Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 5 April 2005
Accepted 7 June 2005

Outlier recognition in crystal-structure least-squares modelling by diagnostic techniques based on leverage analysis

Marcello Merli

Dipartimento di Chimica e Fisica della Terra ed Applicazioni alle Georisorse ed ai Rischi Naturali, Università degli Studi di Palermo, Via Archirafi 36, I-90123 Palermo, Italy. Correspondence e-mail: merli@unipa.it

The identification of the actual outliers in a least-squares crystal-structure model refinement and their subsequent elimination from the data set is a non-trivial task that has to be carried out carefully when a high level of accuracy of the estimates is required. One of the most suitable tools for detecting the influence of each data entry on the regression is the identification of 'leverage points'. On the other hand, the recognition of the actual statistical outliers is effectively possible by using some diagnostics as a function of the leverage, such as Cook's distance, DFFITS and FVARATIO. The evaluation of these estimators makes it possible to achieve a reliable identification of the outliers and the elimination of those that impair the least-squares fit. In this paper, a procedure for filtering data points based on this kind of analysis for crystallographic X-ray data is presented and discussed.

1. List of symbols and abbreviations

I: intensity of a reflection.
σ: standard error associated with a reflection.
s.u: standard uncertainty.
a.d.p: atomic displacement parameters.
F_o, F_c: observed structure factor, calculated structure factor.
w: statistical weight of the least-squares refinement.
n: number of observations.
p: number of variables in the least-squares procedure.
$\mathbf{y}$, $\hat{\mathbf{y}}$: n-length vector of the observations, n-length vector of the calculated reflections.
$\mathbf{A}$: design matrix of the least-squares system.
$\mathbf{W}$: weight matrix of the least-squares system.
$\mathbf{x}$: vector of the solutions of the least-squares system.
$\mathbf{H}$: projection matrix (hat matrix, leverage matrix) of the least-squares system.
h_i: ith diagonal element of the projection matrix.
$F_{p,n}$: Fisher's distribution function with p and n degrees of freedom.
GoF: goodness of fit.
R, R_w: crystallographic discrepancy factor, weighted crystallographic discrepancy factor.
e_i: residual error $y_i - \hat{y}_i$ associated with the ith reflection.
s: estimated error variance.
s'_i: estimated error variance when the ith row of $\mathbf{A}$ and $\mathbf{y}$ have been deleted.
e^*_i: studentized deleted residual.

2. Introduction

Over the years, there has been great interest in the statistical aspects of fitting procedures commonly used in crystallographic practice (see for example the IUCr reports by Schwarzenbach *et al.*, 1989, 1995, and references therein). In particular, customarily adopted fitting of the diffraction data is a crucial process because of the intrinsic noise of experimental data and its usual departure from Gaussian distribution, which makes the crystallographic least-squares procedure a delicate task when great precision of the results is required. A number of critical articles can be found in the crystallographic literature regarding the algebra and the statistical control of diffraction regression data (for example, Watkin, 1994; Harris & Moss, 1992; Pannu & Read, 1996; Spagna & Camalli, 1999; Lunin *et al.*, 2002, and so on). We would also cite Kuntzinger *et al.* (1998) here as their paper deals with the concepts and statistical tools that are summarized and developed in the present paper.

It is well known [see Prince & Boggs (1992) for a discussion of the crystallographic case] that any attempt to fit an outlier in an optimization procedure is a dangerous practice that may affect the estimates of some other data points and the estimates of some other model variables. Thus, a suitable identification of the actual outliers and their consequent elimination (or their appropriate weighting) is a necessary procedure if we are to obtain highly accurate estimations of the variables, instead of an indiscriminate cutting of the reflections based on

doi:10.1107/S010876730501809X

the $I/\sigma(I)$ ratio, resolution *etc.*, which, in most cases, only improve the regression results cosmetically.

Reliable detection of the outliers might pass through the calculation of each point's 'leverage', *i.e.* the diagonal terms of the so-called 'hat matrix' associated with the least-squares system [in Belsey *et al.* (1980), a very extensive review on the matter is presented], first introduced in crystallography by Prince & Nicholson (1985). Leverage analysis itself is a good means of detecting the most influential data points in the regression. Indeed, this approach allowed, for example, Hazen & Finger (1989), Merli *et al.* (2000, 2001) and Merli (2002) to identify some classes of reflections that proved to be particularly influential in the estimation of some specific classes of variables (site occupancies, a.d.p.'s and so on), suggesting the best strategies for collecting and/or treating data in a rigorous way. If the aim of statistical analysis is to detect dangerous outliers of the least-squares procedure, leverage information by itself is not sufficient if we are to identify aberrant data because leverage only indicates the potentially influential data points on the least-squares estimation. The identification of outliers must be carried out by calculating any diagnostic as a function both of leverage and of some measure of the residuals. The aim of this paper is to check the reliability of a filtering procedure based on leverage and its derived diagnostics for crystallographic model refinements.

3. Mathematical analysis

3.1. Theoretical basis

The reader can refer to Prince & Nicholson (1985) and to Prince & Boggs (1992) for details regarding least-squares algebra and leverage definition. Here, we can briefly recall that, given a linear model $\mathbf{y} = \mathbf{Ax}$, the least-squares solution of the system is given by $\mathbf{x} = (\mathbf{A}^{\rm T}\mathbf{WA})^{-1}\mathbf{A}^{\rm T}\mathbf{Wy}$. The so-called 'hat matrix' is written as $\mathbf{H} = \mathbf{A}(\mathbf{A}^{\rm T}\mathbf{WA})^{-1}\mathbf{A}^{\rm T}\mathbf{W}$ and the diagonal elements $0 < h_i < 1$ of this matrix are defined as the leverage of each ith data point. Note that $\mathbf{H}$ depends only on the model $\mathbf{A}$ and the weights $\mathbf{W}$, and not on the observations $\mathbf{y}$.

In a crystallographic case, the refined model is not linear. Nevertheless, it can be shown that the leverage analysis and, consequently, each of the related diagnostics can be extended for non-linear problems as well [see Belsey *et al.* (1980) for further explanation]. Note that $\hat{\mathbf{y}} = \mathbf{Hy}$, where $\hat{\mathbf{y}}$ is the vector of the calculated reflections. As pointed out by Belsey *et al.* (1980), the influence of the response value y_i on the fit is most directly reflected in its impact on the corresponding fitted value $\hat{y}_i$: this information is contained in h_i. Therefore, data with both high leverage and a discrepancy between the observed and the calculated values may be considered as being actual 'outliers' of the refinement, dangerous data points that may affect the estimations of some variables owing to their importance in the least-squares procedure. Moreover, the diagonal element h_i of the hat matrix represents the rate of change in the calculated value of a data point resulting from a change in the observed value: as a consequence, a number of statistical criteria employed to forecast effects on the regression (*i.e.* the change in fit) when an observation is deleted are functions of h_i, as shown below.

3.2. Statistical criteria

Diagnostic techniques for discovering influential reflections can be obtained by combining leverage and some (standardized) measures of the discrepancy e_i. Definitions of the diagnostics used in this work (from Belsey *et al.*, 1980) are given in Appendix *A*.

When there are no outliers that can affect the efficiency of the fit, both in terms of reproduction of the data and the reliability of the estimates of the variables, COVRATIO and FVARATIO will take similar values. The correct combination of COVRATIO and FVARATIO results can lead to a safe control of the data truncation, *i.e.* of the detection of the really dangerous outliers, as shown below.

The values obtained from each diagnostic technique should be interpreted with reference to some important considerations. In general, any diagnostic measure should be based on the choice of a suitable cut-off threshold. As far as the leverage thresholds are concerned, with the assumption that the variables of the refinement are Gaussian, it is straightforward to compute the exact distribution for either h_i or some of their functions.

As pointed out by Rao (1973), Wilks's Λ statistics assume that $(n-p)/(p-1)\{[1-\Lambda(\mathbf{a}_i)]/\Lambda(\mathbf{a})_i\} \sim F_{p-1,n-p}$, where $\mathbf{a}_i$ is the ith row of the design matrix and Λ is defined as $n/(n-1)(1-h_i)$. It follows that $(n-p)[h_i - 1/n]/(1-h_i)(p-1)$ is distributed as F with $p-1$ and $n-p$ degrees of freedom. For example, in large systems, 95% percentile for F distribution is less than 2, so $2p/n$ is a rough cut-off threshold to determine if h_i is an actual 'leverage point' of the refinement (note that p/n, the average leverage, corresponds to the perfectly balanced least-squares system).

As mentioned above, these diagnostics combine with information regarding the influence (represented by the leverage) of the data point and some measures of the (standardized) residual: if scaled by an appropriate standard error, all of these diagnostic tools can be considered as being large if their value is greater than 2 ('absolute cut-off'). For practical reasons, it is useful to deal with cut-offs that represent the influence on the fitting regardless of sample size: Belsey *et al.* (1980) call them 'size-adjusted' cut-offs and, for the diagnostic estimators used in this work, size-adjusted cut-offs have been calculated following Rao's (1973) assumption.

The greater the Gaussian character of the distributions of the residuals, the more effective the cut-offs are. This assumption is far from being satisfied when dealing with crystallographic data: for instance, the use of weighting schemes that are assessed so as to ensure the goodness of fit (hereafter GoF, defined as $\{\sum[w(F_o-F_c)^2]/(n-p)\}^{1/2}$) close to unity can influence the diagnostics. Nevertheless, such approximate diagnostics may still be useful in the actual identification of outliers, as shown in the simulations described below.

Table 1
'True' reflections and modified values introduced in the perovskite data set (absolute scale).

h k l	Calculated F_o^2	Leverage	Modified F_o^2
2 0 2	19314 (20)	0.0904	18500
0 4 0	20545 (210)	0.1091	18500
1 2 1	9858 (31)	0.0489	9100
0 0 4	8576 (29)	0.0609	7750
0 4 2	7396 (20)	0.0440	6950
0 0 2	9687 (20)	0.0563	9000
0 14 0	33 (4)	0.0605	50
2 14 0	58 (4)	0.0492	70
1 16 1	65 (9)	0.0406	80
0 14 2	58 (17)	0.0451	100

Table 2
'True' reflections and modified values introduced in the loganin data set (absolute scale).

h k l	Calculated F_o^2	Leverage	Modified F_o^2
2 0 0	11147 (231)	0.3854	9135
4 1 0	839 (10)	0.3250	625
0 4 0	8753 (58)	0.4232	8475
2 13 1	361 (21)	0.0844	675
$\bar{7}$ 1 2	293 (9)	0.1158	576

Table 3
'True' reflections and modified values introduced in the oxalic acid data set (absolute scale).

h k l	Calculated F_o^2	Leverage	Modified F_o^2
1 0 3	839 (15)	0.3020	790
1 0 $\bar{5}$	1711 (10)	0.2892	1660
2 0 0	643 (3)	0.2726	630
1 1 $\bar{3}$	427 (2)	0.2205	410
3 0 $\bar{3}$	400 (2)	0.2849	380

3.3. Limits for the diagnostics

Once the outliers have been detected by means of one of the above diagnostics, it should be noted that their simultaneous elimination could involve, in certain cases, a loss of some reflections that are not actual outliers of the fit. This is possible because the high level of influence of the extreme outliers on the refinement can cause a misleading identification of 'minor' outliers, which only seem to be leverage points. These data will no longer be identified as outliers when the 'true' outliers are eliminated from the data set. Besides, it may be possible that the outliers found do not constitute all or the only true outliers (this effect has been observed in some cases tested in the present work). It is possible that the simultaneous identification of all outliers and only the 'true' outliers may be feasible if diagnostic estimators of a higher order (for example, Cook's statistics of 2, 3, ..., *k*th rank) are employed. Other effective methods have been described by, for example, Seaver *et al.* (1990) and Gray & Ling (1984) but such methods are not easily applicable to crystallographic problems. Moreover, any feasible diagnostics that take the joint influence of the reflections into account [for instance, MDFFIT as described by Belsey *et al.* (1980)] are too expensive in terms of CPU time because of the great size of the design matrices generally involved in crystallographic least-squares computations. Thus, in this paper, rather than considering the possibility of 'legitimate' simultaneous detection of the actual set of outliers, we have focused on the identification of an improvement in crystal-structure least-squares modelling by means of iterative identification and elimination of one outlier at a time.

4. The experiments: results and discussion

Two main crystal structure typologies have been considered: an inorganic structure, $CaTiO_3$ perovskite (data collected at our laboratory, which are unpublished) and two organic samples, loganin (see the documentation of *SIR2002*, Burla *et al.*, 2003) and oxalic acid dihydrate (see the documentation of *XD*, Koritsanszky *et al.*, 1995).

The process was the same for all structures: after obtaining a reliable structure refinement of the sample, a synthetic data set was calculated on the basis of the model obtained by fitting the experimental data. A random noise, calculated as $\mathbf{e} = (0.5 - r)|\mathbf{y}|/10$ (where r is a random number $0 < r < 1$), was added to the theoretical structure factors.

The use of synthetic data is justified by the need to test data sets with outliers arising only from experimental bias, with the assumption that the model used is the correct one, *i.e.* it is able to reproduce the data perfectly. Further investigations into the possibility of detecting an outlier from an imperfection of the model will be presented elsewhere.

In order to obtain outliers, an extra error was added to some of the reflections of the synthetic data sets (chosen from among the potentially most influential data, *i.e.* the highest leverage reflections). In particular, the strongest high-leverage intensities were lowered by a maximum of 10% and the weakest high-leverage intensities were increased by a maximum of 10% to emulate two negative effects commonly observed in X-ray single-crystal crystallography such as secondary extinction and the Renninger effect: ten reflections were modified in the perovskite data set (Table 1) and five reflections were modified in both of the organic structures (Tables 2 and 3).

The full-matrix refinements on $|F_o|$ were then made by eliminating the most aberrant outlier each time.

In all cases, spherically averaged scattering factors for all the atoms and harmonic second-order thermal tensors were used. The calculation of leverage and the diagnostic measures listed above were performed only after convergence was achieved (mean shift/s.u. < 0.0001).

The criteria followed step by step during the filtering procedure can be summarized as follows:

1. the largest value for |DFFITS| (or alternatively for Cook's distance) represents the most aberrant outlier;

2. COVRATIO values <1 indicate a reflection that can potentially improve the efficiency of the fitting after its elimination; COVRATIO values >1 represent potentially influential reflections; aberrant reflections should lie outside the

range limited by the fixed thresholds, as described in Appendix *A*;

3. a FVARATIO value that is too far from its threshold indicates an outlier whose elimination can improve the refinement, while a FVARATIO value that is close to unity indicates that the refinement is insensitive (or even detrimental to the quality of the estimates) with respect to the elimination of the outlier reflection.

It should be noted that, in all of the cases tested, this procedure was able to detect all the artificially introduced outliers.

The results of each refinement were evaluated by a measure of the discrepancy of all the atomic coordinates of the structure compared with those of the reference model (*i.e.* the refinement on theoretical data with added noise) such as $\sum_j |x_j^{\rm true} - x_j^{\rm calc}|/p$ (where $x_j^{\rm true}$ is the value of the *j*th atomic coordinate of the reference model and $x_j^{\rm calc}$ is the value of the *j*th atomic coordinates of the refinement of the model based on data containing outliers), together with an analysis of the behaviour of the estimate of some selected variables for each case. The history of the iterative elimination of outliers carried out using the above-mentioned statistics (*i.e.* the list of reflections with aberrant |DFFITS|, COVRATIO and FVARATIO found at each cycle) has been deposited.[1]

4.1. The case of the CaTiO$_3$ perovskite

The experimental X-ray data set of CaTiO$_3$ consists of 1737 unique reflections from an orthorhombic crystal [space group *Pnma*, a = 5.4473 (1), b = 7.6477 (1), c = 5.3825 (1) Å]. Initial crystal-structure refinement on $|F_o|$ was carried out using a locally modified version of the program *XD* (Koritsanszky *et al.*, 1995), using all the reflections up to a reciprocal resolution of $[\sin(\theta)/\lambda]_{\max}$ = 1.22 Å^{-1}. The refinement of the model using a theoretical data set with added noise gave final R = 0.0058, R_w = 0.0086, GoF = 1.012.

In Table 1, the indices of the reflections, their original intensities and the new values are reported for perovskite, together with the leverage calculated before altering the data. In the present case, suitable thresholds for COVRATIO are 0.95 and 1.05, and 1.04 for FVARATIO. The thresholds for Cook's distance and |DFFITS| are 0.024 and 0.19, respectively: the reflections showing a |DFFITS| value >0.19 and a COVRATIO value outside the range limited by the thresholds are thus recognized as outliers that can affect the results. |DFFITS| and Cook's distance showed the same behaviour in this case.

Fig. 1 shows |DFFITS| values against the leverage for the inorganic sample. As can be seen, all of the artificially introduced outliers can be recognized when a comparison is made with the thresholds. In this case, all of them have been detected in the first run. In other runs for this structure using different weights, only the strongest artificially modified reflections were recognized as outliers by the estimator, though a number of 'apparent' outliers (for instance, 200, 123, 242), which are still strong reflections, were also found. Moreover, in certain cases the weakest reflections are not recognized as outliers by DFFITS in the first run. This behaviour strongly depends, for instance, on the weighting scheme introduced, since design matrix, variance estimation and outlier diagnostics depend on weights. Moreover, results are obviously influenced by the amount of noise added, by the criterion for choosing the reflection, by the discrepancy between theoretical and modified values and so on. In all of the cases tested for this work, the simultaneous deletion of outliers yielded results that were practically identical to those obtained by iterative elimination of the aberrant reflections. However, this behaviour is probably due to both the nature of synthetic data and the low number of artificial outliers introduced. Therefore, caution in simultaneously eliminating outliers is strongly recommended when dealing with experimental data, since the person carrying out the experiment has no *a priori* knowledge about the behaviour of the system under study.

The iterative elimination of reflections 040, 004, 202, 121, 002, 042, 0,14,2 and 2,14,0 always exhibits the presence of outliers with COVRATIO significantly outside the range limited by cut-offs and FVARATIO > 1.05, whereas the outliers reflection 1,16,1 cannot be considered as an 'extreme'

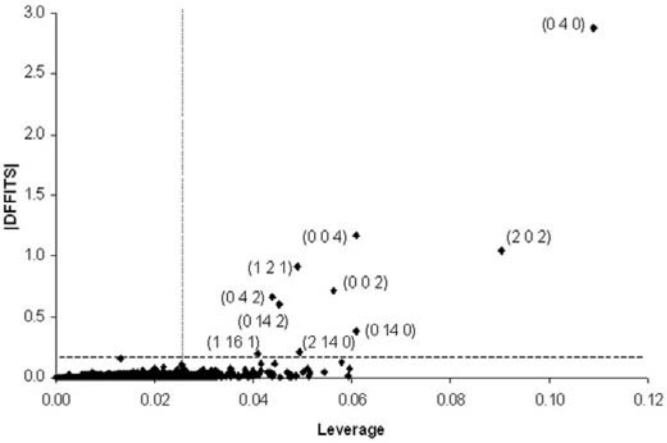

Figure 1
|DFFITS| *versus* leverage for CaTiO$_3$ perovskite. Dotted line = leverage cut-off; dashed line = |DFFITS| cut-off.

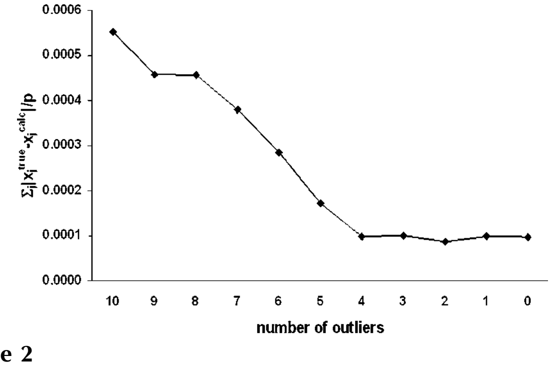

Figure 2
$\sum_j |x_j^{\rm true} - x_j^{\rm calc}|/p$ for CaTiO$_3$ perovskite. The number of outliers present in the data set is indicated on the abscissa.

[1] Output of leverage analysis for perovskite, loganin and oxalic acid are available from the IUCr electronic archives (Reference: SH5030). Services for accessing these data are given at the back of the journal.

outlier, as was the case with the former, since it never exhibits aberrant COVRATIO/FVARATIO values. The elimination of outliers can be stopped once FVARATIO is close to the fixed threshold. Further elimination is worthless and could even be unwise in certain cases.

In Fig. 2, the overall estimator of the discrepancy between the coordinates is plotted for the 10 outliers in the first run in the data set, and for runs ii–xi) after removing the highest outlier (040, 004, 202, 121, 002, 042, 0,14,2, 0,14,0, 2,14,0, 1,16,1, respectively).

As can be noticed, there is a significant improvement in the estimates after deleting the outliers one at a time, in terms of the correctness of the model. Cases (i) to (vi) refer to the strongest aberrant reflection and cases (vii) to (xi) refer to the elimination of the weakest outlier reflections.

In the perovskite case, COVRATIO ranges from 0.400 to 0.945 for the outlier reflections, whereas all of the other values are really close to unity. Such statistics allow us to explain other features of the results, as will be described in the following paragraph.

Fig. 3 depicts the values of U_{11} for the Ti atom. As can be seen, there is an improvement of the estimate of this variable in comparison with the reference model after each iterative elimination of the most aberrant reflection, as indicated by the diagnostics [cases (i) to (vi)]. At the same time, however, the elimination of the weakest outlier reflections does not improve the results [cases (vii) to (xi)]. Similar considerations can be made in the case of an atomic coordinate, *i.e.* the z coordinate of Ca (Fig. 4).

In our experience, in inorganic structure refinements (with a spherical-atom model), the most significant improvement of the estimates, using the diagnostics described above, is for the thermal parameters, owing to the systematically higher leverage of most of the reflections on this kind of variable (Merli *et al.*, 2000).

4.2. The case of loganin

For the loganin structure, the initial crystal structure refinement on $|F_o|$ up to a reciprocal resolution of $[\sin(\theta)/\lambda]_{\max}$ = 0.67 Å^{-1} was carried out on 3498 unique reflections for 340 variables. After adding noise to the theoretical structure factors as described above, the refinement gave final R = 0.0623, R_w = 0.0969, GoF = 1.009.

Suitable cut-offs for this structure were 0.35 for DFFITS, 0.72 and 1.38 for low and high COVRATIO cut-offs, respectively, and 1.20 for FVARATIO.

The history of the iterative elimination has been deposited. After eliminating reflection 410, there were no more reflections with FVARATIO > 1.20. Therefore, any further elimination of reflections was not necessary in terms of overall

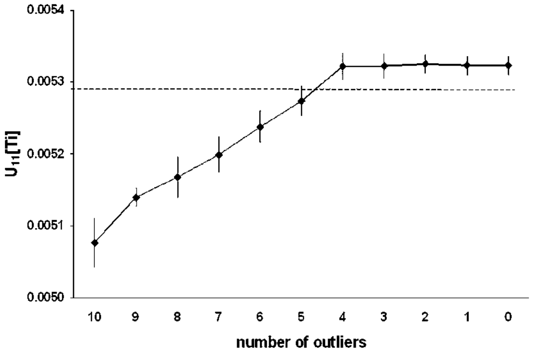

Figure 3
U_{11} (Å^2) of Ti in CaTiO$_3$ perovskite. Dashed line = value of the variable in the model obtained from data with added noise and no outliers present; labels on abscissa as in Fig. 2.

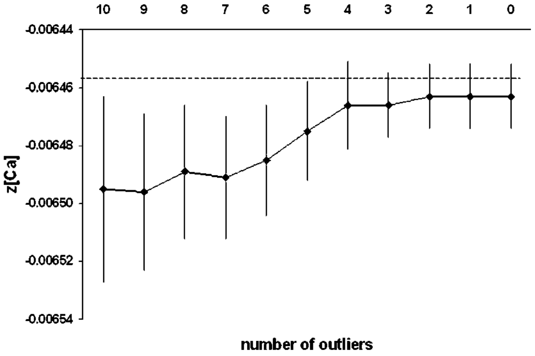

Figure 4
z fractional coordinate of Ca in CaTiO$_3$ perovskite. Dashed line = value of the variable in the model obtained from data with added noise and no outliers present; labels on abscissa as in Fig. 2.

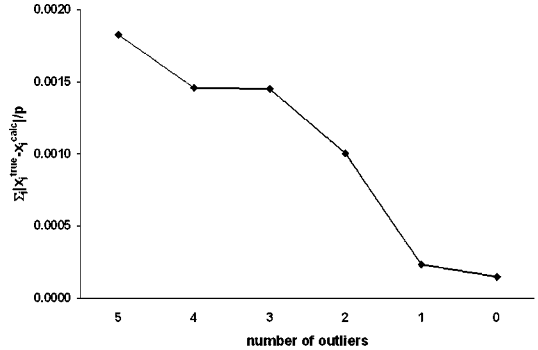

Figure 5
$\sum_j |x_j^{\rm true} - x_j^{\rm calc}|/p$ for loganin *versus* the number of outliers in the data set.

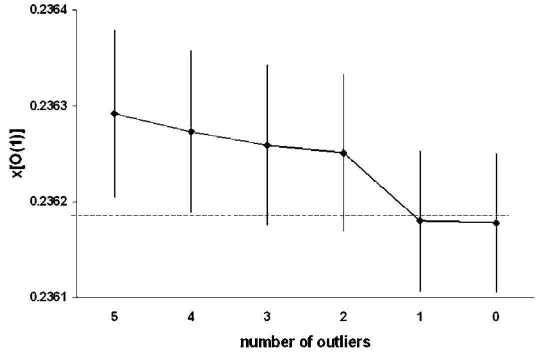

Figure 6
x fractional coordinate of O(1) in loganin. Dashed line = value of the variable in the model obtained from data with added noise and no outliers present; labels on abscissa as in Fig. 5.

improvement of the refinement and could even have slightly impaired some variables.

Similarly to the perovskite case, these results can be visualized in terms of the mean coordinate error (Fig. 5) and with respect to each estimate [as an example, the *x* fractional coordinate of O(1) is plotted in Fig. 6] in comparison with individual iterative elimination of 040, 200, 410, $\bar{7}12$ and 2,13,1, respectively. As can be seen, the overall departure from the reference model gradually decreases after the elimination of the outliers, as evidenced by the estimate of the *x* coordinate of O(1) plotted in Fig. 6.

4.3. The case of oxalic acid dihydrate

In the last test, the oxalic acid dihydrate structure is presented (see *XD* documentation for further details on the structure). In this case, 3 scale factors were refined for 3 subsets of reflections. The initial crystal-structure refinement was carried out on 3498 experimental unique reflections, up to a reciprocal resolution of $[\sin(\theta)/\lambda]_{\max}$ = 0.99 Å^{-1} for 66 variables.

After adding noise to the theoretical structure factors as described above, the structure refinement gave final R = 0.0131, R_w = 0.0146, GoF = 1.088. Up to 5 reflections with medium–high leverage were altered within a range of ±10% as indicated in Table 3. Thresholds for the diagnostics were 0.25 for |DFFITS|, 0.89 and 1.12 for low and high COVRATIO cut-offs, and 1.08 for FVARATIO.

The results of the iterative leverage analysis, following the removal sequence 103, $10\bar{5}$, 200, $11\bar{3}$ and $30\bar{3}$, has been deposited. Even in this case, FVARATIO is the diagnostic tool that can reliably indicate the outliers that are capable of impairing the estimates. It is worth saying that, for this structure, there is no outlier reflection with COVRATIO < 1, *i.e.* there are no reflections for which the results are definitely improved after their elimination: given the noise added to this data set, in this structure, all of the reflections have more or less equal influence on the estimates. After the elimination of the $30\bar{3}$ reflection, the leverage analysis does not indicate further dangerous outliers. Figs. 7 and 8, as well as the previous cases, show the improvement of the refinement by means of the recognition and subsequent elimination of these outliers.

5. Conclusions

It should be remembered that the simulations presented here represent only ‘ideal’ cases with just a few aberrant reflections. Besides, the ‘minor outliers’ introduced in the data sets can only affect the results within the statistical fluctuations, so the presented examples only represent a further exploratory investigation into the subject. In common practice, experimental data sets are affected by a greater number of aberrant reflections and the disparity between observed and calculated data should be ascribed either to experimental bias or to the model’s incongruity, or to both. Further studies about these facts need to be carried out, alongside tests of other diagnostic tools. Nevertheless, it could be stated that the elimination of the outliers of a crystal-structure refinement is a strongly recommended procedure that can improve the reliability of results, where both data fitting and the precision and accuracy of the estimated variables are concerned.

This task has to be accomplished using several sets of statistics, all of them based on leverage. A promising algorithm to detect the outliers one at a time could be the following one.

1. Detect the outliers by means of distance measures, such as Cook’s distance or |DFFITS|, provided that there are suitable ‘size-adjusted’ thresholds.

2. Pick the reflection with the highest Cook’s distance or |DFFITS| as the extreme outlier and check for its FVARATIO values by comparing it to the appropriate thresholds.

3. Delete the reflection if FVARATIO is significantly distant from the threshold. In the presence of a large number of outlier reflections, a simultaneous elimination of them can be roughly adopted when very large values of |DFFITS| or Cook’s distance are observed. In strongly biased experimental data, the elimination of an ‘untrue’ outlier is likely to bias the estimates in a negligible way.

4. Continue until all of the diagnostic tools lie within the fixed thresholds.

These procedures should be carried out routinely to improve the results of refinement, thus avoiding indiscriminate truncation of data and empirical protocols that are not corroborated by the statistics.

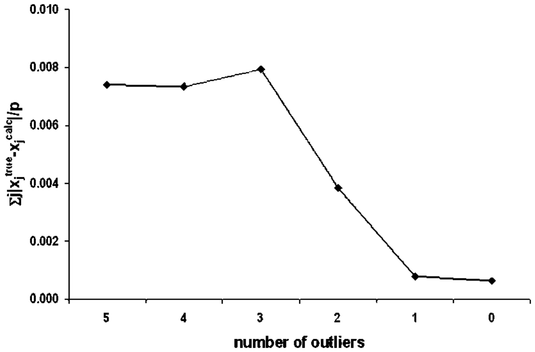

Figure 7
$\sum_j |x_j^{\rm true} - x_j^{\rm calc}|/p$ for oxalic acid *versus* the number of outliers in the data set.

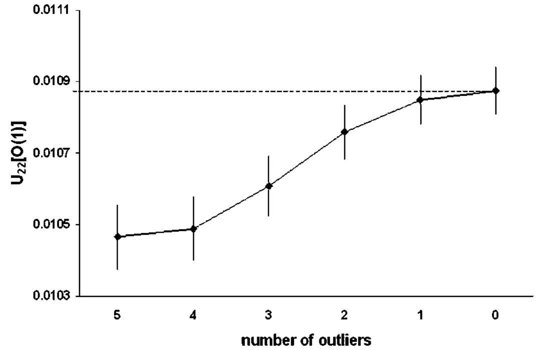

Figure 8
U_{22} (Å^2) of O(1) in oxalic acid. Dashed line = value of the variable in the model obtained from data with added noise and no outliers present; labels on abscissa as in Fig. 7.

APPENDIX *A*
Diagnostic criteria: formulae and thresholds

The reader can refer to Belsey *et al.* (1980) for a further explanation of the expressions reported here.

A way of summarizing coefficient changes and changes in fit when an observation is deleted is given by

$$\mathrm{DFFIT}_i = \frac{h_i e_i}{(1-h_i)}.$$

In this work, a measure of DFFIT scaled by $sh_i^{1/2}$ has been adopted, as defined below:

$$\mathrm{DFFITS}_i = \left(\frac{h_i}{1-h_i}\right)^{1/2} \frac{e_i}{s_i'(1-h_i)^{1/2}}.$$

A suitable adjustable threshold for DFFITS can be $F_{p-1,n-p}(p/n)^{1/2}$.

Note that an equivalent expression for DFFITS is given by

$$\mathrm{DFFITS_i} = |e_i^*|\left(\frac{h_i}{1-h_i}\right)^{1/2},$$

where e_i^* is the studentized deleted residual, defined as

$$e_i^* = \frac{e_i}{s_i'(1-h_i)^{1/2}}.$$

Cook (1977, 1979) proposed a measure of the distance between the observed and the calculated points given by

$$D_i = \frac{(\mathbf{x}_i' - \mathbf{x})^{\mathrm{T}} A^{\mathrm{T}} A (\mathbf{x}_i' - \mathbf{x})}{ps^2} = \frac{1}{p}\frac{e_i^2}{s^2(1-h_i)}\frac{h_i}{(1-h_i)},$$

where $\mathbf{x}_i'$ is the least-squares estimate of $\mathbf{x}$ computed without the ith case and $s^2 = \mathbf{e}^{\mathrm{T}}\mathbf{e}/(n-p)$ is the usual estimate of the variance which can alternatively be substituted by the deleted s_i', as used before. Thus, D_i can be considered as the standardized squared distance between the parameter estimate $\mathbf{x}$ and the parameter estimate $\mathbf{x}_i'$ when the ith case is removed. The threshold used in this work for Cook's distance is $F_{p-1,n-p}(p/n)$.

It should be noted that, in the crystal structure refinement code *MOLLY* (Hansen & Coppens, 1978), the calculation of DFFITS and Cook's distance has been implemented (Kuntzinger *et al.*, 1998).

A measure of the ratio between the variance–covariance matrix when the ith row has been deleted and the covariance matrix using all the data (further explanations can be found in Belsey *et al.*, 1980) is represented by the following criterion:

$$\mathrm{COVRATIO}_i = \frac{1}{\left[\frac{n-p-1}{n-p} + \frac{e_i^{*2}}{n-p}\right]^p (1-h_i)}$$

or, alternatively, by

$$\mathrm{FVARATIO}_i = \frac{(s_i')^2}{s^2(1-h_i)},$$

which is related to COVRATIO since

$$\mathrm{FVARATIO} = \left[\frac{\mathrm{COVRATIO}}{(1-h_i)}\right]^{1/p}(1-h_i).$$

COVRATIO can range in the interval $1/[1+3/(n-p)]^p(1-2p/n)$ to $1/[1+3/(n-p)]^p$, thus reflections with values outside this interval should be considered as outliers. For large systems, the low and high cut-offs for COVRATIO can be approximated as $1 - 3p/n$ and $1 + 3p/n$, respectively, while FVARATIO can range in the interval $1 - 3/n$ to $1 + (2p + 3)/n$.

Financial support was obtained from MIUR (FURD 2004–2006). The author is grateful to Dr M. Boiocchi for his help.

References

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

Burla, M. C., Camalli, M., Carrozzini, B., Cascarano, G. L., Giacovazzo, C., Polidori, G. & Spagna, R. (2003). *J. Appl. Cryst.* **36**, 1103.

Cook, R. D. (1977). *Technometrics*, **19**, 15–18.

Cook, R. D. (1979). *J. Am. Stat. Assoc.* **74**, 169–174.

Gray, J. B. & Ling, R. F. (1984). *Technometrics*, **26**, 305–318.

Hansen, N. K. & Coppens, P. (1978). *Acta Cryst.* A**34**, 909–921.

Harris, G. W. & Moss, D. S. (1992). *Acta Cryst.* A**48**, 42–45.

Hazen, R. M. & Finger, L. W. (1989). *Am. Mineral.* **74**, 352–359.

Koritsanszky, T., Howard, S. T., Richter, T., Mallinson, P. R., Su, Z. & Hansen, N. K. (1995). *XD – a Computer Program Package for Multipole Refinement and Analysis of Charge Density from Diffraction Data.* Free University of Berlin, Germany.

Kuntzinger, S., Ghermani, N. E., Dusausoy, Y. & Lecomte, C. (1998). *Acta Cryst.* B**54**, 819–833.

Lunin, V. Y., Afonine, P. V. & Urzhumtsev, A. G. (2002). *Acta Cryst.* A**58**, 270–282.

Merli, M. (2002). *Z. Kristallogr.* **217**, 103–108.

Merli, M., Oberti, R., Caucia, F. & Ungaretti, L. (2001). *Am. Mineral.* **86**, 55–65.

Merli, M., Ungaretti, L. & Oberti, R. (2000). *Am. Mineral.* **85**, 532–542.

Pannu, N. S. & Read, R. J. (1996). *Acta Cryst.* A**52**, 659–668.

Prince, E. & Boggs, P. T. (1992). *International Tables for Crystallography*, Vol. C, edited by A. J. C. Wilson, p. 594. Dordrecht: Kluwer Academic Publishers.

Prince, E. & Nicholson, W. L. (1985). *Structure and Statistics in Crystallography*, edited by A. J. C. Wilson, p. 183. New York: Adenine Press.

Rao, C. R. (1973). *Linear Statistical Inference and its Applications.* New York: John Wiley and Sons.

Schwarzenbach, D., Abrahams, S. C., Flack, H. D., Gonschorek, W., Hahn, Th., Huml, K., Marsh, R. E., Prince, E., Robertson, B. E., Rollett, J. S. & Wilson, A. J. C. (1989). *Acta Cryst.* A**45**, 63–75.

Schwarzenbach, D., Abrahams, S. C., Flack, H. D., Prince, E. & Wilson, A. J. C. (1995). *Acta Cryst.* A**51**, 565–569.

Seaver, B., Triantis, K. & Reeves, C. (1990). *Technometrics*, **41**, 340–351.

Spagna, R. & Camalli, M. (1999). *J. Appl. Cryst.* **32**, 934–942.

Watkin, D. (1994). *Acta Cryst.* A**50**, 411–437.