can be extended to nonlinear model functions by first finding, by numerical methods, a point in parameter space, $\mathbf{x}_0$, at which the gradient vanishes and then expanding the model functions about that point in Taylor's series, retaining only the linear terms. Equation (8.1.2.4) then becomes

$$\boldsymbol{M}(\mathbf{x}) \approx \boldsymbol{M}(\mathbf{x}_0) + \boldsymbol{A}(\mathbf{x} - \mathbf{x}_0), \tag{8.1.2.10}$$

where $A_{ij} = \partial M_i(\mathbf{x})/\partial x_j$ evaluated at $\mathbf{x} = \mathbf{x}_0$. Because we have already found the least-squares solution, the estimate

$$\begin{aligned}\widehat{\mathbf{x}} &= \mathbf{x}_0 + (\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{W}[\mathbf{y} - \boldsymbol{M}(\mathbf{x}_0)] \\ &= \mathbf{x}_0 + \boldsymbol{H}[\mathbf{y} - \boldsymbol{M}(\mathbf{x}_0)]\end{aligned} \tag{8.1.2.11}$$

reduces to $\widehat{\mathbf{x}} = \mathbf{x}_0$. It is important, however, not to confuse $\mathbf{x}_0$, which is a convenient origin, with $\widehat{\mathbf{x}}$, which is a random variable describable by a joint p.d.f. with mean $\mathbf{x}_0$ and a variance–covariance matrix $\boldsymbol{V}_\mathbf{x} = \boldsymbol{H}\boldsymbol{V}_\mathbf{y}\boldsymbol{H}^T$, reducing to $(\boldsymbol{A}^T\boldsymbol{V}_\mathbf{y}^{-1}\boldsymbol{A})^{-1}$ when $\boldsymbol{W} = \boldsymbol{V}_\mathbf{y}^{-1}$.

This variance–covariance matrix is the one appropriate to the linear approximation given in (8.1.2.10), and it is valid (and the estimate is unbiased) only to the extent that the approximation is a good one. *A* useful criterion for an adequate approximation (Fedorov, 1972) is, for each j and k,

$$\left|\sum_{i=1}^{n} w_i\sigma_i \frac{\partial^2 M_i(\mathbf{x}_0)}{\partial x_j \partial x_k}\right| \ll \left(\left\{\sum_{i=1}^{n} w_i\left[\frac{\partial M_i(\mathbf{x}_0)}{\partial x_j}\right]^2\right\} \times \left\{\sum_{i=1}^{n} w_i\left[\frac{\partial M_i(\mathbf{x}_0)}{\partial x_k}\right]^2\right\}\right)^{1/2}, \tag{8.1.2.12}$$

where σ_i is the estimated standard deviation or *standard uncertainty* (Schwarzenbach, Abrahams, Flack, Prince & Wilson, 1995) of y_i. This criterion states that the curvature of $S(\mathbf{y}, \mathbf{x})$ in a region whose size is of order σ in observation space is small; it ensures that the effect of second-derivative terms in the normal-equations matrix on the eigenvalues and eigenvectors of the matrix is negligible. [For a further discussion and some numerical tests of alternatives, see Donaldson & Schnabel (1986).]

The process of refinement can be viewed as the construction of a conditional p.d.f. of a set of model parameters, $\mathbf{x}$, given a set of observations, $\mathbf{y}$. An important expression for this p.d.f. is derived from two equivalent expressions for the joint p.d.f. of $\mathbf{x}$ and $\mathbf{y}$:

$$\Phi_J(\mathbf{x}, \mathbf{y}) = \Phi_C(\mathbf{x}|\mathbf{y})\Phi_M(\mathbf{y}) = \Phi_C(\mathbf{y}|\mathbf{x})\Phi_M(\mathbf{x}). \tag{8.1.2.13}$$

Provided $\Phi_M(\mathbf{y}) > 0$, the conditional p.d.f. we seek can be written

$$\Phi_C(\mathbf{x}|\mathbf{y}) = \Phi_C(\mathbf{y}|\mathbf{x})\Phi_M(\mathbf{x})/\Phi_M(\mathbf{y}). \tag{8.1.2.14}$$

Here, the factor $[1/\Phi_M(\mathbf{y})]$ is the factor that is required to normalize the p.d.f. $\Phi_C(\mathbf{y}|\mathbf{x})$ is the conditional probability of observing a set of values of $\mathbf{y}$ as a function of $\mathbf{x}$. When the observations have already been made, however, this can also be considered a density function for $\mathbf{x}$ that measures the *likelihood* that those particular values of $\mathbf{y}$ would have been observed for various values of $\mathbf{x}$. It is therefore frequently written $\ell(\mathbf{x}|\mathbf{y})$, and (8.1.2.14) becomes

$$\Phi_C(\mathbf{x}|\mathbf{y}) = c\ell(\mathbf{x}|\mathbf{y})\Phi_M(\mathbf{x}), \tag{8.1.2.15}$$

where $c = [1/\Phi_M(\mathbf{y})]$ is the normalizing constant. $\Phi_M(\mathbf{x})$, the marginal p.d.f. of $\mathbf{x}$ in the absence of any additional information, incorporates all previously available information concerning $\mathbf{x}$, and is known as the *prior p.d.f.*, or, frequently, simply as the *prior* of $\mathbf{x}$. Similarly, $\Phi_C(\mathbf{x}|\mathbf{y})$ is the *posterior p.d.f.*, or the *posterior*, of $\mathbf{x}$. The relation in (8.1.2.14) and (8.1.2.15) was first stated in the eighteenth century by Thomas Bayes, and it is therefore known as Bayes's theorem (Box & Tiao, 1973). Although its validity has never been in serious question, its application has divided statisticians into two vehemently disputing camps, one of which, the frequentists, considers that Bayesian methods give nonobjective results, while the other, the Bayesians, considers that only by careful construction of a 'noninformative' prior can true objectivity be achieved (Berger & Wolpert, 1984).

Diffraction data, in general, contain no phase information, so the likelihood function for the structure factor, F, given a value of observed intensity, will have a value significantly different from zero in an annular region of the complex plane with a mean radius equal to $|F|$. Because this is insufficient information with which to determine a crystal structure, a prior p.d.f. is constructed in one (or some combination) of two ways. Either the prior knowledge that electron density is non negative is used to construct a joint p.d.f. of amplitudes and phases, given amplitudes for all reflections and phases for a few of them (direct methods), or chemical knowledge and intuition are used to construct a trial structure from which structure factors can be calculated, and the phase of $F_{\rm calc}$ is assigned to $F_{\rm obs}$. Both of these procedures can be considered to be applications of Bayes's theorem. In fact, $F_{\rm calc}$ for a refined structure can be considered a Bayesian estimate of F.

8.1.3. Implementation of linear least squares

In this section, we consider in detail numerical methods for solving linear least-squares problems, that is, the situation where (8.1.2.4) and (8.1.2.5) apply exactly.

8.1.3.1. *Use of the QR factorization*

The linear least-squares problem can be viewed geometrically as the problem of finding the point in a p-dimensional subspace, defined as the set of points that can be reached by a linear combination of the columns of $\boldsymbol{A}$, closest to a given point, $\mathbf{y}$, in an n-dimensional observation space. Since this is equivalent to finding the orthogonal projection of point $\mathbf{y}$ into that subspace, it is not surprising that an orthogonal decomposition of $\boldsymbol{A}$ helps to solve the problem. For convenience in this discussion, let us remove the weight matrix from the problem by defining the standardized design matrix by

$$\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{A}, \tag{8.1.3.1}$$

where $\boldsymbol{U}$ is the upper triangular Cholesky factor of $\boldsymbol{W}$.

Consider the least-squares problem with the QR factorization of $\boldsymbol{Z}$, as given in Subsection 8.1.1.1. For $\mathbf{y}' = \boldsymbol{U}(\mathbf{y} - \mathbf{b})$, (8.1.2.5) becomes

$$\begin{aligned}S &= (\mathbf{y}' - \boldsymbol{Z}\mathbf{x})^T(\mathbf{y}' - \boldsymbol{Z}\mathbf{x}) \\ &= [\boldsymbol{Q}^T(\mathbf{y}' - \boldsymbol{Z}\mathbf{x})]^T[\boldsymbol{Q}^T(\mathbf{y}' - \boldsymbol{Z}\mathbf{x})],\end{aligned} \tag{8.1.3.2}$$

which reduces to

$$S = (\boldsymbol{Q}_Z^T\mathbf{y}' - \boldsymbol{R}\mathbf{x})^T(\boldsymbol{Q}_Z^T\mathbf{y}' - \boldsymbol{R}\mathbf{x}) + \mathbf{y}'^T\boldsymbol{Q}_\perp\boldsymbol{Q}_\perp^T\mathbf{y}'. \tag{8.1.3.3}$$

The second term in (8.1.3.3) is independent of $\mathbf{x}$, and is therefore the sum of squared residuals. The first term vanishes if

$$\boldsymbol{R}\mathbf{x} = \boldsymbol{Q}_Z^T\mathbf{y}', \tag{8.1.3.4}$$

which, because $\boldsymbol{R}$ is upper triangular, is easily solved for $\mathbf{x}$. The QR decomposition of $\boldsymbol{Z}$ therefore leads naturally to the following algorithm for solving the linear least-squares problem:

(1) compute the QR factorization of $\boldsymbol{Z}$;
(2) compute $\boldsymbol{Q}_Z^T\mathbf{y}'$;
(3) solve $\boldsymbol{R}\mathbf{x} = \boldsymbol{Q}_Z^T\mathbf{y}'$ for $\mathbf{x}$.
(4) compute the residual sum of squares by $\mathbf{y}'^T\mathbf{y}' - \mathbf{y}'\boldsymbol{Q}_Z\boldsymbol{Q}_Z^T\mathbf{y}'$.
(5) compute the variance–covariance matrix from $\boldsymbol{V}_\mathbf{x} = \boldsymbol{R}^{-1}(\boldsymbol{R}^{-1})^T$.

8.1.3.2. *The normal equations*

Let us now consider the relationship of the QR procedure for solving the linear least-squares problem to the classical method based on the normal equations. The normal equations can be derived by differentiating (8.1.3.2) and equating the result to a null vector. This yields

$$\boldsymbol{Z}^T\boldsymbol{Z}\mathbf{x} = \boldsymbol{Z}^T\mathbf{y}'. \tag{8.1.3.5}$$

The algorithm is therefore to compute the cross-product matrix, $\boldsymbol{B} = \boldsymbol{Z}^T\boldsymbol{Z}$, and the right-hand side, $\mathbf{d} = \boldsymbol{Z}^T\mathbf{y}'$, and to solve the resulting system of equations, $\boldsymbol{B}\mathbf{x} = \mathbf{d}$. This is usually accomplished by computing the Cholesky decomposition of $\boldsymbol{B}$, that is $\boldsymbol{B} = \boldsymbol{C}^T\boldsymbol{C}$, where $\boldsymbol{C}$ is upper triangular, and then solving the two triangular systems $\boldsymbol{C}^T\mathbf{v} = \mathbf{d}$ and $\boldsymbol{C}\mathbf{x} = \mathbf{v}$. Because $\boldsymbol{Z} = \boldsymbol{Q}_Z\boldsymbol{R}$, equation (8.1.3.5) becomes

$$\boldsymbol{R}^T\boldsymbol{Q}_Z^T\boldsymbol{Q}_Z\boldsymbol{R}\mathbf{x} = \boldsymbol{R}^T\boldsymbol{Q}_Z^T\mathbf{y}', \tag{8.1.3.6}$$

or

$$\boldsymbol{R}^T\boldsymbol{R}\mathbf{x} = \boldsymbol{R}^T\boldsymbol{Q}_Z^T\mathbf{y}'. \tag{8.1.3.7}$$

It is clear that $\boldsymbol{R}$ is the Cholesky factor of $\boldsymbol{Z}^T\boldsymbol{Z}$, although it is formed in a different way. This procedure requires of order $(np^2)/2$ operations to form the product $\boldsymbol{Z}^T\boldsymbol{Z}$ and $p^3/3$ operations for the Cholesky decomposition. In some situations, the extra time to compute the QR factorization is justified because of greater stability, as will be discussed below. Most other quantities of statistical interest can be computed directly from the QR factorization.

8.1.3.3. *Conditioning*

The condition number of $\boldsymbol{Z}$, which is defined (Subsection 8.1.1.1) as the square root of the ratio of the largest to the smallest eigenvalue of $\boldsymbol{Z}^T\boldsymbol{Z}$, is an indicator of the effect a small change in an element of $\boldsymbol{Z}$ will have on the elements of $(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}$and of $\hat{\mathbf{x}}$. A large value of the condition number means that small errors in computing an element of $\boldsymbol{Z}$, owing possibly to truncation or roundoff in the computer, can introduce large errors into the elements of the inverse matrix. Also, when the condition number is large, the standard uncertainties of some estimated parameters will be large. A large condition number, as defined in this way, can result from either scaling or correlation or some combination of these. To illustrate this, consider the matrices

$$\boldsymbol{Z}^T\boldsymbol{Z} = \begin{pmatrix} 2+\varepsilon & 0 \\ 0 & \varepsilon \end{pmatrix}$$

and

$$\boldsymbol{Z}^T\boldsymbol{Z} = \begin{pmatrix} 1 & 1-\varepsilon \\ 1-\varepsilon & 1 \end{pmatrix},$$

where ε represents machine precision, which can be defined as the smallest number in machine representation that, when added to 1, gives a result different from 1. By the conventional definition, both of these matrices have a condition number for $\boldsymbol{Z}$ of $[(2+\varepsilon)/\varepsilon]^{1/2}$. Because numbers of order ε can be perfectly well represented, however, the first one can be inverted without loss of precision, whereas an inverse for the second would be totally meaningless. It is good practice, therefore, to factor the design matrix, $\boldsymbol{Z}$, into the form

$$\boldsymbol{Z} = \boldsymbol{T}\boldsymbol{S}, \tag{8.1.3.8}$$

where $\boldsymbol{S}$ is a $p \times p$ diagonal matrix whose elements define some kind of 'natural' unit appropriate to the parameter represented in each column of $\boldsymbol{Z}$. The ideal natural unit would be the standard uncertainty of that parameter, but this is not available until after the calculation has been completed. If correlation is not too severe, suitable values for the elements of $\boldsymbol{S}$, of the same order of magnitude as those derived from the standard uncertainty, are the column Euclidean norms, that is

$$S = \|\mathbf{z}_j\| = (\mathbf{z}_j^T\mathbf{z}_j)^{1/2}, \tag{8.1.3.9}$$

where $\mathbf{z}_j$ denotes the jth column of $\boldsymbol{Z}$. This scaling causes all diagonal elements of $\boldsymbol{Z}^T\boldsymbol{Z}$ to be equal to one, and errors in the elements of $\boldsymbol{Z}$ will have roughly equal effects.

Ill conditioning that results from correlation, as in the second example above, is more difficult to deal with. It is an indication that some linear combination of parameters, some eigenvector of the normal equations matrix, is poorly determined by the available data. Use of the QR factorization of $\boldsymbol{Z}$ to compute the Cholesky factor of $\boldsymbol{Z}^T\boldsymbol{Z}$ may be advantageous, in spite of the additional computation time, because better numerical stability is obtained in marginal situations. As a practical matter, however, it is important to recognize that an ill conditioned matrix is a symptom of a flaw in the model or in the experimental design (or both). Use can be made of the fact that, although determining the entire set of eigenvalues and eigenvectors of a large matrix is computationally an inherently difficult problem, a relatively simple algorithm, known as a *condition estimator* (Anderson *et al.*, 1992), can produce a good approximation to the eigenvector that corresponds to the smallest eigenvalue of a nearly singular matrix. This information can be used in either or both of two ways. First, without any fundamental modification to the model or the experiment, a simple, linear transformation of the parameters so that the problem eigenvector is one of the independent parameters, followed by rescaling, can resolve the numerical difficulties in computing the estimates. A common example is the situation where a phase transition results in the doubling of a unit cell, with pairs of atoms almost but not quite related by a lattice translation. A transformation that makes the estimated parameters the sums and differences of corresponding parameters in related pairs of atoms can make a dramatic improvement in the condition number. Alternatively, the problem eigenvector can be set to some value determined from theory or from some other experiment (see Section 8.3.1), or additional data can be collected that are selected to make that combination of parameters determinate.

8.1.4. Methods for nonlinear least squares

Recall (equation 8.1.2.1) that the general, nonlinear problem can be stated in the form: find the minimum of

$$S(\mathbf{x}) = \sum_{i=1}^{n} w_i[y_i - M_i(\mathbf{x})]^2, \tag{8.1.4.1}$$

where $\mathbf{x}$ is a vector of p parameters, and $\boldsymbol{M}(\mathbf{x})$ represents a set of model functions that predict the values of observations, $\mathbf{y}$. In this section, we discuss two useful ways of solving this problem and consider the relative merits of each. The first is based on iteratively linearizing the functions $M_i(\mathbf{x})$ and approximating