International Tables for Crystallography (2006). Vol. C, Section 8.1.1, pp. 678–680.

8.1. Least squares

BY E. PRINCE AND P. T. BOGGS

The process of arriving at a model for a crystal structure may usefully be considered to consist of two distinct stages. The first, which may be called determination, involves the use of chemical and physical intuition, direct methods, Fourier and Patterson methods, and other techniques to arrive at an approximate model for the structure that incorporates unit-cell dimensions, space group, chemical composition, and information with respect to the immediate environment of each atom. The second stage, which we shall call refinement, involves finding the values of adjustable parameters in the model that give the best fit between the predicted diffraction intensities and those observed in an experiment, in order to extract precise information about interatomic distances and bond angles, thermal motion, site occupancies, electron distribution, and so forth. Although there are several different criteria for the best fit to data, such as maximum likelihood and maximum entropy, one of the most commonly used is the method of least squares. This chapter discusses both numerical and statistical aspects of refinement by the method of least squares. Because both aspects make extensive use of linear algebra, we begin with a summary of definitions and fundamental operations in linear algebra (Stewart, 1973; Prince, 1994) and of basic definitions and concepts in mathematical statistics (Draper & Smith, 1981; Box, Hunter & Hunter, 1978). We then discuss the principles of linear and nonlinear least squares and conclude with an extensive discussion of numerical methods used in practical implementation of the technique.

8.1.1. Definitions

8.1.1.1. *Linear algebra*

A *matrix* is an ordered, rectangular array of numbers, real or complex. Matrices will be denoted by upper-case, bold italic letters, $\boldsymbol{A}$. Their individual elements will be denoted by upper-case, italic letters with subscripts. A_{ij} denotes the element in the ith row and the jth column of $\boldsymbol{A}$. A matrix with only one row is a *row vector*; a matrix with only one column is a *column vector*. Vectors will be denoted by lower-case, bold roman letters, and their elements will be denoted by lower-case, italic letters with single subscripts. Scalar constants will usually be denoted by lower-case, Greek letters.

A matrix with the same number of rows as columns is *square*. If $A_{ij} = 0$ for all $i > j$, $\boldsymbol{A}$ is *upper triangular*. If $A_{ij} = 0$ for all $i < j$, $\boldsymbol{A}$ is *lower triangular*. If $A_{ij} = 0$ for all $i \neq j$, $\boldsymbol{A}$ is *diagonal*. If $A_{ij} = 0$ for all i and j, $\boldsymbol{A}$ is *null*. A matrix, $\boldsymbol{B}$, such that $B_{ij} = A_{ji}$ for all i and j is the *transpose* of $\boldsymbol{A}$, and is denoted by $\mathbf{A}^T$. Matrices with the same dimensions may be added and subtracted: $(\boldsymbol{A}+\boldsymbol{B})_{ij} = A_{ij} + B_{ij}$. A matrix may be multiplied by a scalar: $(\alpha\boldsymbol{A})_{ij} = \alpha A_{ij}$. Multiplication of matrices is defined by $(\boldsymbol{AB})_{ij} = \sum_{k=1}^{m} A_{ik}B_{kj}$, where m is the number of columns of $\boldsymbol{A}$ and the number of rows of $\boldsymbol{B}$ (which must be equal). Addition and multiplication of matrices obey the associative law: $(\boldsymbol{A}+\boldsymbol{B})+\boldsymbol{C} = \boldsymbol{A}+(\boldsymbol{B}+\boldsymbol{C})$; $(\boldsymbol{AB})\boldsymbol{C} = \boldsymbol{A}(\boldsymbol{BC})$. Multiplication of matrices obeys the distributive law: $\boldsymbol{A}(\boldsymbol{B}+\boldsymbol{C}) = \boldsymbol{AB}+\boldsymbol{AC}$. Addition of matrices obeys the commutative law: $\boldsymbol{A}+\boldsymbol{B} = \boldsymbol{B}+\boldsymbol{A}$, but multiplication, except in certain (important) special cases, does not: $\boldsymbol{AB} \neq \boldsymbol{BA}$. The transpose of a product is the product of the transposes of the factors in reverse order: $(\boldsymbol{AB})^T = \boldsymbol{B}^T\boldsymbol{A}^T$.

The *trace* of a square matrix is the sum of its diagonal elements. The *determinant* of an $n \times n$ square matrix, $\boldsymbol{A}$, denoted by $|\boldsymbol{A}|$, is the sum of $n!$ terms, each of which is a product of the diagonal elements of a matrix derived from $\boldsymbol{A}$ by permuting columns or rows (see Stewart, 1973). The *rank* of a matrix (not necessarily square) is the dimension of the largest square submatrix that can be formed from it, by selecting rows and columns, whose determinant is not equal to zero. A matrix has *full column rank* if its rank is equal to its number of columns. A square matrix whose diagonal elements are equal to one and whose off-diagonal elements are equal to zero is an *identity matrix*, denoted by $\boldsymbol{I}$. If $|\boldsymbol{A}| \neq 0$, $\boldsymbol{A}$ is *nonsingular*, and there exists a matrix $\boldsymbol{A}^{-1}$, the *inverse* of $\boldsymbol{A}$, such that $\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}$. If $|\boldsymbol{A}| = 0$, $\boldsymbol{A}$ is *singular*, and has no inverse. The *adjoint*, or *conjugate transpose*, of $\boldsymbol{A}$ is a matrix, $A^\dagger$, such that $A^\dagger_{ij} = A^*_{ji}$, where the asterisk indicates complex conjugate. If $\boldsymbol{A}^\dagger = \boldsymbol{A}^{-1}$, $\boldsymbol{A}$ is *unitary*. If the elements of a unitary matrix are real, it is *orthogonal*. From this definition, if $\boldsymbol{A}$ is orthogonal, it follows that

$$\sum_{i=1}^{n} A_{ij}^2 = 1$$

for all j, and

$$\sum_{i=1}^{n} A_{ij}A_{ik} = 0$$

if $j \neq k$. By analogy, two column vectors, $\mathbf{x}$ and $\mathbf{y}$, are said to be orthogonal if $\mathbf{x}^T\mathbf{y} = 0$.

For any square matrix, $\boldsymbol{A}$, there exists a set of vectors, $\mathbf{x}_i$, such that $\boldsymbol{A}\mathbf{x}_i = \lambda_i\mathbf{x}_i$, where λ_i is a scalar. The values λ_i are the *eigenvalues* of $\boldsymbol{A}$, and the vectors $\mathbf{x}_i$ are the corresponding *eigenvectors*. If $\boldsymbol{A} = \boldsymbol{A}^\dagger$, $\boldsymbol{A}$ is *Hermitian*, and, if the elements are real, $\boldsymbol{A} = \boldsymbol{A}^T$, so that $\boldsymbol{A}$ is *symmetric*. It can be shown (see, for example, Stewart, 1973) that, if $\boldsymbol{A}$ is Hermitian, all eigenvalues are real, and there exists a unitary matrix, $\boldsymbol{T}$, such that $\boldsymbol{D} = \boldsymbol{T}^\dagger\boldsymbol{A}\boldsymbol{T}$ is diagonal, with the elements of $\boldsymbol{D}$ equal to the eigenvalues of $\boldsymbol{A}$, and the columns of $\boldsymbol{T}$ are the eigenvectors. An $n \times n$ symmetric matrix therefore has n mutually orthogonal eigenvectors. If the product $\mathbf{x}^T\boldsymbol{A}\mathbf{x}$ is greater than (or equal to) zero for any non-null vector, $\mathbf{x}$, $\boldsymbol{A}$ is *positive (semi)definite*. Because $\mathbf{x}$ may be, in particular, an eigenvector, all eigenvalues of a positive (semi)definite matrix are greater than (or equal to) zero. Any matrix of the form $\boldsymbol{B}^T\boldsymbol{B}$ is positive semi definite, and, if $\boldsymbol{B}$ has full column rank, $\boldsymbol{A} = \boldsymbol{B}^T\boldsymbol{B}$ is positive definite. If $\boldsymbol{A}$ is positive definite, there exists an upper triangular matrix, $\boldsymbol{R}$, or, equivalently, a lower triangular matrix, $\boldsymbol{L}$, with positive diagonal elements, such that $\boldsymbol{R}^T\boldsymbol{R} = \boldsymbol{L}\boldsymbol{L}^T = \boldsymbol{A}$. $\boldsymbol{R}$, or $\boldsymbol{L}$, is called the *Cholesky factor* of $\boldsymbol{A}$. The *magnitude*, *length* or *Euclidean norm* of a vector, $\mathbf{x}$, denoted by $\|\mathbf{x}\|$, is defined by $\|\mathbf{x}\| = (\mathbf{x}^T\mathbf{x})^{1/2}$. The *induced matrix norm* of a matrix, $\boldsymbol{B}$, denoted $\|\boldsymbol{B}\|$, is defined as the maximum value of $\|\boldsymbol{B}\mathbf{x}\|/\|\mathbf{x}\| = (\mathbf{x}^T\boldsymbol{B}^T\boldsymbol{B}\mathbf{x}/\mathbf{x}^T\mathbf{x})^{1/2}$ for $\|\mathbf{x}\| > 0$. Because $\mathbf{x}^T\boldsymbol{B}^T\boldsymbol{B}\mathbf{x}$ will have its maximum value for a fixed value of $\mathbf{x}^T\mathbf{x}$ when $\mathbf{x}$ is parallel to the eigenvector that corresponds to the largest eigenvalue of $\boldsymbol{B}^T\boldsymbol{B}$, this definition implies that $\|\boldsymbol{B}\|$ is equal to the square root of the largest eigenvalue of $\boldsymbol{B}^T\boldsymbol{B}$. The *condition number* of $\boldsymbol{B}$ is the square root of the ratio of the largest and smallest eigenvalues of $\boldsymbol{B}^T\boldsymbol{B}$. (Other definitions of norms exist, with corresponding definitions of condition number. We shall not be concerned with any of these.)

We shall make extensive use of the so-called QR decomposition, which is defined as follows: For any $n \times p$ ($n \geq p$) real matrix, $\boldsymbol{Z}$, there exists an $n \times n$ orthogonal matrix, $\boldsymbol{Q}$, such that

$$\boldsymbol{Q}^T \boldsymbol{Z} = \begin{pmatrix} \boldsymbol{R} \\ \boldsymbol{O} \end{pmatrix}, \tag{8.1.1.1}$$

where $\boldsymbol{R}$ is a $p \times p$ upper triangular matrix, and $\boldsymbol{O}$ denotes an $(n-p) \times p$ null matrix. Thus, we have

$$\boldsymbol{Z} = \boldsymbol{Q} \begin{pmatrix} \boldsymbol{R} \\ \boldsymbol{O} \end{pmatrix}, \tag{8.1.1.2}$$

which is known as the QR decomposition of $\boldsymbol{Z}$. If we partition $\boldsymbol{Q}$ as $(\boldsymbol{Q}_Z, \boldsymbol{Q}_\perp)$, where $\boldsymbol{Q}_Z$ has dimensions $n \times p$, and $\boldsymbol{Q}_\perp$ has dimensions $n \times (n-p)$, (8.1.1.2) becomes

$$\boldsymbol{Z} = \boldsymbol{Q}_Z \boldsymbol{R}, \tag{8.1.1.3}$$

which is known as the QR factorization. We shall make use of the following facts. First, $\boldsymbol{R}$ is nonsingular if and only if the columns of $\boldsymbol{Z}$ are linearly independent; second, the columns of $\boldsymbol{Q}_Z$ form an orthonormal basis for the range space of $\boldsymbol{Z}$, that is, they span the same space as $\boldsymbol{Z}$; and, third, the columns of $\boldsymbol{Q}_\perp$ form an orthonormal basis for the null space of $\boldsymbol{Z}^T$, that is, $\boldsymbol{Z}^T \boldsymbol{Q}_\perp = \boldsymbol{O}$.

There are two common procedures for computing the QR factorization. The first makes use of *Householder transformations*, which are defined by

$$\boldsymbol{H} = \boldsymbol{I} - 2\mathbf{x}\mathbf{x}^T, \tag{8.1.1.4}$$

where $\mathbf{x}^T\mathbf{x} = 1$. $\boldsymbol{H}$ is symmetric, and $\boldsymbol{H}^2 = \boldsymbol{I}$, so that $\boldsymbol{H}$ is orthogonal. In three dimensions, $\boldsymbol{H}$ corresponds to a reflection in a mirror plane perpendicular to $\mathbf{x}$, because of which Stewart (1973) has suggested the alternative term *elementary reflector*. $\boldsymbol{A}$ vector $\mathbf{v}$ is transformed by $\boldsymbol{H}\mathbf{v}$ into the vector $\|\mathbf{v}\|\mathbf{e}$, where $\mathbf{e}$ represents a vector with $e_1 = 1$, and $e_i = 0$ for $i \neq 1$, if

$$\mathbf{x} = [\mathbf{v} - \|\mathbf{v}\|\mathbf{e}] \big/ \big\|[\mathbf{v} - \|\mathbf{v}\|\mathbf{e}]\big\|. \tag{8.1.1.5}$$

The factorization procedure for an $n \times p$ matrix, $\boldsymbol{A}$ (Stewart, 1973; Anderson *et al.*, 1992), takes as $\mathbf{v}$ in the first step the first column of $\boldsymbol{A}$, and forms $\boldsymbol{A}_1 = \boldsymbol{H}_1\boldsymbol{A}$, which has zeros in all elements of the first column below the diagonal. In the second step, $\mathbf{v}$ has a zero as the first element and is filled out by those elements of the second column of $\boldsymbol{A}_1$ on or below the diagonal. $\boldsymbol{A}_2 = \boldsymbol{H}_2\boldsymbol{A}_1$ then has zeros in all elements below the diagonal in the first two columns. This process is repeated $(p-2)$ more times, after which $\boldsymbol{Q} = \boldsymbol{H}_p \ldots \boldsymbol{H}_2\boldsymbol{H}_1$, and $\boldsymbol{R} = \boldsymbol{A}_p$ is upper triangular.

QR factorization by Householder transformations requires for efficiency that the entire $n \times p$ matrix be stored in memory, and requires of order np^2 operations. A procedure that requires storage of only the upper triangle makes use of *Givens rotations*, which are 2×2 matrices of the form

$$\boldsymbol{G} = \begin{pmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{pmatrix}. \tag{8.1.1.6}$$

Multiplication of a $2 \times m$ matrix, $\boldsymbol{B}$, by $\boldsymbol{G}$ will put a zero in the B_{21} element if $\alpha = \arctan B_{21}/B_{11}$. The factorization of $\boldsymbol{A}$ involves reading, or computing, the rows of $\boldsymbol{A}$ one at a time. In the first step, matrix $\boldsymbol{B}_1$ consists of the first row of $\boldsymbol{R}$ and the current row of $\boldsymbol{A}$, from which the first element is eliminated. In the second step, $\boldsymbol{B}_{21}$ is the second row of $\boldsymbol{R}$ and the $(p-1)$ nonzero elements of the second row of the transformed $\boldsymbol{B}_1$. After the first p rows have been treated, each additional row of $\boldsymbol{A}$ requires $2p(p+1)$ multiplications to fill it with zeros. However, because the operation is easily vectorized, the time required may be a small proportion of the total computing time on a vector oriented computer.

8.1.1.2. *Statistics*

A *probability density function*, which will be abbreviated p.d.f., is a function, $\Phi(x)$, such that the probability of finding the random variable x in the interval $a \leq x \leq b$ is given by

$$p(a \leq x \leq b) = \int_a^b \Phi(x)\,\mathrm{d}x.$$

A p.d.f. has the properties

$$\Phi(x) \geq 0, \qquad -\infty < x < +\infty,$$

and

$$\int_{-\infty}^{+\infty} \Phi(x)\,\mathrm{d}x = 1.$$

A *cumulative distribution function*, which will be abbreviated c.d.f., is defined by

$$\Psi(x) = \int_{-\infty}^{x} \Phi(t)\,\mathrm{d}t.$$

The properties of $\Phi(x)$ imply that $0 \leq \Psi(x) \leq 1$, and $\Phi(x) = \mathrm{d}\Psi(x)/\mathrm{d}x$. The *expected value* of a function, $f(x)$, of random variable x is defined by

$$\big\langle f(x) \big\rangle = \int_{-\infty}^{+\infty} f(x)\Phi(x)\,\mathrm{d}x.$$

If $f(x) = x^n$, $\big\langle f(x) \big\rangle = \langle x^n \rangle$ is the *nth moment* of $\Phi(x)$. The first moment, often denoted by μ, is the *mean* of $\Phi(x)$. The second moment about the mean, $\big\langle (x - \langle x \rangle)^2 \big\rangle$, usually denoted by σ^2, is the *variance* of $\Phi(x)$. The positive square root of the variance is the *standard deviation*.

For a vector, $\mathbf{x}$, of random variables, $x_1, x_2, \ldots, x_n$, the *joint probability density function*, or *joint p.d.f.*, is a function, $\Phi_J(\mathbf{x})$, such that

$$\begin{aligned} &p(a_1 \leq x_1 \leq b_1;\ a_2 \leq x_2 \leq b_2; \ldots;\ a_n \leq x_n \leq b_n) \\ &\quad = \int_{a_1}^{b_1} \int_{a_2}^{b_2} \ldots \int_{a_n}^{b_n} \Phi_J(\mathbf{x})\,\mathrm{d}x_1\,\mathrm{d}x_2 \ldots \mathrm{d}x_n. \end{aligned} \tag{8.1.1.7}$$

The *marginal p.d.f.* of an element (or a subset of elements), x_i, is a function, $\Phi_M(x_i)$, such that

$$\begin{aligned} p(a_i \leq x_i \leq b_i) &= \int_{a_i}^{b_i} \Phi_M(x_i)\,\mathrm{d}x_i \\ &= \int_{-\infty}^{+\infty} \ldots \int_{a_i}^{b_i} \ldots \int_{-\infty}^{\infty} \Phi_J(\mathbf{x})\,\mathrm{d}x_1 \ldots \mathrm{d}x_i \ldots \mathrm{d}x_n. \end{aligned} \tag{8.1.1.8}$$

This is a p.d.f. for x_i alone, irrespective of the values that may be found for any other element of $\mathbf{x}$. For two random variables, x and y (either or both of which may be vectors), the *conditional p.d.f. of x given* $y = y_0$ is defined by

$$\Phi_C(x|y_0) = c\Phi_J(x, y)_{y=y_0},$$

where $c = 1/\Phi_M(y_0)$ is a *renormalizing factor*. This is a p.d.f. for x when it is known that $y = y_0$. If $\Phi_C(x|y) = \Phi_M(x)$ for all y, or, equivalently, if $\Phi_J(x, y) = \Phi_M(x)\Phi_M(y)$, the random variables x and y are said to be *statistically independent*.

Moments may be defined for multivariate p.d.f.s in a manner analogous to the one-dimensional case. The mean is a vector defined by

$$\mu_i = \langle x_i \rangle = \int x_i \Phi(\mathbf{x})\, d\mathbf{x},$$

where the volume of integration is the entire domain of $\mathbf{x}$. The *variance–covariance matrix* is defined by

$$\begin{aligned} V_{ij} &= \big\langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \big\rangle \\ &= \int (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \Phi_J(\mathbf{x})\, d\mathbf{x}. \end{aligned} \tag{8.1.1.9}$$

The diagonal elements of $\boldsymbol{V}$ are the variances of the marginal p.d.f.s of the elements of $\mathbf{x}$, that is, $V_{ii} = \sigma_i^2$. It can be shown that, if x_i and x_j are statistically independent, $V_{ij} = 0$ when $i \neq j$. If two vectors of random variables, $\mathbf{x}$ and $\mathbf{y}$, are related by a linear transformation, $\mathbf{x} = \boldsymbol{B}\mathbf{y}$, the means of their joint p.d.f.s are related by $\mu_x = B\mu_y$, and their variance–covariance matrices are related by $\boldsymbol{V}_\mathbf{x} = \boldsymbol{B}\boldsymbol{V}_\mathbf{y}\boldsymbol{B}^T$.

8.1.2. Principles of least squares

The method of least squares may be formulated as follows: Given a set of n observations, $y_i\ (i = 1, 2, \ldots, n)$, that are measurements of quantities that can be described by differentiable model functions, $M_i(\mathbf{x})$, where $\mathbf{x}$ is a vector of parameters, $x_j (j = 1, 2, \ldots, p)$, find the values of the parameters for which the sum

$$S = \sum_{i=1}^{n} w_i [y_i - M_i(\mathbf{x})]^2 \tag{8.1.2.1}$$

is minimum. Here, w_i represents a weight assigned to the ith observation. The values of the parameters that give the minimum value of S are called *estimates* of the parameters, and a function of the data that locates the minimum is an *estimator*. A necessary condition for S to be a minimum is for the gradient to vanish, which gives a set of simultaneous equations, the *normal equations*, of the form

$$\frac{\partial S}{\partial x_j} = -2 \sum_{i=1}^{n} w_i [y_i - M_i(\mathbf{x})] \frac{\partial M_i(\mathbf{x})}{\partial x_j} = 0. \tag{8.1.2.2}$$

The model functions, $M_i(\mathbf{x})$, are, in general, nonlinear, and there are no direct ways to solve these systems of equations. Iterative methods for solving them are discussed in Section 8.1.4. Much of the analysis of results, however, is based on the assumption that linear approximations to the model functions are good approximations in the vicinity of the minimum, and we shall therefore begin with a discussion of linear least squares.

To express linear relationships, it is convenient to use matrix notation. Let $\boldsymbol{M}(\mathbf{x})$ and $\mathbf{y}$ be column vectors whose ith elements are $M_i(\mathbf{x})$ and y_i. Similarly, let $\mathbf{b}$ be a vector and $\boldsymbol{A}$ be a matrix such that a linear approximation to the ith model function can be written

$$M_i(\mathbf{x}) = b_i + \sum_{j=1}^{p} A_{ij} x_j. \tag{8.1.2.3}$$

Equations (8.1.2.3) can be written, in matrix form,

$$\boldsymbol{M}(\mathbf{x}) = \mathbf{b} + \boldsymbol{A}\mathbf{x}, \tag{8.1.2.4}$$

and, for this linear model, (8.1.2.1) becomes

$$S = [(\mathbf{y} - \mathbf{b}) - \boldsymbol{A}\mathbf{x}]^T \boldsymbol{W} [(\mathbf{y} - \mathbf{b}) - \boldsymbol{A}\mathbf{x}], \tag{8.1.2.5}$$

where $\boldsymbol{W}$ is a diagonal matrix whose diagonal elements are $W_{ii} = w_i$. In this notation, the normal equations (8.1.2.2) can be written

$$\boldsymbol{A}^T \boldsymbol{W} \boldsymbol{A}\mathbf{x} = \boldsymbol{A}^T \boldsymbol{W} (\mathbf{y} - \mathbf{b}), \tag{8.1.2.6}$$

and their solution is

$$\widehat{\mathbf{x}} = (\boldsymbol{A}^T \boldsymbol{W} \boldsymbol{A})^{-1} \boldsymbol{A}^T \boldsymbol{W} (\mathbf{y} - \mathbf{b}). \tag{8.1.2.7}$$

If $W_{ii} > 0$ for all i, and $\boldsymbol{A}$ has full column rank, then $\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{A}$ will be positive definite, and S will have a unique minimum at $\mathbf{x} = \widehat{\mathbf{x}}$. The matrix $\boldsymbol{H} = (\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{W}$ is a $p \times n$ matrix that relates the n-dimensional observation space to the p-dimensional parameter space and is known as the *least-squares estimator*; because each element of $\widehat{\mathbf{x}}$ is a linear function of the observations, it is a *linear estimator*. [Note that, in actual practice, the matrix $\boldsymbol{H}$ is not actually evaluated, except, possibly, in very small problems. Rather, the linear system $\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{A}\mathbf{x} = \boldsymbol{A}^T\boldsymbol{W}(\mathbf{y} - \mathbf{b})$ is solved using the methods of Section 8.1.3.]

The least-squares estimator has some special properties in statistical analysis. Suppose that the elements of $\mathbf{y}$ are experimental observations drawn at random from populations whose means are given by the model, $\boldsymbol{M}(\mathbf{x})$, for some unknown $\mathbf{x}$, which we wish to estimate. This may be written

$$\big\langle \mathbf{y} - \mathbf{b} \big\rangle = \boldsymbol{A}\mathbf{x}. \tag{8.1.2.8}$$

The expected value of the least-squares estimate is

$$\begin{aligned} \big\langle \widehat{\mathbf{x}} \big\rangle &= \big\langle (\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{W}(\mathbf{y} - \mathbf{b}) \big\rangle \\ &= (\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{W}\big\langle \mathbf{y} - \mathbf{b} \big\rangle \\ &= (\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{A}\mathbf{x} \\ &= \mathbf{x}. \end{aligned} \tag{8.1.2.9}$$

If the expected value of an estimate is equal to the variable to be estimated, the estimator is said to be *unbiased*. Equation (8.1.2.9) shows that the least-squares estimator is an unbiased estimator for $\mathbf{x}$, independent of $\boldsymbol{W}$, provided only that $\mathbf{y}$ is an unbiased estimate of $\boldsymbol{M}(\mathbf{x})$, the matrix $\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{A}$ is nonsingular, and the elements of $\boldsymbol{W}$ are constants independent of $\mathbf{y}$ and $\boldsymbol{M}(\mathbf{x})$. Let $\boldsymbol{V}_\mathbf{x}$ and $\boldsymbol{V}_\mathbf{y}$ be the variance–covariance matrices for the joint p.d.f.s of the elements of $\mathbf{x}$ and $\mathbf{y}$, respectively. Then, $\boldsymbol{V}_\mathbf{x} = \boldsymbol{H}\boldsymbol{V}_\mathbf{y}\boldsymbol{H}^T$. Let $\boldsymbol{G}$ be the matrix $(\boldsymbol{A}^T\boldsymbol{V}_\mathbf{y}^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{V}^{-1}$, so that $\widehat{\mathbf{x}} = \boldsymbol{G}(\mathbf{y} - \mathbf{b})$ is the particular least-squares estimate for which $\boldsymbol{W} = \boldsymbol{V}_\mathbf{y}^{-1}$. Then, $\boldsymbol{V}_\mathbf{x} = \boldsymbol{G}\boldsymbol{V}_\mathbf{y}\boldsymbol{G}^T$. If $\boldsymbol{V}_\mathbf{y}$ is positive definite, its lower triangular Cholesky factor, $\boldsymbol{L}$, exists, so that $\boldsymbol{L}\boldsymbol{L}^T = \boldsymbol{V}_\mathbf{y}$. [If $\boldsymbol{V}$ is diagonal, $\boldsymbol{L}$ is also diagonal, with $L_{ii} = (\boldsymbol{V}_\mathbf{y})_{ii}^{1/2}$.] It is readily verified that the matrix product $[(\boldsymbol{H} - \boldsymbol{G})\boldsymbol{L}][(\boldsymbol{H} - \boldsymbol{G})\boldsymbol{L}]^T = \boldsymbol{H}\boldsymbol{V}_\mathbf{y}\boldsymbol{H}^T - \boldsymbol{G}\boldsymbol{V}_\mathbf{y}\boldsymbol{G}^T$, but the diagonal elements of this product are the sums of squares of the elements of rows of $(\boldsymbol{H} - \boldsymbol{G})\boldsymbol{L}$, and are therefore greater than or equal to zero. Therefore, the diagonal elements of $\boldsymbol{V}_\mathbf{x}$, which are the variances of the marginal p.d.f.s of the elements of $\widehat{\mathbf{x}}$, are minimum when $\boldsymbol{W} = \boldsymbol{V}_\mathbf{y}^{-1}$.

Thus, the least-squares estimator is unbiased for any positive-definite weight matrix, $\boldsymbol{W}$, but the variances of the elements of the vector of estimated parameters are minimized if $\boldsymbol{W} = \boldsymbol{V}_\mathbf{y}^{-1}$. [Note also that $\boldsymbol{V}_\mathbf{x} = (\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{A})^{-1}$ if, *and only if*, $\boldsymbol{W} = \boldsymbol{V}_\mathbf{y}^{-1}$.] For this reason, the least-squares estimator with weights proportional to the reciprocals of the variances of the observations is referred to as the *best linear unbiased estimator* for the parameters of a model describing those observations. (These specific results are included in a more general result known as the *Gauss–Markov theorem*.)

The analysis up to this point has assumed that the model is linear, that is that the expected values of the observations can be expressed by $\langle \mathbf{y} \rangle = \mathbf{b} + \boldsymbol{A}\mathbf{x}$, where $\boldsymbol{A}$ is some matrix. In crystallography, of course, the model is highly nonlinear, and this assumption is not valid. The principles of linear least squares