatrix} = \lambda_i \begin{bmatrix} q_{1i} \\ \cdot \\ q_{Ni} \end{bmatrix}, \tag{2.248}$$

$$\mathbf{E} \begin{bmatrix} q_{1i} \\ \cdot \\ q_{Ni} \end{bmatrix} = \lambda_i \begin{bmatrix} q_{N+1,i} \\ \cdot \\ q_{N+M,i} \end{bmatrix}, \quad i = 1, 2, \ldots, M + N. \tag{2.249}$$

Define

$$\mathbf{u}_i = [\, q_{N+1,i} \quad \cdot \quad q_{N+M,i} \,]^{\mathrm{T}}, \ \mathbf{v}_i = [\, q_{1,i} \quad \cdot \quad q_{N,i} \,]^{\mathrm{T}}, \ \mathbf{q}_i = [\, \mathbf{v}_i^{\mathrm{T}} \quad \mathbf{u}_i^{\mathrm{T}} \,]^{\mathrm{T}}, \tag{2.250}$$

that is, defining the first N elements of $\mathbf{q}_i$ to be called $\mathbf{v}_i$ and the last M to be called $\mathbf{u}_i$, the two sets together being the "singular vectors." Then (2.248)–(2.249) are

$$\mathbf{E}\mathbf{v}_i = \lambda_i \mathbf{u}_i, \tag{2.251}$$

$$\mathbf{E}^{\mathrm{T}}\mathbf{u}_i = \lambda_i \mathbf{v}_i. \tag{2.252}$$

If (2.251) is left multiplied by $\mathbf{E}^{\mathrm{T}}$, and using (2.252), one has

$$\mathbf{E}^{\mathrm{T}}\mathbf{E}\mathbf{v}_i = \lambda_i^2 \mathbf{v}_i, \quad i = 1, 2, \ldots, N. \tag{2.253}$$

Similarly, left multiplying (2.252) by $\mathbf{E}$ and using (2.251) produces

$$\mathbf{E}\mathbf{E}^{\mathrm{T}}\mathbf{u}_i = \lambda_i^2 \mathbf{u}_i \quad i = 1, 2, \ldots, M. \tag{2.254}$$

These last two equations show that the $\mathbf{u}_i$, $\mathbf{v}_i$ each separately satisfy two independent eigenvector/eigenvalue problems of the square symmetric matrices $\mathbf{E}\mathbf{E}^{\mathrm{T}}$, $\mathbf{E}^{\mathrm{T}}\mathbf{E}$ and they can be separately given unit norm. The λ_i come in pairs as $\pm\lambda_i$ and the convention is made that only the non-negative ones are retained, as the $\mathbf{u}_i$, $\mathbf{v}_i$ corresponding to the singular values also differ at most by a minus sign, and hence are not independent of the ones retained.[34] If one of M, N is much smaller than the other, only the smaller eigenvalue/eigenvector problem needs to be solved for either of $\mathbf{u}_i$, $\mathbf{v}_i$; the other set is immediately calculated from (2.252) or (2.251). Evidently, in the limiting cases of either a single equation or a single unknown, the eigenvalue/eigenvector problem is purely scalar, no matter how large the other dimension.

In going from (2.248), (2.249) to (2.253), (2.254), the range of the index i has dropped from $M + N$ to M or N. The missing "extra" equations correspond to negative λ_i and carry no independent information.

Example *Consider the non-square, non-symmetric matrix*

$$\mathbf{E} = \left\{\begin{matrix} 0 & 0 & 1 & -1 & 2 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 \\ 1 & 2 & 0 & 0 & 0 \end{matrix}\right\}.$$

Forming the larger matrix $\mathbf{B}$, *solve the eigenvector/eigenvalue problem that produces*

$$\mathbf{Q} = \left\{\begin{matrix} -0.31623 & 0.63246 & -1.1796 \times 10^{-16} & -0.63246 & 0.31623 \\ -0.63246 & -0.31623 & -2.0817 \times 10^{-16} & 0.31623 & 0.63246 \\ 0.35857 & -0.22361 & 0.80178 & -0.22361 & 0.35857 \\ -0.11952 & -0.67082 & -0.26726 & -0.67082 & -0.11952 \\ 0.59761 & 0.00000 & -0.53452 & 0.00000 & 0.59761 \end{matrix}\right\},$$

$$\mathbf{S} = \left\{\begin{matrix} -2.6458 & 0 & 0 & 0 & 0 \\ 0 & -1.4142 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1.4142 & 0 \\ 0 & 0 & 0 & 0 & 2.6458 \end{matrix}\right\},$$

where $\mathbf{Q}$ *is the matrix whose columns are* $\mathbf{q}_i$ *and* $\mathbf{S}$ *is the diagonal matrix whose values are the corresponding eigenvalues. Note that one of the eigenvalues vanishes identically, and that the others occur in positive and negative pairs. The corresponding* $\mathbf{q}_i$ *differ only by sign changes in parts of the vectors, but they are all linearly independent. First define a* $\mathbf{V}$ *matrix from the first two rows of* $\mathbf{Q}$,

$$\mathbf{V} = \left\{\begin{matrix} -0.31623 & 0.63246 & 0 & -0.63246 & 0.31623 \\ -0.63246 & -0.31623 & 0 & 0.31623 & 0.63246 \end{matrix}\right\}.$$

Only two of the vectors are linearly independent (the zero-vector is not physically realizable). Similarly, the last three rows of $\mathbf{Q}$ *define a* $\mathbf{U}$ *matrix,*

$$\mathbf{U} = \left\{\begin{matrix} 0.35857 & -0.22361 & 0.80178 & -0.22361 & 0.35857 \\ -0.11952 & -0.67082 & -0.26726 & -0.67082 & -0.11952 \\ 0.59761 & 0.00000 & -0.53452 & 0.00000 & 0.59761 \end{matrix}\right\},$$

in which only three columns are linearly independent. Retaining only the last two columns of $\mathbf{V}$ *and the last three of* $\mathbf{U}$, *and column normalizing each to unity, produces the singular vectors.*

By convention, the λ_i are ordered in decreasing numerical value. Equations (2.251)–(2.252) provide a relationship between each $\mathbf{u}_i$, $\mathbf{v}_i$ pair. But because $M \neq N$, generally, there will be more of one set than the other. The only way Eqs. (2.251)–(2.252) 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 (2.254) is solved for $\mathbf{u}_i$, $i = 1, 2, \ldots, M$, and (2.252) is used to find the corresponding $\mathbf{v}_i$. There are $N - M$ $\mathbf{v}_i$ not generated this way, but which can be found using the Gram–Schmidt method described on p. 22.

Let there be K non-zero λ_i; then

$$\mathbf{E}\mathbf{v}_i \neq 0, \quad i = 1, 2, \ldots, K. \tag{2.255}$$

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

$$\mathbf{E}\mathbf{v}_i = 0, \quad i = K + 1, \ldots N, \tag{2.256}$$

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

$$\mathbf{E}^{\mathrm{T}}\mathbf{u}_i \neq 0, \quad i = 1, 2, \ldots, K, \tag{2.257}$$

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

$$\mathbf{E}^{\mathrm{T}}\mathbf{u}_i = 0, \quad i = K + 1, \ldots, M, \tag{2.258}$$

the "nullspace vectors of $\mathbf{E}^{\mathrm{T}}$" or the "data, or observation, nullspace vectors." The "nullspace" of $\mathbf{E}$ is spanned by its nullspace vectors, and the "range" of $\mathbf{E}$ 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. We now have two complete orthonormal sets in the two different spaces. Note that Eqs. (2.256) and (2.258) imply that

$$\mathbf{E}\mathbf{v}_i = 0, \; \mathbf{u}_i^{\mathrm{T}}\mathbf{E} = 0, \quad i = K + 1, \ldots, N, \tag{2.259}$$

which expresses hard relationships among the columns and rows of $\mathbf{E}$.

Because the $\mathbf{u}_i$, $\mathbf{v}_i$ are complete in their corresponding spaces, $\mathbf{x}$, $\mathbf{y}$, $\mathbf{n}$ can be expanded without error:

$$\mathbf{x} = \sum_{i=1}^{N} \alpha_i \mathbf{v}_i, \quad \mathbf{y} = \sum_{j=1}^{M} \beta_i \mathbf{u}_i, \quad \mathbf{n} = \sum_{i=1}^{M} \gamma_i \mathbf{u}_i, \tag{2.260}$$

where $\mathbf{y}$ has been measured, so that we know $\beta_j = \mathbf{u}_j^{\mathrm{T}}\mathbf{y}$. To find $\mathbf{x}$, we need α_i, and to find $\mathbf{n}$, we need the γ_i. Substitute (2.260) into (2.87), and using (2.251)–(2.252),

$$\begin{aligned}\sum_{i=1}^{N} \alpha_i \mathbf{E}\mathbf{v}_i + \sum_{i=1}^{M} \gamma_i \mathbf{u}_i &= \sum_{i=1}^{K} \alpha_i \lambda_i \mathbf{u}_i + \sum_{i=1}^{M} \gamma_i \mathbf{u}_i \\ &= \sum_{i=1}^{M} \left(\mathbf{u}_i^{\mathrm{T}}\mathbf{y}\right)\mathbf{u}_i.\end{aligned} \tag{2.261}$$

Notice the differing upper limits on the summations. Because of the orthonormality of the singular vectors, (2.261) can be solved as

$$\alpha_i \lambda_i + \gamma_i = \mathbf{u}_i^{\mathrm{T}}\mathbf{y}, \quad i = 1, 2, \ldots, M, \tag{2.262}$$

$$\alpha_i = \left(\mathbf{u}_i^{\mathrm{T}}\mathbf{y} - \gamma_i\right)/\lambda_i, \quad \lambda_i \neq 0, \; i = 1, 2, \ldots, K. \tag{2.263}$$

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

$$\mathbf{u}_i^{\mathrm{T}}\mathbf{n} = 0, \quad i = 1, 2, \ldots, K, \tag{2.264}$$

should we wish, which will have the effect of making the noise norm as small as possible (there is arbitrariness in this choice, and later γ_i will be chosen differently).

Then (2.263) produces

$$\alpha_i = \frac{\mathbf{u}_i^{\mathrm{T}}\mathbf{y}}{\lambda_i}, \quad i = 1, 2, \ldots, K. \tag{2.265}$$

But, because $\lambda_i = 0$, $i > K$, the only solution to (2.262) for these values of i is $\gamma_i = \mathbf{u}_i^{\mathrm{T}}\mathbf{y}$, and α_i is indeterminate. These γ_i are non-zero, except in the event (unlikely with real data) that

$$\mathbf{u}_i^{\mathrm{T}}\mathbf{y} = 0, \quad i = K+1, \ldots, N. \tag{2.266}$$

This last equation is a solvability condition – in direct analogy to (2.205).

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

$$\tilde{\mathbf{x}} = \sum_{i=1}^{K} \frac{\mathbf{u}_i^{\mathrm{T}}\mathbf{y}}{\lambda_i}\mathbf{v}_i + \sum_{i=K+1}^{N} \alpha_i \mathbf{v}_i, \tag{2.267}$$

$$\tilde{\mathbf{y}} = \mathbf{E}\tilde{\mathbf{x}} = \sum_{i=1}^{K} \left(\mathbf{u}_i^{\mathrm{T}}\mathbf{y}\right)\mathbf{u}_i, \tag{2.268}$$

$$\tilde{\mathbf{n}} = \sum_{i=K+1}^{M} \left(\mathbf{u}_i^{\mathrm{T}}\mathbf{y}\right)\mathbf{u}_i. \tag{2.269}$$

The coefficients of the last $N - K$ of the $\mathbf{v}_i$ in Eq. (2.267), 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 $\mathbf{E}^{\mathrm{T}}$ and will vanish only if $K = M$, or if the solvability conditions are met.

Just as in the simpler 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 gives

$$\tilde{\mathbf{x}} = \sum_{i=1}^{K} \frac{\mathbf{u}_i^{\mathrm{T}}\mathbf{y}}{\lambda_i}\mathbf{v}_i. \tag{2.270}$$

Along with (2.269), this is the "particular-SVD solution." It minimizes the residuals, and simultaneously produces the corresponding $\tilde{\mathbf{x}}$ with the smallest norm. If $\langle \mathbf{n} \rangle = 0$, the bias of (2.270) is evidently

$$\langle \tilde{\mathbf{x}} - \mathbf{x} \rangle = -\sum_{i=K+1}^{N} \alpha_i \mathbf{v}_i. \tag{2.271}$$

The solution uncertainty is

$$\boxed{\mathbf{P} = \sum_{i=1}^{K} \sum_{j=1}^{K} \mathbf{v}_i \frac{\mathbf{u}_i^{\mathrm{T}} \mathbf{R}_{nn} \mathbf{u}_j}{\lambda_i \lambda_j} \mathbf{v}_i^{\mathrm{T}} + \sum_{i=K+1}^{N} \sum_{j=K+1}^{N} \mathbf{v}_i \langle \alpha_i \alpha_j \rangle \mathbf{v}_j^{\mathrm{T}}. \qquad (2.272)}$$

If the noise is white with variance σ_n^2 or, if a row scaling matrix $\mathbf{W}^{-\mathrm{T}/2}$ has been applied to make it so, then (2.272) becomes

$$\mathbf{P} = \sum_{i=1}^{K} \frac{\sigma_n^2}{\lambda_i^2} \mathbf{v}_i \mathbf{v}_i^{\mathrm{T}} + \sum_{i=K+1}^{N} \langle \alpha_i^2 \rangle \mathbf{v}_i \mathbf{v}_i^{\mathrm{T}}, \qquad (2.273)$$

where it was also assumed that $\langle \alpha_i \alpha_j \rangle = \langle \alpha_i^2 \rangle \delta_{ij}$ in the nullspace. The influence of very small singular values on the uncertainty is plain: In the solution (2.267) or (2.270) there are error terms $\mathbf{u}_i^{\mathrm{T}} \mathbf{y} / \lambda_i$ that are greatly magnified by small or nearly vanishing singular values, introducing large terms proportional to σ_n^2 / λ_i^2 into (2.273).

The structures dominating $\tilde{\mathbf{x}}$ are a competition between the magnitudes of $\mathbf{u}_i^{\mathrm{T}} \mathbf{y}$ and λ_i, given by the ratio, $\mathbf{u}_i^{\mathrm{T}} \mathbf{y} / \lambda_i$. Large λ_i can suppress comparatively large projections onto $\mathbf{u}_i$, and similarly, small, but non-zero λ_i may greatly amplify comparatively modest projections. In practice,[35] one is well-advised to study the behavior of both $\mathbf{u}_i^{\mathrm{T}} \mathbf{y}$, $\mathbf{u}_i^{\mathrm{T}} \mathbf{y} / \lambda_i$ as a function of i to understand the nature of the solution.

The decision to omit contributions to the residuals by the range vectors of $\mathbf{E}^{\mathrm{T}}$, as we did in Eqs. (2.264) and (2.269), needs to be examined. Should some other choice be made, the $\tilde{\mathbf{x}}$ norm would decrease, but the residual norm would increase. Determining the desirability of such a trade-off requires an understanding of the noise structure – in particular, (2.264) imposes rigid structures, and hence covariances, on the residuals.

2.5.4 The singular value decomposition

The singular vectors and values have been used to provide a convenient pair of orthonormal bases 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 $\mathbf{v}_i$. As an example, suppose $M = 3$, $N = 4$; then

$$\boldsymbol{\Lambda} = \left\{ \begin{matrix} \lambda_i & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & \lambda_3 & 0 \end{matrix} \right\}.$$

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

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

therefore extending the definition of a diagonal matrix to non-square ones.

Precisely as with matrix $\mathbf{G}$ considered above (see p. 75), the column orthonormality of $\mathbf{U}$, $\mathbf{V}$ implies that these matrices are orthogonal:

$$\mathbf{U}\mathbf{U}^{\mathrm{T}} = \mathbf{I}_M, \qquad \mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}_M, \tag{2.274}$$

$$\mathbf{V}\mathbf{V}^{\mathrm{T}} = \mathbf{I}_N, \qquad \mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}_N. \tag{2.275}$$

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

Equations (2.251)–(2.254) can be written compactly as:

$$\mathbf{E}\mathbf{V} = \mathbf{U}\mathbf{\Lambda}, \qquad \mathbf{E}^{\mathrm{T}}\mathbf{U} = \mathbf{V}\mathbf{\Lambda}^{\mathrm{T}}, \tag{2.276}$$

$$\mathbf{E}^{\mathrm{T}}\mathbf{E}\mathbf{V} = \mathbf{V}\mathbf{\Lambda}^{\mathrm{T}}\mathbf{\Lambda}, \qquad \mathbf{E}\mathbf{E}^{\mathrm{T}}\mathbf{U} = \mathbf{U}\mathbf{\Lambda}\mathbf{\Lambda}^{\mathrm{T}}. \tag{2.277}$$

Left multiply the first relation of (2.276) by $\mathbf{U}^{\mathrm{T}}$ and right multiply it by $\mathbf{V}^{\mathrm{T}}$, and, invoking Eq. (2.275),

$$\mathbf{U}^{\mathrm{T}}\mathbf{E}\mathbf{V} = \mathbf{\Lambda}. \tag{2.278}$$

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

Right multiplying the first relation of (2.276) by $\mathbf{V}^{\mathrm{T}}$ gives

$$\mathbf{E} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^{\mathrm{T}}. \tag{2.279}$$

This last equation represents a product, called the "singular value decomposition" (SVD), of an arbitrary matrix, of two orthogonal matrices, $\mathbf{U}$, $\mathbf{V}$, and a usually non-square diagonal matrix, $\mathbf{\Lambda}$.

There is one further step to take. Notice that for a rectangular $\mathbf{\Lambda}$, as in the examples above, one or more rows or columns must be all zero, depending upon the shape of the matrix. In addition, if any of the $\lambda_i = 0$, $i < \min(M, N)$, the corresponding rows or columns of $\mathbf{\Lambda}$ will be all zeros. Let K be the number of non-vanishing 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 $\mathbf{\Lambda}$ to a $K \times K$ square matrix, then the representation (2.279) remains exact, in the form

$$\mathbf{E} = \mathbf{U}_K \boldsymbol{\Lambda}_K \mathbf{V}_K^{\mathrm{T}} = \lambda_1 \mathbf{u}_1 \mathbf{v}_1^{\mathrm{T}} + \lambda_2 \mathbf{u}_2 \mathbf{v}_2^{\mathrm{T}} + \cdots + \lambda_K \mathbf{u}_K \mathbf{v}_K^{\mathrm{T}}, \tag{2.280}$$

with 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 (2.280) should be compared to (2.226).[36]

The SVD solution can be obtained by matrix manipulation, rather than vector by vector. Consider once again finding the solution to the simultaneous equations (2.87), but first write $\mathbf{E}$ in its reduced SVD,

$$\mathbf{U}_K \boldsymbol{\Lambda}_K \mathbf{V}_K^{\mathrm{T}} \mathbf{x} + \mathbf{n} = \mathbf{y}. \tag{2.281}$$

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

$$\boldsymbol{\Lambda}_K \mathbf{V}_K^{\mathrm{T}} \mathbf{x} + \mathbf{U}_K^{\mathrm{T}} \mathbf{n} = \mathbf{U}_K^{\mathrm{T}} \mathbf{y}. \tag{2.282}$$

The inverse of $\boldsymbol{\Lambda}_K$ is easily computed, and

$$\mathbf{V}_K^{\mathrm{T}} \mathbf{x} + \boldsymbol{\Lambda}_K^{-1} \mathbf{U}_K^{\mathrm{T}} \mathbf{n} = \boldsymbol{\Lambda}_K^{-1} \mathbf{U}_K^{\mathrm{T}} \mathbf{y}. \tag{2.283}$$

But $\mathbf{V}_K^{\mathrm{T}} \mathbf{x}$ is the dot product of the first K of the $\mathbf{v}_i$ with the unknown $\mathbf{x}$. Equation (2.283) thus represents statements about the relationship between 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

$$\mathbf{U}_K^{\mathrm{T}} \mathbf{n} = 0, \tag{2.284}$$

then

$$\mathbf{V}_K^{\mathrm{T}} \mathbf{x} = \boldsymbol{\Lambda}_K^{-1} \mathbf{U}_K^{\mathrm{T}} \mathbf{y}, \tag{2.285}$$

and hence

$$\tilde{\mathbf{x}} = \mathbf{V}_K \boldsymbol{\Lambda}_K^{-1} \mathbf{U}_K^{\mathrm{T}} \mathbf{y}. \tag{2.286}$$

Equation (2.286) is identical to the solution (2.270), and can be confirmed by writing it out explicitly. As with the square symmetric case, the contribution of any structure in $\mathbf{y}$ proportional to $\mathbf{u}_i$ depends upon the ratio of the projection $\mathbf{u}_i^{\mathrm{T}} \mathbf{y}$ to λ_i. Substituting (2.286) into (2.281),

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

or

$$\tilde{\mathbf{n}} = \left(\mathbf{I} - \mathbf{U}_K \mathbf{U}_K^{\mathrm{T}}\right) \mathbf{y}. \tag{2.287}$$

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

$$\mathbf{U} = \{\mathbf{U}_K \quad \mathbf{Q}_u\},\ \mathbf{V} = \{\mathbf{V}_K \quad \mathbf{Q}_v\}, \tag{2.288}$$

where $\mathbf{Q}_u$, $\mathbf{Q}_v$ are the matrices whose columns are the corresponding nullspace vectors. Then

$$\mathbf{E}\tilde{\mathbf{x}} + \tilde{\mathbf{n}} = \mathbf{y},\ \mathbf{E}\tilde{\mathbf{x}} = \tilde{\mathbf{y}}, \tag{2.289}$$

and

$$\tilde{\mathbf{y}} = \mathbf{U}_K\mathbf{U}_K^{\mathrm{T}}\mathbf{y},\ \tilde{\mathbf{n}} = \mathbf{Q}_u\mathbf{Q}_u^{\mathrm{T}}\mathbf{y} = \sum_{j=K+1}^{N}\left(\mathbf{u}_j^{\mathrm{T}}\mathbf{y}\right)\mathbf{u}_j, \tag{2.290}$$

which is identical to (2.268). Note that

$$\mathbf{Q}_u\mathbf{Q}_u^{\mathrm{T}} = \left(\mathbf{I} - \mathbf{U}_K\mathbf{U}_K^{\mathrm{T}}\right),\ \mathbf{Q}_v\mathbf{Q}_v^{\mathrm{T}} = \left(\mathbf{I} - \mathbf{V}_K\mathbf{V}_K^{\mathrm{T}}\right), \tag{2.291}$$

which are idempotent ($\mathbf{V}_K\mathbf{V}_K^{\mathrm{T}}$ is matrix $\mathbf{H}$ of Eq. (2.97)). The two vector sets $\mathbf{Q}_u$, $\mathbf{Q}_v$ span the data and solution nullspaces respectively. The general solution is

$$\tilde{\mathbf{x}} = \mathbf{V}_K\boldsymbol{\Lambda}_K^{-1}\mathbf{U}_K\mathbf{y} + \mathbf{Q}_v\boldsymbol{\alpha}, \tag{2.292}$$

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

The solution uncertainty (2.272) is thus

$$\begin{aligned}\mathbf{P} = \mathbf{V}_K\boldsymbol{\Lambda}_K^{-1}\mathbf{U}_K^{\mathrm{T}}\left\langle\mathbf{n}\mathbf{n}^{\mathrm{T}}\right\rangle\mathbf{U}_K\boldsymbol{\Lambda}_K^{-1}\mathbf{V}_K^{\mathrm{T}}\\ + \mathbf{Q}_v\left\langle\alpha\alpha^{\mathrm{T}}\right\rangle\mathbf{Q}_G^{\mathrm{T}} = \mathbf{C}_{xx} + \mathbf{Q}_v\left\langle\boldsymbol{\alpha}\boldsymbol{\alpha}^{\mathrm{T}}\right\rangle\mathbf{Q}_v^{\mathrm{T}},\end{aligned} \tag{2.293}$$

or

$$\mathbf{P} = \sigma_n^2\mathbf{V}_K\boldsymbol{\Lambda}_K^{-2}\mathbf{V}_K^{\mathrm{T}} + \mathbf{Q}_v\left\langle\boldsymbol{\alpha}\boldsymbol{\alpha}^{\mathrm{T}}\right\rangle\mathbf{Q}_v^{\mathrm{T}}, \tag{2.294}$$

for white noise.

Least-squares solution of simultaneous solutions 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. Unlike the eigenvalue/eigenvector solution for an arbitrary square system, the singular values (eigenvalues) are always non-negative and real, and the singular vectors (eigenvectors) can always be made a complete orthonormal set. Furthermore, the relations (2.276) provide 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.

2.5.5 Some simple examples: algebraic equations

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

$$\mathbf{E} = \left\{1 \quad -2\right\}, \quad \mathbf{U} = \{1\}, \quad \mathbf{V} = \left\{\begin{matrix} 0.447 & -0.894 \\ -0.894 & -0.447 \end{matrix}\right\}, \quad \lambda_1 = 2.23,$$

where the second column of V is the nullspace of E. The general solution is $\tilde{\mathbf{x}} = [0.6, -1.2]^{\mathrm{T}} + \alpha_2 \mathbf{v}_2$. *Because* $K = 1$ *is the only possible choice, this solution satisfies the equation exactly, and a data nullspace is not possible.*

Example *The most elementary overdetermined problem is* 2×1. *Suppose that*

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

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

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

$\mathbf{E} = \{1, 1\}^{\mathrm{T}}, \mathbf{E}^{\mathrm{T}}\mathbf{E}$ *represents the eigenvalue problem of the smaller dimension, again* 1×1, and

$$\mathbf{U} = \left\{\begin{matrix} 0.707 & -0.707 \\ 0.707 & 0.707 \end{matrix}\right\}, \quad \mathbf{V} = \{1\}, \quad \lambda_1 = \sqrt{2},$$

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*

$$\mathbf{E}\tilde{\mathbf{x}} = \begin{bmatrix} 2 & 2 \end{bmatrix}^{\mathrm{T}} = \tilde{\mathbf{y}}.$$

Hence there are residuals $\tilde{\mathbf{n}} = \tilde{\mathbf{y}} - \mathbf{y} = [1, -1]^{\mathrm{T}}$ *that are necessarily proportional to* $\mathbf{u}_2$ *and thus orthogonal to* $\tilde{\mathbf{y}}$. *No other solution can produce a smaller* l_2 *norm residual than this one. The SVD provides a solution that compromises the contradiction between the two original equations.*

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

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

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 value means that the row and column vectors are not linearly independent 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 of the equations is given explicitly as* $\mathbf{u}_3^T\mathbf{E} = 0$.

Example *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, $|\epsilon| > 0$*, to produce*

$$\begin{aligned} x_1 + x_2 - (2+\epsilon)x_3 &= 1, \\ x_1 + x_2 - 2x_3 &= 2, \end{aligned} \tag{2.295}$$

then not only is there a solution, there is an infinite number of them, which can be shown by computing the particular SVD solution and the nullspace. Thus the slightest perturbation in the coefficients has made the system jump from one having no solution to one having an infinite number – a disconcerting situation. The label for 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*. If we set* $\epsilon = 10^{-10}$*, 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 non-vanishing of the second singular value is enough to assure a solution. It is the inevitable slight errors in y that suggest sufficiently small singular values should be treated as though they were actually zero.)*

Example *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

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

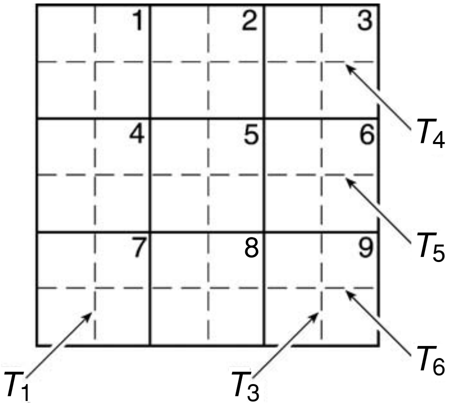

Figure 2.10 Tomographic problem with nine unknowns and only six integral constraints. Box numbers are in the upper right-corner of each, and the T_i are the measured integrals through various combinations of three boxes.

for arbitrarily small ϵ produces a system with no solutions in the conventional mathematical sense, although the SVD will handle the system in a sensible way, which the reader should confirm.

Problems like these are simple examples of the practical issues that arise once one recognizes that, unlike 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.

Example *(The tomographic problem from Chapter 1.) A square box is made up of* 3×3 *unit dimension sub-boxes (Fig. 2.10). All rays are in the r_x or r_y directions.* Therefore, the equations are

$$\left\{\begin{matrix} 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{matrix}\right\} \begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ . \\ x_9 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix},$$

that is, $\mathbf{Ex} = \mathbf{y}$*. There are six integrals (rays) across the nine boxes in which one seeks the corresponding value of x_i.* $\mathbf{y}$ *was calculated by assuming that the "true"* *value is* $x_5 = 1$, $x_i = 0$, $i \neq 5$. *The SVD produces*

$$\mathbf{U} = \left\{\begin{matrix} -0.408 & 0 & 0 & 0.816 & 0 & 0.408 \\ -0.408 & 0.703 & -0.0543 & -0.408 & -0.0549 & 0.408 \\ -0.408 & -0.703 & 0.0543 & -0.408 & 0.0549 & 0.408 \\ -0.408 & -0.0566 & 0.0858 & 0 & -0.81 & -0.408 \\ -0.408 & -0.0313 & -0.744 & 0 & 0.335 & -0.408 \\ -0.408 & 0.0879 & 0.658 & 0 & 0.475 & -0.408 \end{matrix}\right\},$$

$$\boldsymbol{\Lambda} = \mathrm{diag}([2.45 \quad 1.73 \quad 1.73 \quad 1.73 \quad 1.73 \quad 0]),$$

$$\mathbf{V} = \left\{\begin{matrix} -0.333 & -0.0327 & 0.0495 & 0.471 & -0.468 & -0.38 & -0.224 & 0.353 & 0.353 \\ -0.333 & 0.373 & 0.0182 & -0.236 & -0.499 & 0.432 & 0.302 & -0.275 & 0.302 \\ -0.333 & -0.438 & 0.0808 & -0.236 & -0.436 & -0.0515 & -0.0781 & -0.0781 & -0.655 \\ -0.333 & -0.0181 & -0.43 & 0.471 & 0.193 & 0.519 & -0.361 & -0.15 & -0.15 \\ -0.333 & 0.388 & -0.461 & -0.236 & 0.162 & -0.59 & -0.0791 & -0.29 & -0.0791 \\ -0.333 & -0.424 & -0.398 & -0.236 & 0.225 & 0.0704 & 0.44 & 0.44 & 0.229 \\ -0.333 & 0.0507 & 0.38 & 0.471 & 0.274 & -0.139 & 0.585 & -0.204 & -0.204 \\ -0.333 & 0.457 & 0.349 & -0.236 & 0.243 & 0.158 & -0.223 & 0.566 & -0.223 \\ -0.333 & -0.355 & 0.411 & -0.236 & 0.306 & -0.0189 & -0.362 & -0.362 & 0.427 \end{matrix}\right\}.$$

The zeros appearing in $\mathbf{U}$, *and in the last element of* $\mathrm{diag}(\boldsymbol{\Lambda})$, *are actually very small numbers* ($\mathrm{O}(10^{-16})$ or less). *Rank* $K = 5$ *despite there being six equations – a consequence of redundancy in the integrals. Notice that there are four repeated* λ_i, *and the lack of expected simple symmetries in the corresponding* $\mathbf{v}_i$ *is a consequence of a random assignment in the eigenvectors.*

Singular vector $\mathbf{u}_1$ *just averages the right-hand side values, and the corresponding solution is completely uniform, proportional to* $\mathbf{v}_1$. *The average of* $\mathbf{y}$ *is often the most robust piece of information.*

The "right" answer is $\mathbf{x} = [0, 0, 0, 0, 1, 0, 0, 0, 0]^{\mathrm{T}}$. *The rank 5 answer by SVD is* $\tilde{\mathbf{x}} = [-0.1111, \ 0.2222, -0.1111, 0.2222, 0.5556, 0.2222, -0.1111, 0.2222, -0.1111]^{\mathrm{T}}$, *which exactly satisfies the same equations, with* $\tilde{\mathbf{x}}^{\mathrm{T}}\tilde{\mathbf{x}} = 0.556 \ < \mathbf{x}^{\mathrm{T}}\mathbf{x}$. *When mapped into two dimensions,* $\tilde{\mathbf{x}}$ at *rank 5 is*

$$\begin{array}{cc} & r_x\rightarrow \\ r_y\uparrow & \begin{bmatrix} -0.11 & 0.22 & -0.11 \\ 0.22 & 0.56 & 0.22 \\ -0.11 & 0.22 & -0.11 \end{bmatrix} \end{array}, \tag{2.296}$$

and is the minimum norm solution. The mapped $\mathbf{v}_6$, *which belongs in the null-space, is*

$$\begin{array}{cc} & r_x\rightarrow \\ r_y\uparrow & \begin{bmatrix} -0.38 & 0.43 & -0.05 \\ 0.52 & -0.59 & 0.07 \\ -0.14 & 0.16 & -0.02 \end{bmatrix} \end{array},$$

and along with any remaining nullspace vectors produces a zero sum along any of the ray paths. $\mathbf{u}_6$ *is in the data nullspace.* $\mathbf{u}_6^T\mathbf{E} = 0$ *shows that*

$$a\left(y_1 + y_2 + y_3\right) - a\left(y_4 + y_5 + y_6\right) = 0,$$

if there is to be a solution without a residual, or alternatively, that no solution would permit this sum to be non-zero. This requirement is physically sensible, as it says that the vertical and horizontal rays cover the same territory and must therefore produce the same sum travel times. It shows why the rank is 5, and not 6.

There is no noise in the problem as stated. The correct solution and the SVD solution differ by the nullspace vectors. One can easily confirm that $\tilde{\mathbf{x}}$ *is column 5 of* $\mathbf{V}_5\mathbf{V}_5^T$.

Least-squares allows one to minimize (or maximize) anything one pleases. Suppose that for some reason we want the solution that minimizes the differences between the value in box 5 and its neighbors, perhaps as a way of finding a "smooth" solution. Let

$$\mathbf{W} = \left\{\begin{matrix} -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{matrix}\right\}. \tag{2.297}$$

The last row is included to render $\mathbf{W}$ *a full-rank matrix. Then*

$$\mathbf{Wx} = [x_5 - x_1 \quad x_5 - x_2 \quad \ldots \quad x_5 - x_9 \quad x_5]^T, \tag{2.298}$$

and we can minimize

$$J = \mathbf{x}^T\mathbf{W}^T\mathbf{W}\mathbf{x} \tag{2.299}$$

subject to $\mathbf{Ex} = \mathbf{y}$ *by finding the stationary value of*

$$J' = J - 2\boldsymbol{\mu}^T\left(\mathbf{y} - \mathbf{Ex}\right). \tag{2.300}$$

The normal equations are then

$$\mathbf{W}^T\mathbf{W}\mathbf{x} = \mathbf{E}^T\boldsymbol{\mu}, \tag{2.301}$$

$$\mathbf{Ex} = \mathbf{y}, \tag{2.302}$$

and

$$\tilde{\mathbf{x}} = (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{E}^T\boldsymbol{\mu}.$$

Then

$$\mathbf{E}(\mathbf{W}^{\mathrm{T}}\mathbf{W})^{-1}\mathbf{E}^{\mathrm{T}}\boldsymbol{\mu} = \mathbf{y}.$$

The rank of $\mathbf{E}(\mathbf{W}^{\mathrm{T}}\mathbf{W})^{-1}\mathbf{E}^{\mathrm{T}}$ *is* $K = 5 < M = 6$, *and so we need a generalized inverse,*

$$\tilde{\boldsymbol{\mu}} = (\mathbf{E}(\mathbf{W}^{\mathrm{T}}\mathbf{W})^{-1}\mathbf{E}^{\mathrm{T}})^{+}\mathbf{y} = \sum_{j=1}^{5} \mathbf{v}_i \frac{\mathbf{v}_i^{\mathrm{T}}\mathbf{y}}{\lambda_i}.$$

The nullspace of $\mathbf{E}(\mathbf{W}^{\mathrm{T}}\mathbf{W})^{-1}\mathbf{E}^{\mathrm{T}}$ *is the vector*

$$[-0.408 \quad -0.408 \quad -0.408 \quad 0.408 \quad 0.408 \quad 0.408]^{\mathrm{T}}, \tag{2.303}$$

which produces the solvability condition. Here, because $\mathbf{E}(\mathbf{W}^{\mathrm{T}}\mathbf{W})^{-1}\mathbf{E}^{\mathrm{T}}$ *is symmetric, the SVD reduces to the symmetric decomposition.*

Finally, the mapped $\tilde{\mathbf{x}}$ *is*

$$\begin{array}{cc} & r_x\rightarrow \\ r_y\uparrow & \begin{bmatrix} -0.20 & 0.41 & -0.20 \\ 0.41 & 0.18 & 0.41 \\ -0.20 & 0.41 & -0.21 \end{bmatrix} \end{array},$$

and one cannot further decrease the sum-squared differences of the solution elements. One can confirm that this solution satisfies the equations. Evidently, it produces a minimum, not a maximum (it suffices to show that the eigenvalues of $\mathbf{W}^{\mathrm{T}}\mathbf{W}$ *are all non-negative). The addition of any of the nullspace vectors of* $\mathbf{E}$ *to* $\tilde{\mathbf{x}}$ *will necessarily increase the value of* J *and hence there is no bounded maximum. In real tomographic problems, the arc lengths making up matrix* $\mathbf{E}$ *are three-dimensional curves and depend upon the background index of refraction in the medium, which is usually itself determined from observations.*[37] *There are thus errors in* $\mathbf{E}$ *itself, rendering the problem one of non-linear estimation. Approaches to solving such problems are described in Chapter 3.*

Example *Consider the box reservoir problem with two sources ("end members") described in Eqs. (2.144)–(2.145), which was reduced to two equations in two unknowns by dividing through by one of the unknown fluxes,* J_0, *and solving for the ratios* J_1/J_0, J_2/J_0. *Suppose they are solved instead in their original form as two equations in three unknowns:*

$$J_1 + J_2 - J_0 = 0,$$
$$C_1 J_1 + C_2 J_2 - C_0 J_0 = 0.$$

To make it numerically definite, let $C_1 = 1$, $C_2 = 2$, $C_0 = 1.75$. *The SVD produces*

$$\mathbf{U} = \begin{Bmatrix} -0.51 & -0.86 \\ -0.86 & 0.51 \end{Bmatrix}, \; \mathbf{V} = \begin{Bmatrix} -0.41 & -0.89 & 0.20 \\ -0.67 & 0.44 & 0.59 \\ 0.61 & -0.11 & 0.78 \end{Bmatrix}, \; \boldsymbol{\Lambda} = \mathrm{diag}\,([3.3, 0.39, 0])$$

(rounded). As the right-hand side of the governing equations vanishes, the coefficients of the range vectors, $\mathbf{v}_{1,2}$, *must also vanish, and the only possible solution here is proportional to the nullspace vector,* $\alpha_3\mathbf{v}_3$, *or* $[J_1, J_2, J_0] = \alpha_3[0.20, 0.59, 0.78]^{\mathrm{T}}$, *and* α_3 *is arbitrary. Alternatively,* $J_1/J_0 = 0.25$, $J_2/J_0 = 0.75$.

Example *Consider the flow into a four-sided box with missing integration constant as described in Chapter 1. Total mass conservation and conservation of dye is denoted by* C_i. *Let the relative areas of each interface be 1, 2, 3, 1 units respectively. Let the corresponding velocities on each side be* $1, 1/2, -2/3, 0$ *respectively, with the minus sign indicating a flow out. That mass is conserved is confirmed by*

$$1\,(1) + 2\left(\frac{1}{2}\right) + 3\left(\frac{-2}{3}\right) + 1\,(0) = 0.$$

Now suppose that the total velocity is not in fact known, but that an integration constant is missing on each interface, so that

$$1\left(\frac{1}{2} + b_1\right) + 2\,(1 + b_2) + 3\left(\frac{1}{3} + b_3\right) + 1\,(2 + b_4) = 0,$$

where the $b_i = [1/2, -1/2, -1, -2]$, *but are here treated as unknown. Then the above equation becomes*

$$b_1 + 2b_2 + 3b_3 + b_4 = -5.5,$$

or one equation in four unknowns. One linear combination of the unknown b_i *can be determined. We would like more information. Suppose that a tracer of concentration* $C_i = [2, 1, 3/2, 0]$ *is measured at each side, and is believed conserved. The governing equation is*

$$1\left(\frac{1}{2} + b_1\right)2 + 2\,(1 + b_2)\,1 + 3\left(\frac{1}{3} + b_3\right)\frac{3}{2} + 1\,(2 + b_4)\,0 = 0,$$

or

$$2b_1 + 2b_2 + 4.5b_3 + 0b_4 = -4.5,$$

giving a system of two equations in four unknowns:

$$\left\{\begin{matrix} 1 & 2 & 3 & 1 \\ 2 & 2 & 4.5 & 0 \end{matrix}\right\} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix} = \begin{bmatrix} -5.5 \\ -4.5 \end{bmatrix}.$$

The SVD of the coefficient matrix, **E**, *is*

$$\mathbf{E} = \left\{\begin{matrix} -0.582 & -0.813 \\ 0.813 & 0.582 \end{matrix}\right\} \left\{\begin{matrix} 6.50 & 0 & 0 & 0 \\ 0 & 1.02 & 0 & 0 \end{matrix}\right\}$$
$$\times \left\{\begin{matrix} -0.801 & 0.179 & -0.454 & 0.347 \\ 0.009 & 0.832 & 0.429 & 0.340 \\ -0.116 & 0.479 & -0.243 & -0.835 \\ 0.581 & 0.215 & -0.742 & 0.259 \end{matrix}\right\}.$$

The remainder of the solution is left to the reader.

2.5.6 Simple examples: differential and partial differential equations

Example *As an example of the use of this machinery with differential equations, consider*

$$\frac{\mathrm{d}^2 x(r)}{\mathrm{d}r^2} - k^2 x(r) = 0, \qquad (2.304)$$

subject to initial and/or boundary conditions. Using one-sided, uniform discretization,

$$x\left((m+1)\,\Delta r\right) - (2 + k^2(\Delta r)^2)x\,(m\Delta r) + x\left((m-1)\,\Delta r\right) = 0, \qquad (2.305)$$

at all interior points. Take the specific case, with two end conditions, $x\,(\Delta r) = 10,\ x\,(51\Delta r) = 1,\ \Delta r = 0.1$. *The numerical solution is depicted in Fig. 2.11 from the direct (conventional) solution to* $\mathbf{Ax} = \mathbf{y}$. *The first two rows of* **A** *were used to impose the boundary conditions on* $x(\Delta r), x(51\Delta r)$. *The singular values of* **A** *are also plotted in Fig. 2.11. The range is over about two orders of magnitude, and there is no reason to suspect numerical difficulties. The first and last singular vectors* $\mathbf{u}_1,\ \mathbf{v}_1,\ \mathbf{u}_{51}, \mathbf{v}_{51}$ *are also plotted. One infers (by plotting additional such vectors), that the large singular values correspond to singular vectors showing a great deal of small-scale structure, and the smallest singular values correspond to the least structured (largest spatial scales) in both the solution and in the specific corresponding weighted averages of the equations. This result may be counterintuitive. But note that, in this problem, all elements of* **y** *vanish except the first two, which are being used to set the boundary conditions. We know from the analytical*

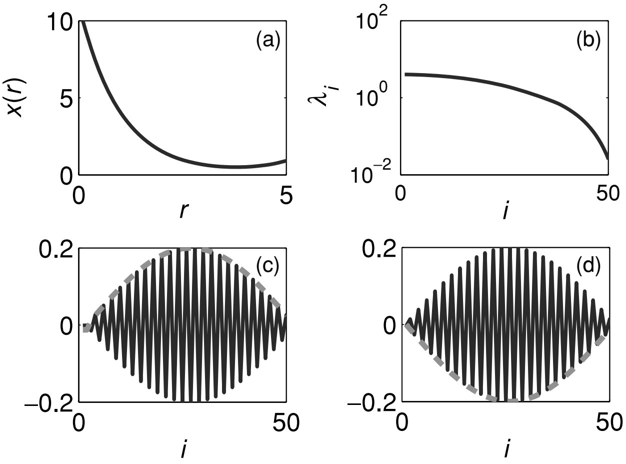

Figure 2.11 (a) $\tilde{\mathbf{x}}$ from Eq. (2.304) by brute force from the simultaneous equations; (b) displays the corresponding singular values; all are finite (there is no nullspace); (c) shows $\mathbf{u}_1$ (solid curve), and $\mathbf{u}_{51}$ (dashed); (d) shows the corresponding $\mathbf{v}_1$, $\mathbf{v}_{51}$. The most robust information corresponds to the *absence* of small scales in the solution.

solution that the true solution is large-scale; most of the information contained in the differential equation (2.304) or its numerical counterpart (2.305) is an assertion that all small scales are absent; this information is the most robust and corresponds to the largest singular values. The remaining information, on the exact nature of the largest scales, which is contained in only two of the 51 equations – given by the boundary conditions, is extremely important, but is less robust than that concerning the absence of small scales. (Less "robust" is being used in the sense that small changes in the boundary conditions will lead to relatively large changes in the large-scale structures in the solution because of the division by relatively small λ_i*.)*

Example *Consider now the classical Neumann problem described in Chapter 1. The problem is to be solved on a* 10×10 *grid as in Eq. (1.17),* $\mathbf{Ax} = \mathbf{b}$. *The singular values of* $\mathbf{A}$ *are plotted in Fig. 2.12; the largest one is* $\lambda_1 = 7.8$, *and the smallest non-zero one is* $\lambda_{99} = 0.08$. *As expected,* $\lambda_{100} = 0$. *The singular vector* $\mathbf{v}_{100}$ *corresponding to the zero singular value is a constant;* $\mathbf{u}_{100}$, *also shown in Fig. 2.12, is not a constant, and has considerable structure – which provides the solvability condition for the Neumann problem,* $\mathbf{u}_{100}^{\mathrm{T}}\mathbf{y} = 0$. *The physical origin of the solvability condition is readily understood: Neumann boundary conditions prescribe boundary flux rates, and the sum of the interior source strengths plus the boundary flux rates must sum to 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* $\mathbf{u}_{100}$

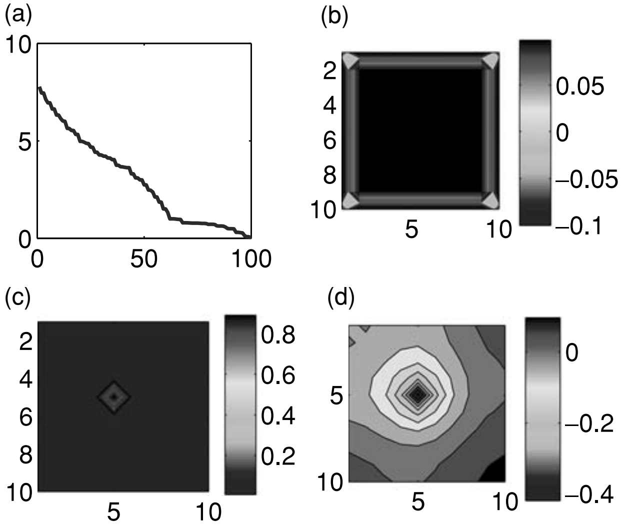

Figure 2.12 (a) Graph of the singular values of the coefficient matrix **A** of the numerical Neumann problem on a 10×10 grid. All λ_i are non-zero except the last one. (b) shows $\mathbf{u}_{100,}$ the nullspace vector of $\mathbf{E}^{\mathrm{T}}$ defining the solvability or consistency condition for a solution through $\mathbf{u}_{100}^{\mathrm{T}}\mathbf{y} = 0$. Plotted as mapped onto the two-dimensional spatial grid (r_x, r_y) with $\Delta x = \Delta y = 1$. The interpretation is that the sum of the influx through the boundaries and from interior sources must vanish. Note that corner derivatives differ from other boundary derivatives by $1/\sqrt{2}$. The corresponding $\mathbf{v}_{100}$ is a constant, indeterminate with the information available, and not shown. (c) A source **b** (a numerical delta function) is present, not satisfying the solvability condition $\mathbf{u}_{100}^{\mathrm{T}}\mathbf{b} = 0$, because all boundary fluxes were set to vanish. (d) The particular SVD solution, $\tilde{\mathbf{x}}$, at rank $K = 99$. One confirms that $\mathbf{A}\tilde{\mathbf{x}} - \mathbf{b}$ is proportional to $\mathbf{u}_{100}$ as the source is otherwise inconsistent with no flux boundary conditions. With **b** a Kronecker delta function at one grid point, this solution is a numerical Green function for the Neumann problem and insulating boundary conditions. (See color figs.)

on the interior grid points is a constant. The Neumann problem is thus a forward one requiring coping with both a solution nullspace and a solvability condition.

2.5.7 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: they *are* the least-squares solution, where it exists. When does the simple least-squares solution exist? Consider first the formally overdetermined problem, $M > N$. The solution (2.95) is meaningful if and only if the matrix inverse exists. Substituting the SVD for **E**, one finds that

$$(\mathbf{E}^{\mathrm{T}}\mathbf{E})^{-1} = \left(\mathbf{V}_N \boldsymbol{\Lambda}_N^{\mathrm{T}} \mathbf{U}_N^{\mathrm{T}} \mathbf{U}_N \boldsymbol{\Lambda}_N \mathbf{V}_N^{\mathrm{T}}\right)^{-1} = \left(\mathbf{V}_N \boldsymbol{\Lambda}_N^2 \mathbf{V}_N^{\mathrm{T}}\right)^{-1}, \tag{2.306}$$

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 non-zero diagonal elements* λ_i^2. The inverse in (2.306) may be found by inspection, using $\mathbf{V}_N^{\mathrm{T}}\mathbf{V}_N = \mathbf{I}_N$,

$$(\mathbf{E}^{\mathrm{T}}\mathbf{E})^{-1} = \mathbf{V}_N \mathbf{\Lambda}_N^{-2} \mathbf{V}_N^{\mathrm{T}}. \tag{2.307}$$

Then the solution (2.95) becomes

$$\tilde{\mathbf{x}} = \left(\mathbf{V}_N \mathbf{\Lambda}_N^{-2} \mathbf{V}_N^{\mathrm{T}}\right) \mathbf{V}_N \mathbf{\Lambda}_N \mathbf{U}_N^{\mathrm{T}} \mathbf{y} = \mathbf{V}_N \mathbf{\Lambda}_N^{-1} \mathbf{U}_N^{\mathrm{T}} \mathbf{y}, \tag{2.308}$$

which is identical to the SVD solution (2.286). If $K < N$, $\mathbf{\Lambda}_N^2$ has at least one zero on the diagonal, there is no matrix inverse, 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 non-vanishing, singular values is obvious.

That the full-rank overdetermined case is unbiassed, as previously asserted (p. 47), can now be seen from

$$\langle \tilde{\mathbf{x}} - \mathbf{x} \rangle = \sum_{i=1}^{N} \frac{\left(\mathbf{u}_i^{\mathrm{T}} \langle \mathbf{y} \rangle\right)}{\lambda_i} \mathbf{v}_i - \mathbf{x} = \sum_{i=1}^{N} \frac{\mathbf{u}_i^{\mathrm{T}} \mathbf{y}_0}{\lambda_i} \mathbf{v}_i - \mathbf{x} = \mathbf{0},$$

with $\mathbf{y} = \mathbf{y}_0 + \mathbf{n}$, if $\langle \mathbf{n} \rangle = \mathbf{0}$, assuming that the correct $\mathbf{E}$ (model) is being used.

Now consider another problem, the conventional purely underdetermined least-squares one, whose solution is (2.167). When does that exist? Substituting the SVD,

$$\begin{aligned} \tilde{\mathbf{x}} &= \mathbf{V}_M \mathbf{\Lambda}_M \mathbf{U}_M^{\mathrm{T}} \left(\mathbf{U}_M \mathbf{\Lambda}_M \mathbf{V}_M^{\mathrm{T}} \mathbf{V}_M \mathbf{\Lambda}_M^{\mathrm{T}} \mathbf{U}_M^{\mathrm{T}}\right)^{-1} \mathbf{y} \\ &= \mathbf{V}_M \mathbf{\Lambda}_M \mathbf{U}_M^{\mathrm{T}} \left(\mathbf{U}_M \mathbf{\Lambda}_M^2 \mathbf{U}_M^{\mathrm{T}}\right)^{-1} \mathbf{y}. \end{aligned} \tag{2.309}$$

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

$$\tilde{\mathbf{x}} = \mathbf{V}_M \mathbf{\Lambda}_M \mathbf{U}_M^{\mathrm{T}} \left(\mathbf{U}_M \mathbf{\Lambda}_M^{-2} \mathbf{U}_M^{\mathrm{T}}\right) \mathbf{y} = \mathbf{V}_M \mathbf{\Lambda}_M^{-1} \mathbf{U}_M^{\mathrm{T}} \mathbf{y}, \tag{2.310}$$

$$\tilde{\mathbf{n}} = 0, \tag{2.311}$$

which is once again the particular-SVD solution (2.286) – with 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 (2.310). The bias of (2.309) is given by the nullspace elements, and its uncertainty arises only from their contribution, because with $\tilde{\mathbf{n}} = \mathbf{0}$, the noise variance vanishes, and the particular-SVD solution covariance $\mathbf{C}_{xx}$ would be zero.

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* (2.264)*, is present in the least-squares solution*. If the system is not of full rank, then the simple least-squares solutions do not exist. *The SVD generalizes these results* by determining what it can: the elements of the solution lying in the range of $\mathbf{E}$, and an explicit structure for the resulting nullspace vectors.

The SVD provides a lot of flexibility. For example, it permits one to modify the simplest underdetermined solution (2.167) to remove its greatest shortcoming, the necessities that $\tilde{\mathbf{n}} = \mathbf{0}$ and that the residuals be orthogonal to all range vectors. One simply truncates the solution (2.270) at $K = K' < M$, thus assigning all vectors $\mathbf{v}_i$, $K' + i = 1, 2, \ldots, K$, to an "effective nullspace" (or substitutes K' for K everywhere). The residual is then

$$\tilde{\mathbf{n}} = \sum_{i=K'+1}^{M} \left(\mathbf{u}_i^{\mathrm{T}}\mathbf{y}\right)\mathbf{u}_i, \tag{2.312}$$

with an uncertainty for $\tilde{\mathbf{x}}$ given by (2.293), but with the upper limit being K' rather than K. Such truncation has the effect of reducing the solution covariance contribution to the uncertainty, but increasing the contribution owing to the nullspace (and increasing the bias). In the presence of singular values small compared to σ_n, the resulting overall reduction in uncertainty may be very great – at the expense of a solution bias.

The general solution now consists of three parts,

$$\tilde{\mathbf{x}} = \sum_{i=1}^{K'} \frac{\mathbf{u}_i^{\mathrm{T}}\mathbf{y}}{\lambda_i}\mathbf{v}_i + \sum_{i=K'+1}^{K} \alpha_i\,\mathbf{v}_i + \sum_{i=K+1}^{N} \alpha_i\mathbf{v}_i, \tag{2.313}$$

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. Resolution and covariance matrices are modified by the substitution of K' for K.

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.

It was already noticed that the simplest form of least-squares does not provide a method to control the ratios of the solution and noise norms. Evidently, truncation of the SVD offers a simple way to do so – 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 (see p. 116).

2.5.8 Pseudo-inverses

Consider an arbitrary $M \times N$ matrix $\mathbf{E} = \mathbf{U}_K \boldsymbol{\Lambda}_K \mathbf{V}_K^{\mathrm{T}}$ and

$$\mathbf{Ex} + \mathbf{n} = \mathbf{y}.$$

Then if $\mathbf{E}$ is full-rank underdetermined, the minimum norm solution is

$$\tilde{\mathbf{x}} = \mathbf{E}^{\mathrm{T}}(\mathbf{E}\mathbf{E}^{\mathrm{T}})^{-1}\mathbf{y} = \mathbf{V}_K \boldsymbol{\Lambda}_K^{-1} \mathbf{U}_K^{\mathrm{T}} \mathbf{y}, \quad K = M,$$

and if it is full-rank overdetermined, the minimum noise solution is

$$\tilde{\mathbf{x}} = (\mathbf{E}^{\mathrm{T}}\mathbf{E})^{-1}\mathbf{E}^{\mathrm{T}}\mathbf{y} = \mathbf{V}_K \boldsymbol{\Lambda}_K^{-1} \mathbf{U}_K^{\mathrm{T}} \mathbf{y}, \quad K = N.$$

The first of these, the Moore–Penrose, or pseudo-inverse, $\mathbf{E}_1^{+} = \mathbf{E}^{\mathrm{T}}(\mathbf{E}\mathbf{E}^{\mathrm{T}})^{-1}$ is sometimes also known as a "right-inverse," because $\mathbf{E}\mathbf{E}_1^{+} = \mathbf{I}_M$. The second pseudo-inverse, $\mathbf{E}_2^{+} = (\mathbf{E}^{\mathrm{T}}\mathbf{E})^{-1}\mathbf{E}^{\mathrm{T}}$, is a "left-inverse" as $\mathbf{E}_2^{+}\mathbf{E} = \mathbf{I}_N$. They can both be represented as $\mathbf{V}_K \boldsymbol{\Lambda}_K^{-1} \mathbf{U}_K^{\mathrm{T}}$, but with differing values of K. If $K < M, N$ neither of the pseudo-inverses exists, but $\mathbf{V}_K \boldsymbol{\Lambda}_K^{-1} \mathbf{U}_K^{\mathrm{T}} \mathbf{y}$ still provides the particular SVD solution. When $K = M = N$, one has a demonstration that the left and right inverses are identical; they are then written as $\mathbf{E}^{-1}$.

2.5.9 Row and column scaling

The effects on the least-squares solutions of the row and column scaling can now be understood. We discuss them in the context of noise covariances, but as always in least-squares, the weight matrices need no statistical interpretation, and can be chosen by the investigator to suit his or her convenience or taste.

Suppose we have two equations written as

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

where $\mathbf{R}_{nn} = \mathbf{I}_2$, $\mathbf{W} = \mathbf{I}_3$. The SVD of $\mathbf{E}$ is

$$\mathbf{U} = \left\{\begin{matrix} 0.7059 & -0.7083 \\ 0.7083 & 0.7059 \end{matrix}\right\}, \qquad \mathbf{V} = \left\{\begin{matrix} 0.5764 & -0.4096 & 0.7071 \\ 0.5793 & 0.8151 & 0.0000 \\ 0.5764 & -0.4096 & -0.7071 \end{matrix}\right\},$$

$$\lambda_1 = 2.4536, \qquad \lambda_2 = 0.0058.$$

The SVD solutions, choosing ranks $K' = 1, 2$ in succession, are very nearly (the numbers having been rounded)

$$\tilde{\mathbf{x}} \approx \left(\frac{y_1 + y_2}{2.45}\right)[0.58 \quad 0.58 \quad 0.58]^{\mathrm{T}}, \tag{2.314}$$
$$\tilde{\mathbf{x}} \approx \left(\frac{y_1 + y_2}{2.45}\right)[0.58 \quad 0.58 \quad 0.58]^{\mathrm{T}} + \left(\frac{y_1 - y_2}{0.0058}\right)[-0.41 \quad 0.82 \quad 0.41]^{\mathrm{T}},$$

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. The uncertainty is given by

$$(\mathbf{E}\mathbf{E}^{\mathrm{T}})^{-1} = \begin{Bmatrix} 1.51 \times 10^4 & -1.50 \times 10^4 \\ -1.50 \times 10^4 & 1.51 \times 10^4 \end{Bmatrix}.$$

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

$$\mathbf{R}_{nn} = \begin{Bmatrix} 1 & 0.999999 \\ 0.999999 & 1 \end{Bmatrix},$$

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

$$\mathbf{W}^{1/2} = \begin{Bmatrix} 1.0000 & 1.0000 \\ 0 & 0.0014 \end{Bmatrix}, \qquad \mathbf{W}^{-\mathrm{T}/2} = \begin{Bmatrix} 1.0000 & 0 \\ -707.1063 & 707.1070 \end{Bmatrix}.$$

The new system to be solved is

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

The SVD is

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

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

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

$$\tilde{\mathbf{x}} \approx \frac{y_2 - y_1}{7.1}[0 \quad 1 \quad 0]^{\mathrm{T}}, \tag{2.315}$$
$$\tilde{\mathbf{x}} \approx \frac{y_2 - y_1}{7.1}[0 \quad 1 \quad 0]^{\mathrm{T}} + \frac{y_1 - 103\,(y_2 - y_1)}{1.4}[0.71 \quad 0 \quad 0.71\,]^{\mathrm{T}},$$

and the rank 1 solution is given by the difference of the observations, in contrast to the unscaled solution. The result is quite sensible – the noise in the two equations is so nearly perfectly correlated that it can be removed by subtraction; the difference $y_2 - y_1$ is a nearly noise-free piece of information and accurately defines the appropriate structure in $\tilde{\mathbf{x}}$. In effect, the information provided in the row scaling with $\mathbf{R}$ 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 (2.314) at rank 1 only by averaging.

At full rank, that is, $K = 2$, it can be confirmed that the solutions (2.314) and (2.315) are identical, as they must be. But the error covariances are quite different:

$$(\mathbf{E}'\mathbf{E}'^{\mathrm{T}})^{-1} = \begin{Bmatrix} 0.5001 & -0.707 \\ -0.707 & 0.5001 \end{Bmatrix}. \tag{2.316}$$

because the imposed covariance permits a large degree of noise suppression.

It was previously asserted (p. 67) 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:

$$\begin{aligned} \tilde{\mathbf{x}} &= \mathbf{E}'^{\mathrm{T}}(\mathbf{E}'\mathbf{E}'^{\mathrm{T}})^{-1}\mathbf{y}' \\ &= \mathbf{E}^{\mathrm{T}}\mathbf{W}^{-1/2}\left(\mathbf{W}^{-\mathrm{T}/2}\mathbf{E}\mathbf{E}^{\mathrm{T}}\mathbf{W}^{-1/2}\right)^{-1}\mathbf{W}^{-\mathrm{T}/2}\mathbf{y} \\ &= \mathbf{E}^{\mathrm{T}}\mathbf{W}^{-1/2}\mathbf{W}^{1/2}(\mathbf{E}\mathbf{E}^{\mathrm{T}})^{-1}\mathbf{W}^{\mathrm{T}/2}\mathbf{W}^{-\mathrm{T}/2}\mathbf{y} \\ &= \mathbf{E}^{\mathrm{T}}(\mathbf{E}\mathbf{E}^{\mathrm{T}})^{-1}\mathbf{y}, \end{aligned} \tag{2.317}$$

but which is true only in the full rank situation.

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

$$\begin{aligned} 10x_1 + 5x_2 + x_3 + n_1 &= 1, \\ 100x_1 + 50x_2 + 10x_3 + n_2 &= 2, \end{aligned} \tag{2.318}$$

after row scaling to make the expected noise variance in each the same. A rank 1 solution to these equations by SVD is $\tilde{\mathbf{x}} = [0.0165, 0.0083, 0.0017]^{\mathrm{T}}$, which produces residuals $\tilde{\mathbf{y}} - \mathbf{y} = [-0.79, 0.079]^{\mathrm{T}}$ – much smaller in the second equation than in the first one.

Consider that the second equation is ten times the first one – in effect saying that a measurement of ten 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 represents a much more accurate determination of this linear combination and the equation should be given much more weight in determining the unknowns – and both 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 $(\sum_j E_{ij}^2)^{1/2}$. But there will be a contradiction with any assertion that the noise in all equations was the same to begin with. Such row scaling is best regarded as non-statistical in nature.

An example of this situation is readily apparent in the box balances discussed in Chapter 1. Equations such as (1.32) could have row norms much larger than those (1.31) for the corresponding mass balance, simply because the tracer is measured by convention in its own units. If the tracer is, e.g., oceanic salt, values are, by convention, measured on the Practical Salinity Scale, and are near 35 (but are dimensionless). Because there is nothing fundamental about the choice of units, it seems unreasonable to infer that the requirement of tracer balance has an expected error 35 times smaller than for mass. One usually proceeds in the obvious way by dividing the tracer equations by their row norms as the first step. (This approach need have no underlying statistical validity, but is often done simply on the assumption that salt balance equations are unlikely to be 35 times more accurate than the mass ones.) 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.

Column scaling

In the least-squares problem, we formally introduced a "column scaling" matrix $\mathbf{S}$. 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 (and row) scaled system, the solution is for the scaled unknown $\mathbf{x}'$, and the physical solution is

$$\tilde{\mathbf{x}} = \mathbf{S}^{\mathrm{T}/2}\tilde{\mathbf{x}}'. \qquad (2.319)$$

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

Example *Suppose we have one equation in two unknowns:*

$$10x_1 + 1x_2 = 3. \qquad (2.320)$$

The particular-SVD solution produces $\tilde{\mathbf{x}} = [0.2970, 0.0297]^{\mathrm{T}}$, *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 automatically finds the exact solution, subject to making the solution norm as small as possible. Because the coefficient of x_1 in (2.320) is ten times that of x_2, it is 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, the numerical values appearing in the coefficient matrix* $\mathbf{E}$ *are an "accident." In the box-balance example of Chapter 1, the distances defining the interfaces of the boxes are a consequence of the spatial distance between measurements. Unless one believed that velocities should be larger where the distances are greater or the fluid depth was greater, then the solutions may behave unphysically.*[38] *Indeed, in some situations the velocities are expected to be inverse to the fluid 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 rigid. In particular, the equations themselves may preclude the proportionality.) Take a positive definite, diagonal matrix* $\mathbf{S}$, *and rewrite (2.87) as*

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

Then

$$\mathbf{E}'\mathbf{x}'+\mathbf{n}=\mathbf{y},\ \mathbf{E}'=\mathbf{E}\mathbf{S}^{\mathrm{T}/2},\ \mathbf{x}'=\mathbf{S}^{-\mathrm{T}/2}\mathbf{x}.$$

Solving

$$\tilde{\mathbf{x}}'=\mathbf{E}'^{\mathrm{T}}(\mathbf{E}'\mathbf{E}'^{\mathrm{T}})^{-1}\mathbf{y},\ \tilde{\mathbf{x}}=\mathbf{S}^{\mathrm{T}/2}\tilde{\mathbf{x}}'. \tag{2.321}$$

How should $\mathbf{S}$ *be chosen? Apply the recipe (2.321) for the one equation example of (2.320), with*

$$\mathbf{S}=\begin{Bmatrix}1/a^2 & 0\\ 0 & 1/b^2\end{Bmatrix},$$

$$\mathbf{E}'=\begin{Bmatrix}10/a & 1/b\end{Bmatrix},\mathbf{E}'\mathbf{E}'^{\mathrm{T}}=\frac{100}{a^2}+\frac{1}{b^2}, \tag{2.322}$$

$$(\mathbf{E}'\mathbf{E}'^{\mathrm{T}})^{-1}=\frac{a^2b^2}{100b^2+a^2}, \tag{2.323}$$

$$\tilde{\mathbf{x}}'=\begin{Bmatrix}10/a\\ 1/b\end{Bmatrix}\frac{a^2b^2}{100b^2+a^2}3, \tag{2.324}$$

$$\tilde{\mathbf{x}}=\mathbf{S}^{\mathrm{T}/2}\tilde{\mathbf{x}}'=\begin{Bmatrix}10/a^2\\ 1/b^2\end{Bmatrix}\frac{a^2b^2}{100b^2+a^2}3. \tag{2.325}$$

The relative magnitudes of the elements of $\tilde{\mathbf{x}}$ *are proportional to* $10/a^2$, $1/b^2$. *To make the numerical values identical, choose* $a^2=10$, $b^2=1$, *that is, 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 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* $\mathbf{S}^{\mathrm{T}/2}$ *in (2.321). This form of column scaling should be regarded as "non-statistical," in that it is based upon inferences about 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. In general, the elements of* $\tilde{\mathbf{x}}$ *will not prove to be equal – because the equations themselves do not permit it.*

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

$$\begin{aligned}\tilde{\mathbf{x}}' &= (\mathbf{E}'^{\mathrm{T}}\mathbf{E}')^{-1}\mathbf{E}'^{\mathrm{T}}\mathbf{y}\\ \tilde{\mathbf{x}} &= \mathbf{S}^{\mathrm{T}/2}(\mathbf{S}^{1/2}\mathbf{E}^{\mathrm{T}}\mathbf{E}\mathbf{S}^{\mathrm{T}/2})^{-1}\mathbf{S}^{1/2}\mathbf{E}^{\mathrm{T}}\mathbf{y}\\ &= \mathbf{S}^{\mathrm{T}/2}\mathbf{S}^{-\mathrm{T}/2}(\mathbf{E}^{\mathrm{T}}\mathbf{E})^{-1}\mathbf{S}^{-1/2}\mathbf{S}^{1/2}\mathbf{E}^{\mathrm{T}}\mathbf{y} = (\mathbf{E}^{\mathrm{T}}\mathbf{E})^{-1}\mathbf{E}^{\mathrm{T}}\mathbf{y}.\end{aligned} \tag{2.326}$$

Usually row scaling is done prior to column scaling so that the row norms have a simple physical interpretation, but one can row normalize in the column normalized space.

2.5.10 Solution and observation resolution: data ranking

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 (2.187). It follows immediately from Eqs. (2.188) that the particular-SVD solution is

$$\tilde{\mathbf{x}} = \mathbf{V}_K\mathbf{V}_K^{\mathrm{T}}\mathbf{x} = \mathbf{T}_v\mathbf{x}, \tag{2.327}$$

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

$$\tilde{\mathbf{y}} = \mathbf{U}_K\mathbf{U}_K^{\mathrm{T}}\mathbf{y} \; = \mathbf{T}_u\mathbf{y}. \tag{2.328}$$

It is convenient therefore, to define the solution and observation resolution matrices,

$$\mathbf{T}_v = \mathbf{V}_K\mathbf{V}_K^{\mathrm{T}}, \qquad \mathbf{T}_u = \mathbf{U}_K\mathbf{U}_K^{\mathrm{T}}. \tag{2.329}$$

The interpretation of the solution resolution matrix is identical to that in the square-symmetric case (p. 77).

Interpretation of the data resolution matrix is slightly subtle. Suppose an element of $\mathbf{y}$ was fully resolved, that is, some column, j_0, of $\mathbf{U}_K\mathbf{U}_K^{\mathrm{T}}$ 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

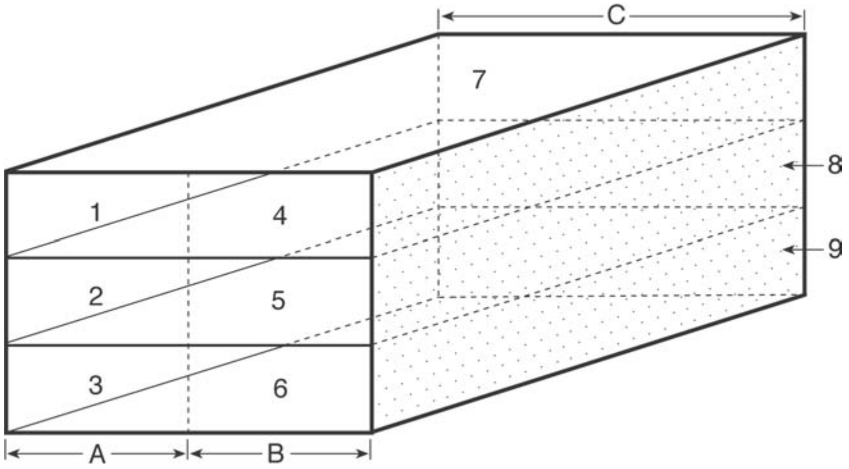

Figure 2.13 Box model with mass fluxes across the bounding interfaces $i = 1, \ldots, 9$. Mass conservation equations are written for the total volume, for specifying the flux across the southern and northern boundaries separately, and for the mass balance in the three layers shown.

resolved, the system lacks adequate information to distinguish equation i from a linear dependence on one or more other equations.[39]

One can use these ideas to construct quantitative statements of which observations are the most important ("data ranking"). From (2.190), 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 (2.318) without row weighting. At rank 1,

$$\mathbf{T}_u = \begin{Bmatrix} 0.0099 & 0.099 \\ 0.099 & 0.9901 \end{Bmatrix},$$

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. The elements of $\mathbf{T}_u$ can be used to rank the data in order of importance to the final solution. All of the statements about the properties of resolution matrices made above apply to both $\mathbf{T}_u$, $\mathbf{T}_v$.

Example *A fluid flows into a box as depicted in Fig. 2.13, which is divided into three layers. The box is bounded on the left and right by solid walls. At its southern boundary, the inward (positive directed) mass flow has six unknown values, three each into the layers on the west (region A, q_i, $i = 1, 2, 3$), and three each into the* *layers on the east (region B, q_i, $i = 4, 5, 6$). At the northern boundary, there are only three unknown flow fields, for which a positive value denotes a flow outward (region C, $i = 7, 8, 9$). We write seven mass conservation equations. The first three represent mass balance in each of the three layers. The second three fix the mass transports across each of the sections A, B, C. The final equation asserts that the sum of the three mass transports across the sections must vanish (for overall mass balance). Note that the last equation is a linear combination of equations 4 to 6. Then the equations are*

$$\left\{\begin{matrix} 1 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & -1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 \end{matrix}\right\} \mathbf{x} + \mathbf{n} = \mathbf{y}, \tag{2.330}$$

which is 7×9. Inspection, or the singular values, demonstrate that the rank of the coefficient matrix, $\mathbf{E}$, is $K = 5$ because of the linear dependencies noted. Figures 2.14 and 2.15 display the singular vectors and the rank 5 resolution matrices for Eq. (2.330). Notice that the first pair of singular vectors, $\mathbf{u}_1$, $\mathbf{v}_1$ show that the mean mass flux (accounting for the sign convention making the northern flow positive outwards) is determined by a simple sum over all of the equations. Other linear combinations of the mass fluxes are determined from various sums and differences of the equations. It is left to the reader to determine whether they make physical sense. The resolution matrices show that none of the equations nor elements of the mass flux are fully resolved, and that all the equations contribute roughly equally.

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

$$\mathbf{T}_v = \mathbf{S}^{\mathrm{T}/2}\mathbf{T}_{v'}\mathbf{S}^{-\mathrm{T}/2}, \tag{2.331}$$

$$\mathbf{T}_u = \mathbf{W}^{\mathrm{T}/2}\mathbf{T}_{u'}\mathbf{W}^{-\mathrm{T}/2}, \tag{2.332}$$

so that

$$\tilde{\mathbf{x}} = \mathbf{T}_v\mathbf{x}, \qquad \tilde{\mathbf{y}} = \mathbf{T}_u\mathbf{y}, \tag{2.333}$$

where $\mathbf{T}_{v'}$, $\mathbf{T}_{u'}$ are the expressions Eq. (2.329) in the scaled space. Some insight into the effects of weighting can be obtained by applying row and column scaling

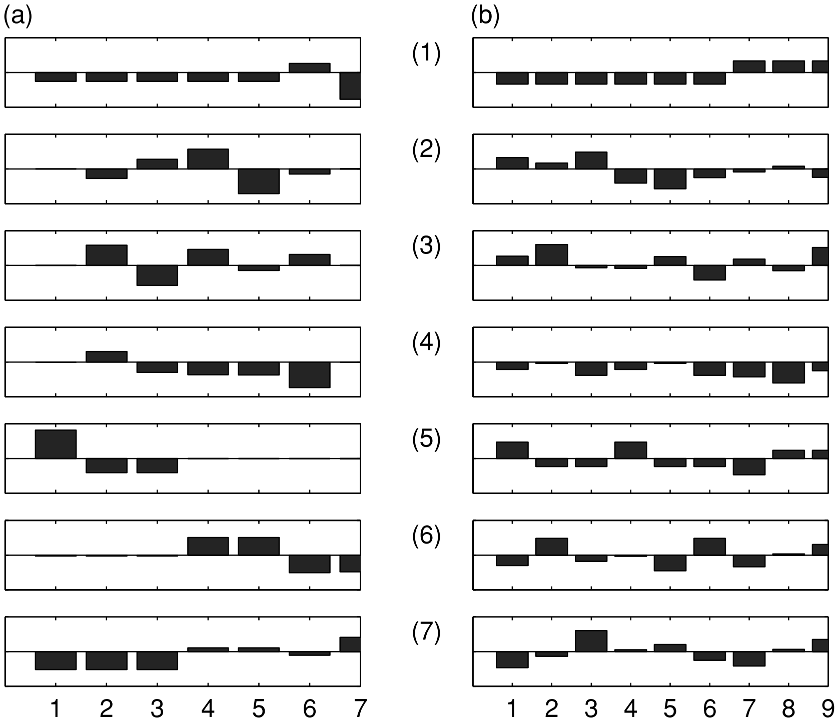

Figure 2.14 Singular vectors of the box flow coefficient matrix. Column (a) shows the $\mathbf{u}_i$, the column (b) the $\mathbf{v}_i$. The last two null space vectors $\mathbf{v}_8$, $\mathbf{v}_9$ are not shown, but $\mathbf{u}_6$, $\mathbf{u}_7$ do lie in the data nullspace, and $\mathbf{v}_6$, $\mathbf{v}_7$ are also in the solution nullspace. All plots have a full scale of ± 1.

to the simple physical example of Eq. (2.330). The uncertainty in the new space is $\mathbf{P} = \mathbf{S}^{1/2}\mathbf{P}'\mathbf{S}^{\mathrm{T}/2}$ where $\mathbf{P}'$ is the uncertainty in the scaled space.

We have seen an interpretation of three matrices obtained from the SVD: $\mathbf{V}_K\mathbf{V}_K^{\mathrm{T}}$, $\mathbf{U}_K\mathbf{U}_K^{\mathrm{T}}$, $\mathbf{V}_K\boldsymbol{\Lambda}_K^{-2}\mathbf{V}_K^{\mathrm{T}}$. 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 $\mathbf{U}_K\boldsymbol{\Lambda}_K^{-2}\mathbf{U}_K^{\mathrm{T}}$? To understand its use, recall the normal equations (2.163, 2.164) that emerged from the constrained objective function (2.149). They become, using the SVD for $\mathbf{E}$,

$$\mathbf{V}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{T}}\boldsymbol{\mu} = \mathbf{x}, \tag{2.334}$$

$$\mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^{\mathrm{T}}\mathbf{x} = \mathbf{y}. \tag{2.335}$$

The pair of equations is always square, of dimension $M + N$. These equations show that $\mathbf{U}\boldsymbol{\Lambda}^2\mathbf{U}^{\mathrm{T}}\boldsymbol{\mu} = \mathbf{y}$. The particular SVD solution is

$$\boldsymbol{\mu} = \mathbf{U}_K\boldsymbol{\Lambda}_K^{-2}\mathbf{U}_K^{\mathrm{T}}\mathbf{y}, \tag{2.336}$$

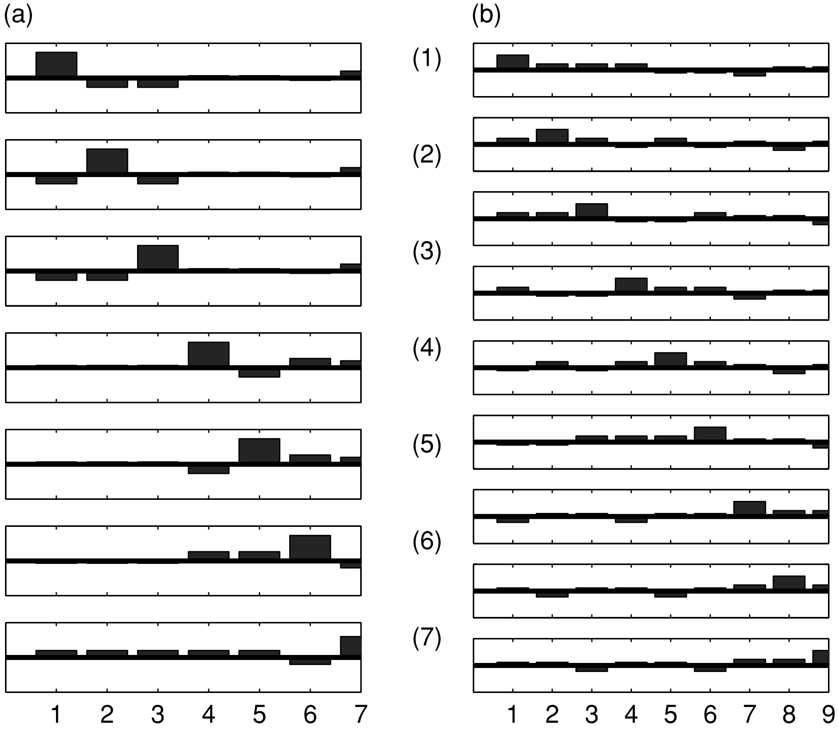

Figure 2.15 Rank $K = 5$ $\mathbf{T}_u$ (a) and $\mathbf{T}_v$ (b) for the box model coefficient matrix. Full scale is ±1 in all plots.

involving the "missing" fourth matrix. Thus,

$$\frac{\partial J}{\partial \mathbf{y}} = 2\mathbf{U}_K \boldsymbol{\Lambda}_K^{-2} \mathbf{U}_K^{\mathrm{T}} \mathbf{y},$$

and, taking the second derivative,

$$\frac{\partial^2 J}{\partial \mathbf{y}^2} = 2\mathbf{U}_K \boldsymbol{\Lambda}_K^{-2} \mathbf{U}_K^{\mathrm{T}}, \tag{2.337}$$

which is the Hessian of J with respect to the data. If any of the λ_i become very small, the objective function will be extremely sensitive to small perturbations in $\mathbf{y}$ – producing an effective nullspace of the problem. Equation (2.337) supports the suggestion that perfect constraints can lead to difficulties.

2.5.11 Relation to tapered and weighted least-squares

In using least-squares, a shift was made from the simple objective functions (2.89) and (2.149) to the more complicated ones in (2.114) or (2.125). 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 (2.114), but the SVD is actually useful in understanding the effect of such functions.

Assume that any necessary $\mathbf{W}, \mathbf{S}$ have been applied. Then the full SVD, including zero singular values and corresponding singular vectors, is substituted into (2.116),

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

and

$$\begin{aligned}\tilde{\mathbf{x}} &= \mathbf{V}(\boldsymbol{\Lambda}^{\mathrm{T}}\boldsymbol{\Lambda} + \gamma^2\mathbf{I})^{-1}\mathbf{V}^{\mathrm{T}}\mathbf{V}\boldsymbol{\Lambda}^{\mathrm{T}}\mathbf{U}^{\mathrm{T}}\mathbf{y} \\ &= \mathbf{V}\,\mathrm{diag}\left(\lambda_i^2 + \gamma^2\right)^{-1}\boldsymbol{\Lambda}^{\mathrm{T}}\mathbf{U}^{\mathrm{T}}\mathbf{y},\end{aligned} \tag{2.338}$$

or

$$\tilde{\mathbf{x}} = \sum_{i=1}^{N} \frac{\lambda_i\left(\mathbf{u}_i^{\mathrm{T}}\mathbf{y}\right)}{\lambda_i^2 + \gamma^2}\mathbf{v}_i. \tag{2.339}$$

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 $\mathbf{v}_i$ by the presence of γ^2 from the values they would have in the "pure" SVD. In particular, the guarantee that matrices like $(\mathbf{E}^{\mathrm{T}}\mathbf{E} + \gamma^2\mathbf{I})$ always have an inverse despite vanishing singular values, is seen to follow because the presence of $\gamma^2 > 0$ assures that the inverse of the sum always exists, irrespective of the rank of $\mathbf{E}$. The simple addition of a positive constant to the diagonal of a singular matrix is a well-known ad hoc method for giving it an approximate inverse. Such methods are a form of what is usually known as "regularization," and are procedures for suppressing nullspaces. Note that the coefficients of $\mathbf{v}_i$ vanish with λ_i and a solution nullspace still exists.

The residuals of the tapered least-squares solution can be written in various forms. Eqs. (2.117) are

$$\begin{aligned}\tilde{\mathbf{n}} &= \gamma^2\mathbf{U}(\gamma^2\mathbf{I} + \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathrm{T}})^{-1}\mathbf{U}^{\mathrm{T}}\mathbf{y} \\ &= \sum_{i=1}^{M}\frac{\left(\mathbf{u}_i^{\mathrm{T}}\mathbf{y}\right)\gamma^2}{\lambda_i^2 + \gamma^2}\mathbf{u}_i, \quad \gamma^2 > 0,\end{aligned} \tag{2.340}$$

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 $\mathbf{E}^{\mathrm{T}}$ is being attributed to noise and it is no longer true that the residuals are subject to the rigid requirement (2.264) 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 (2.339) becomes overly large – values we can control by varying γ^2. The covariance of this solution about its mean (Eq. (2.118)) is readily rewritten as

$$\begin{aligned}\mathbf{C}_{xx} &= \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{\lambda_i \lambda_j \mathbf{u}_i^{\mathrm{T}} \mathbf{R}_{nn} \mathbf{u}_j^{\mathrm{T}}}{\left(\lambda_i^2+\gamma^2\right)\left(\lambda_j^2+\gamma^2\right)} \mathbf{v}_i \mathbf{v}_j^{\mathrm{T}} \\ &= \sigma_n^2 \sum_{i=1}^{N} \frac{\lambda_i^2}{\left(\lambda_i^2+\gamma^2\right)^2} \mathbf{v}_i \mathbf{v}_i^{\mathrm{T}} \\ &= \sigma_n^2 \mathbf{V}(\boldsymbol{\Lambda}^{\mathrm{T}}\boldsymbol{\Lambda}+\gamma^2\mathbf{I}_N)^{-1}\boldsymbol{\Lambda}^{\mathrm{T}}\boldsymbol{\Lambda}(\boldsymbol{\Lambda}^{\mathrm{T}}\boldsymbol{\Lambda}+\gamma^2\mathbf{I}_N)^{-1}\mathbf{V}^{\mathrm{T}},\end{aligned} \tag{2.341}$$

where the second and third lines are again the special case of white noise. The role of γ^2 in controlling the solution variance, as well as the solution size, should be plain. The tapered least-squares solution is biassed – but the presence of the bias can greatly reduce the solution variance. Study of the solution as a function of γ^2 is known as "ridge regression." Elaborate techniques have been developed for determining the "right" value of γ^2.[40]

The uncertainty, **P**, is readily found as

$$\begin{aligned}\mathbf{P} &= \gamma^2 \sum_{i=1}^{N} \frac{\mathbf{v}_i\mathbf{v}_i^{\mathrm{T}}}{(\lambda_i^2+\gamma^2)^2} + \sigma_n^2 \sum_{i=1}^{N} \frac{\lambda_i^2 \mathbf{v}_i\mathbf{v}_i^{\mathrm{T}}}{(\lambda_i^2+\gamma^2)^2} \\ &= \gamma^2\mathbf{V}(\boldsymbol{\Lambda}^{\mathrm{T}}\boldsymbol{\Lambda}+\gamma^2\mathbf{I})^{-2}\mathbf{V}^{\mathrm{T}} + \sigma_n^2\mathbf{V}(\boldsymbol{\Lambda}^{\mathrm{T}}\boldsymbol{\Lambda}+\gamma^2\mathbf{I})^{-1}\boldsymbol{\Lambda}^{\mathrm{T}}\boldsymbol{\Lambda}(\boldsymbol{\Lambda}^{\mathrm{T}}\boldsymbol{\Lambda}+\gamma^2\mathbf{I})^{-1}\mathbf{V}^{\mathrm{T}},\end{aligned} \tag{2.342}$$

showing the variance reduction possible for finite γ^2 (reduction of the second term), and the bias error incurred in compensation in the first term.

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 the purpose stated above – to extend ordinary least-squares in such a way that one can control the relative noise and solution 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: the basic 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: then both solution and residuals contain some contribution from both formal range and effective nullspaces ($0 < \gamma^2$).

We have already seen several times that preventing $\tilde{\mathbf{n}}$ from having any contribution from the range of $\mathbf{E}^{\mathrm{T}}$ introduces covariances into the residuals, with a consequent inability to produce values that are strictly white noise in character (although it is only a real issue as the number of degrees of freedom, $M - K$, goes toward zero). In the tapered form of least-squares, or the equivalent tapered SVD, contributions from the range vectors $\mathbf{u}_i$, $i = 1, 2, \ldots, K$, are permitted, and a potentially more realistic residual estimate is obtained. (There is usually no good reason why $\tilde{\mathbf{n}}$ is expected to be orthogonal to the range vectors.)

2.5.12 Resolution and variance of tapered solutions

The tapered least-squares solutions have an implicit nullspace, arising both from the terms corresponding to zero singular values, or from values small compared to γ^2. To obtain a measure of solution resolution when the $\mathbf{v}_i$ vectors have not been computed, 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 correct value of $\mathbf{y}$ would be

$$\mathbf{E}\mathbf{x}_{j_0} = \mathbf{y}_{j_0}, \tag{2.343}$$

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

Assuming all covariance matrices have been applied and suppressing any primes, tapered least-squares (Eq. (2.120)) produces

$$\tilde{\mathbf{x}}_{j_0} = \mathbf{E}^{\mathrm{T}}(\mathbf{E}\mathbf{E}^{\mathrm{T}} + \gamma^2\mathbf{I})^{-1}\mathbf{y}_{j_0} = \mathbf{E}^{\mathrm{T}}(\mathbf{E}\mathbf{E}^{\mathrm{T}} + \gamma^2\mathbf{I})^{-1}\mathbf{E}\mathbf{x}_{j_0}, \tag{2.344}$$

which is row (or column) j_0 of

$$\mathbf{T}_v = \mathbf{E}^{\mathrm{T}}(\mathbf{E}\mathbf{E}^{\mathrm{T}} + \gamma^2\mathbf{I})^{-1}\mathbf{E}. \tag{2.345}$$

Thus we can interpret any row or column of $\mathbf{T}_v$ as the solution for one in which a Kronecker delta was the underlying correct one. It is an easy matter, using the SVD of $\mathbf{E}$ and letting $\gamma^2 \to 0$, to show that (2.345) reduces to $\mathbf{V}\mathbf{V}^{\mathrm{T}}$, if $K = M$. These expressions apply in the row- and column-scaled space and are suitably modified to take account of any $\mathbf{W}$, $\mathbf{S}$ which may have been applied, as in Eqs. (2.331) and (2.332). An obvious variant of (2.345) follows from the alternative least-squares solution (2.127), with $\mathbf{W} = \gamma^2\mathbf{I}$, $\mathbf{S} = \mathbf{I}$,

$$\mathbf{T}_v = (\mathbf{E}^{\mathrm{T}}\mathbf{E} + \gamma^2\mathbf{I})^{-1}\mathbf{E}^{\mathrm{T}}\mathbf{E}. \tag{2.346}$$

Data resolution matrices are obtained similarly. Let $y_j = \delta_{jj_1}$. Equation (2.135) produces

$$\tilde{\mathbf{x}}_{j_1} = \mathbf{E}^{\mathrm{T}}(\mathbf{E}\mathbf{E}^{\mathrm{T}} + \gamma^2\mathbf{I})^{-1}\mathbf{y}_{j_1}, \tag{2.347}$$

which if substituted into the original equations is

$$\mathbf{E}\tilde{\mathbf{x}}_{j_1} = \mathbf{E}\mathbf{E}^{\mathrm{T}}(\mathbf{E}\mathbf{E}^{\mathrm{T}} + \gamma^2\mathbf{I})^{-1}\mathbf{y}_{j_1}. \tag{2.348}$$

Thus,

$$\mathbf{T}_u = \mathbf{E}\mathbf{E}^{\mathrm{T}}(\mathbf{E}\mathbf{E}^{\mathrm{T}} + \gamma^2\mathbf{I})^{-1}. \tag{2.349}$$

The alternate form is

$$\mathbf{T}_u = \mathbf{E}(\mathbf{E}^{\mathrm{T}}\mathbf{E} + \gamma^2\mathbf{I})^{-1}\mathbf{E}^{\mathrm{T}}. \tag{2.350}$$

All of the resolution matrices reduce properly to either $\mathbf{U}\mathbf{U}^{\mathrm{T}}$ or $\mathbf{V}\mathbf{V}^{\mathrm{T}}$ as $\gamma^2 \to 0$ when the SVD for $\mathbf{E}$ is substituted, and $K = M$ or N as necessary.

2.6 Combined least-squares and adjoints

2.6.1 Exact constraints

Consider now a modest generalization of the constrained problem Eq. (2.87) in which the unknowns $\mathbf{x}$ are also meant to satisfy some constraints exactly, or nearly so, for example,

$$\mathbf{A}\mathbf{x} = \mathbf{b}. \tag{2.351}$$

In some contexts, (2.351) is referred to as the "model," a term also employed, confusingly, for the physics defining $\mathbf{E}$ and/or the statistics assumed to describe $\mathbf{x}$, $\mathbf{n}$. We will temporarily refer to Eq. (2.351) as "perfect constraints," as opposed to those involving $\mathbf{E}$, which generally always have a non-zero noise element.

An example of a model in these terms occurs in acoustic tomography (Chapter 1), where measurements exist of both density and velocity fields, and they are connected by dynamical relations; the errors in the relations are believed to be so much smaller than those in the data, that for practical purposes, the constraints (2.351) might as well be treated as though they are perfect.[41] But otherwise, the distinction between constraints (2.351) and the observations is an arbitrary one, and the introduction of an error term in the former, no matter how small, removes any particular reason to distinguish them: $\mathbf{A}$ may well be some subset of the rows of $\mathbf{E}$. 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. 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, and which may dominate the results. Nonetheless, carrying one or more perfect constraints does produce some insight into how the system is behaving.

Several approaches are possible. Consider, for example, the objective function

$$J = (\mathbf{Ex} - \mathbf{y})^{\mathrm{T}}(\mathbf{Ex} - \mathbf{y}) + \gamma^2(\mathbf{Ax} - \mathbf{b})^{\mathrm{T}}(\mathbf{Ax} - \mathbf{b}), \tag{2.352}$$

where $\mathbf{W}, \mathbf{S}$ have been previously applied if necessary, and γ^2 is retained as a trade-off parameter. This objective function corresponds to the requirement of a solution of the combined equation sets,

$$\begin{Bmatrix} \mathbf{E} \\ \mathbf{A} \end{Bmatrix} \mathbf{x} + \begin{bmatrix} \mathbf{n} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{y} \\ \mathbf{b} \end{bmatrix}, \tag{2.353}$$

in which $\mathbf{u}$ is the model noise, and the weight given to the model is $\gamma^2\mathbf{I}$. For any finite γ^2, the perfect constraints are formally "soft" because they are being applied only as a minimized sum of squares. The solution follows immediately from (2.95) with

$$\mathbf{E} \longrightarrow \begin{Bmatrix} \mathbf{E} \\ \gamma\mathbf{A} \end{Bmatrix}, \qquad \mathbf{y} \longrightarrow \begin{Bmatrix} \mathbf{y} \\ \gamma\mathbf{b} \end{Bmatrix},$$

assuming the matrix inverse exists. As $\gamma^2 \to \infty$, the second set of equations is being imposed with arbitrarily great accuracy, and, barring numerical issues, becomes as close to exactly satisfied as one wants.

Alternatively, the model can be imposed as a hard constraint. All prior covariances and scalings having been applied, and Lagrange multipliers introduced, the problem is one with an objective function,

$$J = \mathbf{n}^{\mathrm{T}}\mathbf{n} - 2\boldsymbol{\mu}^{\mathrm{T}}(\mathbf{Ax} - \mathbf{b}) = (\mathbf{Ex} - \mathbf{y})^{\mathrm{T}}(\mathbf{Ex} - \mathbf{y}) - 2\boldsymbol{\mu}^{\mathrm{T}}(\mathbf{Ax} - \mathbf{b}), \tag{2.354}$$

which is a variant of (2.149). But now, Eq. (2.351) is to be exactly satisfied, and the observations only approximately so.

Setting the derivatives of J with respect to $\mathbf{x}$, $\boldsymbol{\mu}$ to zero, gives the following normal equations:

$$\mathbf{A}^{\mathrm{T}}\boldsymbol{\mu} = \mathbf{E}^{\mathrm{T}}(\mathbf{Ex} - \mathbf{y}), \tag{2.355}$$

$$\mathbf{Ax} = \mathbf{b}. \tag{2.356}$$

Equation (2.355) represents the adjoint, or "dual" model, for the adjoint or dual solution $\boldsymbol{\mu}$, and the two equation sets are to be solved simultaneously for $\mathbf{x}$, $\boldsymbol{\mu}$. They are again $M + N$ equations in $M + N$ unknowns (M of the μ_i, N of the x_i), but need not be full-rank. The first set, sometimes referred to as the "adjoint model," determines $\boldsymbol{\mu}$ from the *difference between* $\mathbf{Ex}$ and $\mathbf{y}$. The last set is just the exact constraints.

We can most easily solve two extreme cases in Eqs. (2.355) and (2.356) – 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,

$$\tilde{\mathbf{x}} = \mathbf{A}^{-1}\mathbf{b}, \tag{2.357}$$

and

$$\tilde{\boldsymbol{\mu}} = \mathbf{A}^{-\mathrm{T}}(\mathbf{E}^{\mathrm{T}}\mathbf{E}\mathbf{A}^{-1} - \mathbf{E}^{\mathrm{T}})\mathbf{b}. \tag{2.358}$$

Here, the values of $\tilde{\mathbf{x}}$ are completely determined by the full-rank, perfect constraints and the minimization of the deviation from the observations is passive. The Lagrange multipliers or adjoint solution, however, are useful in providing the sensitivity information, $\partial J/\partial\mathbf{b} = 2\boldsymbol{\mu}$, as already discussed. The uncertainty of this solution is zero because of the full-rank perfect model assumption (2.356).

In the second case, from (2.355),

$$\tilde{\mathbf{x}} = (\mathbf{E}^{\mathrm{T}}\mathbf{E})^{-1}[\mathbf{E}^{\mathrm{T}}\mathbf{y} + \mathbf{A}^{\mathrm{T}}\boldsymbol{\mu}] \equiv \tilde{\mathbf{x}}_u + (\mathbf{E}^{\mathrm{T}}\mathbf{E})^{-1}\mathbf{A}^{\mathrm{T}}\boldsymbol{\mu},$$

where $\tilde{\mathbf{x}}_u = (\mathbf{E}^{\mathrm{T}}\mathbf{E})^{-1}\mathbf{E}^{\mathrm{T}}\mathbf{y}$ is the ordinary, unconstrained least-squares solution. Substituting into (2.356) produces

$$\tilde{\boldsymbol{\mu}} = [\mathbf{A}(\mathbf{E}^{\mathrm{T}}\mathbf{E})^{-1}\mathbf{A}^{\mathrm{T}}]^{-1}(\mathbf{b} - \mathbf{A}\tilde{\mathbf{x}}_u), \tag{2.359}$$

and

$$\tilde{\mathbf{x}} = \tilde{\mathbf{x}}_u + (\mathbf{E}^{\mathrm{T}}\mathbf{E})^{-1}\mathbf{A}^{\mathrm{T}}[\mathbf{A}(\mathbf{E}^{\mathrm{T}}\mathbf{E})^{-1}\mathbf{A}^{\mathrm{T}}]^{-1}(\mathbf{b} - \mathbf{A}\tilde{\mathbf{x}}_u), \tag{2.360}$$

assuming $\mathbf{A}$ is full-rank underdetermined. The perfect constraints are underdetermined; their 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 as[42]

$$\begin{aligned}\mathbf{P} &= D^2(\tilde{\mathbf{x}} - \mathbf{x}) \\ &= \sigma^2\{(\mathbf{E}^{\mathrm{T}}\mathbf{E})^{-1} - (\mathbf{E}^{\mathrm{T}}\mathbf{E})^{-1}\mathbf{A}^{\mathrm{T}}[\mathbf{A}(\mathbf{E}^{\mathrm{T}}\mathbf{E})^{-1}\mathbf{A}^{\mathrm{T}}]^{-1}\mathbf{A}(\mathbf{E}^{\mathrm{T}}\mathbf{E})^{-1}\},\end{aligned} \tag{2.361}$$

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 of $\mathbf{A}^{-1}$ in these solutions is a manifestation of the warning about the possible introduction of components dependent upon small eigenvalues of $\mathbf{A}$. If neither $\mathbf{E}^{\mathrm{T}}\mathbf{E}$ nor $\mathbf{A}$ is of full-rank one can use, e.g., the SVD with the above solution; the combined $\mathbf{E}$, $\mathbf{A}$ may be rank deficient, or just-determined.

Example *Consider the least-squares problem of solving*

$$x_1 + n_1 = 1,$$
$$x_2 + n_2 = 1,$$
$$x_1 + x_2 + n_3 = 3,$$

with uniform, uncorrelated noise of variance 1 *in each of the equations. The least-squares solution is then*

$$\tilde{\mathbf{x}} = [1.3333 \quad 1.3333]^{\mathrm{T}},$$

with uncertainty

$$\mathbf{P} = \begin{Bmatrix} 0.6667 & -0.3333 \\ -0.333 & 0.6667 \end{Bmatrix}.$$

But suppose that it is known or desired that $x_1 - x_2 = 1$. *Then* (2.360) *produces* $\tilde{\mathbf{x}} = [1.8333 \quad 0.8333]^{\mathrm{T}}$, $\boldsymbol{\mu} = 0.5$, $J' = 0.8333$, *with 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* $\tilde{\mathbf{x}} = [1.8833 \quad 0.7833]^{\mathrm{T}}$ *and the new objective function is* $J' = 0.9383$, *consistent with the sensitivity deduced from* μ.

A more generally useful case occurs when the errors normally expected to be present in the supposedly exact constraints are explicitly acknowledged. If the exact constraints have errors either in the "forcing," **b**, or in a mis-specification of **A**, then we write

$$\mathbf{Ax} = \mathbf{b} + \boldsymbol{\Gamma}\mathbf{u}, \tag{2.362}$$

assuming that $\langle \mathbf{u} \rangle = 0, \ \langle \mathbf{u}\mathbf{u}^{\mathrm{T}} \rangle = \mathbf{Q}$. $\boldsymbol{\Gamma}$ is a known coefficient matrix included for generality. If, for example, the errors were thought to be the same in all equations, we could write $\boldsymbol{\Gamma} = [1, 1, \ldots 1]^{\mathrm{T}}$, and then **u** would be just a scalar. Let the dimension of **u** be $P \times 1$. Such representations are not unique and more will be said about them in Chapter 4. A hard constraint formulation can still be used, in which (2.362) is to be exactly satisfied, imposed through an objective function of form

$$J = (\mathbf{Ex} - \mathbf{y})^{\mathrm{T}}\mathbf{R}_{nn}^{-1}(\mathbf{Ex} - \mathbf{y}) + \mathbf{u}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{u} - 2\boldsymbol{\mu}^{\mathrm{T}}(\mathbf{Ax} - \mathbf{b} - \boldsymbol{\Gamma}\mathbf{u}). \tag{2.363}$$

Here, the noise error covariance matrix has been explicitly included. Finding the normal equations by setting the derivatives with respect to $(\mathbf{x}, \mathbf{u}, \boldsymbol{\mu})$ to zero produces

$$\mathbf{A}^{\mathrm{T}}\boldsymbol{\mu} = \mathbf{E}^{\mathrm{T}}\mathbf{R}_{nn}^{-1}(\mathbf{Ex} - \mathbf{y}), \tag{2.364}$$

$$\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\mu} = \mathbf{Q}^{-1}\mathbf{u}, \tag{2.365}$$

$$\mathbf{Ax} + \boldsymbol{\Gamma}\mathbf{u} = \mathbf{b}. \tag{2.366}$$

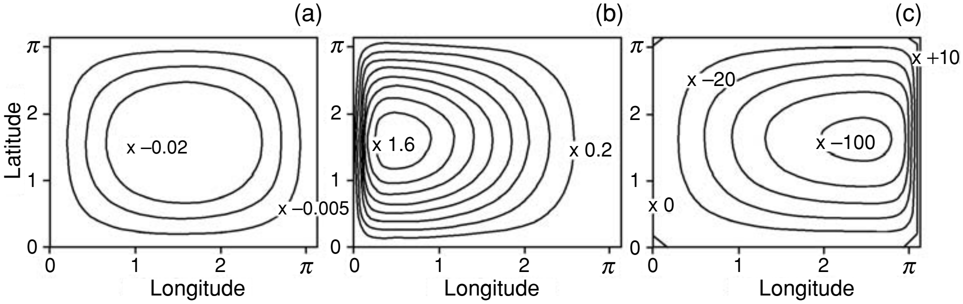

Figure 2.16 Numerical solution of the partial differential equation, Eq. (2.370). Panel (a) shows the imposed symmetric forcing $-\sin x \sin y$. (b) Displays the solution ϕ, and (c) shows the Lagrange multipliers, or adjoint solution, μ, whose structure is a near mirror image of ϕ. (Source: Schröter and Wunsch, 1986)

This system is $(2N + P)$ equations in $(2N + P)$ unknowns, where the first equation is again the adjoint system, and dependent upon $\mathbf{Ex} - \mathbf{y}$. Because $\mathbf{u}$ is a simple function of the Lagrange multipliers, the system is easily reduced to

$$\mathbf{A}^{\mathrm{T}}\boldsymbol{\mu} = \mathbf{E}^{\mathrm{T}}\mathbf{R}_{nn}^{-1}(\mathbf{Ex} - \mathbf{y}), \tag{2.367}$$

$$\mathbf{Ax} + \boldsymbol{\Gamma}\mathbf{Q}\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\mu} = \mathbf{b}, \tag{2.368}$$

which is now $2N \times 2N$, the $\mathbf{u}$ having dropped out. If all matrices are full-rank, the solution is immediate; otherwise the SVD can be used.

To use a soft constraint methodology, write

$$J = (\mathbf{Ex} - \mathbf{y})^{\mathrm{T}}\mathbf{R}_{nn}^{-1}(\mathbf{Ex} - \mathbf{y}) + (\mathbf{Ax} - \mathbf{b} - \boldsymbol{\Gamma}\mathbf{u})^{\mathrm{T}}\mathbf{Q}^{-1}(\mathbf{Ax} - \mathbf{b} - \boldsymbol{\Gamma}\mathbf{u})^{\mathrm{T}}, \tag{2.369}$$

and find the normal equations. It is again readily confirmed that the solutions using (2.352) or (2.363) are identical, and the hard/soft distinction is seen again to be artificial. The soft constraint method can deal with perfect constraints, by letting $\|\mathbf{Q}^{-1}\| \to 0$ but stopping when numerical instability sets in. The resulting numerical algorithms fall under the general subject of "penalty" and "barrier" methods.[43] Objective functions like (2.363) and (2.369) will be used extensively in Chapter 4.

Example *Consider the partial differential equation*

$$\epsilon\nabla^2\phi + \frac{\partial\phi}{\partial x} = -\sin x \sin y. \tag{2.370}$$

A code was written to solve it by finite differences for the case $\epsilon = 0.05$ *and* $\phi = 0$ *on the boundaries,* $0 \leq x \leq \pi$, $0 \leq y \leq \pi$, *as depicted in Fig. 2.16. The discretized form of the model is then the perfect* $N \times N$ *constraint system*

$$\mathbf{Ax} = \mathbf{b}, \quad \mathbf{x} = \{\phi_{ij}\}, \tag{2.371}$$

and $\mathbf{b}$ *is equivalently discretized* $-\sin x \sin y$. *The theory of partial differential equations shows that this system is full-rank and generally well-behaved. But let us pretend that this information is unknown to us, and seek the values of* $\mathbf{x}$ *that make the objective function,*

$$J = \mathbf{x}^{\mathrm{T}}\mathbf{x} - 2\boldsymbol{\mu}^{\mathrm{T}}(\mathbf{A}\mathbf{x} - \mathbf{b}), \tag{2.372}$$

stationary with respect to $\mathbf{x}$, $\boldsymbol{\mu}$, *that is the Eqs. (2.355) and (2.356) with* $\mathbf{E} = \mathbf{I}$, $\mathbf{y} = \mathbf{0}$. *Physically,* $\mathbf{x}^{\mathrm{T}}\mathbf{x}$ *is identified with the solution potential energy. The solution* $\boldsymbol{\mu}$, *corresponding to the solution of Fig. 2.16(b) is shown in Fig. 2.16(c). What is the interpretation? The Lagrange multipliers represent the sensitivity of the solution potential energy to perturbations in the forcing field. The sensitivity is greatest in the right-half of the domain, and indeed displays a boundary layer character. A physical interpretation of the Lagrange multipliers can be inferred, given the simple structure of the governing equation* (2.370), *and the Dirichlet boundary conditions. This equation is not self-adjoint; the adjoint partial differential equation is of form*

$$\epsilon \nabla^2 \nu - \frac{\partial \nu}{\partial x} = d, \tag{2.373}$$

where d is a forcing term, subject to mixed boundary conditions, and whose discrete form is obtained by taking the transpose of the $\mathbf{A}$ *matrix of the discretization (see Appendix 2 to this chapter). The forward solution exhibits a boundary layer on the left-hand wall, while the adjoint solution has a corresponding behavior in the dual space on the right-hand wall. The structure of the* $\boldsymbol{\mu}$ *would evidently change if J were changed.*[44]

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 virtual forces (forces required to enforce the constraints). 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.[45] In an economics/management context, the multipliers are usually called "shadow prices" as they are intimately related to the question of how much profit will change with a shift in the availability or cost of a product ingredient. The terminology "cost function" is a sensible substitute for what we call the "objective function."

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. An elegant Hamiltonian formulation of the material is possible.

2.6.2 Relation to Green functions

Consider any linear set of simultaneous equations, involving an arbitrary matrix, **A**,

$$\mathbf{A}\mathbf{x} = \mathbf{b}. \tag{2.374}$$

Write the adjoint equations for an arbitrary right-hand side,

$$\mathbf{A}^{\mathrm{T}}\mathbf{z} = \mathbf{r}. \tag{2.375}$$

Then the scalar relation

$$\mathbf{z}^{\mathrm{T}}\mathbf{A}\mathbf{x} - \mathbf{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{z} = 0 \tag{2.376}$$

(the "bilinear identity") implies that

$$\mathbf{z}^{\mathrm{T}}\mathbf{b} = \mathbf{x}^{\mathrm{T}}\mathbf{r}. \tag{2.377}$$

In the special case, $\mathbf{r} = \mathbf{0}$, we have

$$\mathbf{z}^{\mathrm{T}}\mathbf{b} = 0, \tag{2.378}$$

that is, **b**, the right-hand side of the original equations (2.374), must be orthogonal to any solution of the homogeneous adjoint equations. (In SVD terms, this result is the solvability condition Eq. (2.266).) If **A** is of full rank, then there is no non-zero solution to the homogeneous adjoint equations.

Now assume that **A** is $N \times N$ of full rank. Add a single equation to (2.374) of the form

$$x_p = \alpha_p, \tag{2.379}$$

or

$$\mathbf{e}_p^{\mathrm{T}}\mathbf{x} = \alpha_p, \tag{2.380}$$

where $\mathbf{e}_p = \delta_{ip}$ and α_p is unknown. We also demand that Eq. (2.374) should remain exactly satisfied. The combined system of (2.374) and (2.380), written as

$$\mathbf{A}_1\mathbf{x} = \mathbf{b}_1, \tag{2.381}$$

is overdetermined. If it is to have a solution without any residual, it must still be orthogonal to any solution of the homogeneous adjoint equations,

$$\mathbf{A}_1^{\mathrm{T}}\mathbf{z} = \mathbf{0}. \tag{2.382}$$

There is only one such solution (because there is only one vector, $\mathbf{z} = \mathbf{u}_{N+1}$, in the null space of $\mathbf{A}_1^{\mathrm{T}}$). Write $\mathbf{u}_{N+1} = [\mathbf{g}_p, \gamma]^{\mathrm{T}}$, separating out the first N elements of

$\mathbf{u}_{N+1}$, calling them $\mathbf{g}_p$, and calling the one remaining element, γ. Thus Eq. (2.377) is

$$\mathbf{u}_{N+1}^{\mathrm{T}}\mathbf{b}_1 = \mathbf{g}_p^{\mathrm{T}}\mathbf{b} + \gamma\alpha_p = 0. \tag{2.383}$$

Choose $\gamma = -1$ (any other choice can be absorbed into $\mathbf{g}_p$). Then

$$\alpha_p = \mathbf{g}_p^{\mathrm{T}}\mathbf{b}. \tag{2.384}$$

If $\mathbf{g}_p$ were known, then α_p in (2.384) would be the only value consistent with the solutions to (2.374), and would be the correct value of $x_{p.}$ But (2.382) is the same as

$$\mathbf{A}^{\mathrm{T}}\mathbf{g}_p = \mathbf{e}_p. \tag{2.385}$$

Because *all* elements x_p are needed, Eq. (2.385) is solved for all $p = 1, 2, \ldots, N$, that is,

$$\mathbf{A}^{\mathrm{T}}\mathbf{G} = \mathbf{I}_N, \tag{2.386}$$

which is N separate problems, each for the corresponding column of $\mathbf{G} = \{\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_N\}$. Here, $\mathbf{G}$ is the Green function. With $\mathbf{G}$ known, we have immediately that

$$\mathbf{x} = \mathbf{G}^{\mathrm{T}}\mathbf{b}, \tag{2.387}$$

(from Eq. (2.384)). The Green function is an inverse to the adjoint equations; its significance here is that it generalizes in the continuous case to an operator inverse.[46]

2.7 Minimum variance estimation and simultaneous equations

The fundamental objective for least-squares is minimization of the noise norm (2.89), 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 (2.95) or indirectly through the vector representations of the SVD, are fundamentally deterministic. 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, very different, approach to obtaining estimates of the solution to equation sets like (2.87), directed more clearly toward the physical goal: to find an estimate $\tilde{\mathbf{x}}$ that deviates as little as possible in the *mean-square* from the true solution. That is, we wish to minimize the statistical quantities $\langle(\tilde{x}_i - x_i)^2\rangle$ for all i. The next section is devoted to finding 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 seriously confuse the methodologies, and therefore the interpretation of the result.

2.7.1 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 is either constant or varying with space and time. 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, specifying its mean and second moments,

$$\langle \mathbf{x} \rangle = \mathbf{0}, \qquad \langle \mathbf{x}(\mathbf{r}_i)\mathbf{x}(\mathbf{r}_j)^{\mathrm{T}} \rangle = \mathbf{R}_{xx}(\mathbf{r}_i,\ \mathbf{r}_j). \tag{2.388}$$

To make this problem concrete, one might think of $\mathbf{x}$ as being the temperature anomaly (about the mean) at a fixed depth in a fluid (a scalar) and $\mathbf{r}_j$ a vector of horizontal positions; or conductivity in a well, where $\mathbf{r}_j$ would be the depth coordinate, and $\mathbf{x}$ is the vector of scalars at any location, $\mathbf{r}_p$, $x_p = x(\mathbf{r}_p)$. Alternatively, $\mathbf{x}$ might be the temperature at a fixed point, with r_j being the scalar of time. 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, all elements of $\mathbf{x}$ are written as scalars.

Now suppose that we have some observations, y_i, as a function of the same coordinate $\mathbf{r}_i$, with a known, zero mean, and second moments

$$\mathbf{R}_{yy}(\mathbf{r}_i, \mathbf{r}_j) = \langle \mathbf{y}(\mathbf{r}_i)\mathbf{y}(\mathbf{r}_j)^{\mathrm{T}} \rangle, \quad \mathbf{R}_{xy}(\mathbf{r}_i,\ \mathbf{r}_j) = \langle \mathbf{x}(\mathbf{r}_i)\mathbf{y}(\mathbf{r}_j)^{\mathrm{T}} \rangle, \quad i, j = 1, 2, \ldots, M. \tag{2.389}$$

(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 scalar observations are assumed.) Could the measurements be used to make an estimate of $\mathbf{x}$ at a point $\tilde{\mathbf{r}}_\alpha$ where no measurement is available? Or could many measurements be used to obtain a better estimate even at points where there exists a measurement? 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 temperature estimates are sought at different locations, perhaps on a regular grid $\tilde{\mathbf{r}}_\alpha$,

$\alpha = 1, 2, \ldots, N$. 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 assumed known). Alternatively, and somewhat more interesting, perhaps the measurements are more indirect, with $y(r_i)$ representing a velocity field component at depth in a fluid and believed connected, through a differential equation, to the temperature field. We might want to estimate the temperature from measurements of the velocity.

Given the previous statistical discussion (p. 31), it is reasonable to ask for 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,

$$P\left(\tilde{\mathbf{r}}_\alpha, \tilde{\mathbf{r}}_\alpha\right) = \langle(\tilde{x}(\tilde{\mathbf{r}}_\alpha) - x(\tilde{\mathbf{r}}_\alpha))(\tilde{x}(\tilde{\mathbf{r}}_\beta) - x(\tilde{\mathbf{r}}_\beta))\rangle|_{\tilde{\mathbf{r}}_\alpha = \tilde{\mathbf{r}}_\beta}$$

is to be minimized. If an estimate is needed at more than one point, $\tilde{\mathbf{r}}_\alpha$, the covariance of the errors in the different estimates would usually be required, too. Form a vector of values to be estimated, $\{\tilde{x}(\mathbf{r}_\alpha)\} \equiv \tilde{\mathbf{x}}$, and their uncertainty is

$$\begin{aligned}\mathbf{P}(\tilde{\mathbf{r}}_\alpha, \ \tilde{\mathbf{r}}_\beta) &= \langle(\tilde{x}(\tilde{\mathbf{r}}_\alpha) - x(\tilde{\mathbf{r}}_\alpha))(\tilde{x}(\tilde{\mathbf{r}}_\beta) - x(\tilde{\mathbf{r}}_\beta))\rangle \\ &= \langle(\tilde{\mathbf{x}} - \mathbf{x})(\tilde{\mathbf{x}} - \mathbf{x})^{\mathrm{T}}\rangle, \quad \alpha, \beta = 1, 2, \ldots, N,\end{aligned} \tag{2.390}$$

where the *diagonal* elements, $\mathbf{P}(\tilde{\mathbf{r}}_\alpha, \ \tilde{\mathbf{r}}_\alpha)$, are to be *individually* minimized (not in the sum of squares). Thus a solution with *minimum variance about the correct value* is sought.

What should the relationship be between data and estimate? At least initially, a linear combination of data is a reasonable starting point,

$$\tilde{x}(\tilde{\mathbf{r}}_\alpha) = \sum_{j=1}^{M} B(\tilde{\mathbf{r}}_\alpha, \ \mathbf{r}_j) y(\mathbf{r}_j), \tag{2.391}$$

for all α, which makes the diagonal elements of $\mathbf{P}$ in (2.390) as small as possible. By letting $\mathbf{B}$ be an $N \times M$ matrix, all of the points can be handled at once:

$$\tilde{\mathbf{x}}\left(\tilde{\mathbf{r}}_\alpha\right) = \mathbf{B}(\tilde{\mathbf{r}}_\alpha, \ \mathbf{r}_j)\mathbf{y}(\mathbf{r}_j). \tag{2.392}$$

(This notation is redundant. Equation (2.392) is a shorthand for (2.391), 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, $\tilde{\mathbf{r}}_\alpha$, but it is automatically accounted for by the usual matrix multiplication convention. It suffices to write $\tilde{\mathbf{x}} = \mathbf{B}\mathbf{y}$.)

An important result, often called the "Gauss–Markov theorem," produces the values of $\mathbf{B}$ that will minimize the diagonal elements of $\mathbf{P}$.[47] Substituting (2.392) into (2.390) and expanding,

$$\begin{aligned}\mathbf{P}(\tilde{\mathbf{r}}_\alpha,\ \tilde{\mathbf{r}}_\beta) &= \langle(\mathbf{B}(\tilde{\mathbf{r}}_\alpha,\ \mathbf{r}_j)\mathbf{y}(\mathbf{r}_j) - \mathbf{x}(\tilde{\mathbf{r}}_\alpha))(\mathbf{B}(\tilde{\mathbf{r}}_\beta,\ \mathbf{r}_l)\mathbf{y}\,(\mathbf{r}_l) - \mathbf{x}(\tilde{\mathbf{r}}_\beta))^{\mathrm{T}}\rangle \\ &\equiv \langle(\mathbf{By} - \mathbf{x})(\mathbf{By} - \mathbf{x})^{\mathrm{T}}\rangle \\ &= \mathbf{B}\langle\mathbf{yy}^{\mathrm{T}}\rangle - \langle\mathbf{xy}^{\mathrm{T}}\rangle\mathbf{B}^{\mathrm{T}} - \mathbf{B}\langle\mathbf{yx}^{\mathrm{T}}\rangle + \langle\mathbf{xx}^{\mathrm{T}}\rangle.\end{aligned} \tag{2.393}$$

Using $\mathbf{R}_{xy} = \mathbf{R}_{yx}^{\mathrm{T}}$, Eq. (2.393) is

$$\mathbf{P} = \mathbf{BR}_{yy}\mathbf{B}^{\mathrm{T}} - \mathbf{R}_{xy}\mathbf{B}^{\mathrm{T}} - \mathbf{BR}_{xy}^{\mathrm{T}} + \mathbf{R}_{xx}. \tag{2.394}$$

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 completing the square (Eq. (2.37)) (by adding and subtracting $\mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{R}_{xy}^{\mathrm{T}}$), (2.394) becomes

$$\mathbf{P} = (\mathbf{B} - \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1})\mathbf{R}_{yy}(\mathbf{B} - \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1})^{\mathrm{T}} - \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{R}_{xy}^{\mathrm{T}} + \mathbf{R}_{xx}. \tag{2.395}$$

Setting $\tilde{\mathbf{r}}_\alpha = \tilde{\mathbf{r}}_\beta$ so that (2.395) is the variance of the estimate at point $\tilde{\mathbf{r}}_\alpha$ about its true value, and noting that all three terms in Eq. (2.395) are positive definite, minimization of any diagonal element of $\mathbf{P}$ is obtained by choosing $\mathbf{B}$ so that the first term vanishes, or

$$\mathbf{B}(\tilde{\mathbf{r}}_\alpha,\ \mathbf{r}_j) = \mathbf{R}_{xy}\,(\tilde{\mathbf{r}}_\alpha, \mathbf{r}_i)\,\mathbf{R}_{yy}(\mathbf{r}_i, \mathbf{r}_j)^{-1} = \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}. \tag{2.396}$$

(The diagonal elements of $(\mathbf{B} - \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1})\mathbf{R}_{yy}(\mathbf{B} - \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1})^{\mathrm{T}}$ need to be written out explicitly to see that Eq. (2.396) is necessary. Consider the 2 × 2 case: the first term of Eq. (2.395) is of the form

$$\begin{Bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{Bmatrix} \begin{Bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{Bmatrix} \begin{Bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{Bmatrix}^{\mathrm{T}},$$

where $\mathbf{C} = \mathbf{B} - \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}$. Then, one has the diagonal of

$$\begin{Bmatrix} C_{11}^2 R_{11} + C_{12}C_{11}\,(R_{21} + R_{12}) + C_{12}^2 R_{22} & \cdot \\ \ddots & C_{21}^2 R_{11} + C_{21}C_{22}\,(R_{21} + R_{12}) + C_{22}^2 R_{22} \end{Bmatrix},$$

and these diagonals vanish (with $R_{11}, R_{22} > 0$, only if $C_{11} = C_{12} = C_{21} = C_{22} = 0$). Thus the minimum variance estimate is

$$\tilde{\mathbf{x}}(\tilde{\mathbf{r}}_\alpha) = \mathbf{R}_{xy}\,(\tilde{\mathbf{r}}_\alpha, \mathbf{r}_i)\,\mathbf{R}_{yy}^{-1}(\mathbf{r}_i, \mathbf{r}_j)\mathbf{y}(\mathbf{r}_j), \tag{2.397}$$

and the minimum of the diagonal elements of $\mathbf{P}$ is found by substituting back into (2.394), producing

$$\mathbf{P}(\tilde{\mathbf{r}}_\alpha,\ \tilde{\mathbf{r}}_\beta) = \mathbf{R}_{xx}(\tilde{\mathbf{r}}_\alpha,\ \tilde{\mathbf{r}}_\beta) - \mathbf{R}_{xy}(\tilde{\mathbf{r}}_\alpha,\ \mathbf{r}_j)\mathbf{R}_{yy}^{-1}(\mathbf{r}_j,\ \mathbf{r}_k)\mathbf{R}_{xy}^{\mathrm{T}}(\tilde{\mathbf{r}}_\beta,\ \mathbf{r}_k). \tag{2.398}$$

The bias of (2.398) is

$$\langle \tilde{\mathbf{x}} - \mathbf{x} \rangle = \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\langle \mathbf{y} \rangle - \mathbf{x}. \tag{2.399}$$

If $\langle \mathbf{y} \rangle = \mathbf{x} = 0$, the estimator is unbiassed, and called a "best linear unbiassed estimator," or "BLUE"; otherwise it is biassed. The whole development here began with the assumption that $\langle \mathbf{x} \rangle = \langle \mathbf{y} \rangle = 0$; what is usually done is to remove the *sample* mean from the observations $\mathbf{y}$, and to ignore the difference between the true and sample means. An example of using this machinery for mapping purposes will be seen in Chapter 3. Under some circumstances, this approximation is unacceptable, and the mapping error introduced by the use of the sample mean must be found. A general approach falls under the label of "kriging," which is also briefly discussed in Chapter 3.

2.7.2 Linear algebraic equations

The result shown in (2.396)–(2.398) is the abstract general case and is deceptively simple. Invocation of the physical problem of interpolating temperatures, etc., is not necessary: the only information actually used is that there are finite covariances between $\mathbf{x}$, $\mathbf{y}$, $\mathbf{n}$. Although mapping will be explicitly explored in Chapter 3, suppose instead that 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 covariance, $\mathbf{R}_{yy}$, can then be computed directly as

$$\mathbf{R}_{yy} = \langle (\mathbf{Ex} + \mathbf{n})(\mathbf{Ex} + \mathbf{n})^{\mathrm{T}} \rangle = \mathbf{E}\mathbf{R}_{xx}\mathbf{E}^{\mathrm{T}} + \mathbf{R}_{nn}. \tag{2.400}$$

The unnecessary, but simplifying and often excellent, assumption was made that the cross-terms of form

$$\mathbf{R}_{xn} = \mathbf{R}_{nx}^{\mathrm{T}} = \mathbf{0}, \tag{2.401}$$

so that

$$\mathbf{R}_{xy} = \langle \mathbf{x}(\mathbf{Ex} + \mathbf{n})^{\mathrm{T}} \rangle = \mathbf{R}_{xx}\mathbf{E}^{\mathrm{T}}, \tag{2.402}$$

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, Eqs. (2.397) and (2.398) take on the following form:

$$\boxed{\begin{aligned} \tilde{\mathbf{x}} &= \mathbf{R}_{xx}\mathbf{E}^{\mathrm{T}}(\mathbf{E}\mathbf{R}_{xx}\mathbf{E}^{\mathrm{T}} + \mathbf{R}_{nn})^{-1}\mathbf{y}, && (2.403)\\ \tilde{\mathbf{n}} &= \mathbf{y} - \mathbf{E}\tilde{\mathbf{x}}, && (2.404)\\ \mathbf{P} &= \mathbf{R}_{xx} - \mathbf{R}_{xx}\mathbf{E}^{\mathrm{T}}(\mathbf{E}\mathbf{R}_{xx}\mathbf{E}^{\mathrm{T}} + \mathbf{R}_{nn})^{-1}\mathbf{E}\mathbf{R}_{xx}. && (2.405) \end{aligned}}$$

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 both the noise and the solution. Notice that they are *identical to the least-squares expression* (2.135) *if* $\mathbf{S} = \mathbf{R}_{xx}$, $\mathbf{W} = \mathbf{R}_{nn}$, except that there the uncertainty was estimated about the mean solution; here it is taken about the true one. As is generally true of all linear methods, the uncertainty, $\mathbf{P}$, is independent of the actual data, and can be computed in advance should one wish.

From the matrix inversion lemma, Eqs. (2.403)–(2.405) can be rewritten as

$$\tilde{\mathbf{x}} = \left(\mathbf{R}_{xx}^{-1} + \mathbf{E}^{\mathrm{T}}\mathbf{R}_{nn}^{-1}\mathbf{E}\right)^{-1}\mathbf{E}^{\mathrm{T}}\mathbf{R}_{nn}^{-1}\mathbf{y}, \tag{2.406}$$

$$\tilde{\mathbf{n}} = \mathbf{y} - \mathbf{E}\tilde{\mathbf{x}}, \tag{2.407}$$

$$\mathbf{P} = \left(\mathbf{R}_{xx}^{-1} + \mathbf{E}^{\mathrm{T}}\mathbf{R}_{nn}^{-1}\mathbf{E}\right)^{-1}. \tag{2.408}$$

Although these alternate forms are algebraically and numerically identical to Eqs. (2.403)–(2.405), 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 more preferable to the other. Finally, (2.408) has an important interpretation that we will discuss 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. Note that in the limit of complete a priori ignorance of the solution, $\|\mathbf{R}_{xx}^{-1}\| \to 0$, Eqs. (2.406) and (2.408) reduce to

$$\tilde{\mathbf{x}} = \left(\mathbf{E}^{\mathrm{T}}\mathbf{R}_{nn}^{-1}\mathbf{E}\right)^{-1}\mathbf{E}^{\mathrm{T}}\mathbf{R}_{nn}^{-1}\mathbf{y},$$
$$\mathbf{P} = (\mathbf{E}^{\mathrm{T}}\mathbf{R}_{nn}^{-1}\mathbf{E})^{-1},$$

the conventional weighted least-squares solution, now with $\mathbf{P} = \mathbf{C}_{xx.}$ More generally, the presence of finite $\mathbf{R}_{xx}^{-1}$ introduces a bias into the solution so that $\langle\tilde{\mathbf{x}}\rangle \neq \mathbf{x}$, which, however, produces a smaller solution variance than in the unbiassed solution.

The solution shown in Eqs. (2.403)–(2.405) and (2.406)–(2.408) is an "estimator"; it was found by demanding a solution with the minimum dispersion about the true solution and it is found, surprisingly, identical to the tapered, weighted least-squares solution when $\mathbf{S} = \mathbf{R}_{xx},\ \mathbf{W} = \mathbf{R}_{nn}$, the least-squares objective function weights are chosen. 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 which were anything we wished – as long as appropriate inverses existed.

The correspondence of least-squares with what is usually known as 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 by minimizing a summation which is a sample *estimate* of the Gauss–Markov objective function when **S**, **W** are chosen properly.

2.7.3 Testing after the fact

As with any statistical estimator, an essential step after an apparent solution has been found is the testing 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 a-posteriori checks are both necessary 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 covariance 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. It is also true that the solutions tend to be somewhat insensitive to the details of the prior covariances 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 covariances, true ignorance meaning that arbitrarily large or small numerical values of x_i, n_i would be acceptable. In the box inversions of Chapter 1 (to be revisited in Chapter 5), solution velocities of order 1000 cm/s might be regarded as 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(10cm/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 already listed 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, **E**, and the assumed statistics of solution and/or noise.

Example *The underdetermined system*

$$\begin{Bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 1 \end{Bmatrix} \mathbf{x} + \mathbf{n} = \begin{bmatrix} 1 \\ -1 \end{bmatrix},$$

with noise variance $\langle \mathbf{n}\mathbf{n}^{\mathrm{T}} \rangle = 0.01\mathbf{I}$, *has a solution, if* $\mathbf{R}_{xx} = \mathbf{I}$, *of*

$$\tilde{\mathbf{x}} = \mathbf{E}^{\mathrm{T}}(\mathbf{E}\mathbf{E}^{\mathrm{T}} + 0.01\mathbf{I})^{-1}\mathbf{y} = [0 \quad 0.4988 \quad 0.4988 \quad 0]^{\mathrm{T}},$$

$$\tilde{\mathbf{n}} = [0.0025, -0.0025]^{\mathrm{T}}.$$

If the solution was thought to be large scale and smooth, one might use the covariance

$$\mathbf{R}_{xx} = \begin{Bmatrix} 1 & 0.999 & 0.998 & 0.997 \\ 0.999 & 1 & 0.999 & 0.998 \\ 0.998 & 0.999 & 1 & 0.999 \\ 0.997 & 0.998 & 0.999 & 1 \end{Bmatrix},$$

which produces a solution

$$\tilde{\mathbf{x}} = [0.2402 \pm 0.028 \quad 0.2595 \pm 0.0264 \quad 0.2595 \pm 0.0264 \quad 0.2402 \pm 0.0283]^{\mathrm{T}},$$
$$\tilde{\mathbf{n}} = [0.0006 \quad -0.9615]^{\mathrm{T}},$$

having the desired large-scale property. (One might worry a bit about the structure of the residuals, but two equations are inadequate to draw any conclusions.)

2.7.4 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, Chebyschev polynomials, ordinary polynomials, etc. One might write

$$x(r_i) = \sum_{j=1}^{L} \alpha_j F_j(r_i),$$

or

$$\mathbf{x} = \mathbf{F}\boldsymbol{\alpha}, \quad \mathbf{F} = \left\{\begin{matrix} F_1(r_1) & F_2(r_1) & \cdots & F_L(r_1) \\ F_1(r_2) & F_2(r_2) & \cdots & F_L(r_2) \\ \cdot & \cdot & \cdot & \cdot \\ F_1(r_N) & F_2(r_N) & \cdots & F_L(r_N) \end{matrix}\right\}, \quad \boldsymbol{\alpha} = [\alpha_1 \cdots \alpha_L]^{\mathrm{T}},$$

which, when substituted into (2.87), produces

$$\mathbf{T}\boldsymbol{\alpha} + \mathbf{n} = \mathbf{y}, \quad \mathbf{T} = \mathbf{EF}. \tag{2.409}$$

If $L < M < N$, one can convert an underdetermined system into one which is formally overdetermined and, of course, the reverse is possible as well. It should be apparent, however, that the solution to (2.409) will have a covariance structure dictated in large part by that contained in the basis functions chosen, and 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

$$\mathbf{P} = \mathbf{F}\mathbf{P}_{\alpha\alpha}\mathbf{F}^{\mathrm{T}}, \tag{2.410}$$

is the uncertainty of $\tilde{\mathbf{x}}$. 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.

Example *If, in the last example, one attempts a solution as a first-order polynomial,*

$$x_i = a + br_i, \quad r_1 = 0,\ r_2 = 1,\ r_3 = 2, \ldots,$$

the system will become two equations in the two unknowns a, b:

$$\mathbf{EF}\begin{bmatrix} a \\ b \end{bmatrix} = \begin{Bmatrix} 4 & 6 \\ 0 & 0 \end{Bmatrix}\begin{bmatrix} a \\ b \end{bmatrix} + \mathbf{n} = \begin{bmatrix} 1 \\ -1 \end{bmatrix},$$

and if no prior information about the covariance of a, b is provided,

$$[\tilde{a},\ \tilde{b}] = [0.0769,\ 0.1154],$$
$$\tilde{\mathbf{x}} = \begin{bmatrix} 0.0769 \pm 0.0077 & 0.1923 \pm 0.0192 & 0.3076 \pm 0.0308 & 0.4230 \pm 0.0423 \end{bmatrix}^{\mathrm{T}},$$
$$\tilde{\mathbf{n}} = [0.0002,\ -1.00]^{\mathrm{T}},$$

which is also large scale and smooth, but different than that obtained above. 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.

2.7.5 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, θ_i. Then

$$m + \theta_i = y_i, \quad i = 1, 2, \ldots, M, \tag{2.411}$$

or

$$\mathbf{D}m + \boldsymbol{\theta} = \mathbf{y}, \quad \mathbf{D} = [1 \quad 1 \quad 1 \quad \cdots \quad 1]^{\mathrm{T}}, \tag{2.412}$$

and we seek a best estimate, $\tilde{m}$, of m. In (2.411) or (2.412) 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 $\mathbf{R}_{nn} = \langle \boldsymbol{\theta}\boldsymbol{\theta}^{\mathrm{T}} \rangle$. If we were estimating a large-scale mean temperature in a fluid flow filled with smaller-scale 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 $\mathbf{R}_{xx} = \langle m^2 \rangle =$

m_0^2, and that, from (2.406),

$$\begin{aligned}\tilde{m} &= \left\{\frac{1}{m_0^2} + \mathbf{D}^{\mathrm{T}}\mathbf{R}_{nn}^{-1}\mathbf{D}\right\}^{-1}\mathbf{D}^{\mathrm{T}}\mathbf{R}_{nn}^{-1}\mathbf{y} \\ &= \frac{1}{1/m_0^2 + \mathbf{D}^{\mathrm{T}}\mathbf{R}_{nn}^{-1}\mathbf{D}}\,\mathbf{D}^{\mathrm{T}}\mathbf{R}_{nn}^{-1}\mathbf{y},\end{aligned} \tag{2.413}$$

where $\mathbf{D}^{\mathrm{T}}\mathbf{R}_{nn}^{-1}\mathbf{D}$ is a scalar.[48] The expected uncertainty of this estimate is (2.408),

$$P = \left\{\frac{1}{m_0^2} + \mathbf{D}^{\mathrm{T}}\mathbf{R}_{nn}^{-1}\mathbf{D}\right\}^{-1} = \frac{1}{1/m_0^2 + \mathbf{D}^{\mathrm{T}}\mathbf{R}_{nn}^{-1}\mathbf{D}}, \tag{2.414}$$

(also a scalar).

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 (2.413) is

$$\tilde{m} = \frac{1}{\left(1/m_0^2 + M/\sigma^2\right)\sigma^2}\sum_{i=1}^{M} y_i = \frac{m_0^2}{\sigma^2 + M m_0^2}\sum_{i=1}^{M} y_i, \tag{2.415}$$

where the relations $\mathbf{D}^{\mathrm{T}}\mathbf{D} = M$, $\mathbf{D}^{\mathrm{T}}\mathbf{y} = \sum_{i=1}^{M} y_i$ were used. The expected value of the estimate is

$$\langle\tilde{m}\rangle = \frac{m_0^2}{\sigma^2 + M m_0^2}\sum_{i}^{M}\langle y_i\rangle = \frac{m_0^2}{\sigma^2 + M m_0^2} M m \neq m, \tag{2.416}$$

that is, it is biassed, as inferred above, unless $\langle y_i\rangle = 0$, implying $m = 0$. $\mathbf{P}$ becomes

$$P = \frac{1}{1/m_0^2 + M/\sigma^2} = \frac{\sigma^2 m_0^2}{\sigma^2 + M m_0^2}. \tag{2.417}$$

Under the further assumption that $m_0^2 \to \infty$,

$$\tilde{m} = \frac{1}{M}\sum_{i=1}^{M} y_i, \tag{2.418}$$

$$P = \sigma^2/M, \tag{2.419}$$

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; recall Eq. (2.42)); $\langle\tilde{m}\rangle$ in (2.418) is readily seen to be the true mean – this estimate has become unbiassed. However, the magnitude of (2.419) always exceeds that of (2.417) – acceptance of bias in the estimate (2.415) reduces the uncertainty of the result – a common trade-off.

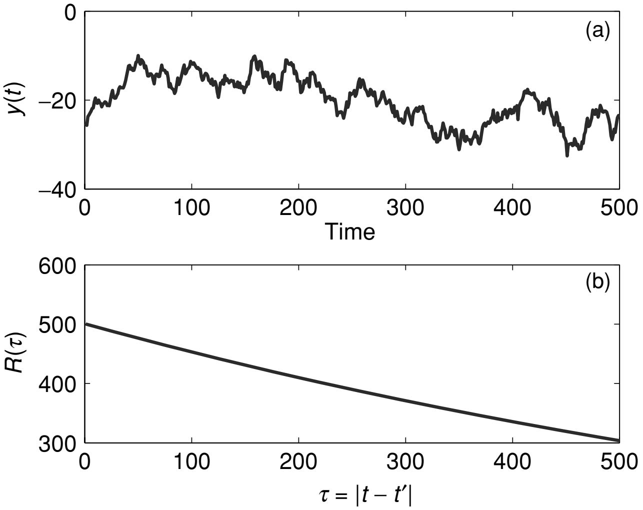

Figure 2.17 (a) Time series $y(t)$ whose mean is required. (b) The autocovariance $\langle y(t)y(t')\rangle$ as a function of $|t - t'|$ (in this special case, it does not depend upon t, t' separately.) The true mean of $y(t)$ is zero by construction.

Equations (2.413) and (2.414) are the more general estimation rule – accounting through $\mathbf{R}_{nn}$ for correlations in the observations and their irregular distribution. Because samples may not be independent, (2.419) may be extremely optimistic. Equation (2.414) gives one 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).

Example *The mean is needed for the $M = 500$ values of the measured time series, shown in Fig. 2.17. If one calculates the ordinary average, $\tilde{m} = -20.0$, the standard error, treating the measurements as uncorrelated, by Eq. (2.419) is ± 0.31. If, on the other hand, one uses the covariance function displayed in Fig. 2.17, and Eqs. (2.413) and (2.414) with $m_0^2 \to \infty$, one obtains $\tilde{m} = -23.7$, with a standard error of ± 20. The true mean of the time series is actually zero (it was generated that way), and one sees the dire effects of assuming uncorrelated measurement noise, when the correlation is actually very strong. Within two standard deviations (a so-called 95% confidence interval for the mean, if n is Gaussian), one finds, correctly, that the sample mean is indistinguishable from zero, whereas the mean assuming uncorrelated noise would appear to be very well determined and markedly different from zero.[49] (One might be tempted to apply a transformation to render the observations uncorrelated before averaging, and so treat the result as having M degrees-of-freedom. But recall, e.g., that for Gaussian variables (p. 39),* *the resulting numbers will have different variances, and one would be averaging apples and oranges.)*

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 (Eq. (2.418) is finite). It is is merely a statement that there is no information whatsoever, before we start, as to the magnitude of the true average – it could be arbitrarily large (or small and of either sign) and if it came out that way, it would be acceptable. 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). 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 very important not to be tempted into making a first estimate of m_0^2 by using (2.418), substituting into (2.415), 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.

2.8 Improving recursively

2.8.1 Least-squares

A common situation arises that one has a solution $\tilde{\mathbf{x}}$, $\tilde{\mathbf{n}}$, $\mathbf{P}$, and more information becomes available, often in the form of further noisy linear constraints. One way of using the new information is to combine the old and new equations into one larger system, and re-solve. This approach may well be the best one. Sometimes, however, perhaps because the earlier equations have been discarded, or for reasons of storage or both, one prefers to retain the information from the previous solution without having to re-solve the entire system. So-called recursive methods, in both least-squares and minimum variance estimation, provide the appropriate recipes.

Let the original equations be re-labeled so that we can distinguish them from those that come later, in the form

$$\mathbf{E}(1)\mathbf{x} + \mathbf{n}(1) = \mathbf{y}(1), \tag{2.420}$$

where the noise $\mathbf{n}(1)$ has zero mean and covariance matrix $\mathbf{R}_{nn}(1)$. Let the estimate of the solution to (2.420) from one of the estimators be written as $\tilde{\mathbf{x}}(1)$, with uncertainty $\mathbf{P}(1)$. To be specific, suppose (2.420) is full-rank overdetermined, and was solved using row-weighted least-squares, as

$$\tilde{\mathbf{x}}(1) = [\mathbf{E}(1)^{\mathrm{T}}\mathbf{R}_{nn}(1)^{-1}\mathbf{E}(1)]^{-1}\mathbf{E}(1)^{\mathrm{T}}\mathbf{R}_{nn}(1)^{-1}\mathbf{y}(1), \tag{2.421}$$

with corresponding $\mathbf{P}(1)$ (column weighting is redundant in the full-rank fully-determined case).

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 for the unknown $\mathbf{x}$ is

$$\left\{\begin{matrix}\mathbf{E}(1)\\ \mathbf{E}(2)\end{matrix}\right\}\mathbf{x}+\begin{bmatrix}\mathbf{n}(1)\\ \mathbf{n}(2)\end{bmatrix}=\begin{bmatrix}\mathbf{y}(1)\\ \mathbf{y}(2)\end{bmatrix}, \tag{2.422}$$

where $\mathbf{x}$ is the same unknown. We assume $\langle\mathbf{n}(2)\rangle=\mathbf{0}$ and

$$\langle\mathbf{n}(1)\mathbf{n}(2)^{\mathrm{T}}\rangle=\mathbf{0}, \tag{2.423}$$

that is, *no correlation of the old and new measurement errors*. A solution to (2.422) should give a better estimate of $\mathbf{x}$ than (2.420) alone, because more observations are available. It is sensible to row weight the concatenated set to

$$\begin{bmatrix}\mathbf{R}_{nn}(1)^{-\mathrm{T}/2}\mathbf{E}(1)\\ \mathbf{R}_{nn}(2)^{-\mathrm{T}/2}\mathbf{E}(2)\end{bmatrix}\mathbf{x}+\begin{bmatrix}\mathbf{R}_{nn}(1)^{-\mathrm{T}/2}\mathbf{n}(1)\\ \mathbf{R}_{nn}(2)^{-\mathrm{T}/2}\mathbf{n}(2)\end{bmatrix}=\begin{bmatrix}\mathbf{R}_{nn}(1)^{-\mathrm{T}/2}\mathbf{y}(1)\\ \mathbf{R}_{nn}(2)^{-\mathrm{T}/2}\mathbf{y}(2)\end{bmatrix}. \tag{2.424}$$

"Recursive weighted least-squares" seeks the solution to (2.424) without inverting the new, larger matrix, by taking advantage of the existing knowledge of $\tilde{\mathbf{x}}(1)$, $\mathbf{P}(1)$ – however they might actually have been obtained. The objective function corresponding to finding the minimum weighted error norm in (2.424) is

$$\begin{aligned}J&=(\mathbf{y}(1)-\mathbf{E}(1)\mathbf{x})^{\mathrm{T}}\mathbf{R}_{nn}(1)^{-1}(\mathbf{y}(1)-\mathbf{E}(1)\mathbf{x})\\ &\quad+(\mathbf{y}(2)-\mathbf{E}(2)\mathbf{x})^{\mathrm{T}}\mathbf{R}_{nn}(2)^{-1}(\mathbf{y}(2)-\mathbf{E}(2)\mathbf{x}).\end{aligned} \tag{2.425}$$

Taking the derivatives with respect to $\mathbf{x}$, the normal equations produce a new solution,

$$\begin{aligned}\tilde{\mathbf{x}}(2)&=\{\mathbf{E}(1)^{\mathrm{T}}\mathbf{R}_{nn}(1)^{-1}\mathbf{E}(1)+\mathbf{E}(2)^{\mathrm{T}}\mathbf{R}_{nn}(2)^{-1}\mathbf{E}(2)\}^{-1}\\ &\quad\times\{\mathbf{E}(1)^{\mathrm{T}}\mathbf{R}_{nn}(1)^{-1}\mathbf{y}(1)+\mathbf{E}(2)^{\mathrm{T}}\mathbf{R}_{nn}(2)^{-1}\mathbf{y}(2)\}.\end{aligned} \tag{2.426}$$

This is the result from the brute-force re-solution. But one can manipulate (2.426) into[50] (see Appendix 3 to this chapter):

$$\begin{aligned}\tilde{\mathbf{x}}(2)&=\tilde{\mathbf{x}}(1)+\mathbf{P}(1)\mathbf{E}(2)^{\mathrm{T}}[\mathbf{E}(2)\mathbf{P}(1)\mathbf{E}(2)^{\mathrm{T}}+\mathbf{R}_{nn}(2)]^{-1}[\mathbf{y}(2)-\mathbf{E}(2)\tilde{\mathbf{x}}(1)]\\ &=\tilde{\mathbf{x}}(1)+\mathbf{K}(2)[\mathbf{y}(2)-\mathbf{E}(2)\tilde{\mathbf{x}}(1)],\end{aligned} \tag{2.427}$$

$$\mathbf{P}(2)=\mathbf{P}(1)-\mathbf{K}(2)\mathbf{E}(2)\mathbf{P}(1), \tag{2.428}$$

$$\mathbf{K}(2)=\mathbf{P}(1)\mathbf{E}(2)^{\mathrm{T}}[\mathbf{E}(2)\mathbf{P}(1)\mathbf{E}(2)^{\mathrm{T}}+\mathbf{R}_{nn}(2)]^{-1}. \tag{2.429}$$

An alternate form for $\mathbf{P}(2)$, found from the matrix inversion lemma, is

$$\mathbf{P}(2)=[\mathbf{P}(1)^{-1}+\mathbf{E}(2)^{\mathrm{T}}\mathbf{R}_{nn}(2)^{-1}\mathbf{E}(2)]^{-1}. \tag{2.430}$$

A similar alternate for $\tilde{\mathbf{x}}(2)$, involving different dimensions of the matrices to be inverted, is also available from the matrix inversion lemma, but is generally less useful. (In some large problems, however, matrix inversion can prove less onerous than matrix multiplication.)

The solution (2.427) 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 $\tilde{\mathbf{x}}(1)$, and $\mathbf{P}(1)$, the particular methodology originally employed for obtaining them is irrelevant: they might even have been obtained from an educated guess, or from some previous calculation of arbitrary complexity. If the initial set of equations (2.420) 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. Similarly, the dimensionality and rank of $\mathbf{E}\,(2)$ is arbitrary, as long as the matrix inverse exists.

Example *Suppose we have a single measurement of a scalar,* x*, so that* $x + n\,(1) = y(1)$, $\langle n(1)\rangle = 0$, $\langle n(1)^2\rangle = R(1)$*. Then an estimate of* x *is* $\tilde{x}(1) = y(1)$*, with uncertainty* $P(1) = R(1)$*. A second measurement then becomes available,* $x + n(2) = y(2)$, $\langle n(2)\rangle = 0$, $\langle n(2)^2\rangle = R(2)$*. By Eq. (2.427), an improved solution is*

$$\tilde{x}(2) = y(1) + R(1)/(R(1) + R(2))(y(1) - y(2)),$$

with uncertainty by Eq. (2.430),

$$P\,(2) = 1/\,(1/R\,(1) + 1/R(2)) = R(1)R(2)/(R\,(1) + R\,(2)).$$

If $R\,(1) = R\,(2) = R$*, we have* $\tilde{x}\,(2) = (y\,(1) + y\,(2))\,/2$, $P\,(2) = R/2$*. If there are* M *successive measurements all with the same error variance,* R*, one finds the last estimate is*

$$\begin{aligned}\tilde{x}\,(M) &= \tilde{x}\,(M-1) + R/\,(M-1)\,(R/\,(M-1) + R)^{-1}\,y\,(M)\\ &= \tilde{x}\,(M-1) + \frac{1}{M}y\,(M)\\ &= \frac{1}{M}\,(y\,(1) + y\,(2) + \cdots + y\,(M))\,,\end{aligned}$$

with uncertainty

$$P\,(M) = \frac{1}{((M-1)\,/R + 1/R)} = \frac{R}{M},$$

the conventional average and its variance. Note that each new measurement is given a weight $1/M$ *relative to the average,* $\tilde{x}(M-1)$, *already computed from the previous* $M-1$ *data points.*

The structure of the improved solution (2.427) is also 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 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, $\mathbf{P}(1) = \Delta^2\mathbf{I}$. Then,

$$\mathbf{K}(2) = \mathbf{E}(2)^{\mathrm{T}}[\mathbf{E}(2)\mathbf{E}(2)^{\mathrm{T}} + \mathbf{R}_{nn}(2)/\Delta^2]^{-1}. \tag{2.431}$$

If the new observations are extremely accurate, the norm of $\mathbf{R}_{nn}(2)/\Delta^2$ is small, and if the second set of observations is full-rank underdetermined,

$$\mathbf{K}(2) \longrightarrow \mathbf{E}(2)^{\mathrm{T}}(\mathbf{E}(2)\mathbf{E}(2)^{\mathrm{T}})^{-1}$$

and

$$\begin{aligned}\tilde{\mathbf{x}}(2) &= \tilde{\mathbf{x}}(1) + \mathbf{E}(2)^{\mathrm{T}}(\mathbf{E}(2)\mathbf{E}(2)^{\mathrm{T}})^{-1}[\mathbf{y}(2) - \mathbf{E}(2)\tilde{\mathbf{x}}(1)]\\ &= [\mathbf{I} - \mathbf{E}(2)^{\mathrm{T}}(\mathbf{E}(2)\mathbf{E}(2)^{\mathrm{T}})^{-1}\mathbf{E}(2)]\tilde{\mathbf{x}}(1) + \mathbf{E}(2)^{\mathrm{T}}(\mathbf{E}(2)\mathbf{E}(2)^{\mathrm{T}})^{-1}\mathbf{y}(2).\end{aligned} \tag{2.432}$$

Now, $[\mathbf{I} - \mathbf{E}(2)^{\mathrm{T}}(\mathbf{E}(2)\mathbf{E}(2)^{\mathrm{T}})^{-1}\mathbf{E}(2)] = \mathbf{I}_N - \mathbf{V}\mathbf{V}^{\mathrm{T}} = \mathbf{Q}_v\mathbf{Q}_v^{\mathrm{T}}$, where $\mathbf{V}$ is the full-rank singular vector matrix for $\mathbf{E}(2)$, and it spans the nullspace of $\mathbf{E}(2)$ (see Eq. (2.291)). The update thus replaces, in the first estimate, all the structures given perfectly by the second set of observations, but retains those structures from the first estimate about which the new observations say nothing – a sensible result. (Compare Eq. (2.427) with Eq. (2.360).) At the opposite extreme, when the new observations are very noisy compared to the previous ones, $\|\mathbf{R}_{nn}/\Delta^2\| \to \infty$, $\|\mathbf{K}(2)\| \to 0$, and the first estimate is left unchanged.

The general case represents a weighted average of the previous estimate with elements found from the new data, with the weighting depending both upon the relative noise in each, and upon the structure of the observations relative to the structure of $\mathbf{x}$ as represented in $\mathbf{P}(1)$, $\mathbf{R}_{nn}(2)$, $\mathbf{E}\,(2)$. The matrix being inverted in (2.429) 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 γ be the error component in $\tilde{\mathbf{x}}(1) = \mathbf{x}\,(1) + \gamma$, which by definition has covariance $\langle\gamma\gamma^{\mathrm{T}}\rangle = \mathbf{P}(1)$. Then the expected covariance of the error of prediction is $\langle \mathbf{E}(1)\gamma\gamma^{\mathrm{T}}\mathbf{E}\,(1)^{\mathrm{T}}\rangle = \mathbf{E}(1)\mathbf{P}(1)\mathbf{E}(1)^{\mathrm{T}}$, which appears in $\mathbf{K}(2)$. Because of the assumptions (2.423), and $\langle \gamma\,(1)\,\mathbf{x}\,(1)^{\mathrm{T}}\rangle = \mathbf{0}$, it follows that

$$\langle \mathbf{y}\,(1)\,(\mathbf{y}\,(2) - \mathbf{E}\,(2)\,\tilde{\mathbf{x}}\,(1))\rangle = \mathbf{0}. \tag{2.433}$$

That is, the prediction error or "innovation," $\mathbf{y}\,(2) - \mathbf{E}(2)\tilde{\mathbf{x}}(1)$, is uncorrelated with the previous measurement.

The possibility of a recursion based on Eqs. (2.427) and (2.428) (or (2.430)) is obvious – all subscript 1 variables being replaced by subscript 2 variables, which in turn are replaced by subscript 3 variables, etc. The general form would be:

$$\boxed{\begin{aligned} \tilde{\mathbf{x}}(m) &= \tilde{\mathbf{x}}(m-1) + \mathbf{K}(m)[\mathbf{y}(m) - \mathbf{E}(m)\tilde{\mathbf{x}}(m-1)], && (2.434)\\ \mathbf{K}(m) &= \mathbf{P}(m-1)\mathbf{E}(m)^{\mathrm{T}}[\mathbf{E}(m)\mathbf{P}(m-1)\mathbf{E}(m)^{\mathrm{T}} + \mathbf{R}_{nn}(m)]^{-1}, && (2.435)\\ \mathbf{P}(m) &= \mathbf{P}(m-1) - \mathbf{K}(m)\mathbf{E}(m)\mathbf{P}(m-1), \quad m = 1, 2, \ldots, && (2.436) \end{aligned}}$$

where m conventionally starts with 0. An alternative form for Eq. (2.436) is, from (2.430),

$$\mathbf{P}\,(m) = [\mathbf{P}\,(m-1)^{-1} + \mathbf{E}\,(m)^{\mathrm{T}}\,\mathbf{R}_{nn}\,(m)^{-1}\,\mathbf{E}\,(m)]^{-1}. \tag{2.437}$$

The computational load of the recursive solution needs to be addressed. A least-squares solution does *not* require one to calculate the uncertainty $\mathbf{P}$ (although the utility of $\tilde{\mathbf{x}}$ without such an estimate is unclear). But to use the recursive form, one must have $\mathbf{P}\,(m-1)$, otherwise the update step, Eq. (2.434) cannot be used. In very large problems, such as appear in oceanography and meteorology (Chapter 6), the computation of the uncertainty, from (2.436) or (2.437), can become prohibitive. In such a situation, one might simply store all the data, and do one large calculation – if this is feasible. Normally, it will involve less pure computation than will the recursive solution which must repeatedly update $\mathbf{P}(m)$.

The comparatively simple interpretation of the recursive, weighted least-squares problem will be used in Chapter 4 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.

2.8.2 Minimum variance recursive estimates

The recursive least-squares result is identical to a recursive estimation procedure, if appropriate least-squares weight matrices were used. To see this result, suppose there exist two *independent* 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 unbiassed, or have the same mean, $\langle \tilde{\mathbf{x}}_a \rangle = \langle \tilde{\mathbf{x}}_b \rangle = \mathbf{x}_B$. How should the two be combined to give a third estimate $\tilde{\mathbf{x}}^+$ with minimum error variance?

Try a linear combination,

$$\tilde{\mathbf{x}}^+ = \mathbf{L}_a \tilde{\mathbf{x}}_a + \mathbf{L}_b \tilde{\mathbf{x}}_b. \tag{2.438}$$

If the new estimate is to be unbiassed, or is to retain the prior bias (that is, the same mean), it follows that,

$$\langle \tilde{\mathbf{x}}^+ \rangle = \mathbf{L}_a \langle \tilde{\mathbf{x}}_a \rangle + \mathbf{L}_b \langle \tilde{\mathbf{x}}_b \rangle, \tag{2.439}$$

or

$$\mathbf{x}_B = \mathbf{L}_a \mathbf{x}_B + \mathbf{L}_b \mathbf{x}_B, \tag{2.440}$$

or

$$\mathbf{L}_b = \mathbf{I} - \mathbf{L}_a. \tag{2.441}$$

Then the uncertainty is

$$\begin{aligned}\mathbf{P}^+ &= \langle (\tilde{\mathbf{x}}^+ - \mathbf{x})(\tilde{\mathbf{x}}^+ - \mathbf{x})^{\mathrm{T}} \rangle = \langle (\mathbf{L}_a \tilde{\mathbf{x}}_a + (\mathbf{I} - \mathbf{L}_a)\tilde{\mathbf{x}}_b)(\mathbf{L}_a \tilde{\mathbf{x}}_a + (\mathbf{I} - \mathbf{L}_a)\tilde{\mathbf{x}}_b)^{\mathrm{T}} \rangle \\ &= \mathbf{L}_a \mathbf{P}_a \mathbf{L}_a^{\mathrm{T}} + (\mathbf{I} - \mathbf{L}_a)\mathbf{P}_b(\mathbf{I} - \mathbf{L}_a)^{\mathrm{T}},\end{aligned} \tag{2.442}$$

where the independence assumption has been used to set $\langle (\tilde{\mathbf{x}}_a - \mathbf{x})(\tilde{\mathbf{x}}_b - \mathbf{x}) \rangle = \mathbf{0}$. $\mathbf{P}^+$ is positive definite; minimizing its diagonal elements with respect to $\mathbf{L}_a$ yields (after writing out the diagonal elements of the products)

$$\mathbf{L}_a = \mathbf{P}_b(\mathbf{P}_a + \mathbf{P}_b)^{-1}, \qquad \mathbf{L}_b = \mathbf{P}_a(\mathbf{P}_a + \mathbf{P}_b)^{-1}.$$

(Blithely differentiating and setting to zero produces the correct answer:

$$\frac{\partial(\operatorname{diag} \mathbf{P}^+)}{\partial \mathbf{L}_a} = \operatorname{diag}\left(\frac{\partial \mathbf{P}^+}{\partial \mathbf{L}_a}\right) = \operatorname{diag}\left[2\mathbf{P}_a \mathbf{L}_a - \mathbf{P}_b(\mathbf{I} - \mathbf{L}_a)\right] = 0,$$

or $\mathbf{L}_a = \mathbf{P}_b(\mathbf{P}_a + \mathbf{P}_b)^{-1}$.) The new combined estimate is

$$\tilde{\mathbf{x}}^+ = \mathbf{P}_b(\mathbf{P}_a + \mathbf{P}_b)^{-1}\tilde{\mathbf{x}}_a + \mathbf{P}_a(\mathbf{P}_a + \mathbf{P}_b)^{-1}\tilde{\mathbf{x}}_b. \tag{2.443}$$

This last expression can be rewritten by adding and subtracting $\tilde{\mathbf{x}}_a$ as

$$\begin{aligned}\tilde{\mathbf{x}}^+ &= \tilde{\mathbf{x}}_a + \mathbf{P}_b(\mathbf{P}_a + \mathbf{P}_b)^{-1}\tilde{\mathbf{x}}_a \\ &\quad + \mathbf{P}_a(\mathbf{P}_a + \mathbf{P}_b)^{-1}\tilde{\mathbf{x}}_b - (\mathbf{P}_a + \mathbf{P}_b)(\mathbf{P}_a + \mathbf{P}_b)^{-1}\tilde{\mathbf{x}}_a \\ &= \tilde{\mathbf{x}}_a + \mathbf{P}_a(\mathbf{P}_a + \mathbf{P}_b)^{-1}(\tilde{\mathbf{x}}_b - \tilde{\mathbf{x}}_a).\end{aligned} \tag{2.444}$$

Notice, in particular, the re-appearance of a predictor-corrector form relative to $\tilde{\mathbf{x}}_a$.

The uncertainty of the estimate (2.444) is easily evaluated as

$$\mathbf{P}^{+} = \mathbf{P}_a - \mathbf{P}_a(\mathbf{P}_a + \mathbf{P}_b)^{-1}\mathbf{P}_a. \tag{2.445}$$

or, by straightforward application of the matrix inversion lemma,

$$\mathbf{P}^{+} = \left(\mathbf{P}_a^{-1} + \mathbf{P}_b^{-1}\right)^{-1}. \tag{2.446}$$

The uncertainty is again independent of the observations. Equations (2.444)–(2.446) are the general rules for combining two estimates with uncorrelated errors.

Now suppose that $\tilde{\mathbf{x}}_a$ and its uncertainty are known, but that instead of $\tilde{\mathbf{x}}_b$ there are measurements,

$$\mathbf{E}(2)\mathbf{x} + \mathbf{n}(2) = \mathbf{y}(2), \tag{2.447}$$

with $\langle\mathbf{n}(2)\rangle = 0$, $\langle\mathbf{n}(2)\mathbf{n}(2)^{\mathrm{T}}\rangle = \mathbf{R}_{nn}(2)$. From this second set of observations, we *estimate* the solution, using the minimum variance estimator (2.406, 2.408) with no use of the solution variance; that is, let $\|\mathbf{R}_{xx}^{-1}\| \to 0$. The reason for suppressing $\mathbf{R}_{xx}$, which logically could come from $\mathbf{P}_a$, is to maintain the independence of the previous and the new estimates. Then

$$\tilde{\mathbf{x}}_b = [\mathbf{E}(2)^{\mathrm{T}}\mathbf{R}_{nn}(2)^{-1}\mathbf{E}(2)]^{-1}\mathbf{E}(2)^{\mathrm{T}}\mathbf{R}_{nn}(2)^{-1}\mathbf{y}(2), \tag{2.448}$$

$$\mathbf{P}_b = [\mathbf{E}(2)^{\mathrm{T}}\mathbf{R}_{nn}(2)^{-1}\mathbf{E}(2)]^{-1}. \tag{2.449}$$

Substituting (2.448), (2.449) into (2.444), (2.445), and using the matrix inversion lemma, (see Appendix 3 to this chapter) gives

$$\tilde{\mathbf{x}}^{+} = \tilde{\mathbf{x}}_a + \mathbf{P}_a\mathbf{E}(2)^{\mathrm{T}}[\mathbf{E}(2)\mathbf{P}_a\mathbf{E}(2)^{\mathrm{T}} + \mathbf{R}_{nn}(2)]^{-1}(\mathbf{y}(2) - \mathbf{E}(2)\tilde{\mathbf{x}}_a), \tag{2.450}$$

$$\mathbf{P}^{+} = [\mathbf{P}_a^{-1} + \mathbf{E}(2)^{\mathrm{T}}\mathbf{R}_{nn}(2)^{-1}\mathbf{E}(2)]^{-1}, \tag{2.451}$$

which is the same as (2.434), (2.437) and thus *a recursive minimum variance estimate coincides with a corresponding weighted least-squares recursion*. The new covariance may also be confirmed to be that in either of Eqs. (2.436) or (2.437). Notice that if $\tilde{\mathbf{x}}_a$ was itself estimated from an earlier set of observations, then those data have disappeared from the problem, with all the information derived from them contained in $\tilde{\mathbf{x}}_a$ and $\mathbf{P}_a$. Thus, again, earlier data can be wholly discarded after use. It does not matter where $\tilde{\mathbf{x}}_a$ originated, whether from over- or underdetermined equations or a pure guess – as long as $\mathbf{P}_a$ is realistic. Similarly, expression (2.450) remains valid whatever the dimensionality or rank of $\mathbf{E}\,(2)$ as long as the inverse matrix exists. The general implementation of this sequence for a continuing data stream corresponds to Eqs. (2.434)–(2.437).

2.9 Summary

This chapter has not exhausted the possibilities for inverse methods, and the techniques will be extended in several directions in the next chapters. Given the lengthy nature of the discussion so far, however, some summary of what has been accomplished may be helpful.

The focus is on making inferences about parameters or fields, $\mathbf{x}$, $\mathbf{n}$ satisfying linear relationships of the form

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

Such equations arise as we have seen, from both "forward" and "inverse" problems, but the techniques for estimating $\mathbf{x}$, $\mathbf{n}$ and their uncertainty are useful whatever the physical origin of the equations. Two methods for estimating $\mathbf{x}$, $\mathbf{n}$ have been the focus of the chapter: least-squares (including the singular value decomposition) and the Gauss–Markov or minimum variance technique. Least-squares, in any of its many guises, is a very powerful method – but its power and ease of use have (judging from the published literature) led many investigators into serious confusion about what they are doing. This confusion is compounded by the misunderstandings about the difference between an inverse problem and an inverse method.

An attempt is made therefore, to emphasize the two distinct roles of least-squares: as a method of *approximation*, and as a method of *estimation*. It is only in the second formulation that it can be regarded as an inverse method. A working definition of an inverse method is a technique able to estimate unknown parameters or fields of a model, while producing an estimate of the uncertainties of the results. Solution plus uncertainty are the fundamental requirements. There are many desirable additional features of inverse methods. Among them are: (1) separation of nullspace uncertainties from observational noise uncertainties; (2) the ability to rank the data in its importance to the solution; (3) the ability to use prior statistical knowledge; (4) understanding of solution structures in terms of data structure; (5) the ability to trade resolution against variance. (The list is not exhaustive. For example, we will briefly examine in Chapter 3 the use of inequality information.) As with all estimation methods, one also trades computational load against the need for information. (The SVD, for example, is a powerful form of least-squares, but requires more computation than do other forms.) The Gauss–Markov approach has the strength of forcing explicit use of prior statistical information and is directed at the central goal of obtaining $\mathbf{x}$, $\mathbf{n}$ with the smallest mean-square error, and for this reason might well be regarded as the default methodology for linear inverse problems. It has the added advantage that we know we can obtain precisely the same result with appropriate versions of least-squares, including the SVD, permitting the use of least-squares algorithms, but at the risk of losing sight of what we are actually attempting. A limitation is that the underlying probability densities of solution and noise have to be unimodal (so that a minimum variance estimate makes sense). If unimodality fails, one must look to other methods.

The heavy emphasis here on noise and uncertainty may appear to be a tedious business. But readers of the scientific literature will come to recognize how qualitative much of the discussion is – the investigator telling a story about what he or she thinks is going on with no estimate of uncertainties, and no attempt to resolve quantitatively differences with previous competing estimates. In a quantitative subject, such vagueness is ultimately intolerable.

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. Because, in the presence of noise one is dealing with a statistical estimation problem, there is no single "best" answer, 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 is a good first choice – it is easily and efficiently computed and coincides with the Gauss–Markov and tapered SVD solutions, if the weight matrices are the appropriate covariances. Sparse matrix methods exist for its solution should that be necessary.[51] Coincidence with the Gauss–Markov solution means one can reinterpret it as a minimum-variance or maximum-likelihood solution (see Appendix 1 to this chapter) should one wish.

It is a comparatively easy matter to vary the trade-off parameter, γ^2, to explore the consequences of any errors in specifying the noise and solution variances. Once a value for γ^2 is chosen, the tapered SVD can then 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 about 500 × 500), 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 easily be overwhelmed by the available information, particularly if a range of solutions must be examined. The SVD has a flexibility beyond even what we have discussed – one could, for example, change the degree of tapering in each of the terms of (2.338) and (2.339) 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 – should the investigator know enough to justify it.

To the extent that either or both of **x**, **n** have expected structures expressible through covariance matrices, these structures can be removed from the problem through the various weight matrix and/or the Cholesky decomposition. The resulting problem is then one in completely unstructured (equivalent to white noise) elements **x**, **n**. In the resulting scaled and rotated systems, one can use the simplest of all objective functions. Covariance, resolution, etc., in the original spaces of **x**, **n** is readily recovered by appropriately applying the weight matrices to the results of the scaled/rotated space.

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. In particular, the truncated SVD does not produce a minimum variance estimate the way the tapered version can. The tapered SVD (along with the Gauss–Markov estimate, or the tapered least-squares solutions) produce the minimum variance property by tolerating a bias in the solution. Whether the bias is more desirable than a larger uncertainty is a decision the user must make. But the reader is warned against the belief that there is any single best method.

Appendix 1. Maximum likelihood

The estimation procedures used in this book are based primarily upon the idea of minimizing the variance of the estimate about the true value. Alternatives exist. For example, given a set of observations with known joint probability density, one can use a principle of "maximum likelihood." This very general and powerful principle attempts to find those estimated parameters that render the actual observations the most likely to have occurred. By way of motivation, consider the simple case of uncorrelated jointly normal stationary time series, x_i, where

$$\langle x_i \rangle = m,\ \langle (x_i - m)(x_j - m) \rangle = \sigma^2 \delta_{ij}.$$

The corresponding joint probability density for $\mathbf{x} = [x_1, x_2, \ldots, x_N]^{\mathrm{T}}$ can be written as

$$\begin{aligned} p_{\mathbf{x}}(\mathbf{X}) = {} & \frac{1}{(2\pi)^{N/2}\,\sigma^N} \\ & \times \exp\left\{ -\frac{1}{2\sigma^2}[(X_1 - m)^2 + (X_2 - m)^2 + \cdots + (X_N - m)^2] \right\}. \end{aligned} \tag{2.452}$$

Substitution of the observed values, $X_1 = x_1, X_2 = x_2, \ldots,$ into Eq. (2.452) permits evaluation of the probability that these particular values occurred. Denote the corresponding probability density as L. One can demand those values of m, σ rendering the value of L as large as possible. L will be a maximum if $\log(L)$ is as large as possible: that is, we seek to maximize

$$\log(L) = -\frac{1}{2\sigma^2}[(x_1-m)^2 + (x_2-m)^2 + \cdots + (x_N-m)^2] + N\log(\sigma) + \frac{N}{2}\log(2\pi),$$

with respect to m, σ. Setting the corresponding partial derivatives to zero and solving produces

$$\tilde{m} = \frac{1}{N}\sum_{i=1}^{M} x_i,\ \tilde{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \tilde{m})^2.$$

That is, the usual sample mean, and biassed sample variance maximize the probability of the observed data actually occurring. A similar calculation is readily carried out using correlated normal, or any random variables with a different probability density.

Likelihood estimation, and its close cousin, Bayesian methods, are general powerful estimation methods that can be used as an alternative to almost everything covered in this book.[52] Some will prefer that route, but the methods used here are adequate for a wide range of problems.

Appendix 2. Differential operators and Green functions

Adjoints appear prominently in the theory of differential operators and are usually discussed independently of any optimization problem. Many of the concepts are those used in defining Green functions.

Suppose we want to solve an ordinary differential equation,

$$\frac{du(\xi)}{d\xi} + \frac{d^2u(\xi)}{d\xi^2} = \rho(\xi), \tag{2.453}$$

subject to boundary conditions on $u(\xi)$ at $\xi = 0, L$. To proceed, seek first a solution to

$$\alpha\frac{\partial v(\xi,\xi_0)}{\partial\xi} + \frac{\partial^2 v(\xi,\xi_0)}{\partial\xi^2} = \delta(\xi_0 - \xi), \tag{2.454}$$

where α is arbitrary for the time being. Multiply (2.453) by v, and (2.454) by u, and subtract:

$$v(\xi,\xi_0)\frac{du(\xi)}{d\xi} + v(\xi,\xi_0)\frac{d^2u(\xi)}{d\xi^2} - u(\xi)\alpha\frac{\partial v(\xi,\xi_0)}{\partial\xi} - u(\xi)\frac{\partial^2 v(\xi,\xi_0)}{\partial\xi^2} = v(\xi,\xi_0)\rho(\xi) - u(\xi)v(\xi,\xi_0). \tag{2.455}$$

Integrate this last equation over the domain,

$$\int_0^L \left\{ v(\xi,\xi_0)\frac{\mathrm{d}u\,(\xi)}{d\xi} + v(\xi,\xi_0)\frac{\mathrm{d}^2u\,(\xi)}{d\xi^2} - u(\xi)\alpha\frac{\partial v(\xi,\xi_0)}{\partial\xi} - u\,(\xi)\,\frac{\partial^2 v(\xi,\xi_0)}{\partial\xi^2} \right\} \mathrm{d}\xi \tag{2.456}$$

$$= \int_0^L \{v(\xi,\xi_0)\rho(\xi) - u\,(\xi)\,\delta(\xi_0-\xi)\}\mathrm{d}\xi, \tag{2.457}$$

or

$$\int_0^L \frac{\mathrm{d}}{\mathrm{d}\xi}\left\{ v\frac{\mathrm{d}u}{\mathrm{d}\xi} - \alpha u\frac{\mathrm{d}v}{\mathrm{d}\xi} \right\}\mathrm{d}\xi + \int_0^L \left\{ u\frac{\partial^2 v(\xi,\xi_0)}{\partial\xi^2} - u\frac{\partial^2 v(\xi,\xi_0)}{\partial\xi^2} \right\}\mathrm{d}\xi$$
$$= \int_0^L v(\xi,\xi_0)\rho\,(\xi)\,\mathrm{d}\xi - u(\xi_0). \tag{2.458}$$

If we choose $\alpha = -1$, then the first term on the left-hand side is integrable, as

$$\int_0^L \frac{\mathrm{d}}{\mathrm{d}\xi}\{uv\}\,\mathrm{d}\xi = uv|_0^L, \tag{2.459}$$

as is the second term on the left,

$$\int_0^L \frac{\mathrm{d}}{\mathrm{d}\xi}\left\{ u\frac{\partial v}{\partial\xi} - v\frac{\partial u}{\partial\xi} \right\}\partial\xi = \left[u\frac{\partial v}{\partial\xi} - v\frac{\partial u}{\partial\xi} \right]_0^L, \tag{2.460}$$

and thus,

$$u(\xi_0) = \int_0^L v(\xi,\xi_0)\rho\,(\xi)\,\mathrm{d}\xi + uv|_0^L + \left[u\frac{\mathrm{d}v}{\mathrm{d}\xi} - v\frac{\mathrm{d}u}{\mathrm{d}\xi} \right]_0^L. \tag{2.461}$$

Because the boundary conditions on v were not specified, we are free to choose them such that $v = 0$ on $\xi = 0, L$, e.g., the boundary terms reduce simply to $[u\mathrm{d}v/\mathrm{d}\xi]_0^L$, which is then known.

Here, v is the adjoint solution to Eq. (2.454), with $\alpha = -1$, defining the adjoint equation to (2.453); it was found by requiring that the terms on the left-hand side of Eq. (2.458) should be exactly integrable. v is also the problem Green function (although the Green function is sometimes defined so as to satisfy the forward operator, rather than the adjoint one). Textbooks show that for a general differential operator, $\mathcal{L}$, the requirement that v should render the analogous terms integrable is that

$$u^{\mathrm{T}}\mathcal{L}v = v^{\mathrm{T}}\mathcal{L}^{\mathrm{T}}u, \tag{2.462}$$

where here the superscript T denotes the adjoint. Equation (2.462) defines the adjoint operator (compare to (2.376)).

Appendix 3. Recursive least-squares and Gauss–Markov solutions

The recursive least-squares solution Eq. (2.427) is appealingly simple. Unfortunately, obtaining it from the concatenated least-squares form (2.426),

$$\begin{aligned}\tilde{\mathbf{x}}(2) = {}& \{\mathbf{E}(1)^{\mathrm{T}}\mathbf{R}_{nn}(1)^{-1}\mathbf{E}(1) + \mathbf{E}(2)^{\mathrm{T}}\mathbf{R}_{nn}(2)^{-1}\mathbf{E}(2)\}^{-1}\\ & \times \{\mathbf{E}(1)^{\mathrm{T}}\mathbf{R}_{nn}(1)^{-1}\mathbf{y}(1) + \mathbf{E}(2)^{\mathrm{T}}\mathbf{R}_{nn}(2)^{-1}\mathbf{y}\}\end{aligned}$$

is not easy. First note that

$$\begin{aligned}\tilde{\mathbf{x}}(1) &= [\mathbf{E}(1)^{\mathrm{T}}\mathbf{R}_{nn}(1)^{-1}\mathbf{E}(1)]^{-1}\mathbf{E}(1)^{\mathrm{T}}\mathbf{R}_{nn}(1)^{-1}\mathbf{y}(1)\\ &= \mathbf{P}(1)\mathbf{E}(1)^{\mathrm{T}}\mathbf{R}_{nn}(1)^{-1}\mathbf{y}(1),\end{aligned} \tag{2.463}$$

where

$$\mathbf{P}(1) = [\mathbf{E}(1)^{\mathrm{T}}\mathbf{R}_{nn}(1)^{-1}\mathbf{E}(1)]^{-1},$$

are the solution and uncertainty of the overdetermined system from the first set of observations alone. Then

$$\begin{aligned}\tilde{\mathbf{x}}(2) = {}& \{\mathbf{P}(1)^{-1} + \mathbf{E}(2)^{\mathrm{T}}\mathbf{R}_{nn}(2)^{-1}\mathbf{E}(2)\}^{-1}\\ & \times \{\mathbf{E}(1)^{\mathrm{T}}\mathbf{R}_{nn}(1)^{-1}\mathbf{y}(1) + \mathbf{E}(2)^{\mathrm{T}}\mathbf{R}_{nn}(2)^{-1}\mathbf{y}(2)\}.\end{aligned}$$

Apply the matrix inversion lemma, in the form Eq. (2.35), to the first bracket (using $\mathbf{C} \rightarrow \mathbf{P}(1)^{-1}, \mathbf{B} \rightarrow \mathbf{E}(2), \mathbf{A} \rightarrow \mathbf{R}_{nn}(2)$),

$$\begin{aligned}\tilde{\mathbf{x}}(2) &= \{\mathbf{P}(1) - \mathbf{P}(1)\mathbf{E}(2)^{\mathrm{T}}[\mathbf{E}(2)\mathbf{P}(1)\mathbf{E}(2)^{\mathrm{T}} + \mathbf{R}_{nn}(2)]^{-1}\mathbf{E}(2)\mathbf{P}(1)\}\\ &\quad \times \{\mathbf{E}(1)^{\mathrm{T}}\mathbf{R}_{nn}(1)^{-1}\mathbf{y}(1)\} + \{\mathbf{P}(1) - \mathbf{P}(1)\mathbf{E}(2)^{\mathrm{T}}[\mathbf{E}(2)\mathbf{P}(1)\mathbf{E}(2)^{\mathrm{T}}\\ &\quad + \mathbf{R}_{nn}(2)]^{-1}\mathbf{E}(2)\mathbf{P}(1)\}\{\mathbf{E}(2)^{\mathrm{T}}\mathbf{R}_{nn}(2)^{-1}\mathbf{y}(2)\}\\ &= \tilde{\mathbf{x}}(1) - \mathbf{P}(1)\mathbf{E}(2)^{\mathrm{T}}[\mathbf{E}(2)\mathbf{P}(1)\mathbf{E}(2)^{\mathrm{T}} + \mathbf{R}_{nn}(2)]^{-1}\mathbf{E}(2)\tilde{\mathbf{x}}(1)\\ &\quad + \mathbf{P}(1)\mathbf{E}(2)^{\mathrm{T}}\{\mathbf{I} - [\mathbf{E}(2)\mathbf{P}(1)\mathbf{E}(2)^{\mathrm{T}} + \mathbf{R}_{nn}(2)]^{-1}\mathbf{E}(2)\mathbf{P}(1)\mathbf{E}(2)^{\mathrm{T}}\}\\ &\quad \times \mathbf{R}_{nn}(2)^{-1}\mathbf{y}(2),\end{aligned}$$

using (2.463) and factoring $\mathbf{E}(2)^{\mathrm{T}}$ in the last line. Using the identity

$$[\mathbf{E}(2)\mathbf{P}(1)\mathbf{E}(2)^{\mathrm{T}} + \mathbf{R}_{nn}(2)]^{-1}[\mathbf{E}(2)\mathbf{P}(1)\mathbf{E}(2)^{\mathrm{T}} + \mathbf{R}_{nn}(2)] = \mathbf{I},$$

and substituting for $\mathbf{I}$ in the previous expression, factoring, and collecting terms, gives

$$\tilde{\mathbf{x}}(2) = \tilde{\mathbf{x}}(1) + \mathbf{P}(1)\mathbf{E}(2)^{\mathrm{T}}[\mathbf{E}(2)\mathbf{P}(1)\mathbf{E}(2)^{\mathrm{T}} + \mathbf{R}_{nn}(2)]^{-1}[\mathbf{y}(2) - \mathbf{E}(2)\tilde{\mathbf{x}}(1)], \tag{2.464}$$

which is the desired expression. The new uncertainty is given by (2.428) or (2.430).

Manipulation of the recursive Gauss–Markov solution (2.443) or (2.444) is similar, involving repeated use of the matrix inversion lemma. Consider Eq. (2.443) with $\mathbf{x}_b$ from Eq. (2.448),

$$\tilde{\mathbf{x}}^{+} = \left(\mathbf{E}\left(2\right)^{\mathrm{T}}\mathbf{R}_{nn}^{-1}\mathbf{E}\left(2\right)\right)^{-1}\left[\mathbf{P}_a + \mathbf{E}\left(2\right)^{\mathrm{T}}\mathbf{R}_{nn}^{-1}\mathbf{E}\left(2\right)\right]^{-1}\tilde{\mathbf{x}}_a$$
$$+\ \mathbf{P}_a\left[\mathbf{P}_a + \mathbf{E}\left(2\right)^{\mathrm{T}}\mathbf{R}_{nn}^{-1}\mathbf{E}\left(2\right)\right]^{-1}\left(\mathbf{E}\left(2\right)^{\mathrm{T}}\mathbf{R}_{nn}\mathbf{E}\left(2\right)\right)^{-1}\mathbf{E}\left(2\right)^{\mathrm{T}}\mathbf{R}_{nn}^{-1}\mathbf{y}(2).$$

Using Eq. (2.36) on the first term (with $\mathbf{A} \rightarrow (\mathbf{E}(2)^{\mathrm{T}}\mathbf{R}_{nn}^{-1}\mathbf{E}(2))^{-1}$, $\mathbf{B} \rightarrow \mathbf{I}$, $\mathbf{C} \rightarrow \mathbf{P}_a$), and on the second term with $\mathbf{C} \rightarrow (\mathbf{E}(2)\mathbf{R}_{nn}^{-1}\mathbf{E}(2))$, $\mathbf{A} \rightarrow \mathbf{P}_a$, $\mathbf{B} \rightarrow \mathbf{I}$, this last expression becomes

$$\tilde{\mathbf{x}}^{+} = \left[\mathbf{P}_a^{-1} + \mathbf{E}\left(2\right)^{\mathrm{T}}\mathbf{R}_{nn}^{-1}\mathbf{E}\left(2\right)\right]^{-1}\left[\mathbf{P}_a^{-1}\tilde{\mathbf{x}}_a + \mathbf{E}\left(2\right)^{\mathrm{T}}\mathbf{R}_{nn}^{-1}\mathbf{y}\left(2\right)\right],$$

yet another alternate form. By further application of the matrix inversion lemma,[53] this last expression can be manipulated into Eq. (2.450), which is necessarily the same as (2.464).

These expressions have been derived assuming that matrices such as $\mathbf{E}\left(2\right)^{\mathrm{T}}\mathbf{R}_{nn}^{-1}\mathbf{E}\left(2\right)$ are non-singular (full-rank overdetermined). If they are singular, they can be inverted using a generalized inverse, but taking care that $\mathbf{P}(1)$ includes the nullspace contribution (e.g., from Eq. (2.272)).

Notes

1 Noble and Daniel (1977), Strang (1988).
2 Lawson and Hanson (1995).
3 "Positive definite" will be defined below. Here it suffices to mean that $\mathbf{c}^{\mathrm{T}}\mathbf{W}\mathbf{c}$ should never be negative, for any $\mathbf{c}$.
4 Golub and Van Loan (1996).
5 Haykin (2002).
6 Lawson and Hanson (1995), Golub and van Loan (1996), Press *et al.* (1996), etc.
7 Determinants are used only rarely in this book. Their definition and properties are left to the references, as they are usually encountered in high school mathematics.
8 Rogers (1980) is an entire volume of matrix derivative identities, and many other useful properties are discussed by Magnus and Neudecker (1988).
9 Magnus and Neudecker (1988, p. 183).
10 Liebelt (1967, Sections 1–19).
11 The history of this not-very-obvious identity is discussed by Haykin (2002).
12 A good statistics text such as Cramér (1946), or one on regression such as Seber and Lee (2003), should be consulted.
13 Feller (1957) and Jeffreys (1961) represent differing philosophies. Jaynes (2003) forcefully and colorfully argues the case for so-called Bayesian inference (following Jeffreys), and it seems likely that this approach to statistical inference will ultimately become the default method; Gauch (2003) has a particularly clear account of Bayesian methods. For most of the methods in this book, however, we use little more than the first moments of probability distributions, and hence can ignore the underlying philosophical debate.
14 It follows from the Cauchy–Schwarz inequality: Consider $\langle (ax' + y')^2 \rangle = a\langle x'^2 \rangle + \langle y'^2 \rangle + 2a\langle x'y' \rangle \geq 0$ for any constant a. Choose $a = -\langle x'y' \rangle / \langle x'^2 \rangle$, and one has $-\langle x'y' \rangle^2 / \langle x'^2 \rangle + \langle y'^2 \rangle \geq 0$, or $1 \geq \langle x'y' \rangle^2 / (\langle x'^2 \rangle \langle y'^2 \rangle)$. Taking the square root of both sides, the required result follows.
15 Draper and Smith (1998), Seber and Lee (2003).

16 Numerical schemes for finding $\mathbf{C}_{\xi\xi}^{1/2}$ are described by Lawson and Hanson (1995) and Golub and Van Loan (1996)

17 Cramér (1946) discusses what happens when the determinant of $\mathbf{C}_{\xi\xi}$ vanishes, that is, if $\mathbf{C}_{\xi\xi}$ is singular.

18 Bracewell (2000).

19 Cramér (1946).

20 In problems involving time, one needs to be clear that "stationary" is not the same idea as "steady."

21 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.

22 The terminology "least-squares" is reserved in this book, conventionally, for the minimization of discrete sums such as Eq. (2.89). This usage contrasts with that of Bennett (2002) who applies it to continuous integrals, such as $\int_a^b (u(q)-r(q))^2\, \mathrm{d}q$, leading to the calculus of variations and Euler–Lagrange equations.

23 Box *et al.* (1994), Draper and Smith (1998), or Seber and Lee (2003), are all good starting points.

24 Draper and Smith (1998, Chapter 3) and the references given there.

25 Gill *et al.* (1986).

26 Wunsch and Minster (1982).

27 Morse and Feshbach (1953, p. 238), Strang (1988).

28 See Sewell (1987) for an interesting discussion.

29 But the matrix transpose is not what the older literature calls the "adjoint matrix," which is quite different. In the more recent literature the latter has been termed the "adjugate" matrix to avoid confusion.

30 In the meteorological terminology of Sasaki (1970) and others, exact relationships are called "strong" constraints, and those imposed in the mean-square are "weak" ones.

31 Claerbout (2001) displays more examples, and Lanczos (1961) gives a very general discussion of operators and their adjoints, Green functions, and their adjoints. See also the appendix to this chapter.

32 Wiggins (1972).

33 Brogan (1991) has a succinct discussion.

34 Lanczos (1961, pp. 117–18), sorts out the sign dependencies.

35 Lawson and Hanson (1995).

36 The singular value decomposition for arbitrary non-square matrices is apparently due to the physicist-turned-oceanographer Carl Eckart (Eckart and Young, 1939; see the discussion in Klema and Laub, 1980; Stewart, 1993; or Haykin, 2002). A particularly lucid account is given by Lanczos (1961) who, however, fails to give the decomposition a name. Other references are Noble and Daniel (1977), Strang (1988), and many recent books on applied linear algebra. The crucial role it plays in inverse methods appears to have been first noticed by Wiggins (1972).

37 Munk *et al.* (1995).

38 In physical oceanography, the distance would be that traveled by a ship between stops for measurement, and the water depth is clearly determined by the local topography.

39 Menke (1989).

40 Hansen (1992), or Lawson and Hanson (1995). Hansen's (1992) discussion is particularly interesting because he exploits the "generalized SVD," which is used to simultaneously diagonalize two matrices.

41 Munk and Wunsch (1982).

42 Seber and Lee (2003).

43 Luenberger (2003).

44 In oceanographic terms, the exact constraints describe the Stommel Gulf Stream solution. The eastward intensification of the adjoint solution corresponds to the change in sign of β in the adjoint model. See Schröter and Wunsch (1986) for details and an elaboration to a non-linear situation.

45 Lanczos (1960) has a good discussion.
46 See Lanczos (1961, Section 3.19).
47 The derivation follows Liebelt (1967).
48 Cf. Bretherton *et al*. (1976).
49 The time series was generated as $y_t = 0.999 y_{t-1} + \theta_t$, $\langle \theta_t \rangle = 0$, $\langle \theta_t^2 \rangle = 1$, a so-called autoregressive process of order 1 (AR(1)). The covariance $\langle y_i y_j \rangle$ can be determined analytically; see Priestley (1982, p. 119). Many geophysical processes obey similar rules.
50 Stengel (1986); Brogan (1991).
51 Paige and Saunders (1982).
52 See especially, van Trees (2001).
53 Liebelt (1967, p. 164).