Acta Crystallographica Section A
Foundations and Advances
ISSN 2053-2733

Received 20 December 2013
Accepted 4 June 2014

More about systematic errors in charge-density studies

Julian Henn[a]* and Kathrin Meindl[b]

[a]Emil-Warburg-Weg 6, 95447 Bayreuth, Germany, and [b]Instituto de Biología Molecular de Barcelona (IBMB-CSIC), Barcelona Science Park, Baldiri Reixach 15, 08028 Barcelona, Spain. Correspondence e-mail: julian.henn@uni-bayreuth.de

In order to detect and graphically visualize the absence or presence of systematic errors in fit data, conditional probabilities are employed to analyze the statistical independence or dependence of fit residuals. This concept is completely general and applicable to all scientific fields in which model parameters are fitted to experimental data. The applications presented in this work refer to published charge-density data.

1. Introduction

In two preceding publications, the theoretical R value (Henn & Schönleber, 2013) and a meta residual factor $R^{\rm meta}$ (Henn & Meindl, 2014), which employs the theoretical R value, were developed. The concept of a theoretical R value and a meta residual factor is completely general and applicable to all fields where least-squares fits are conducted.[1] Application to crystallographic standard structures and high-resolution charge-density studies revealed that residual distributions reminiscent of a Gaussian only rarely appear. The application of the meta residual factor in Henn & Meindl (2014) focused on the experimental standard uncertainties (s.u.'s) of the crystallographic data, *i.e.* the s.u.'s of the reflection file were used. These are known to be often not very accurate; therefore an error model may be used, for example with the help of a weighting scheme (see *e.g.* Waterman & Evans, 2010). This should result in more appropriate values for the measurement errors. We abbreviate the estimations for the measurement errors derived from an error model with $\hat{\sigma}$ to indicate the difference to statistical weights σ. In the present study we repeat this analysis with respect to $\hat{\sigma}$. Furthermore, we present a tool that we have developed for the visualization and analysis of residual distributions: a Gaussian distribution of residuals is a necessary requirement for the validity of a least-squares fit, but it is not sufficient to prove that the refinement is without systematic errors. To prove this, further tools are needed, which are developed in the publication at hand.

2. $R^{\rm meta}$ for weighted residuals

In this section the meta residual value is discussed for the experimental data sets (1–23) that were introduced with references in Henn & Meindl (2014) together with the artificial data sets (24–29) that correspond to refinements with (set Nos. 25, 27, 29) and without (set Nos. 24, 26, 28) cutoff $I_{\rm o} > 0$ for increasingly noisy data. For more details about the artificial data sets see Henn & Meindl (2014). A weighting scheme was employed according to the corresponding cif files in data sets 1, 2, 8–13, 17–19 and 21–23. The other data sets used either weights $w = 1/\sigma^2(F_{\rm o}^2)$ or $w = 1/\sigma^2(F_{\rm o})$ corresponding to $\hat{\sigma}(F_{\rm o}^2) = \sigma(F_{\rm o}^2)$ and $\hat{\sigma}(F_{\rm o}) = \sigma(F_{\rm o})$, respectively.

A standard weighting scheme applied in charge-density studies has the form $\hat{\sigma}(F_{\rm o}^2) = [\sigma^2(F_{\rm o}^2) + (aP)^2 + bP]^{1/2}$ with $\sigma^2(F_{\rm o}^2)$ from the reflection file and the free parameters a and b as well as $P = fF_{\rm o}^2 + (1-f)F_{\rm c}^2$, where the free parameter f determines to what extent the error model refers to the calculated intensities $F_{\rm c}^2$ (Volkov *et al.*, 2006).

Fig. 1 shows the difference in percentage points between actual and predicted weighted ($\hat{\sigma}$-based) R values (Fig. 1*a*) and the corresponding values of $R^{\rm meta}$ (Fig. 1*b*) in red. Additionally, the corresponding σ-based values are shown in gray and are connected with a dashed line.

The absolute difference between actual and predicted $\hat{\sigma}$-based R values (Fig. 1*a*) is positive for all experimental data sets, as was the case in the σ-based analysis. This may be a hint that the $\hat{\sigma}$ values are still too small or that systematic errors are present. The artificial data sets Nos. 25, 27 and 29 with cutoff $I_{\rm o} > 0$ show slightly negative values, indicating overfitting, whereas those without cutoff virtually lead to meta residual values of zero. The strongest absolute decrease in actual and predicted R values is for data set 13, the anharmonic nuclear motion multipole model refinement of the 298 K data set of the explosive RDX. The relative change is distinct for the whole series 8–13 of anharmonic and harmonic refinements at different temperatures. These small absolute and large relative changes have a strong impact on the meta residual value for data sets 8–13 ($R^{\rm meta}$ = 12.4, 9.9, 10.3, 13.1, 16.3 and 5.4%), which are now closer to $R^{\rm meta}$ = 8.7% from data set 2. Also data sets 17–19, which are all from one publication, show a distinct decrease in $R^{\rm meta}$ from 91.4, 82.0 and 80.9% to 57.8, 16.3 and 16.0%, respectively. The $R^{\rm meta}$ values are much lower

[1] Actually, applicability is given in all cases where a Gaussian distribution of fit residuals is expected. This applies also for example to certain maximum-entropy methods. The concept of theoretical R values is furthermore easily extended to cases in which another distribution is expected, provided the distribution is known in advance.

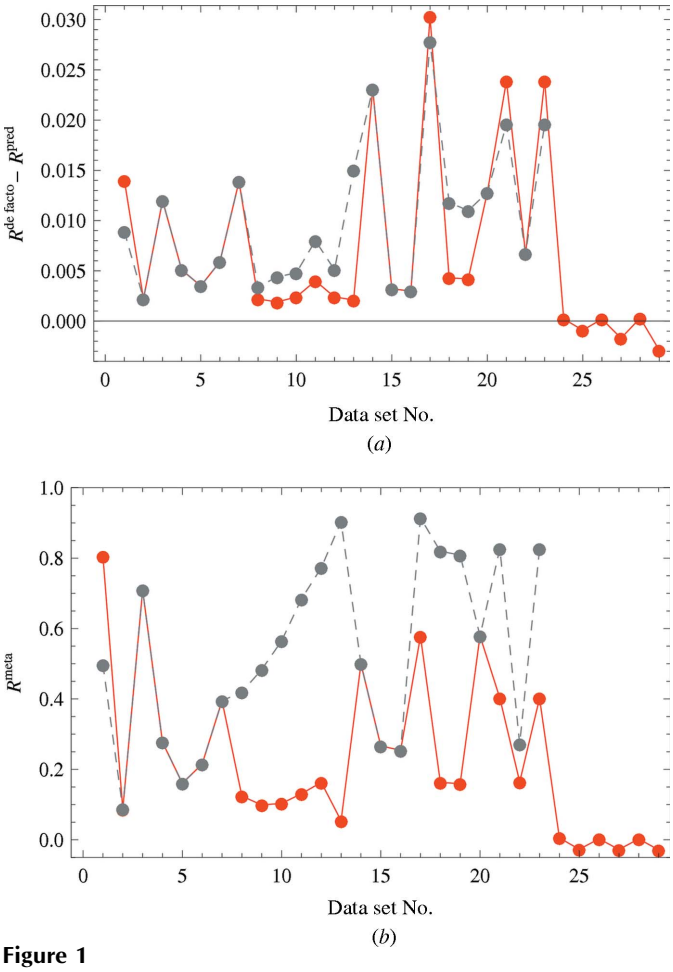

Figure 1
(*a*) Difference in percentage points between actual and predicted weighted R values for $\hat{\sigma}$-based R values (red) and s.u.-based R values (gray). (*b*) Systematic error as given by R^{meta} for $\hat{\sigma}$-based R values (red) and s.u.-based R values (gray).

for the multipole models (sets 18 and 19), indicating a lower degree of contamination with systematic errors as compared to the IAM (independent-atom model) (set 17). The decrease is more prominent for the multipole-model refinements than for the IAM refinement. The same error model was used for sets 17–19.

3. Statistical independence of residuals

After a successful model refinement without systematic errors, the residuals are true random numbers; therefore they are not systematically related to the calculated intensities (I_c), to the standard uncertainties (s.u.'s), the resolution, or to the diffractometer the data were acquired from. This innocuous statement turns into a powerful tool when it is taken as stimulation to *search* for systematic connections. If a systematic relation can be established, this disproves the residuals from being statistically independent. If no relation is found this just proves that with the chosen method a relation cannot be established. When no search for systematic relations is conducted, there will probably be no systematic connections detected. This is the state of affairs in current charge-density studies. As was found earlier, even the easy-to-apply normal probability plots – a minimum requirement for the evaluation of the fit quality – are only rarely used (Henn & Meindl, 2014). In the following paragraphs the connection(s) between residuals $\zeta = (I_o - I_c)/\hat{\sigma}(I_o)$ and observed intensities, calculated intensities, reduced standard uncertainties $\hat{\sigma}$ and resolution are examined with the help of conditional probabilities.

4. Conditional probabilities

The question to be answered is: are the residuals statistically independent of the calculated intensities, the resolution and s.u.'s? As mentioned above, statistical independence implies that any (equally strong and sufficiently large) independent subsets do not show a systematic relation. This is further investigated in the following.

The conditional probability $p_B(A)$ of observing property A, given the condition B with non-vanishing probability $p(B) > 0$, is given by Bayes theorem:

$$p_B(A) = \frac{p_A(B)p(A)}{p(B)}, \tag{1}$$

where the term in the enumerator represents the probability of observing A and B simultaneously $p(AB) = p_A(B)p(A)$.

When the properties A and B are independent, the conditional probability of observing property B, given that property A has already been observed, is just the probability of observing B

$$p_A(B) = p(B), \tag{2}$$

and *vice versa*:

$$p_B(A) = p(A). \tag{3}$$

In this case of independence, the probability that A and B are observed simultaneously, $p(AB) = p_A(B)p(A)$ factorizes in the individual probabilities of observing A and of observing B. This is then also equal to the probability of observing simultaneously B and A:

$$p(AB) = p(A)p(B) = p(BA). \tag{4}$$

This last equation becomes particularly simple when A and B are chosen from percentiles. For example, what is the probability of finding a residual from the last decile (which just means a residual larger than 90% of all residuals), given the s.u. is also from the largest decile of standard uncertainties? The individual probabilities are in both cases 0.1 (as deciles were chosen); therefore, if their combined probability is significantly different from 0.01, it is concluded that they are not statistically independent. If their combined probability is close to 0.01, it *cannot* be concluded that they are statistically independent, as this value may appear accidentally and it may be different for other parts of the probability space, for example when the fourth decile of residuals and the eighth decile of s.u.'s were chosen. In order for A and B to be statistically independent, they have to have a value close to 0.01 for *any* combination of deciles. Deviations from this value are then pure statistical fluctuations.

In this case of statistical independence, large residuals do *not* tend to stem from strong observations, but they appear equally likely from strong and from weak reflections. The reader who wants to object here that large residuals may only appear for large values of the intensity is reminded that the difference $(I_o - I_c)$ is scaled by the respective s.u.: $\zeta = (I_o - I_c)/\text{s.u.}$ Here, the term s.u. refers to what has been used in the least-squares refinement, either $\hat{\sigma}(I_o)$ or $\sigma(I_o)$. Therefore, when the s.u.'s are accurate in relation to each other and the model is fully adequate, the magnitude of the residuals will not show any trend, they will be smoothly distributed over the whole range of *e.g.* calculated intensities, s.u.'s, or the resolution.

These conditional probabilities can be visualized as a plot in the unit square: every point in the unit square represents the appearance of a certain combination of residual value and, *e.g.*, s.u. value. In the case of statistical independence, no combination is preferred over the other and the density of points for every sufficiently large area is approximately the same for all areas. The density of points gives the conditional probability. We call these plots *Bay*esian *Con*ditional probability plots, in short BayCoN plots. The principle is explained in the following paragraph for a BayCoN plot of residuals ζ *versus* s.u.'s, in short notation BayCoN(ζ, s.u.) or just (ζ, s.u.).

In order to construct this plot, a list of N_{ref} residuals and corresponding s.u.'s was sorted in ascending order of s.u.'s. Each s.u. value was then replaced by its ranking number, starting with one for the smallest s.u. and ending with N_{ref} for the largest s.u. The list was then sorted in ascending order of the residuals, and the actual residual values were replaced by their ranking number. Both columns of integer positive ranking numbers were divided by the number of reflections N_{ref}. In this way, N_{ref} pairs of numbers, each between 0 and 1, are obtained. This pair of numbers is interpreted as a pair of coordinates and plotted in a unit square. From this construction it follows that the same number of points will be found in each horizontal or vertical strip of the same width. For example in a strip between 0.0 and 0.1, which corresponds to the lowest decile, there will be $N_{ref}/10$ points. If there is no systematic interrelation between the residuals and the s.u.'s, the plot will show the same density of points in each sufficiently large chosen region. As a consequence of using ranking numbers, the plots are invariant under transformations which do not change the ranking number, for example a plot of residuals *versus* s.u.'s, (ζ, s.u.), will – apart from possible numerical issues – look exactly the same as a plot of residuals *versus* variances [ζ, (s.u.)^2], or a plot (ζ, $\sin\theta/\lambda$) will look equal to a plot [ζ, $(\sin\theta/\lambda)^2$].

A short note about the uniqueness of the results is appropriate: unambiguous results are obtained only in the case that the numbers of each subset are all different, *i.e.* all values of ζ are different from each other and all values of s.u.'s are different from each other as only in this case a unique order is established. The more identical numbers that are found in each column, the higher the number of possible outcomes, and each different result is equally valid. In this case of repeated values the result is also dependent on the sequence of ordering: a BayCoN plot (ζ, s.u.) may then look fundamentally different from a plot (s.u., ζ), which is not desirable. Problems of this kind may occur when the s.u. values in the reflection file are given only to two digits and when the reflection file is large. We regard these effects as a technical and numerical problem, and not as a methodological problem, as in crystallographic applications real random numbers appear in the intensities and s.u.'s due to the quantum nature of the beam. Therefore, if these numbers are given to a sufficient accuracy it is almost certain that they are all different.

Techniques based on ranking numbers have been known for a long time. In the literature for instance it is described how to generate a rank order correlation scatter diagram with punched cards (Bradley, 1963). It is, however, also important to note that the conditional probability approach as proposed here has to be seen in a more general theoretical framework.

4.1. Application to artificial data

The first application is on the artificial data set No. 24 in order to test the hypothesis of uniform distributions in the case of random residuals. The data correspond to a type 1 consistency data set with 1% noise in the intensities and additional background noise, $p_1 = 0.01$, $p_2 = 1.5$ in s.u. = $p_1 \times I_o + p_2$ with $I_o = F_o^2$ [see equation (22) and text in Henn & Meindl (2014)]. All reflections were used in the refinement. Figs. 2(*a*), 2(*b*) and 2(*d*) depict the (ζ, s.u.), (ζ, I_c) and (ζ, $\sin\theta/\lambda$) plots; they appear to be uniform by visual inspection, which is confirmed by a χ^2_S test, whereas plot Fig. 2(*c*), (ζ, I_o), obviously does not show a uniform distribution of points in the unit square.

To quantify the degree of uniformity of the distributions, each of these was subjected to a χ^2 test against the hypothesis of a uniform distribution. The respective data were each collected into 100 bins and the points in each bin were counted, yielding N_i points for the ith bin. As in the case of statistical independence, the points are equally distributed, the expectation value for the number of points in the ith bin is $n_i = N_{ref}/100$. For each data set the corresponding sum

$$\chi^2_S = \sum_{i=1}^{100} \frac{(N_i - n_i)^2}{n_i}$$

was calculated. The minimum requirement for good test statistics is $n_i \geq 5$ (Semendjajew *et al.*, 2012). The threshold value for rejecting the hypothesis that the points are uniformly distributed at a 0.001 level of significance is approximately 149. The sums χ^2_S are 116.04 (*a*), 118.49 (*b*), 2705.82 (*c*) and 66.95 (*d*) (see the supporting information[2] for a list of χ^2_S values for data sets 1–29). The assumption of a uniform distribution over the unit square at the given level of significance is not rejected in the cases Fig. 2(*a*), 2(*b*) and 2(*d*), but it must be rejected in the case of Fig. 2(*c*).

[2] Supporting information for this paper is available from the IUCr electronic archives (Reference: EO5030).

The interpretation of Fig. 2(*c*) is as follows. The weakest (decile of) observed intensities, which are negative and are found in a horizontal stripe at the bottom of the plot, do not contribute to the positive residuals that are found in a vertical stripe in the right part of the figure. This is because for negative observed intensities the residuals are necessarily also negative. The second decile of intensities does contribute to positive and negative residuals, but not to the whole range of positive residuals. It is only after approximately the fourth decile of intensities that the whole positive and negative range is covered. In the case of (only) random errors and in the absence of a cutoff, the conditional probabilities of finding, for example, a large residual, given that, for example, the s.u. is large, are the same as finding a small residual given a large s.u. This probability is equal to the probability of finding a large residual given a small s.u. It is also equal to the probability of finding a small residual given a small s.u., that is to say, the distributions of the residuals and

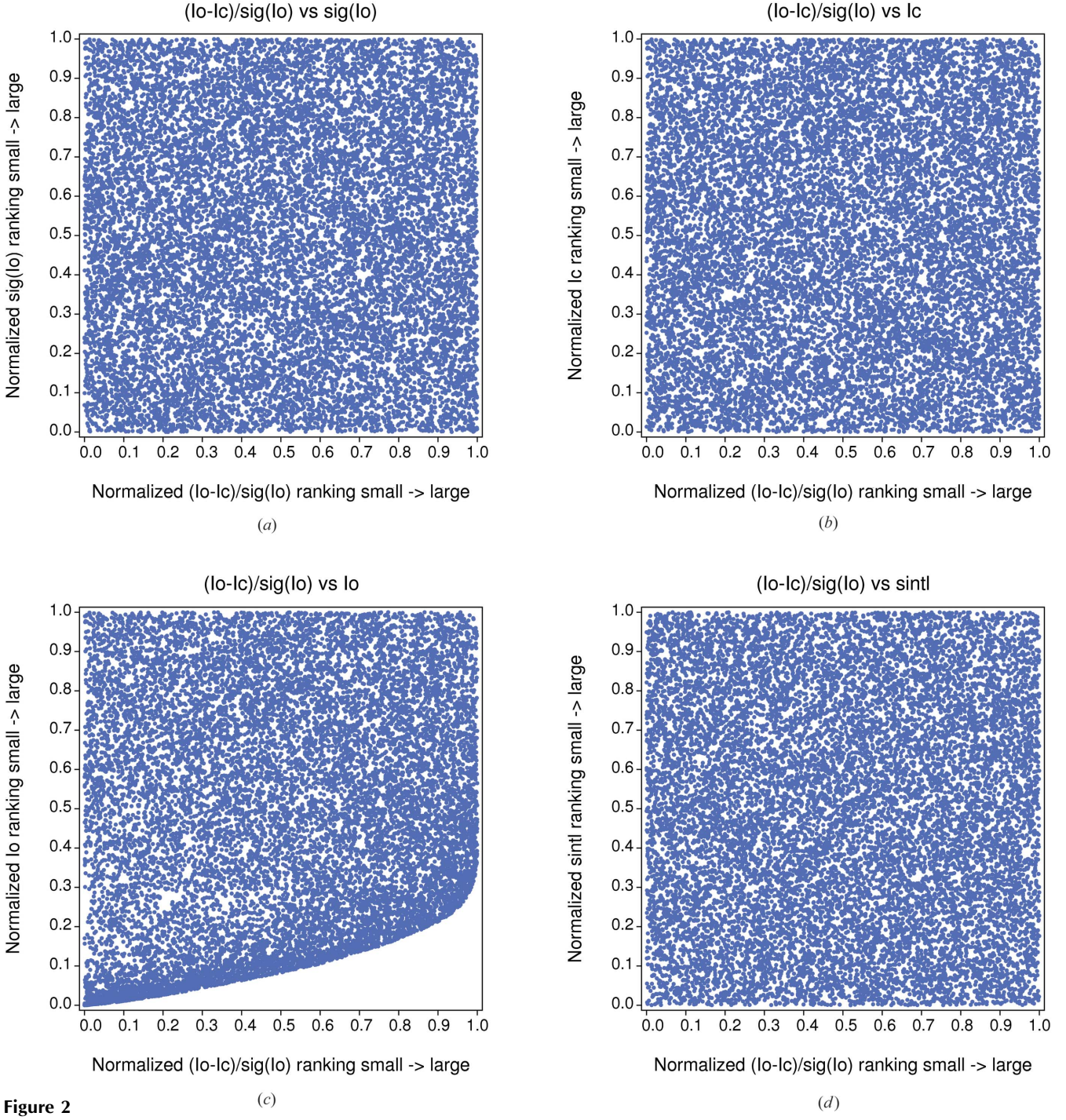

Figure 2
BayCoN plots for the artificial data set No. 24: residuals (*a*) *versus* standard uncertainties, (*b*) *versus* calculated intensities, (*c*) *versus* observed intensities and (*d*) *versus* resolution.

the s.u.'s can be regarded as independent at the given level of significance. If, in the last sentences, the symbol 's.u.'s' is replaced by the symbol 'I_c' or by '$\sin\theta/\lambda$', everything still holds, but if it was replaced by 'I_o', this does not hold. Instead of using the words 'small' and 'large', one could also substitute the terms 'the first (second, third) decile' and 'the eighth, ninth, tenth decile' or similar with median, tercile, quartiles, quintiles and so forth.

In short, under ideal conditions of a refinement against unbiased observed intensities uniform distributions are to be expected for BayCoN plots of (ζ, s.u.), (ζ, I_c) and (ζ, $\sin\theta/\lambda$), but not for (ζ, I_o).

What happens when a cutoff is applied? It was published long ago (Hirshfeld & Rabinovich, 1973), and has been observed in experimental data recently (Henn & Meindl, 2014), that observation and significance cutoffs introduce bias

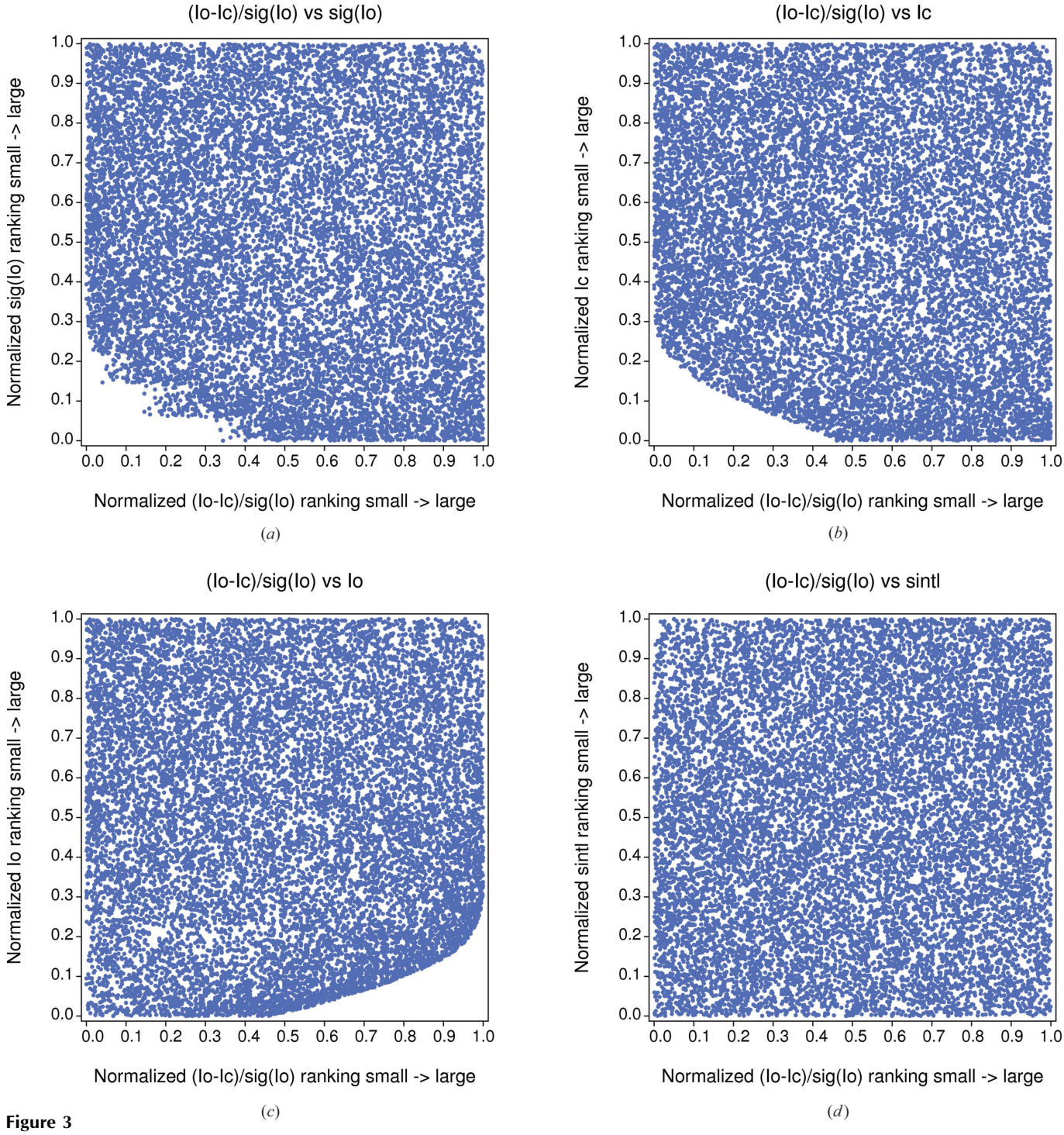

Figure 3
BayCoN plots for the artificial data set No. 25 with cutoff $I_o > 0$: residuals (*a*) *versus* standard uncertainties, (*b*) *versus* calculated intensities, (*c*) *versus* observed intensities and (*d*) *versus* resolution.

and lead to distorted model parameter values. The artificial data set No. 25 results from a refinement of an IAM against exactly the same set of I_o, $\sigma(I_o)$, as in the case of data set No. 24 but with application of an intensity cutoff $I_o > 0$, leading to a rejection of 1219 from originally 14 604 reflections. The corresponding BayCoN plots are shown in Fig. 3.

The (ζ, s.u.) ($\chi^2_S = 827.35$, Fig. 3*a*) and (ζ, I_c) ($\chi^2_S = 928.46$, Fig. 3*b*) distributions show features in the bottom left, whereas the (ζ, I_o) plot ($\chi^2_S = 955.16$, Fig. 3*c*) shows the area of zero point density in the bottom right. Only the distribution of (ζ, $\sin\theta/\lambda$) ($\chi^2_S = 135.25$, Fig. 3*d*) appears to be uniform. Compared to Fig. 2(*c*), the area of zero point density in Fig. 3(*c*) appears reduced. From the corresponding χ^2_S values the assumption of a uniform distribution must be rejected in the three cases (*a*), (*b*) and (*c*).

Fig. 3(*a*) shows that the conditional probability of finding an s.u. value from the lowest decile given that the residual is from the lowest decile (and *vice versa*) is zero. This follows from the area of zero point density in the conditional probability space spanned by the range between 0.0 and 0.1 of the *x* and *y* axes, whereas the product of probabilities is $0.1 \times 0.1 = 0.01$ according to equation (4).

In analogy, Fig. 3(*b*) shows that the larger half of the I_c contributes to all residuals whereas the smaller half of the I_c contributes more to the positive residuals than to the negative residuals due to the area of zero point density, where no contributions are found. As the lowest possible I_c values are zero and the negative observations were omitted, there are no lower I_o values for the lowest I_c values and consequently negative residuals cannot appear for those. Conversely, the positive residuals are obtained from the whole range of I_c, small and large, whereas the negative residuals are obtained with a preference from larger I_c values. Fig. 3(*c*) shows that all observed intensities contribute now to the negative residuals, which are to be found in the left part of the plot, whereas the positive residuals, which are found in the right part, still have a preference for larger observed intensities: the largest decile of residuals receives no contribution from the weakest decile of observed intensities.

When the same model is refined with a significance cutoff $I_o > 3\sigma$, which reduces the number of reflections used in the refinement from 14 604 to 9217, the following results are obtained.

The (ζ, s.u.) (Fig. 4*a*) and (ζ, I_c) (Fig. 4*b*) distributions again show a zero conditional probability area for the lowest deciles, similar to the application of an intensity cutoff $I_o > 0$ in Fig. 3, but in the latter case the *x* axis was met in the middle, whereas now the area of zero point density extends to the far right. This point where the area of plots meets the *x* axis depends on the cutoff chosen: for $I_o > 0$ it is always in the middle (because to the left of this point are only negative residuals and to the right of this point are only positive residuals) and for increasing significance cutoffs this point shifts more and more to the right (data not shown). The (ζ, I_o) (Fig. 4*c*) distribution appears to be uniform, in contrast to the case of an intensity cutoff. The corresponding χ^2_S values are given in Table 1 (second last row). Despite the apparently uniform distributions in Figs. 4(*c*) and 4(*d*), the hypothesis of uniform distributions must be rejected at the 0.001 level of significance.

To summarize the effects studied so far: when no cutoff is applied, the resulting distributions (ζ, s.u.), (ζ, I_c), (ζ, $\sin\theta/\lambda$) are uniform whereas the (ζ, I_o) distribution is not. It was pointed out that this is expected, as for any value of I_c positive and negative residuals may exist, whereas for the negative values of I_o only negative residuals exist; in this respect I_o and I_c behave differently due to the noise that is only part of I_o, but not of I_c. Application of an intensity cutoff leads to non-uniform distributions for (ζ, s.u.) and (ζ, I_c); however, the χ^2_S value for (ζ, I_o) is *reduced* and the value for (ζ, $\sin\theta/\lambda$) increases. Finally, application of a significance cutoff also leads to non-uniformity of (ζ, s.u.) and (ζ, I_c) distributions, whereas the χ^2_S value for (ζ, I_o) is still more reduced. Put simply, the (ζ, I_o) distribution appears to be even *more* uniform than in the cases before.

The χ^2_S values are similar for (ζ, s.u.) and (ζ, I_c) in each of the cases presented in Table 1.

4.2. Application to modified artificial data

What happens when the realized random error in the data is not adequately described by the s.u. values? What happens when other systematic errors are present? Questions of this type can be studied with artificial data that were further modified.

4.2.1. Large standard uncertainty values too small. As an example, we take the case in which the s.u. values of the strong intensities are underestimated. For this purpose the s.u. values of data set 24 are changed according to the transformation

$$\sigma \rightarrow \frac{\sigma}{0.5\,\sigma + 1}, \qquad (5)$$

which leaves very small s.u.'s unchanged and damps larger s.u. values down quickly. The I_o values of data set 24 and the modified s.u. values are combined to data set 30. This data set was used for a refinement with application of a significance cutoff $I_o > 3\sigma(I_o)$. The resulting plots are depicted in Fig. 5.

The transformation [equation (5)] has a strong influence as can be seen from the additional structures in the plots in Fig. 5. The χ^2_S values are 2321.80 (*a*), 2450.17 (*b*), 1383.09 (*c*) and 763.59 (*d*).

An interesting side effect of the transformation [equation (5)] is that the predicted and factual *R* values both *decrease*. In this example, the *de facto* $wR|_{w=1/\sigma^2}$ decreases from approximately 3.6% for the true s.u.'s (set No. 24) to approximately 1% for the modified s.u.'s [and still to 2.1% for choosing a factor of 0.01 instead of 0.5 in the denominator of equation (5)]. The flexibility of the model serves now to describe those reflections with too small s.u.'s. The reduction of *R* values is, however, accompanied by an increase in the goodness of fit, GoF = 3.0, which indicates that (some of) the s.u.'s are too small, and by a distinct increase in $R^{\rm meta}$ from 0.5% (set No. 24) to 67.3% (set No. 30). Multiplication of all s.u. values with a factor of approximately $3^{1/2}$ would increase the predicted *R* value to the level of the *de facto R* value whereas the *de facto R* value itself would remain unchanged as

the refinement was performed with weights $w = 1/\sigma^2$ and the corresponding $wR|_{w=1/\sigma^2}$ factor does not change under such a transformation (Henn & Schönleber, 2013). As a result one would obtain $R^{\rm meta} \simeq 0$. This transformation would also lead to a reduced significance of the data and the s.u.'s of the model parameter would increase. The important point, however, is that this multiplication of s.u. values would not change the χ^2_S values and the plots in Figs. 5(*a*)–5(*d*), as these are based on *ranking numbers* of the analyzed entities and the ranking does not change either under a positive multiplicative transformation. Also, the scaling of s.u. values would not turn a non-Gaussian distribution of residuals (see supporting information, Fig. 2, last row) into a Gaussian distribution of residuals.

It is concluded that the systematic error induced by underestimating the strong s.u. values is severe and is seen in the χ^2_S values, residual distribution, $R^{\rm meta}$ and BayCoN plots.

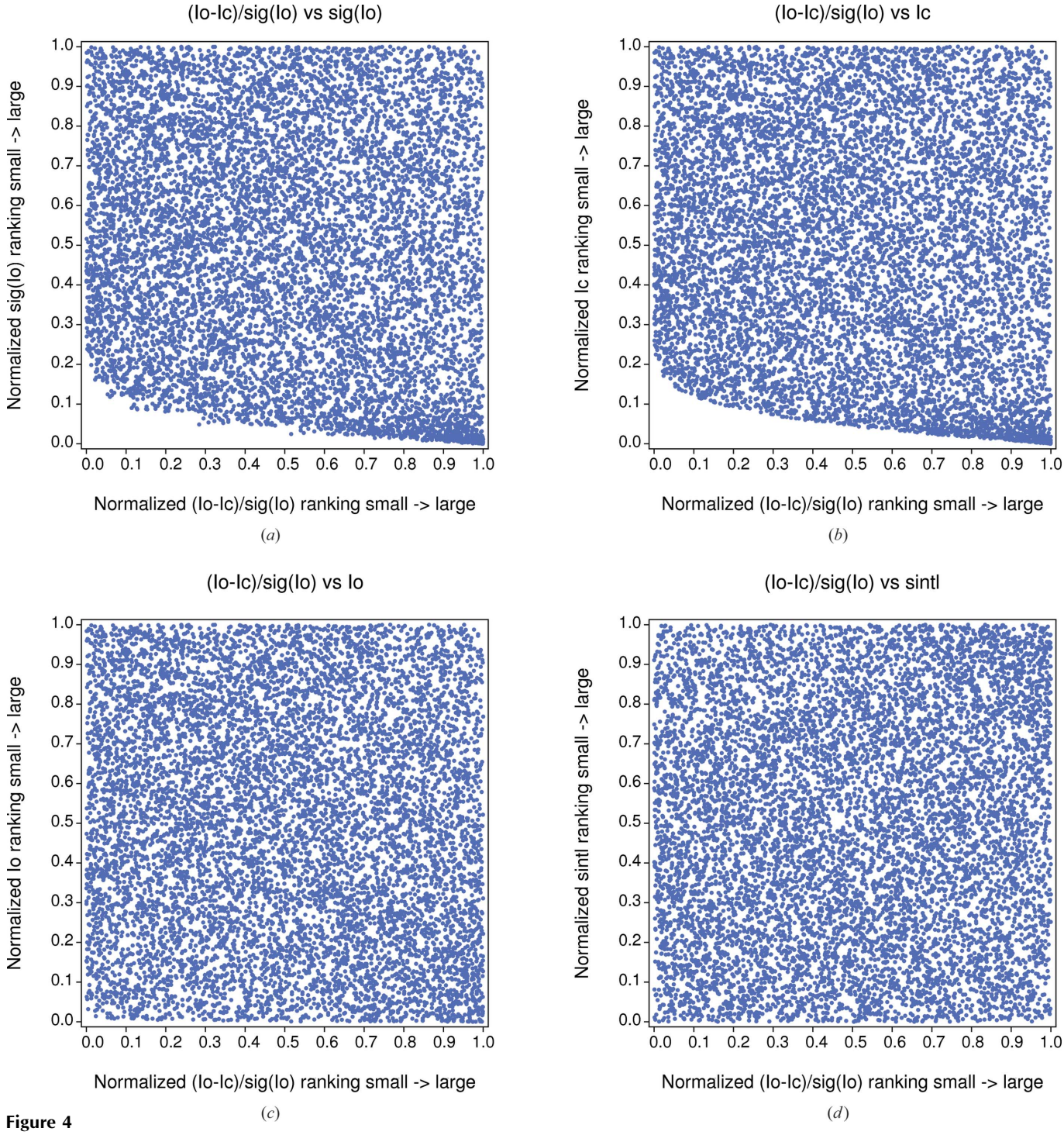

Figure 4
BayCoN plots for the artificial data set No. 24 with application of a significance cutoff $I_o > 3\sigma$: residuals (*a*) *versus* standard uncertainties, (*b*) *versus* calculated intensities, (*c*) *versus* observed intensities and (*d*) *versus* resolution.

These changes also appear for a more moderate transformation of σ values, *e.g.* when a factor of only 0.01 instead of 0.5 is chosen in the denominator of equation (5) (data not shown). However, for illustration purposes the factor 0.5 was chosen. Plots similar to those in Fig. 5 are seen in the application to experimental data, for example for data sets Nos. 3 and 7 that also show the characteristic increase of density of points in the upper corners for the (ζ, I_o), (ζ, I_c) and $(\zeta$, s.u.) plots and in the lower corners for the $(\zeta, \sin\theta/\lambda)$ plot. Also data set No. 9 shows similar features.

Fig. 5(*b*) should be compared to the (ζ, I_c) plots in Fig. 2(*b*), Fig. 3(*b*) and Fig. 4(*b*). The last plot already explains the zero point density area in the bottom of Fig. 5(*b*). There are two areas of higher density of points in the upper left and upper right corner, and in the horizontal stripe connecting these corners the density of points is reduced. This horizontal stripe

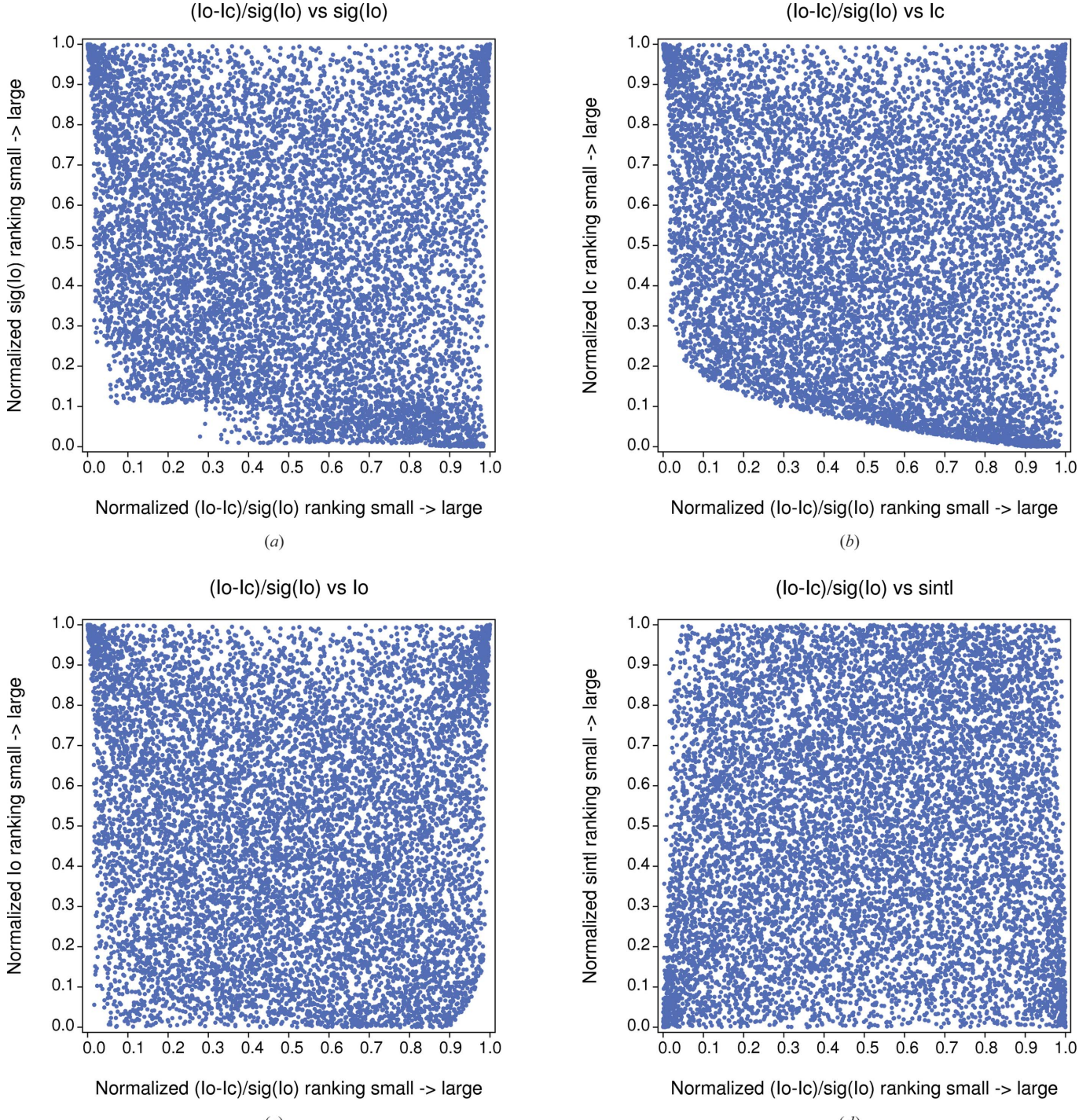

Figure 5
BayCoN plots for the artificial data set No. 30 with s.u. values transformed according to equation (5), which leads to an underestimation of large s.u. values: residuals (*a*) *versus* standard uncertainties, (*b*) *versus* calculated intensities, (*c*) *versus* observed intensities and (*d*) *versus* resolution.

Table 1
Data set with no systematic error: effects of the application of intensity and significance cutoffs on the uniformity of conditional probability distributions shared by the residuals and standard uncertainties, calculated intensities, observed intensities and resolution.

The uniformity of the distribution is tested by a χ^2 test against uniformity. The threshold value is approximately 149, so that for χ^2_S values larger than 149 the hypothesis of a uniform distribution must be rejected.

Cutoff	χ^2_S (ζ, s.u.)	χ^2_S (ζ, I_c)	χ^2_S (ζ, I_o)	χ^2_S (ζ, $\sin\theta/\lambda$)	Fig.
None	116.04	118.49	2705.82	66.95	2
$I_o > 0$	827.35	928.46	955.16	135.25	3
$I_o > 3\sigma(I_o)$	883.74	883.37	160.01	149.14	4

Table 2
Data set with systematic error (large s.u.'s too small): effects of the application of intensity and significance cutoffs to the uniformity of conditional probability distributions shared by the residuals and standard uncertainties, calculated intensities, observed intensities and resolution.

The uniformity of the distribution is checked by a χ^2 test. The threshold value is approximately 149, so that for χ^2_S values larger than 149 the hypothesis of a uniform distribution must be rejected.

Cutoff	χ^2_S (ζ, s.u.)	χ^2_S (ζ, I_c)	χ^2_S (ζ, I_o)	χ^2_S (ζ, $\sin\theta/\lambda$)	Fig.
None	1392.64	1390.63	3954.97	756.89	
$I_o > 0$	2022.39	2250.54	2320.59	897.92	
$I_o > 3\sigma(I_o)$	2321.80	2450.17	1383.09	763.59	5

at the top of the plot corresponds to the largest (*e.g.* decile of) I_c values. These contribute more strongly to the weakest (most negative, in the left part of the plot) and strongest (most positive, in the right part of the plot) residuals. In other words, the extreme positive and negative residuals are caused more frequently by the strongest I_c. This is because the largest absolute differences $(I_o - I_c)$ are still found for the largest values of I_c (and of I_o), whereas the corresponding σ values are damped down to lower values [equation (5)] so that the absolute values of residuals $(I_o - I_c)/\sigma(I_o)$ now tend to become larger for strong intensities. These areas of high density in the top corners are also seen in Figs. 5(*a*), 5(*c*), to which a similar interpretation applies. In Fig. 5(*d*) the areas of increased density of points appear in the bottom corners. A horizontal stripe in the bottom corresponds to the lowest resolution shell, in which the strongest intensities are to be found. The strongest reflections prefer the extreme residuals as just discussed; this carries over to the resolution by finding the extreme residuals in the lowest resolution shell.

The presence of a systematic error in the s.u.'s increases all χ^2_S values substantially, as can be seen from comparison of Tables 1 and 2. This is not surprising as it has already been seen that the respective conditional probability plots show patterns and are not uniform. A side observation is that in this example of a systematic error the χ^2_S (ζ, s.u.) and (ζ, I_c) values are similar in the case of no cutoff and for a significance cutoff $I_o > 3\sigma(I_o)$, and in both cases the values are different from the respective χ^2_S (ζ, I_o) value. But in the case $I_o > 0$ all three values are similar with χ^2_S (ζ, I_c) now being closer to χ^2_S (ζ, I_o) rather than to χ^2_S (ζ, s.u.). A similar tendency will appear in the analysis of published experimental data.

4.2.2. Large standard uncertainty values too large. In the following, the case is investigated in which the strong s.u.'s are overestimated rather than underestimated. For this a transformation

$$\sigma \rightarrow 2\sigma - \frac{\sigma}{0.5\,\sigma + 1} \qquad (6)$$

was applied to the σ values of data set 24, resulting in data set 31. This led to an increase of the large s.u.'s equal to the decrease in equation (5). The ratio of *de facto* ($R^{de\,facto} = 0.0412$) and predicted ($R^{pred} = 0.0605$) R values was at 0.68 identical to the GoF. The corresponding negative value $R^{meta} = -0.46$ also indicates overfitting like the GoF (Henn & Schönleber, 2013). But now it has become apparent from the χ^2_S values that overfitting took place in response to too large s.u. values: the $\chi^2_S(\zeta, \text{s.u.}) = 240.25$ and $\chi^2_S(\zeta, I_c) = 236.40$ are one order of magnitude smaller, whereas $\chi^2_S(\zeta, I_o) = 2925.31$ has more than doubled and $\chi^2_S(\zeta, \sin\theta/\lambda) = 129.09$ is now even within acceptance of the hypothesis of uniformity at the 0.001 level of significance (Fig. 6).

In this case, an *a posteriori* transformation of s.u. values is also feasible; however, this is not the most important point here. From the comparison of the last two examples it is concluded that overestimation of strong s.u. values is less harmful with respect to the statistical independence of residuals than underestimation of s.u. values. However, the mean significance of the data will decrease and the s.u.'s of the model parameters will appear to be larger.

4.3. Application to experimental data

For an overview, all 23 experimental data sets and additionally the six artificial data sets 24–29 were analyzed by application of the test on uniformity with the χ^2_S values. Fig. 7 shows the corresponding χ^2_S values and the threshold value 149 as a blue line. As the χ^2_S values differ over a large range, a logarithmic scale was chosen.

General observations are that:

(*a*) Most of the experimental χ^2_S values are larger than the threshold value.

(*b*) In the case of experimental data, the χ^2_S values tend to be similar on a logarithmic scale for (ζ, I_o) and (ζ, I_c), when no weighting scheme is applied (set Nos. 3–7, 14–16, 20).

(*c*) In the case of experimental data, the χ^2_S values tend to be smaller for (ζ, s.u.) compared to (ζ, I_c) and (ζ, I_o) with exceptions for (ζ, I_o) (sets 2, 12, 13, 21, 23).

(*d*) In the case of artificial data with cutoff $I_o > 0$ (Nos. 25, 27, 29) the χ^2_S values tend to be similar on a logarithmic scale for (ζ, I_c), (ζ, I_o) and (ζ, s.u.) and far above the threshold value.

(*e*) In the case of artificial data with no cutoff (24, 26, 28), the χ^2_S values tend to be similar on a logarithmic scale and below the threshold value for (ζ, I_c) and (ζ, s.u.) and much lower than the corresponding values of (ζ, I_o). The comparison with the previous finding demonstrates the effect of an intensity cutoff.

(*f*) In the case of artificial data with no systematic error the χ^2_S values for (ζ, s.u.) and (ζ, I_c) are always very close to each other irrespective of intensity and significance cutoffs (data with significance cutoff not shown). This is contrary to most of the experimental data.

(*g*) χ^2_S values below the threshold value appear simultaneously for (ζ, I_c) and (ζ, s.u.) only for artificial data with no cutoff (Nos. 24, 26, 28).

More individual observations are that:

(*h*) In the case of experimental data only the χ^2_S values of sets Nos. 4 and 20 meet in the same place as was the case for the artificial data with applied cutoff.

(*i*) The χ^2_S value for (ζ, I_o) of sets 2, 12, 13, 21 and 23 is lower than the corresponding χ^2_S value for (ζ, s.u.).

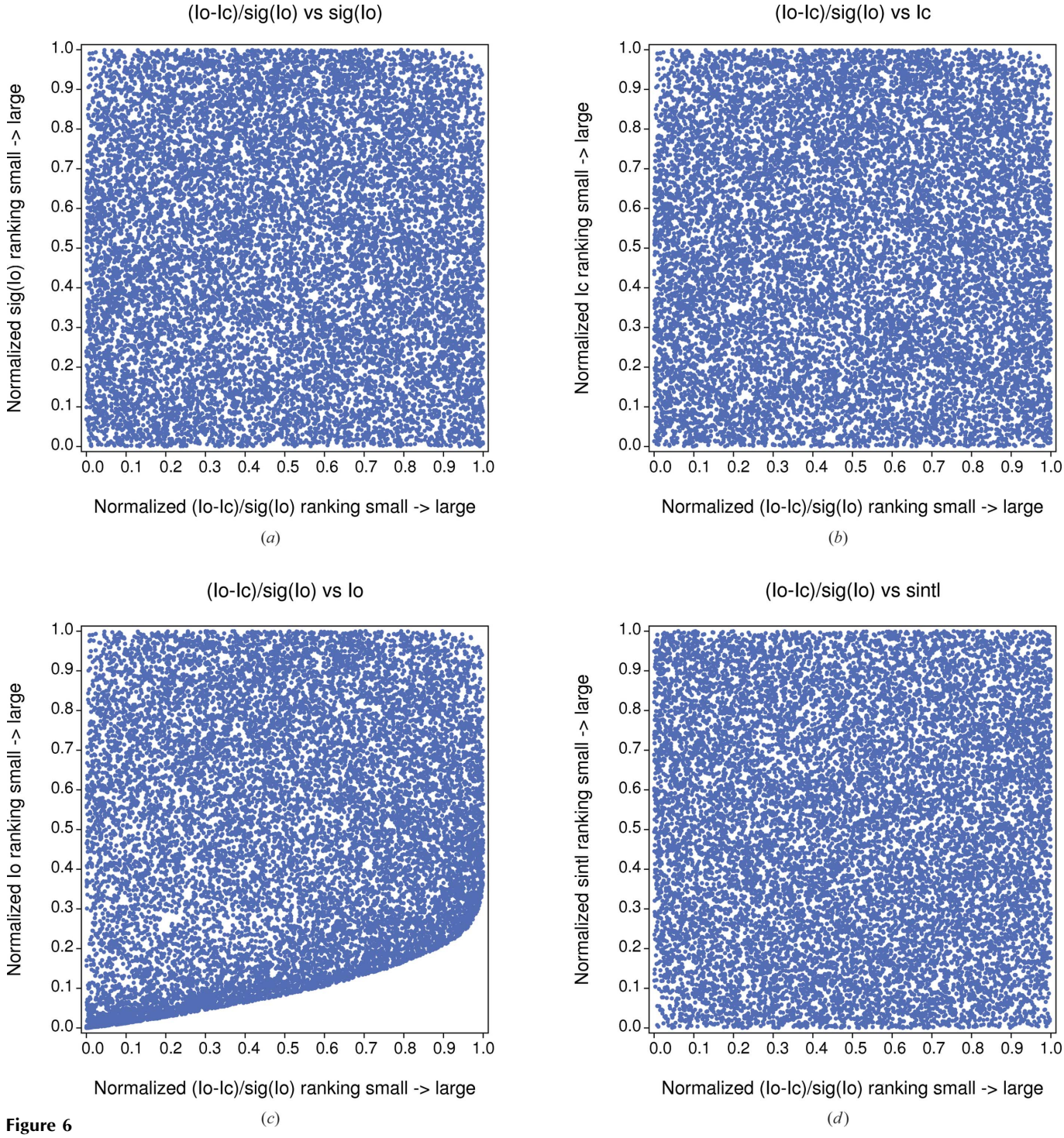

Figure 6
BayCoN plots for the artificial data set No. 31 with s.u. values transformed according to equation (6) in which large s.u. values are overestimated: residuals (*a*) *versus* standard uncertainties, (*b*) *versus* calculated intensities, (*c*) *versus* observed intensities and (*d*) *versus* resolution.

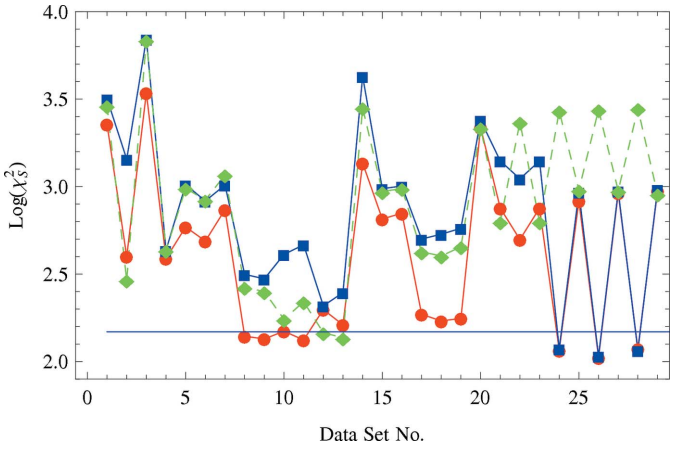

Figure 7
Logarithm of χ^2_S values of the experimental data sets 1–23 and artificial data sets 24–29 for residuals ζ *versus* s.u. (red circles), residuals *versus* I_c (blue squares) and residuals *versus* I_o (green diamonds connected by a dashed line). The threshold value of 149 is shown as a blue line.

(*j*) In the case of experimental data sets 8–11 they achieve a χ^2_S value for (ζ, s.u.) smaller than the threshold value and sets 12 and 13 for (ζ, I_o).

Not all of these observations can be addressed or explained in the publication at hand. However, it is demonstrated how application of an intensity cutoff increases the corresponding χ^2_S values of (ζ, I_c) and (ζ, s.u.), leading to non-uniform distributions, whereas the χ^2_S value of (ζ, I_o) is decreased. In the following some experimental findings are discussed in more detail and corresponding BayCoN plots of the data sets with highest and lowest χ^2_S values are shown.

4.4. Residuals and weighted standard uncertainties

The BayCoN plot of residuals $(I_o - I_c)/\hat{\sigma}(I_o)$ *versus* $\hat{\sigma}$ should be uniform, as deviations from uniformity indicate preference (increasing density of points) or avoidance (low and zero density of points) of certain combinations of residual and $\hat{\sigma}$ values. Statistical independence implies that all combinations are realized with the same frequency, apart from statistical fluctuations. The lowest experimental χ^2 sums for (ζ, $\hat{\sigma}$) are obtained for data sets Nos. 9 ($\chi^2_S = 129.68$, R^{meta} = 9.9%, see Fig. 8*a*) and 11 ($\chi^2_S = 132.79$, R^{meta} = 13.1%, see Fig. 8*b*), which correspond to the harmonic nuclear motion multipole model refinements at 20 K and at 120 K.

The largest χ^2 sums are obtained for data sets Nos. 1 ($\chi^2_S = 2272.82$, see Fig. 8*c*) and 20 ($\chi^2_S = 2180.90$, see Fig. 8*d*). A large coumarin crystal of size 0.33 × 0.65 × 0.91 mm was used in data set 1 together with a beam collimated to 0.6 mm, and data were collected from three different detector positions. Different detector positions were also used in data sets 11 and 9; therefore the high degree of non-uniformity of residuals may be connected to data-processing errors and/or to an incomplete correction of effects of the large crystal size. Data set 1 has the largest value of R^{meta} = 80.5%, which is artificially high due to a weighting scheme $w = 7/[\sigma^2(F_o)]$ which corresponds to $\hat{\sigma}^2(F_o) = (1/7)\sigma^2(F_o)$, *i.e.* the experimental σ values that are known to be too small in most cases are further diminished. When the original experimental s.u. values are used instead, this reduces the systematic error to R^{meta} = 49.7%. Data set 20 has R^{meta} = 58.0%

4.5. Residuals and calculated intensities

These distributions should be uniform as was shown in §4.1. The lowest experimental (ζ, I_c) χ^2_S values are obtained for data sets Nos. 12 ($\chi^2_S = 205.69$, R^{meta} = 16.3%, see Fig. 9*a*) and 13 ($\chi^2_S = 243.59$, R^{meta} = 5.4%, see Fig. 9*b*), which correspond to anharmonic and harmonic nuclear motion multipole refinement of the same experimental data.

Data sets 12 and 13 do not show any notable differences, although differences might be expected as sets 12 and 13 correspond to anharmonic and harmonic nuclear motion refinement of the same experimental data. It is also surprising that the *harmonic* nuclear motion multipole model has the *lower* R^{meta} = 5.4% (set No. 13), thus indicating *less* systematic errors compared to set 12 (with R^{meta} = 16.3%) which included anharmonic nuclear motion modeling, despite a quite large difference in the final $wR(F^2)|_{1/\hat{\sigma}^2}$ values that were 1.47% (data set 12) and 3.91% (data set 13). An explanation of this unexpected behavior may be that different parameters for the weighting schemes were used (data set 12 $a = 0.005$, $b = 0.006$; data set 13 $a = 0.015$, $b = 0.056$). If the weighting scheme is only used to correct the s.u. values from the reflection file, both anharmonic and harmonic refinements will be performed with the same weighting scheme. Exploring these interesting details goes beyond the scope of the present work. Both data sets still belong to those with a low degree of contamination with systematic errors. Data sets 14–16 are from the same publication. Set 14 is a conventional IAM refinement against charge-density data; therefore systematic errors (R^{meta} = 50.0%) and non-uniform BayCoN plots are expected. Employing a multipole model reduces the systematic errors to R^{meta} = 26.7% (data set 15) and to R^{meta} = 25.4% (data set 16) and results in smoother BayCoN plots. From these numbers and the corresponding plots (see supporting information), however, it is still obvious that further systematic errors are present.

4.6. Residuals and resolution

The lowest values χ^2_S for the conditional probability distribution (ζ, $\sin\theta/\lambda$) are obtained for data sets 12 ($\chi^2_S = 144.35$), 9 ($\chi^2_S = 155.65$), 8 ($\chi^2_S = 170.39$) and 11 ($\chi^2_S = 181.77$). The largest values are obtained for data sets 3 ($\chi^2_S = 10\,673.50$) and 20 ($\chi^2_S = 3283.80$).

Data set No. 3 corresponds to a charge-density study of roxythromycin, a large organic molecule, that additionally showed disorder, measured with synchrotron radiation (0.56000 Å) at 100 K. Statistical weights were used. The plot Fig. 10(*c*) shows different features in the low- and high-resolution parts. The low-resolution part (deciles 1–3) consists of two horizontally organized stripe-like structures, which end close to the third decile as indicated by a small sharp increase of the initially linearly increasing white gap to the right vertical axis. This may mark the end of the overlap region

between the two detector positions at $2\theta = 0^\circ$ and $2\theta = -40^\circ$. This low-resolution part can be further divided into the lowest resolution decile, that clearly shows a symmetric polarization towards the extremes of residuals. Too small s.u. values for the strong intensities of the lowest resolution shell lead to such a polarization of residuals as can be seen from Fig. 5. In the horizontal stripe corresponding to the second and third resolution deciles, however, positive residuals appear much more frequently than negative ones. There is an additional shift of residuals to negative values for increasing resolution for the last seven deciles in resolution shells, *i.e.* I_o tends to be increasingly systematically smaller than I_c for the high-resolution batch. This slow shift compensates for the distinct tendency of large positive residuals in the second and third deciles. It remains unclear whether data-processing errors, *e.g.* in scaling, merging and assigning s.u. values of low- and high-

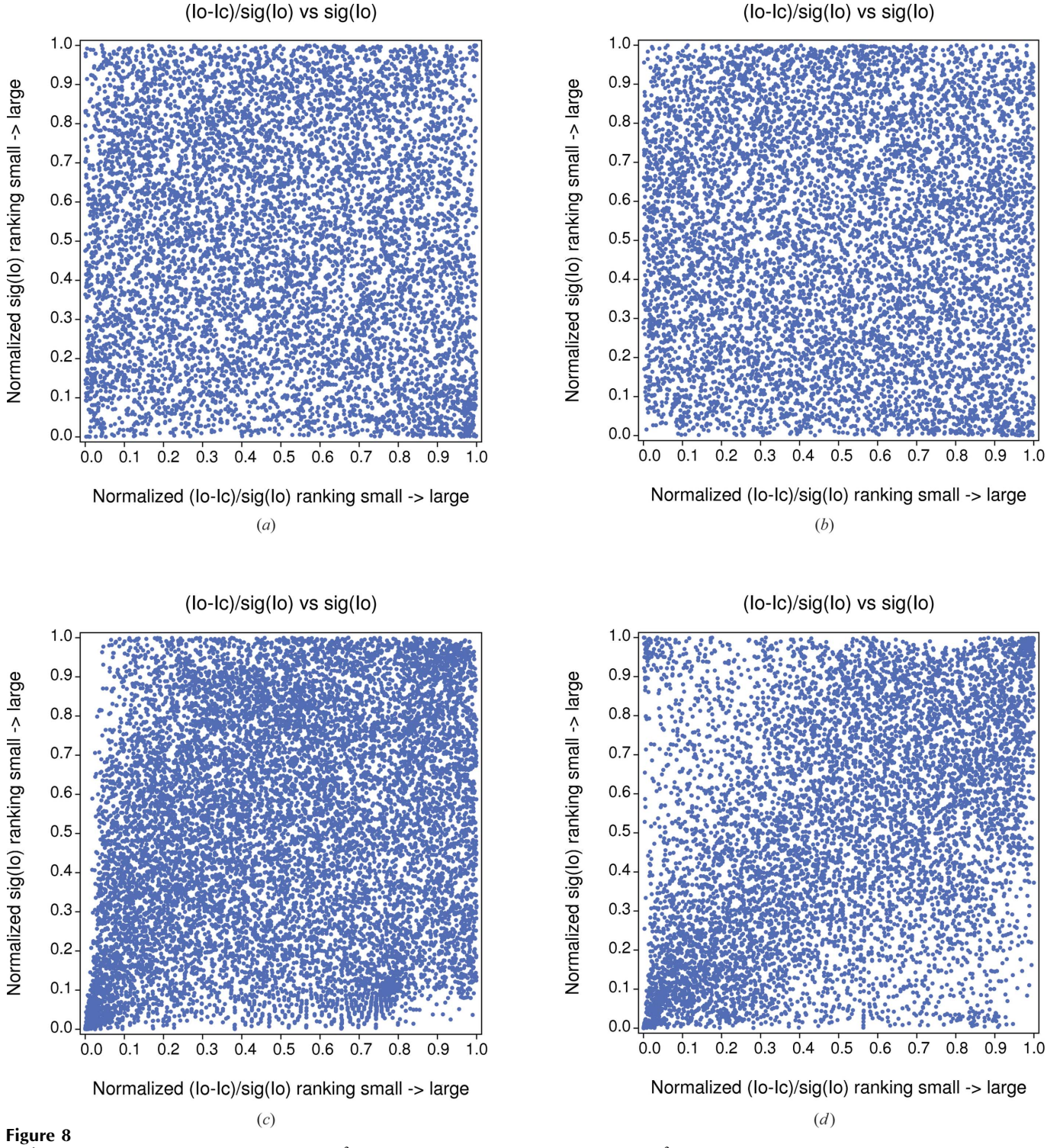

Figure 8
$(\zeta, \hat{\sigma})$ plots for data sets with lowest values of χ^2_S, No. 9 (*a*) and No. 11 (*b*), and for highest χ^2_S values, No. 1 (*c*) and No. 20 (*d*).

resolution batches, or model errors such as disorder are responsible for these shifts. The (ζ, $\sin\theta/\lambda$) plot for data set 20 is similar to that of data set 3 in the respect that there seems to be a symmetric low-resolution stripe at the first decile with a high frequency of extreme positive and negative residuals, a stripe around the second decile with a higher frequency of only one extreme of residuals, in this case negative ones, and a large upper part, again with a slow tendency to produce shifted residuals. In this case the residuals from the highest resolution shell tend to be shifted towards positive values. This implies that the I_o values tend to be larger than I_c for the highest resolution shell, whereas at lower resolution they tend to be smaller, and in the lowest resolution shell they tend to be frequently smaller and larger but not alike. Also in this study statistical weights were used. It is interesting to see that those five data sets with the largest

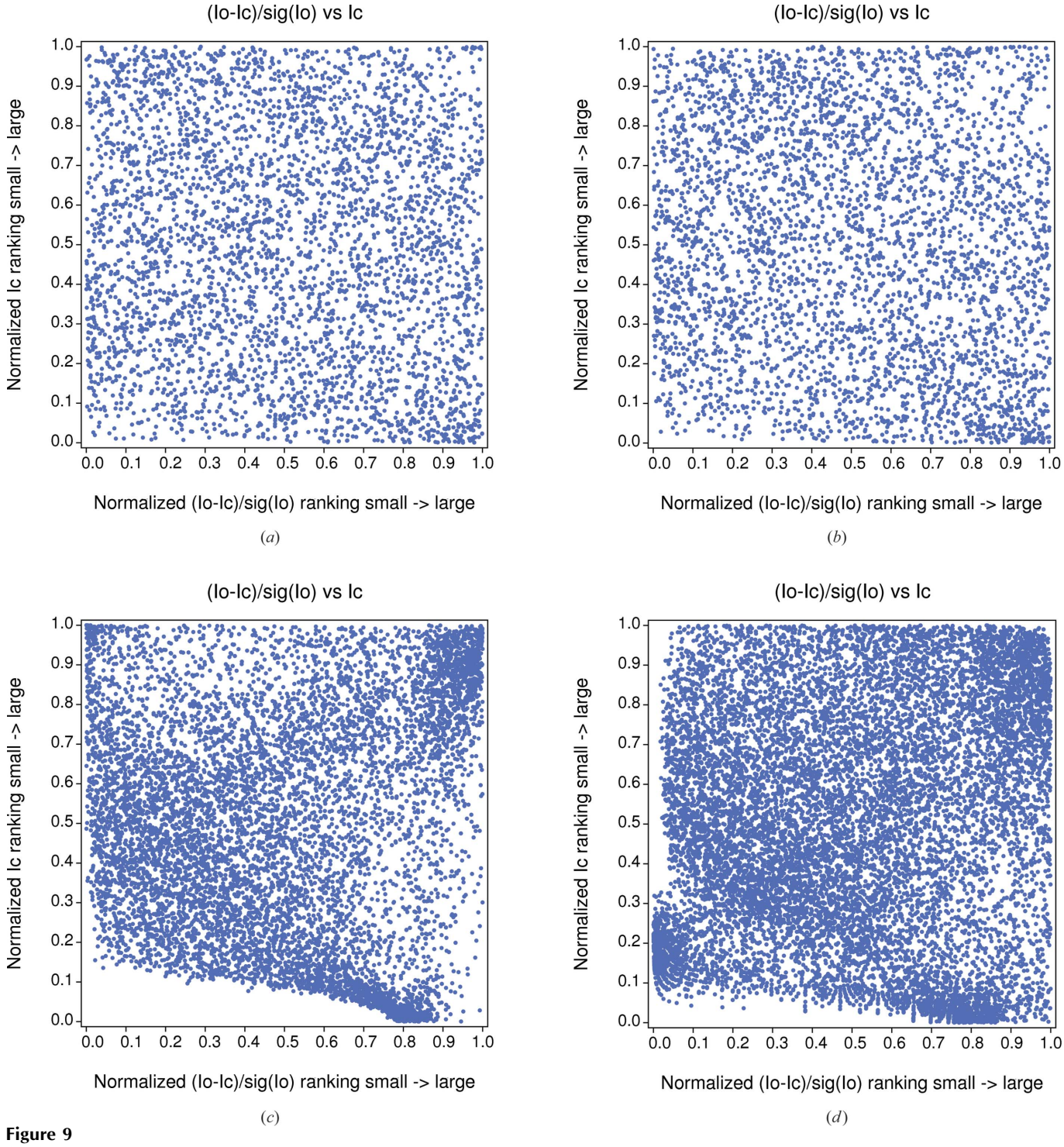

Figure 9
(ζ, I_c) plots for data sets Nos. 12 (*a*) and 13 (*b*), Nos. 14 (*c*) and 1 (*d*).

systematic errors (data set/R^{meta}: 1/80.5%; 3/71.0%; 20/57.9%; 17/57.58%; 14/59.9%) all use statistical weights (or weights in direct proportion to statistical weights) with the exception of data set 17, in which an IAM was refined against high-resolution data. These sets also showed highly non-uniform distributions of residuals *versus* resolution, when the corresponding information was available.

5. Summary

The meta residual value was calculated with respect to reduced residuals $\hat{\sigma}$ for 23 experimental data sets, not all of which employed a weighting scheme. The data sets with lowest meta residual value range between 5 and 10%, but most show a much higher value. The concept of conditional probabilities was used to construct scatter plots, which demonstrate nonlinear connections between the residuals and observed

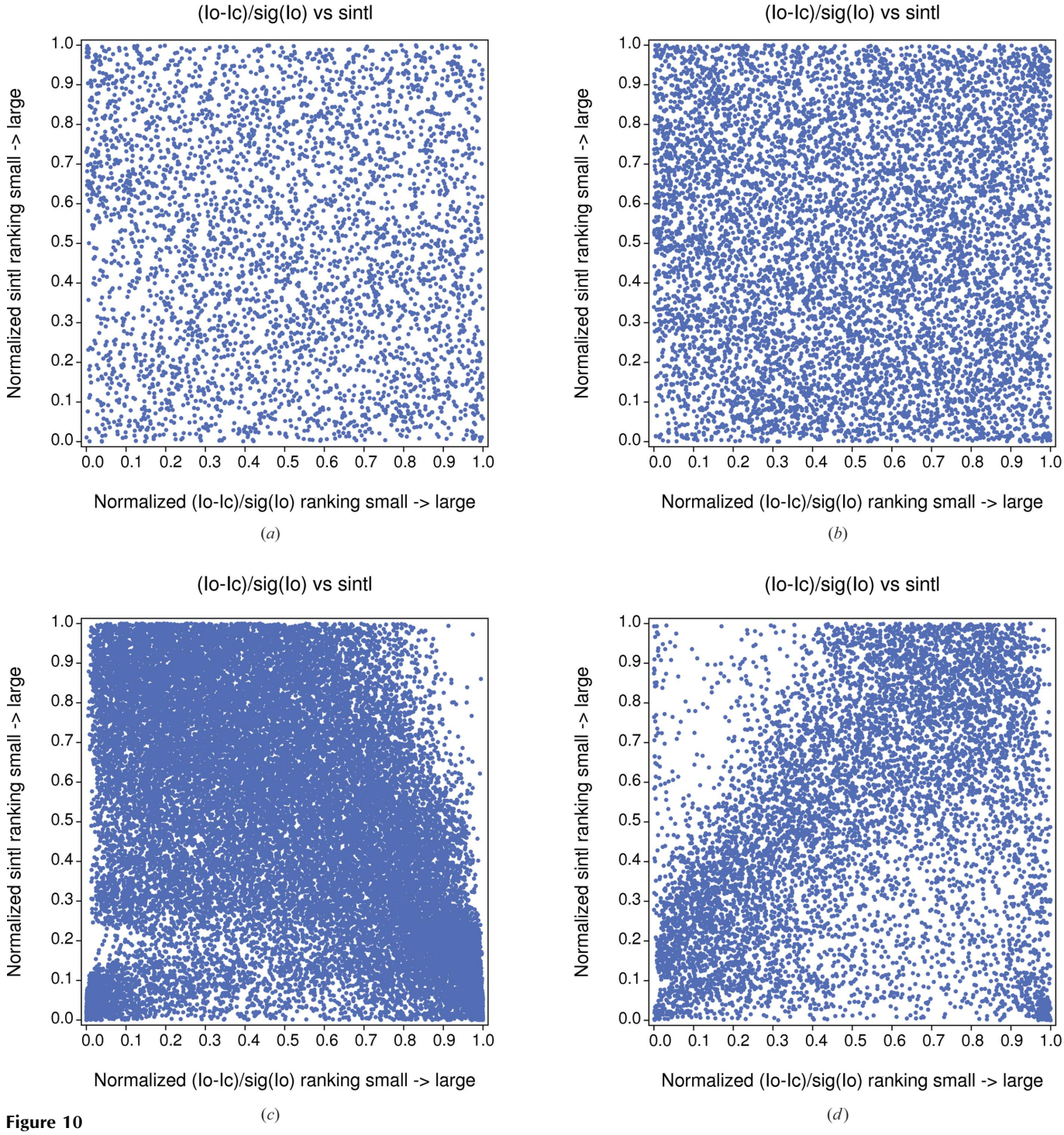

Figure 10
(ζ, $\sin\theta/\lambda$) plots for data sets Nos. 12 (*a*) and 9 (*b*) as well as 3 (*c*) and 20 (*d*).

and calculated intensities as well as s.u.'s and resolution. It was stressed that uniform plots are expected for (ζ, I_c), $(\zeta$, s.u.) and $(\zeta, \sin\theta/\lambda)$ plots when the data are free of systematic errors. It was shown how intensity and significance cutoffs introduce systematic errors in terms of deviation from uniformity of the respective distributions. Application of an intensity cutoff $I_o > 0$ leads to overfitting. The deviations are visualized by the scatter plots and quantified by the χ^2_S values. The cases of too small and too large s.u. values were studied with the help of artificial data, showing that too small s.u. values affect the distribution of residuals more strongly. Applications to experimental data showed that low meta residual values are accompanied by uniform scatter plots. Distinct dependencies of the residuals from the resolution were observed. It was suggested that these are not only caused by model deficiencies, but also by data-processing steps. The main purpose of this work is to develop concepts for the proof of existence of systematic errors, not to identify these. The important and helpful existing tools, like the normal probability plots, are unfortunately only rarely used. We hope that the visualization of systematic errors in the form of scatter plots helps to quickly identify and eliminate many sources of errors that otherwise might pass unnoticed and that this will help to focus research activities and attention on this important topic. The software package *BayCoN*, which is capable of calculating the conditional probability distributions, χ^2_S values and theoretical R values from 'xd.fco' or from 'fcf' files, is available from the authors.

The authors thank the referees for valuable suggestions. Thanks are also due to Professor Pinkerton for discussion, suggestions and data, and to Professor Jelsch for data. The authors are also grateful to Denise Kelk-Huth and John Kelk for support.

References

Bradley, J. V. (1963). *Am. Stat.* **17**, 14–15.
Henn, J. & Meindl, K. (2014). *Acta Cryst.* A**70**, 248–256.
Henn, J. & Schönleber, A. (2013). *Acta Cryst.* A**69**, 549–558.
Hirshfeld, F. L. & Rabinovich, D. (1973). *Acta Cryst.* A**29**, 510–513.
Semendjajew, K. A., Bronstein, I. N., Musiol, G. & Mühlig, H. (2012). *Taschenbuch der Mathematik*. Frankfurt am Main: Harri Deutsch.
Volkov, A., Macchi, P., Farrugia, L. J., Gatti, C., Mallinson, P. R., Richter, T. & Koritsanszky, T. (2006). *XD2006. A Computer Program Package for Multipole Refinement, Topological Analysis of Charge Densities and Evaluation of Intermolecular Energies from Experimental and Theoretical Structure Factors*.
Waterman, D. & Evans, G. (2010). *J. Appl. Cryst.* **43**, 1356–1371.