Unit cell refinement from powder diffraction data: the use of regression diagnostics

T. J. B. HOLLAND AND S. A. T. REDFERN

Deptartment of Earth Sciences, University of Cambridge, Downing Street, Cambridge, CB2 3EQ, UK

Abstract

We discuss the use of regression diagnostics combined with nonlinear least-squares to refine cell parameters from powder diffraction data, presenting a method which minimizes residuals in the experimentally-determined quantity (usually $2\theta_{hkl}$ or energy, E_{hkl}). Regression diagnostics, particularly deletion diagnostics, are invaluable in detection of outliers and influential data which could be deleterious to the regressed results. The usual practice of simple inspection of calculated residuals alone often fails to detect the seriously deleterious outliers in a dataset, because bare residuals provide no information on the leverage (sensitivity) of the datum concerned. The regression diagnostics which predict the change expected in each cell constant upon deletion of each observation (*hkl* reflection) are particularly valuable in assessing the sensitivity of the calculated results to individual reflections. A new computer program, implementing nonlinear regression methods and providing the diagnostic output, is described.

KEYWORDS: powder diffraction, regression diagnostics, lattice parameters, computer program.

Introduction

THE determination of the lattice (or cell) parameters of crystalline materials from powder diffraction data is a very common task in mineralogical and petrological research. Bearing in mind the prevalent nature of this task, it is somewhat surprising to discover that it is very often carried out using a method that could easily be improved upon. The approach that is commonly employed follows that first adopted by Cohen (1935) to refine cell parameters from diffraction data by iterative least-squares refinement of trial cell parameters, using the minimization of the sums of squares of residuals in $Q = d_{hkl}^{-2}$). This is largely a matter of convenience, because the most compact and elegant expression for the dependence of the spacing of the (*hkl*) lattice planes, d_{hkl}, in terms of the unknown cell parameters is given by

$$Q_{hkl} = d_{hkl}^{-2} = h^2a^{*2} + k^2b^{*2} + l^2c^{*2} + 2klb^*c^*\cos\alpha^* + 2lhc^*a^*\cos\beta^* + 2hka^*b^*\cos\gamma^* \quad (1)$$

The values of the reciprocal constants (a^*, b^*, c^*, α^*, β^*, and γ^*) are usually found by fitting the expression above to values of Q_{hkl} (found from measurements of $2\theta_{hkl}$) by a non-linear least-squares procedure. The real space unit cell parameters are then determined from these reciprocal constants with their uncertainties calculated by error propagation.

It is surprising that iterative non-linear refinement is the most common method used for cell parameter determination from powder diffraction data, given that the equation above is actually linear in six parameters which may be readily determined by the much simpler method of linear least-squares. This fact was noted and discussed by Kelsey (1964) who outlined the method of error propagation for the expression for Q_{hkl} recast as

$$Q_{hkl} = h^2x_1 + k^2x_2 + l^2x_3 + klx_4 + lhx_5 + hkx_6 \quad (2)$$

The advantages of this approach are that it is direct and fast, using standard least-squares procedures, and that no initial guesses are required for the cell parameters. The disadvantages are that the last three unknowns x_4, x_5 and x_6 are made up from various combinations of the cell parameters and are not independent of the first three parameters. Large correlations among the various parameters might cause rounding error, reducing the accuracy with which α, β and γ can be determined. Furthermore, equation (2) above is only linear in parameters x_1 ... x_6 when written in terms of Q_{hkl}. If we wish to minimize residuals in another dependent variable, such as (the most usually measured) $2\theta_{hkl}$ or d_{hkl}, then the expression becomes non-linear in the cell parameters and simple linear least-squares cannot be used.

Rather than minimizing residuals in Q, in which case direct linear methods such as those of Kelsey (1964) might be used, it is usually more appropriate to use the experimentally measured quantity (such as $2\theta_{hkl}$ or E_{hkl}) as the dependent variable for *minimization*. Below, we discuss the advantages of this approach. Additionally, we draw attention to the advantages of using regression diagnostics as a tool in detecting not only outliers in measurements of diffraction data but also those diffraction peaks which are most influential in determining the fitted cell parameters.

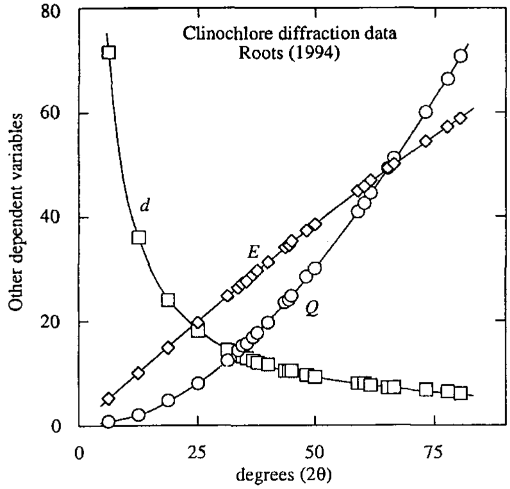

FIG. 1. A Plot of d-spacing, energy and Q against 2θ for a typical set of measured X-ray reflections of chlorite (taken from Roots, 1994). The values for d and Q have been multiplied by 5 and 100, respectively, to scale them to those for E in keV. The nonlinearity between 2θ and d becomes particularly important for materials with large d-spacings, such as the chlorite represented here.

Choice of dependent variable

In many regression problems there exists a choice of which variable to use as the dependent variable. This often turns out to be an important choice since it usually affects the magnitudes of the determined parameters. Most familiar is the question in simple straight line relationships involving two variables (e.g. $y = a + bx$) of whether to regress y on x or x on y. All error is usually placed on the dependent variable (say y) and it is assumed that it is y which we wish to estimate from known values of x using the parameters of the regression equation. In the present situation the choice would appear clear — the values of h,k,l are known (if the indexing has been done correctly) and so Q must be the dependent variable to use. Uncertainties in each Q_{hkl} value are not generally known, however, and generally each is assigned its own weight. This is because it is not usually Q which has been measured but some other experimentally determined value such as the angle ($2\theta_{hkl}$) or energy (E_{hkl}) of a Bragg reflection, depending on the nature of the diffraction experiment. Clearly it would be more satisfactory to minimize the residuals in the experimental observables during the regression. Because Q_{hkl}, E_{hkl} and d_{hkl} do not vary linearly with 2θ (see Fig. 1), the regression results will depend on which one we choose to be the dependent variable. This is, however, a consequence of using unweighted least-squares. With non-linear least-squares methods, any of the possibilities ($2\theta_{hkl}$, Q_{hkl}, E_{hkl} and d_{hkl}) can be easily used as the variable whose residuals are to be minimized and the most reasonable choice must be the one which was measured in the particular diffraction experiment, unless particular care is taken over weighting the data points to compensate. These advantages of reformulating the theory of refinement as a non-linear least-squares procedure rather than a linear least-squares procedure have been recognized previously (Hart *et al.*, 1990; Toraya, 1993). Figure 1, a typical dataset involving reflections in the range 6–80 °2θ, suggests that regressing with d-spacing as the dependent variable will place increasingly excessive weight on low angle reflections, thus seriously biasing the results on the basis of arguably the lowest resolution reflections. This effect becomes particularly significant in materials with large d-spacings, such as the chlorite from which the data of Fig. 1 were obtained. Likewise, use of Q as the dependent variable would place too low a weight on low angle reflections but would begin to place too large a weight on the very high angle reflections when compared with the experimentally determined variables E and 2θ. A strategy that has been employed to overcome this functional bias is to weight the data in Q to compensate, an approach that indeed provides an adequate (if piecemeal) solution to this aspect of having chosen the incorrect dependent variable. Weights may, however, also be needed to account for the *variation in quality of each peak position* measurement. It is known, for example, that the standard deviation of the measured position (in, say, 2θ) is inversely proportional to the square root of the peak intensity (Wilson, 1967). If we wish to weight the observations to take account of this or some other judgement of individual datum quality, further adjustments must be made to those weights which have already been applied to correct for the functional bias of Q. The preferred approach is to carry out the initial nonlinear least-squares on the basis of regression of the measured quantity (2θ or E) rather than Q, and then weights can be applied as necessary to take account of experimental judgements of each datum. Indeed, this has been adopted by previous workers who modified existing methods (Hart *et al.*, 1990).

As an illustration of the potential weakness of performing unweighted regression on Q, we compare the results of regressing the data for Monte Somma anorthite from Redfern and Salje (1987), details given below, using $2\theta_{hkl}$, Q_{hkl}, d_{hkl} and E_{hkl} as the dependent variable. To simulate an energy-dispersive synchrotron experiment, we have assumed a beam 2θ of 10° to calculate an energy spectrum from the original data. The differences in cell parameters, although small, can be as large as the individual estimated uncertainties. Figure 2 shows the differences in the volume and lengths of the cell edges using these four dependent parameters and shows clearly that Q and d yield the most extreme values. Although not shown in Fig. 2, the cell angles α, β and γ all have similar strong dependence on regression variable. In carrying out these regressions we employed a two step approach. First we used linear least-squares of Q_{hkl} to obtain starting guesses for the cell parameters, and then we used nonlinear least-squares of the measured variable of choice to obtain the refined parameters. This approach not only allows the correct regression variable to be selected, it also means that initial guesses at the starting cell

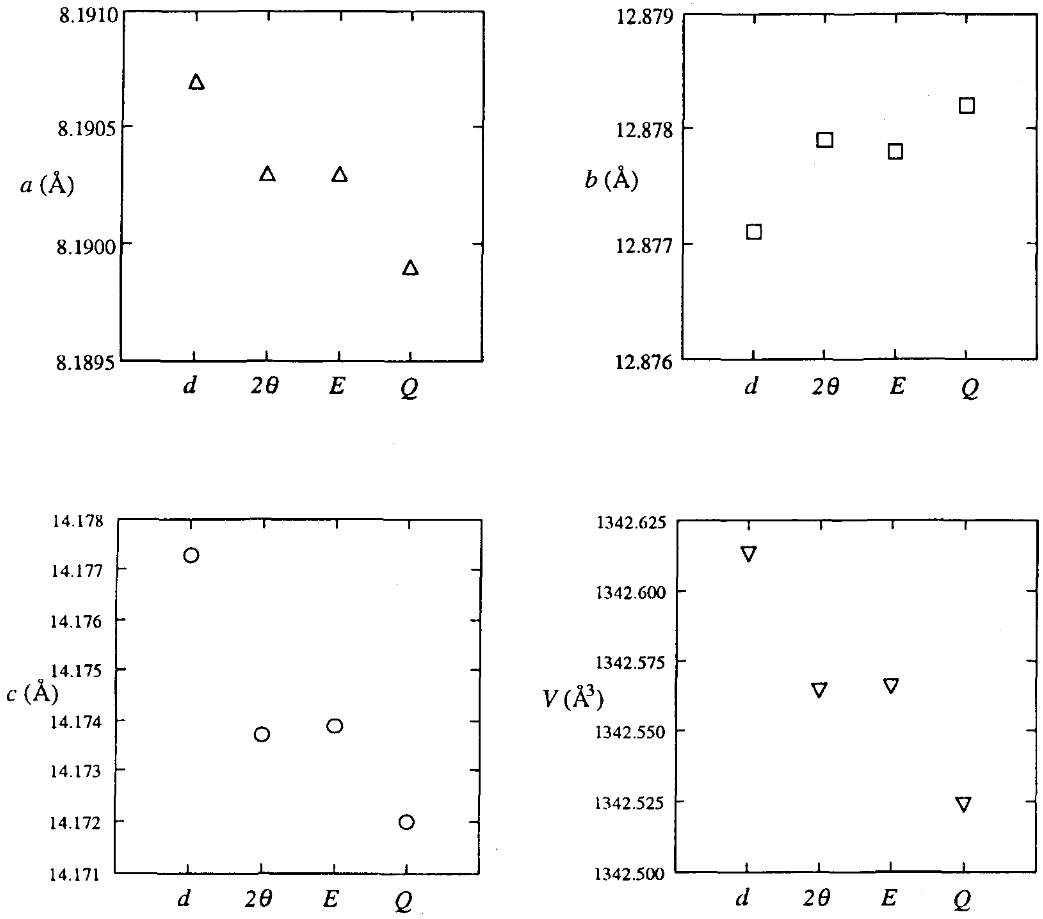

FIG. 2. The effect on the cell dimensions of changing the dependent variable (Q, 2θ, E or d) in the refinement of the anorthite data (see Tables 1 and 2). Note that Q and d typically provide extreme values for the cell constants.

parameters are not required (only indexed peaks and a specification of the crystal system).

Regression diagnostics

As an aid in fitting cell parameters to diffraction data, it is extremely useful to calculate several so-called regression diagnostics along with all the other parameters during the regression in order to identify possible outliers in the data. Regression diagnostics are discussed in some detail in the work of Belsley *et al.* (1980) and Powell (1985) with respect to linear regression analysis where their value is demonstrated in helping identify which data points in a set are outliers and which data are potentially dangerous because they have very high influence on the calculated results (leverage). Although these diagnostics only apply strictly .to linear problems, by linearizing the function at the solution we may use all the machinery of the linear situation. The assumption is that for small errors, the function we are fitting is reasonably linear — an assumption we have to make anyway, in determining the magnitudes of the uncertainties on fitted parameters. We will now introduce five important diagnostic parameters and explain their use.

Typically, the only diagnostic used during refinement of cell parameters is the difference between the observed and calculated values (the residuals) of the data. We shall see that this $2\theta_{\mathrm{obs}}-2\theta_{\mathrm{calc}}$ value can be misleading, and the use of regression diagnostics provides a far superior method for identifying poor data points resulting from measurement or indexing errors. Regression diagnostics provide a useful method for confirming the correct indexing of peaks. It should be noted at the outset, however, that these are single-observation diagnostics — based on the influence a single data point may have, and as such the method cannot detect deleterious effects arising from several observations acting together, since these may mask one another.

(1) Hat. One of the most important diagnostics in helping detect influential data is the Hat matrix **H**, so called because it puts the Hat on **y**, being a projection matrix relating calculated and observed values for the vector of y values, $\hat{\mathbf{y}} = \mathbf{Hy}$. The diagnostics of value are the diagonal elements h_i referring to each observation i and these can take on values from $h_i = 0$, indicating that observation i has no influence on the fit, to $h_i = 1$, indicating extreme influence such that observation i is fixing one of the parameters. The Hat values are related to the distance of any point from the centre of the data spread, so that points lying at the extremities of data space are very influential in determining the values of one or more parameters, whereas data lying in the middle of the spread exert little influence on the calculated parameters. The Hat values sum to the number of parameters in the regression, $\sum_{i=1}^{n} h_i = p$ and the average value of h_i is therefore given by $\frac{p}{n}$ where p is the number of parameters and n is the number of observations. Observations with high leverage are influential, and are flagged by Hat values in excess of a cut-off of $\frac{2p}{n}$ (Belsley *et al.*, 1980). High leverage simply flags the very influential data and does not in itself imply that such data are harmful. Other diagnostics must be used in conjunction with the Hat values in helping to assess the data.

In linear least-squares problems, a solution **b** which minimizes the residuals in **y** for the equations $\mathbf{y} = \mathbf{Xb}$ is found by solving the normal equations, which may be expressed in terms of matrix algebra as $(\mathbf{X}^{\mathrm{T}}\mathbf{X})\mathbf{b} = \mathbf{X}^{\mathrm{T}}\mathbf{y}$, where $\mathbf{X}^{\mathrm{T}}$ is the transpose of **X**. The Hat matrix is then defined as $\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$.

A good non-linear least-squares method for optimizing cell parameters is that of Marquardt (as detailed, for example, by Bevington, 1969) in which the final stage is a Gauss-Newton step to finding solutions **b** to the equations $(\mathbf{J}^{\mathrm{T}}\mathbf{J})\mathbf{b} = \mathbf{J}^{\mathrm{T}}\mathbf{e}$ where **J** is the Jacobian of partial derivatives of the fitting function with respect to the cell parameters **a**, **b** is the vector of increments to the cell parameter estimates, and **e** is the vector of residuals ($y_i - y_i^{\mathrm{calc}}$). Linearizing the fitting function at the solution allows an estimate of the Hat values from $h_i = \mathbf{H}_{ii} = \mathbf{j}_i(\mathbf{J}^{\mathrm{T}}\mathbf{J})^{-1}\mathbf{j}_i^{\mathrm{T}}$, where $\mathbf{j}_i$ is the ith row of **J**, i.e. $[\frac{\partial y_i}{\partial a_1}, \frac{\partial y_i}{\partial a_2} \ldots \frac{\partial y_i}{\partial a_n}]$.

(2) Sigma(i). The standard error of the residuals, σ_y, is a useful measure of the spread of the calculated y values, and a drop in this diagnostic signals a better overall fit to the data. The definition of σ_y is given by $\sigma_y^2 = \frac{\mathbf{e}^{\mathrm{T}}\mathbf{e}}{n-p}$ where **e** is the vector of residuals, and if this value falls significantly upon deletion of an observation i, it points to that observation being potentially deleterious to the fit. The deletion diagnostic $\sigma_y(i)$, calculated for each observation, is the value of σ_y which would result if the observation i were to be deleted from the dataset. Scanning down the list of calculated $\sigma_y(i)$ for values significantly smaller than the overall σ_y highlights observations which might be harmful to the fit.

(3) Rstudent. The use of simple residuals $e_i = y_i - y_i^{\mathrm{calc}}$ are of relatively little diagnostic value where some observations are very much more influential than others, as very influential data are generally associated with small residuals. An adapted form of residual, Rstudent, in which the residual has been normalized by division by $\sqrt{1-h_i}$, allows for the effects of leverage. It is defined as (Belsley *et al.*, 1980)

$$\mathrm{Rstudent}_i = \frac{e_i}{\sigma_y(i)\sqrt{1-h_i}} \qquad (3)$$

Rstudent may be used as a diagnostic parameter since it is expected to be less than 2.0 at the 95%

confidence level, and so values of Rstudent for an i which are in excess of 2.0 suggest that the data point (or in this case observed diffraction vector of the ith hkl reflection) should be treated with suspicion.

(4) DfFits. DfFits is another important deletion diagnostic which gives the change in the predicted value $\hat{y}_i$ upon deletion of the ith observation as a multiple of the standard deviation of $\hat{y}_i$. When this diagnostic is large, the predicted value of y, corresponding to observation i, would change substantially if the regression were to be rerun without that observation. It is therefore a measure of the influence of each observation i on its own calculated position.

$$\mathrm{DfFits}_i = \frac{h_i e_i}{1 - h_i} \cdot \frac{1}{\sigma_y(i)\sqrt{h_i}} \tag{4}$$

Observations with DfFits greater than a cutoff value of $2\sqrt{p/n}$ should be considered as potential outliers in a dataset.

(5) DfBetas. The final diagnostic which may be used to assess and improve cell parameter refinement is called DfBetas, defined as

$$\mathrm{DfBetas}_{ij} = \frac{\beta j - \beta_j(i)}{\sigma_{\beta_j}} = \frac{(\mathbf{J}^{\mathrm{T}}\mathbf{J})^{-1}\mathbf{j}_i^{\mathrm{T}} e_i}{\sigma_{\beta_j}(1 - h_i)} \tag{5}$$

This provides a measure of how much the calculated value of each refined parameter β_j would change if the regression were rerun without using observation i. It may conveniently be expressed as a percentage of the standard deviation of that parameter. Rather than use one of the heuristic cutoff values suggested by Belsley *et al.* (1980) we suggest that observations which would change a parameter by more than 33% of its standard deviation should be flagged as potentially deleterious to the refined results. This diagnostic is particularly valuable in assessing which of the observed reflections has most influence on the calculated cell parameters, and can be used in conjunction with the other diagnostics in weeding out possibly deleterious data as well as assessing the individual sources of error in a cell parameter refinement.

Examples

Refinement of data collected as a function of 2θ

The usefulness of these regression diagnostics, which we have incorporated into a non-linear least-squares refinement procedure to determine unit cell parameters, is best illustrated by taking a closer look at some specific examples. Taking first of all the typical case of cell parameter refinement from diffraction data collected as a function of scattering angle, 2θ, we shall explain the use of the associated regression diagnostics using the dataset for an anorthite measured at high-temperature. The experimental arrangements for the data collection and the scientific significance of the data are described by Redfern and Salje (1987) and Redfern *et al.* (1988). We have refined this example dataset minimizing the square of the residuals in 2θ as well as the residuals in Q, and are thus able to compare directly the differences between the two methods for a real set of data. Computed results from these two refinements of the anorthite data are shown in Tables 1 and 2.

(a) Minimizing residuals in 2θ. Many commonly-used cell parameter refinement programs provide lists of observed and calculated 2θ or d-spacings for each of the observed reflections. The usual measure of the quality of each data point and test of whether a reflection is correctly indexed is taken as the difference between these observed and calculated values. For the Monte Somma anorthite dataset here, therefore, those reflections which show differences between the observed and calculated values which are greater than twice the average absolute deviation are flagged by a bullet in the Tables. This would be the usual limit of diagnostic information available to the experimentalist. Tables 1 and 2, however, also include the regression diagnostics referred to above, which provide an additional invaluable aid to the critical analysis and evaluation of each measured data point as well as providing a check of the accuracy of peak indexing. For example, for the anorthite data refined on the residuals in 2θ (Table 1) we see that both the 224 and $\bar{2}28$ reflections have values of Hat greater than the $\frac{2p}{n}$ cutoff explained above. Thus these reflections are particularly influential. Looking at their values of DfFits and Rstudent, however, we observe that while they are influential, they are not as detrimental to the overall fit as some of the other reflections. The 064 reflection, on the other hand, is not quite as influential (its value of Hat is just less than the cutoff) but it does appear deleterious to the fit, since both DfFits and Rstudent lie well above their limits. Indeed, we see that if this reflection were removed from the dataset it would lower the overall value of sigmafit from its current value of 0.0107 to 0.0100 (as given by the parameter *sigma*(i), in the fourth column of the regression diagnostics), a percentage change in sigmafit of −6.0% (as shown in the final column). This is a case where the experimenter might wish to look again at the measurement of 2θ for the 064 and 224 reflections and assess whether there may be some error of either measurement, indexing, calibration, or a problem of overlap with a strong reflection of another phase if the data are obtained from a mixed-phase sample. If a data point is an outlier, as the regression diagnostics imply, then the only option may be to remove one or

TABLE 1. Nonlinear least squares regression and regression diagnostics of the Monte Somma anorthite data of Redfern and Salje (1987) and Redfern *et al.* (1988) using 2θ as the regressed variable

Refined parameter	Value	Sigma	95% conf
a	8.1903	0.0011	0.0022
b	12.8779	0.0015	0.0031
c	14.1737	0.0019	0.0038
α	93.0933	0.0122	0.0249
β	115.7632	0.0108	0.0221
γ	91.3315	0.0118	0.0242
Volume	1342.5642	0.2121	0.4342

Standard deviation (2θ) = 0.0097
Average deviation (2θ) = 0.0081
Maximum deviation (2θ) = 0.0180
Sigmafit = 0.0107
Students t = 2.05

Observed and fitted results (dependent-variable residuals > twice absolute average deviation are bulleted)

h	*k*	*l*	*d*(obs)	d(calc)	*res*(d)	2θ obs	2θ calc	*res*(2θ)
−1	1	0	6.53418	6.53781	−0.00363	13.540	13.532	0.008
0	−2	2	4.69145	4.69074	0.00071	18.900	18.903	−0.003
−2	0	2	4.04597	4.04404	0.00193	21.950	21.961	−0.011
1	−1	2	3.92420	3.92377	0.00043	22.640	22.642	−0.002
−1	3	0	3.78727	3.78746	−0.00019	23.470	23.469	0.001
1	1	2	3.76198	3.76278	−0.00080	23.630	23.625	0.005
1	3	0	3.62015	3.62029	−0.00015	24.570	24.569	0.001
1	1	−4	3.45951	3.45941	0.00010	25.730	25.731	−0.001
2	2	−2	3.40489	3.40382	0.00107	26.150	26.158	−0.008
−1	1	4	3.36694	3.36476	0.00218	26.450	26.467	−0.017●
−2	2	0	3.26989	3.26890	0.00098	27.250	27.258	−0.008
0	4	0	3.21096	3.21044	0.00052	27.760	27.765	−0.005
−2	0	4	3.19629	3.19566	0.00063	27.890	27.896	−0.006
0	0	4	3.18287	3.18311	−0.00024	28.010	28.008	0.002
2	2	0	3.12603	3.12554	0.00049	28.530	28.535	−0.005
1	−3	2	3.04459	3.04592	−0.00133	29.310	29.297	0.013
0	−4	2	2.95208	2.95226	−0.00017	30.250	30.248	0.002
0	−2	4	2.93597	2.93624	−0.00027	30.420	30.417	0.003
2	2	−4	2.89236	2.89150	0.00086	30.890	30.899	−0.009
1	3	2	2.82985	2.83050	−0.00065	31.590	31.583	0.007
1	3	−4	2.80305	2.80272	0.00034	31.900	31.904	−0.004
−1	3	4	2.65657	2.65742	−0.00085	33.710	33.699	0.011
2	−2	2	2.56369	2.56477	−0.00108	34.970	34.955	0.015
−2	4	2	2.52868	2.52902	−0.00034	35.470	35.465	0.005
2	4	−2	2.49938	2.50010	−0.00072	35.900	35.889	0.011
−3	3	2	2.32705	2.32755	−0.00049	38.660	38.652	0.008
2	−4	2	2.14744	2.14711	0.00033	42.040	42.047	−0.007
0	6	0	2.14016	2.14029	−0.00014	42.190	42.187	0.003
1	5	2	2.09566	2.09650	−0.00083	43.130	43.112	0.018●
−4	0	4	2.02135	2.02202	−0.00068	44.800	44.784	0.016
−4	2	4	1.93561	1.93521	0.00041	46.900	46.910	−0.010
−2	6	0	1.89416	1.89373	0.00043	47.990	48.002	−0.012
2	2	4	1.88200	1.88139	0.00060	48.320	48.337	−0.017●
0	6	4	1.72089	1.72035	0.00054	53.180	53.198	−0.018●
−2	2	8	1.68200	1.68238	−0.00038	54.510	54.497	0.013

TABLE 1. (*Contd.*)

Potentially deleterious or influential observations affecting the fit

h	*k*	*l*	Hat	dfFits	Rstudt	sigma[i]	Δ(sig)%
1	5	2	0.206	1.014	1.989	0.0101	−4.7
−4	0	4	0.319	1.281	1.872	0.0102	−4.1
2	2	4	0.406	−1.760	−2.130	0.0101	−5.6
0	6	4	0.325	−1.519	−2.188	0.0100	−6.0
−2	2	8	0.465	1.653	1.772	0.0103	−3.5
Cut-off			0.343	0.828	2.000		

Observations most strongly affecting the parameter values (shown as % of their standard deviations)

h	*k*	*l*	Δa	Δb	Δc	Δalpha	Δbeta	Δgamma	ΔV
−1	1	4	7	8	−49	18	−18	−14	−20
−1	3	4	−9	7	20	−34	12	22	10
2	−2	2	21	−16	−4	14	−58	16	34
2	4	−2	23	33	−1	37	17	−55	33
1	5	2	1	50	−19	−26	−16	−34	36
−4	0	4	119	−25	11	5	53	−26	60
−4	2	4	−65	13	−4	24	−33	−23	−30
−2	6	0	7	−54	13	5	2	−43	−22
2	2	4	−35	50	−4	22	123	30	−71
0	6	4	29	−69	−1	87	3	−15	−38
−2	2	8	−20	−23	138	−49	49	39	56

TABLE 2. Nonlinear least squares regression and regression diagnostics of the Monte Somma anorthite data of Redfern and Salje (1987) and Redfern *et al.* (1988) using Q as the regressed variable

Refined parameter	Value	Sigma	95% conf
a	8.1899	0.0010	0.0021
b	12.8782	0.0015	0.0030
c	14.1720	0.0019	0.0038
α	93.0864	0.0131	0.0268
β	115.7514	0.0108	0.0222
γ	91.3386	0.0126	0.0258
Volume	1342.5232	0.2229	0.4564

Standard deviation (Q) = 0.000087
Average deviation (Q) = 0.000071
Maximum deviation (Q) = 0.000205
Sigmafit = 0.000096
Students t = 2.05

Observed and fitted results (dependent-variable residuals > twice absolute average deviation are bulleted)

h	k	l	d(obs)	d(calc)	*res*(d)	2θ obs	2θ calc	*res*(2θq)	*res*(Q)
−1	1	0	6.53418	6.53831	−0.00413	13.540	13.531	0.009	0.00003
0	−2	2	4.69145	4.69057	0.00089	18.900	18.904	−0.004	−0.00002
−2	0	2	4.04597	4.04372	0.00226	21.950	21.962	−0.012	−0.00007
1	−1	2	3.92420	3.92406	0.00014	22.640	22.641	−0.001	−0.00000
−1	3	0	3.78727	3.78771	−0.00044	23.470	23.467	0.003	0.00002
1	1	2	3.76198	3.76308	−0.00110	23.630	23.623	0.007	0.00004
1	3	0	3.62015	3.62034	−0.00019	24.570	24.569	0.001	0.00001
1	1	−4	3.45951	3.45894	0.00057	25.730	25.734	−0.004	−0.00003
2	2	−2	3.40489	3.40347	0.00142	26.150	26.161	−0.011	−0.00007
−1	1	4	3.36694	3.36453	0.00241	26.450	26.469	−0.019	−0.00013
−2	2	0	3.26989	3.26915	0.00073	27.250	27.256	−0.006	−0.00004
0	4	0	3.21096	3.21054	0.00042	27.760	27.764	−0.004	−0.00003
−2	0	4	3.19629	3.19519	0.00110	27.890	27.900	−0.010	−0.00007
0	0	4	3.18287	3.18306	−0.00019	28.010	28.008	0.002	0.00001
2	2	0	3.12603	3.12561	0.00042	28.530	28.534	−0.004	−0.00003
1	−3	2	3.04459	3.04607	−0.00148	29.310	29.295	0.015	0.00010
0	−4	2	2.95208	2.95222	−0.00014	30.250	30.249	0.001	0.00001
0	−2	4	2.93597	2.93612	−0.00015	30.420	30.418	0.002	0.00001
2	2	−4	2.89236	2.89101	0.00135	30.890	30.905	−0.015	−0.00011
1	3	2	2.82985	2.83068	−0.00083	31.590	31.580	0.010	0.00007
1	3	−4	2.80305	2.80238	0.00067	31.900	31.908	−0.008	−0.00006
−1	3	4	2.65657	2.65744	−0.00087	33.710	33.699	0.011	0.00009
2	−2	2	2.56369	2.56499	−0.00130	34.970	34.952	0.018	0.00015●
−2	4	2	2.52868	2.52914	−0.00045	35.470	35.463	0.007	0.00006
2	4	−2	2.49938	2.49992	−0.00055	35.900	35.892	0.008	0.00007
−3	3	2	2.32705	2.32762	−0.00057	38.660	38.650	0.010	0.00009
2	−4	2	2.14744	2.14727	0.00018	42.040	42.044	−0.004	−0.00004
0	6	0	2.14016	2.14036	−0.00020	42.190	42.186	0.004	0.00004
1	5	2	2.09566	2.09661	−0.00094	43.130	43.110	0.020	0.00020●
−4	0	4	2.02135	2.02186	−0.00051	44.800	44.788	0.012	0.00012
−4	2	4	1.93561	1.93514	0.00048	46.900	46.912	−0.012	−0.00013
−2	6	0	1.89416	1.89385	0.00031	47.990	47.998	−0.008	−0.00009
2	2	4	1.88200	1.88154	0.00046	48.320	48.332	−0.012	−0.00014
0	6	4	1.72089	1.72044	0.00045	53.180	53.195	−0.015	−0.00018●
−2	2	8	1.68200	1.68227	−0.00027	54.510	54.501	0.009	0.00011

TABLE 2. (*Contd.*)

Potentially deleterious or influential observations affecting the fit

h	*k*	*l*	Hat	dfFits	Rstudt	sigma[i]	Δ(sig)%
2	−2	2	0.181	0.872	1.854	0.0001	−4.0
1	5	2	0.229	1.456	2.673	0.0001	−9.2
−4	0	4	0.403	1.420	1.728	0.0001	−3.3
−4	2	4	0.389	−1.456	−1.823	0.0001	−3.8
−2	6	0	0.375	−0.938	−1.212	0.0001	−0.8
2	2	4	0.572	−2.707	−2.340	0.0001	−6.9
0	6	4	0.447	−2.445	−2.717	0.0001	−9.5
−2	2	8	0.681	3.194	2.186	0.0001	−5.9
Cut-off			0.343	0.828	2.000		

Observations most strongly affecting the parameter values (shown as % of their standard deviations)

h	*k*	*l*	Δa	Δb	Δc	Δalpha	Δbeta	Δgamma	ΔV
2	2	−4	−18	−17	−36	−42	−32	40	−35
2	−2	2	20	−15	1	16	−64	18	38
2	4	−2	20	28	6	33	17	−46	30
1	5	2	13	66	−31	−22	−14	−54	45
−4	0	4	132	−18	12	13	59	−41	68
−4	2	4	−116	24	3	44	−53	−36	−43
−2	6	0	10	−61	12	2	6	−48	−25
2	2	4	−60	79	0	19	190	38	−101
0	6	4	37	−96	24	123	3	−21	−32
−2	2	8	−44	−42	253	−48	87	56	92

both from the refinement. The DfBetas parameters for these reflections provide a prior indication of the effect of such a course of action, since they show how the removal of a datapoint affects the refined cell parameters. Thus, in the case of removal of the 064 reflection, one would see a change of $+0.29\sigma_a$ (0.0003 Å) on the a cell edge, $-0.69\sigma_b$ (0.0010 Å) on b, hardly any change on c, but most significantly an increase of $+0.87\sigma_\alpha$ (0.0106 °) on the α angle. The overall effect is to increase the value of the cell volume by some 0.081 $Å^3$, which is only 38% of its standard deviation (and thus of arguable significance). The limited effect of removal of this datapoint on the refined parameters might have been expected since it is a further confirmation that while both DfFits and Rstudent lie above their limits for this reflection, and its removal can improve the overall quality of the fit, the reflection does not have as strong an influence on the derived cell parameters as, for example, the 224 reflection (since its Hat is lower than that for 224). Indeed, removal of 224 would lead to a reduction in the cell volume by 71% of its standard deviation (0.151 $Å^3$), and a large increase of the β angle by 123% of its standard deviation.

It is interesting to compare the use of the regression diagnostics explained above with the procedure of simple selection of potential outliers on the basis of the differences between observed and calculated 2θ positions (as might usually be performed in the course of a cell parameter refinement). Four reflections show deviations between calculated and observed values greater than 2σ. Of these, the greatest is for the 064 reflection and the 152 reflection. From the values of Hat, however, we see that 152 does not have a strong influence on the refined parameters, neither does its removal lower the overall standard deviation by as much as, say, the removal of 224 (which has a smaller deviation between observed and calculated 2θ). Furthermore, we have seen that while the removal of the 064 reflection gives the greatest reduction in sigmafit, this peak does not influence the refined cell parameters as much as the (almost as poorly fitting) 224 reflection. While removal of 152 might be suggested from the differences between observed and calculated 2θ positions, therefore, consideration of the deletion diagnostics, provided as a computational by-product of the regression, indicates that the 152 reflection is not a priority, and attention should first be paid to the 224 and 064 reflections.

(b) Minimizing residuals in Q. The same dataset for anorthite has been refined by minimizing the residuals in Q, rather than in 2θ. This simulates the operation of a standard linear least-squares refinement of the data, as might be performed by any one of the many public domain programs available for the purpose. The output is shown in Table 2. As expected, the refined cell parameters differ from those obtained by refinement minimizing residuals in 2θ using exactly the same data. In particular the β cell angle is more than one standard deviation smaller. Furthermore, we see that the residuals on individual observations are now quite misleading if employed as a mechanism for detecting outliers. The greatest variation between observed and calculated 2θ is shown by the 152 reflection. If this was all that was known, then the first course of action in an attempt to improve the refinement might be to eliminate this reflection, or at least measure it again to attempt to improve the fit. We showed above, however, that this reflection is not as detrimental to the non-linear least-squares refinement as either 064 or 224, and that 224 was the peak which is most influential on the refined parameters when the data are handled correctly (refining on 2θ, the measured observation). Disturbingly, the 224 reflection shows what would probably be regarded as a perfectly acceptable value of $2\theta_{obs}-2\theta_{calc}$ when the data are fitted by refining residuals in Q, and there is no indication that this is the most deleterious outlier. The regression diagnostics obtained by refining residuals in Q also highlight a number of other peaks (such as $2\bar{2}2$, $\bar{4}24$, and $\bar{2}60$), which we know are not significant outliers from our refinement of the data on the basis of minimising residuals in 2θ. Comparison of the computed results for this dataset using the two methods of refinement highlights both the importance of refining the data on the basis of the observed quantities (rather than a derived function, such as Q), as well as the utility of deletion diagnostics in identifying outliers (compared with simpler yet less robust criteria such as the values of $2\theta_{obs}-2\theta_{calc}$ for individual reflections).

Refinement of high-pressure energy-dispersive data.

The provision of regression diagnostics becomes particularly useful when dealing with data that are inherently low quality. One particular field where this applies is in the analysis of high-pressure powder diffraction data collected in an energy-dispersive experiment. In a typical experiment a beam of white (usually synchrotron) radiation impinges on an extremely small amount of sample, often held static in a diamond anvil cell. Poor sample statistics, collection of a limited part of the diffraction cone, interference with diffraction from internal pressure standards, and the relatively low resolution of solid-state energy-dispersive detectors all conspire against the experimenter and mean that data must be handled and interpreted with care to obtain the best results.

We illustrate the use of regression diagnostics in refining energy-dispersive data using the example of epidote in Table 3. First of all we note that the Bragg

Table 3. Nonlinear least squares regression and regression diagnostics of energy-dispersive epidote data collected by the authors using E as the regressed variable

Refined parameter	Value	Sigma	95% conf
a	8.8820	0.0046	0.0100
b	5.6439	0.0036	0.0078
c	10.1556	0.0059	0.0129
β	115.4220	0.0657	0.1434
Volume	459.7987	0.3391	0.7400

Standard deviation (E) = 0.0255
Average deviation (E) = 0.0214
Maximum deviation (E) = 0.0434
Sigmafit = 0.02945
Students t = 2.18

Observed and fitted results (dependent-variable residuals > twice absolute average deviation are bulleted)

h	k	l	d(obs)	d(calc)	$res(d)$	E obs	E calc	res(E)
2	0	0	4.00491	4.01098	−0.00606	17.499	17.473	0.026
−1	1	2	3.75493	3.74921	0.00572	18.664	18.692	−0.028
−1	1	3	2.90182	2.90002	0.00180	24.151	24.166	−0.015
3	0	0	2.67386	2.67398	−0.00012	26.210	26.209	0.001
−3	1	1	2.60006	2.59823	0.00183	26.954	26.973	−0.019
0	2	2	2.40237	2.40341	−0.00104	29.172	29.159	0.013
1	1	3	2.29950	2.30073	−0.00123	30.477	30.461	0.016
−2	2	2	2.29950	2.30264	−0.00313	30.477	30.436	0.041
−4	0	1	2.16423	2.16247	0.00176	32.382	32.408	−0.026
−1	2	3	2.16423	2.16633	−0.00210	32.382	32.351	0.031
−2	2	3	2.11237	2.10960	0.00276	33.177	33.220	−0.043●
−4	1	2	2.06628	2.06626	0.00002	33.917	33.917	−0.000
−2	2	4	1.87596	1.87460	0.00135	37.358	37.385	−0.027
−4	1	4	1.87596	1.87808	−0.00213	37.358	37.316	0.042
−1	0	6	1.63651	1.63641	0.00010	42.824	42.827	−0.003
−3	3	1	1.58288	1.58257	0.00031	44.275	44.284	−0.009

Potentially deleterious or influential observations affecting the fit

h	k	l	Hat	dfFits	Rstudt	sigma[i]	Δ(sig)%
−4	1	4	0.529	2.663	2.516	0.0245	−16.8
−1	0	6	0.651	−0.195	−0.143	0.0307	4.4
Cut-off			0.500	1.000	2.000		

Observations most strongly affecting the parameter values (shown as % of their standard deviations)

h	k	l	Δa	Δb	Δc	Δbeta	ΔV
1	1	3	−12	7	6	−47	36
−2	2	2	3	54	−8	14	33
−4	0	1	−61	15	5	11	−34
−1	2	3	−13	37	8	−3	31
−2	2	3	−6	−41	−19	−29	−33
−4	1	4	163	−79	140	189	17

maxima were measured as a function of energy, and the cell parameters were determined from refinement of the same measured quantity. The list of observed-calculated peak positions shows that the measured position of the $\bar{2}23$ reflection displays the greatest deviation from the best fit calculated value. A typical strategy might be to assume that the $\bar{2}23$ reflection position was spurious and to recalculate the cell parameters on the basis of a dataset without this data point. The regression diagnostics, however, indicate that rather than $\bar{2}23$, it is the $\bar{4}14$ and $\bar{1}06$ reflections that require closer examination. Values of DfFits and Rstudent for the $\bar{4}14$ peak are substantially greater than the cutoff. We also see that the $\bar{1}06$ reflection has a large Hat, but this need not worry us as the values of DfFits and Rstudent are within their limits, and the value of *sigma*(i) shows that the regression would become worse, not better, if this reflection was omitted. On the other hand, *sigma*(i) for the $\bar{4}14$ reflection is substantially (around 17%) lower than sigmafit for the whole refinement, and the omission of this peak will improve the whole fit. Inspection of the DfBetas given in part (b) reveals that omission of the $\bar{4}14$ reflection from the refinement will increase a, c, and the β angle by more than one standard deviation. It is interesting to note that removal of the $\bar{2}23$ reflection, as might initially have been indicated simply by considering $E_{\rm obs}-E_{\rm calc}$, would not alter any of the cell parameters by more than half their individual standard deviations. Furthermore, although the residual of this point is relatively large, the statistical tests show that it is not significant. By inspection of the diagnostics for every observation rather than just those above the critical cutoffs, we find that the measured observation of the $\bar{2}23$ reflection gives values of -0.708 and -1.720 for DfFits and Rstudent respectively (well below the cutoff), and furthermore has a small value of Hat (0.145) so is in any case insignificant. On the other hand, Table 3 shows that the $\bar{4}14$ reflection is a true outlier in the dataset and identifies this peak as the reflection which should be checked if the refinement is to be improved. Once again, we see that the values of $E_{\rm obs}-E_{\rm calc}$ are not always good indicators of the statistical quality of individual reflections.

Implications of the use of regression diagnostics in cell refinement

We have shown the efficacy of computing essential diagnostic information required for careful cell parameter refinement, and that such diagnostics present a considerable improvement on those procedures of weeding out reflections based purely on the individual deviations between observed and calculated data. There is no reason why the identification of outliers in datasets and subsequent improvement of refinements should not become a routine precursor to the publication and use of powder diffraction data. Indeed, Smith (1989) has already pointed out that any laboratory planning to prepare data for publication or for inclusion in databases such as the Powder Diffraction File or the NIST Crystal Data File should screen their data for errors and poorly-fitting values. The statistical tests described here provide a simple mechanism for carrying out such screening, using regression diagnostics for the first time. It also very important that the refinement is based on minimization of the differences between the true measured quantity and its calculated value (rather than a linearized derived function). R.C. Jenkins (1995, pers. comm.) recently pointed out that of the 5000 powder diffraction datasets culled from the published literature each year, the ICDD find that only 1000 or so are acceptable for inclusion in the Powder Diffraction File. With the early use of regression diagnostics provided here this 'hit-rate' could be significantly improved.

The program UnitCell, which implements the regressions (with regression diagnostics) discussed above, is available free to users from non-profit-making institutions. The executable code (for Macintosh or Windows) may be obtained by anonymous ftp from rock.esc.cam.ac.uk, where it resides in directory pub/minp/UnitCell/. Download the file README for further instructions. The programs and further details may be obtained from the appropriate part of the World Wide Web server at Department of Earth Sciences, Cambridge University (http://www.esc.cam.ac.uk).

Acknowledgements

We wish to express our thanks to Anne Graeme-Barber for her patience in testing the program UnitCell and for her comments and advice on improvements to it, and to Roger Powell who originally introduced TJBH to the subject of regression diagnostics.

References

Belsley, D.A., Kuh, E. and Welsh, R.E. (1980) *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. John Wiley, New York.

Bevington, P.R. (1969) *Data Reduction and Error Analysis for the Physical Sciences*. McGraw-Hill, New York, 336 pp.

Cohen, M.U. (1935) Precision lattice constants from X-ray powder photographs. *Review of Scientific Instruments*, **6**, 68–74.

Hart, M., Cernik, R.J., Parrish, W. and Toraya, H. (1990) Lattice parameter determination for powders using synchrotron radiation. *J. Appl. Crystallogr.*, **23**, 286–91.

Kelsey, C.H. (1964) The calculation of errors in a least squares estimate of unit-cell dimensions. *Mineral. Mag.*, **33**, 809–12.

Powell, R. (1985) Regression diagnostics and robust regression in geothermometer/geobarometer calibration: the garnet-clinopyroxene geothermometer revisited. *J. Met. Geol.*, **3**, 231–43.

Redfern, S.A.T. and Salje, E. (1987) Thermodynamics of plagioclase II: Temperature evolution of the spontaneous strain at the $I\bar{1}$–$P\bar{1}$ phase transition in anorthite. *Phys. Chem. Minerals*, **14**, 189–95.

Redfern, S.A.T., Graeme-Barber, A. and Salje, E. (1988) Thermodynamics of plagioclase III: Spontaneous strain at the $I\bar{1}$–$P\bar{1}$ phase transition in Ca-rich plagioclase. *Phys. Chem. Minerals*, **16**, 157–63.

Roots, M. (1994) Molar volumes on the clinochlore-amesite binary: some new data. *Euro. J. Mineral.*, **6**, 279–83.

Smith, D.K. (1989) Computer analysis of diffraction data. In *Modern Powder Diffraction* (D.L. Bish and J.E. Post, eds) Mineralogical Society of America, *Reviews in Mineralogy*, **20**, 183–216.

Toraya, H. (1993) The determination of unit-cell parameters from Bragg reflection data using a standard reference materials but without a calibration curve. *J. Appl. Crystallogr.*, **26**, 583–90.

Wilson, A.J.C. (1967) Statistical variance of line-profile parameters. Measures of intensity, location and dispersion. *Acta Crystallogr.*, **23**, 888–98.

[*Manuscript received 25 March 1996:
revised 19 July 1996*]