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

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

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

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

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

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

The expected value of the least-squares estimate is

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

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

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: