International Tables for Crystallography (2006). Vol. C, Section 8.5.2, pp. 707–708.

8.5. Detection and treatment of systematic error

By E. Prince and C. H. Spiegelman

8.5.1. Accuracy

Chapter 8.4 discusses statistical tests for goodness of fit between experimental observations and the predictions of a model with adjustable parameters whose values have been estimated by least squares or some similar procedure. In addition to the estimates of parameter values, one can also make estimates of the uncertainties in those values, estimates that are usually expressed in terms of an *estimated standard deviation* or, according to recommended usage (ISO, 1993), a *standard uncertainty*. A standard deviation is a measure of *precision*, that is, a measure of the width of a confidence interval that results from random fluctuations in the measurement process. What the experimenter who collected the data wants to know about, of course, is *accuracy*, a measure of the location of a region within which nature's 'correct' value lies, as well as its width (Prince, 1994). In performing a refinement, we have assumed implicitly that the observations have been drawn at random from a population the mean of whose p.d.f. is given by a model when all of its parameters have those unknown, correct values. If this assumption is incorrect, the expected value of the estimate may no longer be near to the correct value, and the estimate contains *bias*, or *systematic error*. An accurate measurement is one that not only is precise but also has small bias. In this chapter, we shall discuss various criteria by which the results of a refinement may be judged in order to determine whether they are free of systematic error, and thus whether they may be considered accurate.

8.5.2. Lack of fit

We saw in Section 8.4.2 that the sum of squared residuals from an ideally weighted, least-squares fit to a correct model is a sum of terms that has expected value $(n-p)$ and is distributed as χ^2 with $\nu = n-p$ degrees of freedom. Further, the residuals have a distribution with zero mean. A value for the sum that exceeds $(n-p)$ by an amount that is improbably large is an indication of lack of fit, which may be due to an incorrect model for the mean or to nonideal weighting or both. {The sum, S, may be considered to be improbably large when the value of the χ^2 cumulative distribution function, $\Psi_{\chi^2}[S,(n-p)]$, is close to 1.0. A value for the sum that is substantially *less* than $(n-p)$ may also be an indication that the model contains more parameters than can be justified by the data set. Note also that a reasonable value for the sum of squared residuals does *not* prove that the model is correct. It indicates that the model adequately describes the data, but it in no way rules out the existence of alternative models that describe the data equally well.} If the sum of squares is greater than $(n-p)$, it is commonly assumed that the mean model is correct, and that the weights have appropriate relative values, although their absolute values may be too large. If $w = k^2/u_i^2$, where k is some number greater than one, and u_i is the standard uncertainty of the ith observation, the goodness-of-fit parameter,

$$G = \left\{\sum_{i=1}^{n} w_i[y_i - M_i(\widehat{\mathbf{x}})]^2/(n-p)\right\}^{1/2}, \qquad (8.5.2.1)$$

is taken to be an estimate of k, and all elements of the inverse of the normal-equations matrix are multiplied by k^2 to obtain the estimated variance–covariance matrix

$$\widehat{\mathbf{V}}_{\mathbf{x}} = G^2(\mathbf{A}^T\mathbf{W}\mathbf{A})^{-1}. \qquad (8.5.2.2)$$

Frequently, however, there is some other, independent estimate of the variance of the observation, σ_i^2, derived, for example, from counting statistics or from the observed scatter among symmetry-equivalent reflections. If this estimate is inconsistent with the hypothesis that all data points have been overweighted by a constant factor, then the assumption that the parameter estimates are unbiased but less precise than the original weights would indicate must be discarded. Instead, it must be assumed that the model is incorrect, or at least incomplete. A systematic error may be considered to cause the model to be incomplete, and may introduce bias into some or all of the refined parameters. (Note that in many standard statistical texts it is implicitly assumed, without so stating, that the data have already been scaled by a set of correct, relative weights. It is thus easy for the unwary reader to make the error of assuming that the practice of multiplying by the goodness-of-fit parameter is a well established procedure.)

The use of (8.5.2.2) to compute estimated variances and standard uncertainties assumes implicitly that the effect of lack of fit on parameter estimates is random, and applies equally to all parameters, even though different types of parameter may have very different mathematical relations in the model. With a model as complex as the crystallographic structure-factor formula, this assumption is certainly questionable.

Information about the nature of the model inadequacies can be obtained by examining the residuals (Belsley, Kuh & Welsch, 1980; Belsley, 1991). The standardized residuals, $R_i = [y_i - M_i(\widehat{\mathbf{x}})]/u_i'$, where $\widehat{\mathbf{x}}$ is the least-squares estimate of the parameters, should be randomly distributed, with zero mean, not only for the data set as a whole but also for subsets of the data that are chosen in a manner that depends only on the model and not on the observed values of the data. Here, u_i' is the standard uncertainty of the *residual* and is related to u_i, the standard uncertainty of the observation, by $u_i' = u_i(1-P_{ii})$, where P_{ii} is a diagonal element of the projection matrix (Section 8.4.4). A scatter plot, in which the residuals are plotted against some control variable, such as $|F_{\rm calc}|$, $\sin\theta/\lambda$, or one of the Miller indices, should reveal no general trends. The existence of any such trend may indicate a systematic effect that depends on the corresponding variable. The model may then be modified by inclusion of a factor that is proportional to that variable, and the refinement repeated. An examination of the shifts in the other parameters, and of the new row or column of the variance–covariance matrix, will then reveal which of the parameters in the unmodified model are likely to have been biased by the systematic effect. When this procedure has been followed, it is extremely important to consider carefully the nature of the additional effect and determine whether it is plausible in terms of physics and chemistry.

Another procedure for detecting systematic lack of fit makes use of the fact that, if the model is correct, and the error distribution is approximately normal, or Gaussian, the distribution of residuals will also be approximately normal. A large sample may be checked for normality by means of a quantile–quantile, or *Q–Q*, plot (Abrahams & Keve, 1971; Kafadar & Spiegelman, 1986). To make such a plot, the residuals are first sorted in ascending order of magnitude. If there are n points in the data set, the value of the ith sorted residual should be close to the value, x_i, for which

$$\Psi(x_i) = (2i - 1)/2n, \qquad (8.5.2.3)$$

where $\Psi(x)$ is the cumulative distribution function for the normal p.d.f. A plot of R_i against x_i should be a straight line with zero intercept and unit slope. A straight line with a slope greater than one suggests that the model is satisfactory, but that the variances of the data points have been systematically underestimated. Lack of fit is suggested if the curve has a higher slope near the ends, indicating that large residuals occur with greater frequency than would be predicted by the normal p.d.f.

The sorted residuals tend to be strongly correlated. A positive displacement from a smooth curve tends to be followed by another positive displacement, and a negative one by another negative one, which gives the *Q–Q* plot a wavy appearance, and it may be difficult to decide whether it is a straight line or not. Because of this, a useful alternative to the *Q–Q* plot is the conditional *Q–Q* plot (Kafadar & Spiegelman, 1986), so called because the abscissa for plotting the *i*th sorted residual is the mean of a conditional p.d.f. for that residual given the observed values of all the others. To construct a conditional *Q–Q* plot, first transform the distribution to a uniform p.d.f. by

$$U_i = \Psi(R_i, \mu, \sigma), \qquad (8.5.2.4)$$

where μ and σ are resistant estimates (Section 8.2.2) of the mean and standard deviation of the p.d.f., such as the median and 0.75 times the interquartile range, and Ψ represents the cumulative distribution function. Letting $U_0 = 0$ and $U_{n+1} = 1$, the expected value of U_i, given all the others, is

$$\langle U_i \rangle = (U_{i-1} + U_{i+1})/2. \qquad (8.5.2.5)$$

The *i*th abscissa for the *Q–Q* plot is then

$$x_i = \Psi^{-1}(\langle U_i \rangle, \mu, \sigma), \qquad (8.5.2.6)$$

where $\Psi^{-1}(y, \mu, \sigma)$ is a *per cent point function*, or p.p.f., the value of x for which $\Psi(x, \mu, \sigma) = y$.

Q–Q plots for subsets of the data can reveal, by nonzero intercepts, that those subsets are subject to a systematic bias. Because of its property of removing short-range kinks in the curve, the conditional *Q–Q* plot can be particularly useful in this application. The values of μ and σ used for the transformation to a uniform distribution, as in (8.5.2.4), should be those determined from the entire data set.

A *Q–Q* plot will reveal data points that are in poor agreement with the model, but that do not belong to any easily identifiable subset. Because of the central limit theorem (Section 8.4.1), however, the least-squares method tends to force the distribution of the residuals toward a normal distribution, and the discrepant points may not be clearly evident. A robust/resistant procedure (see Section 8.2.2), because it reduces the influence of strongly discrepant data points, helps to separate them from the body of the data. Therefore, if a data set contains discrepant points, a *Q–Q* plot of the residuals from a robust/resistant fit will tend to have greater curvature at the extremes than one from a corresponding least-squares fit. If the discrepant data points that are thus identified have a pattern, this information may enable a systematic error to be characterized.

8.5.3. Influential data points

Section 8.4.4 discusses the influence of individual data points on the estimation of parameters and how to identify the data points that should be measured with particular care in order to make the most precise estimates of particular parameters. The same properties that cause these influential data points to be most effective in reducing the uncertainty of a parameter estimate when the model is a correct predictor for the observations also cause them to have the greatest potential for introducing bias if there is a flaw in the model or, correspondingly, if they are subject to systematic error. Reviews of procedures for studying the effects of influential data points and outliers have been given by Beckman & Cook (1983), by Chatterjee & Hadi (1986), and by Belsley (1991).

The effects of possible systematic error can be studied by identifying influential data points and then observing the effects of deleting them one by one from the refinement. The deletion of a data point should affect the standard uncertainty of an estimate, but should not cause a shift in its mean that is more than a small multiple of the resulting standard uncertainty. As in Section 8.4.4, we define the design matrix, $\boldsymbol{A}$, by

$$A_{ij} = \partial M_i(\mathbf{x})/\partial x_j, \qquad (8.5.3.1)$$

where $M_i(\mathbf{x})$ is the model function for the *i*th data point, and $\mathbf{x}$ is a vector of adjustable parameters. Let $\boldsymbol{R}$ be the upper triangular Cholesky factor of the weight matrix, so that $\boldsymbol{W} = \boldsymbol{R}^T\boldsymbol{R}$, and define the weighted design matrix by $\boldsymbol{Z} = \boldsymbol{RA}$ and the weighted vector of observations by $\mathbf{y}' = \boldsymbol{R}\mathbf{y}$. The least-squares estimate of $\mathbf{x}$ is then

$$\widehat{\mathbf{x}} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{y}', \qquad (8.5.3.2)$$

and the vector of predicted values is

$$\widehat{\mathbf{y}}' = \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{y}' = \boldsymbol{P}\mathbf{y}', \qquad (8.5.3.3)$$

where $\boldsymbol{P}$ is the projection, or hat, matrix. A diagonal element, P_{ii}, of $\boldsymbol{P}$ is a measure of the leverage, that is of the relative influence, of the *i*th data point, and therefore of the sensitivity of the estimates of the elements of $\mathbf{x}$ to an error in the measurement of that data point. P_{ii} lies in the range $0 \le P_{ii} \le 1$, and it has average value p/n, so that data points with values of P_{ii} greater than $2p/n$ can be considered particularly influential.

Let $\boldsymbol{H} = \boldsymbol{Z}^T\boldsymbol{Z}$ be the normal-equations matrix, let $\boldsymbol{V} = \boldsymbol{H}^{-1}$ be the estimated variance–covariance matrix, and let $\mathbf{q} = \boldsymbol{Z}^T\mathbf{y}'$, so that $\widehat{\mathbf{x}} = \boldsymbol{V}\mathbf{q}$. Let $\mathbf{z}_i$ be the *i*th row of $\mathbf{Z}$, and denote by $\boldsymbol{Z}^{(i)}$, $\boldsymbol{H}^{(i)}$, $\boldsymbol{V}^{(i)}$, $\mathbf{q}^{(i)}$, and $\widehat{\mathbf{x}}^{(i)}$ the respective matrices and vectors computed with the *i*th data point deleted from the data set. We wish to find large values of $|\widehat{x}_j - \widehat{x}_j^{(i)}|/[V_{jj}^{(i)}]^{1/2}$, so we need to compute $\boldsymbol{V}^{(i)}$ and $\mathbf{x}^{(i)}$. With a derivation similar to that for (8.4.4.7), it can be shown (Fedorov, 1972; Prince & Nicholson, 1985) that

$$\boldsymbol{V}^{(i)} = \boldsymbol{V} + \frac{\boldsymbol{V}\mathbf{z}_i^T\mathbf{z}_i\boldsymbol{V}}{(1 - \mathbf{z}_i\boldsymbol{V}\mathbf{z}_i^T)} = \boldsymbol{V} + \frac{\boldsymbol{V}\mathbf{z}_i^T\mathbf{z}_i\boldsymbol{V}}{(1 - P_{ii})}. \qquad (8.5.3.4)$$

Note that, if $P_{ii} = 1$, all elements of $\boldsymbol{V}^{(i)}$ become infinite, implying that $\boldsymbol{H}^{(i)}$ is singular. Thus, if such a data point is deleted, the solution is no longer determinate. Now,

$$\widehat{\mathbf{x}}^{(i)} = \boldsymbol{V}^{(i)}\mathbf{q}^{(i)} \qquad (8.5.3.5)$$

and

$$\mathbf{q}^{(i)} = \mathbf{q} - y_i'\mathbf{z}_i^T, \qquad (8.5.3.6)$$

so that, when $\boldsymbol{V}$ and $\widehat{\mathbf{x}}$ have been computed once, it is a straightforward and inexpensive additional computation to determine whether any parameter has been strongly influenced, and therefore potentially biased, by the inclusion of any data point in the refinement. If there is any reason to be concerned about possible systematic error, the leverage of every data point included in the refinement should be computed, and the effects of deletion of all of those with leverage greater than $2p/n$ should be observed.

references