Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 25 August 2003
Accepted 13 May 2004

Determination of standard uncertainties in fits to pair distribution functions

Brian H. Toby[a]* and Simon J. L. Billinge[b]

[a]NIST Center for Neutron Research, National Institute of Standards and Technology, Gaithersburg, MD 20899-8562, USA, and [b]Department of Physics and Astronomy, Michigan State University, East Lansing, MI 48824, USA. Correspondence e-mail: brian.toby@nist.gov

Methods are developed to establish the limits of statistical uncertainty for values computed from a pair distribution function (PDF) or models fit to a PDF. This is done by computation of a variance–covariance matrix for the PDF from the uncertainties in the diffraction intensities. The application of this variance–covariance matrix also offers optimal weighting for least-squares refinement.

1. Introduction

Crystallographic analysis provides structural models as well as *estimates for the statistical uncertainty in the structural parameters*. These uncertainty estimates, known as standard uncertainties (s.u.), do not include all possible error sources, but are extremely valuable nonetheless as they allow scientists to estimate when differences are too small to be taken seriously.

For materials that lack long-range order, or where short-range structure is not reflected in the long-range order of the crystal, an alternative structural analysis approach is used. In this method, sometimes called the real-space structure determination method, the pair distribution function (PDF) is modeled in real space, rather than the reciprocal-space powder diffraction data (Egami & Billinge, 2003). The PDF reflects the short-range ordering in a material. This approach has been widely used for studying the structures of glasses and liquids since the 1930s (Debye & Menke, 1930; Warren, 1969; Bowron & Finney, 2003). More recently, it has been applied to disordered crystalline and partially crystallized materials (Egami & Billinge, 2003). Quantitative structural information on nanometer length scales can be obtained by fitting a model directly to the PDF (Dmowski *et al.*, 1988; Proffen & Billinge, 1999; Petkov *et al.*, 2002). It has been shown that, when there are no short-range deviations from the average structure, the PDF agrees well with the interatomic distances computed from a crystallographic model (Toby & Egami, 1992; Peterson *et al.*, 2003). This real-space method is one of a very small number of experimental techniques that can be used to probe structure on the nanometer length scale, when this local structure is not consistent with the long-range globally averaged structure (Proffen *et al.*, 2003).

In a previous paper, it was shown how to derive the s.u. for the PDF values (Toby & Egami, 1992). That work, however, does not treat the case when the PDF is interrogated to determine structural parameters, since neighboring points in the PDF are not statistically independent. In this paper, we extend this derivation to s.u. for parameters fitted to the PDF. This allows, for the first time, statistically justified s.u. and meaningful χ^2 values to be determined from real-space fits.

2. Background

2.1. Terminology

In the real-space method, powder diffraction data, $I(Q)$, are collected over a wide range of Q ($Q = 4\pi \sin\theta/\lambda$). Typically, data are used up to Q of >25 Å^{-1} and in some cases in excess of 50 Å^{-1}. Following the notations of Lovesey (1984) and Warren (1969), the raw intensity values, $I(Q)$, are converted to scaled and corrected intensity values, $S(Q)$, by compensation for experimental artifacts such as multiple scattering, inelastic scattering, container and sample self-shielding, spectral flux, and polarization (Egami & Billinge, 2003).

The PDF [$\rho(r)$ or $G(r)$] is computed from the Fourier transform of $Q[S(Q) - 1]$:

$$G(r_k) = \frac{2}{\pi} \sum_j Q_j [S(Q_j) - 1] \sin(Q_j r_k) \Delta Q_j$$

and

$$\rho(r_k) = \rho_0 + \frac{1}{2\pi^2 r_k} \sum_j Q_j [S(Q_j) - 1] \sin(Q_j r_k) \Delta Q_j.$$

The PDF is similar to a spherically averaged Patterson function but incorporates both periodic and aperiodic information owing to use of both Bragg and diffuse scattering information.

2.2. Propagation of errors in least-squares refinement

The notation $\sigma[p_j]$ will be used to indicate the s.u. in quantity p_j, $\sigma[p_j] = \langle \{ \langle p_j \rangle - p_j \}^2 \rangle^{1/2}$, where $\langle\ \rangle$ indicates the expectation value for a quantity. Our model will be represented as $M(\mathbf{p})$, with ν parameters, p_j, and n independent observables, y_i. The notation $M_i(\mathbf{p})$ indicates the computed value corresponding to y_i. Following the derivations of Prince (1994), the method of non-linear least-squares refinement determines the p_j values that minimize $\sum_{i=1}^{n} \sigma_i^{-2} [M_i(\mathbf{p}) - y_i]^2$,

 DOI: 10.1107/S0108767304011754

where the s.u. values, σ_i, are used as weighting factors, by iteratively solving the equation $A^TWA\mathbf{p} = A^TW\mathbf{y}$. Here the $n \times \nu$ design matrix A is $A_{ij} = \partial M_i(\mathbf{p})/\partial p_j$ and W is the $n \times n$ weight matrix. Any values may be used for least-squares weighting, but the smallest uncertainties for fitted parameters will be obtained when W is diagonal and matrix elements reflect the uncertainties in the observations: $W_{ii} = 1/\{\sigma[y_i]\}^2$.

The product matrix, A^TWA, is called the Hessian and its inverse, $V = (A^TWA)^{-1}$, is called the variance–covariance matrix. If a good fit is obtained to the data, as noted by a value of χ^2 near unity $\{\chi^2 = \sum_{i=1}^{n} w_i^2[M_i(\mathbf{p}) - y_i]^2/[n - \nu]\}$, and errors in the data follow a normal distribution, then the s.u. for the fitted parameters are obtained from the variance–covariance matrix, $\sigma[p_j] = (V_{jj})^{1/2}$. When $\chi^2 \gg 1$, it is common to increase s.u. estimates by a factor of $(\chi^2)^{1/2}$, but this can either underestimate or overestimate the actual uncertainties, depending on the type of correlation between the omitted parameters and the fitted values.

2.3. Fitting to correlated values

The least-squares method can also be applied to fit a model to correlated values. In this case, the weight matrix is no longer diagonal. Rather, it is the inverse of the variance–covariance matrix for the 'input' parameters. An example of this is where a model is fitted to previously fitted parameters, p_α. The weight matrix for the second fit, W_β, will be $W_\beta = V_\alpha^{-1} = A_\alpha^T W_\alpha A_\alpha$. It has been shown that the parameters and uncertainties for a model fitted to intermediate results is equivalent to those fitted directly to the original data, provided that correlation is properly treated *via* the weight matrix (Prince, 1981).

2.4. Applications of the variance–covariance matrix

The variance–covariance matrix may be used to estimate uncertainty even after a coordinate transformation. If we use a transformation matrix, T, to transform a series of parameters from one basis to another, $\mathbf{q} = T^T\mathbf{p}$, the variance–covariance matrix for $\mathbf{q}$, V_q, is given by $V_q = T^TV_pT$, where V_p is the the variance–covariance in the initial basis (Prince, 1994). Finally, the variance–covariance matrix is also used to find the s.u. of a linear combination of the fitted parameters. If we define a linear function of the parameters, $f(\mathbf{p}) = \mathbf{f}^T\mathbf{p}$, where $\mathbf{f}$ is a vector, then the s.u. for this function is given by $\sigma[f(\mathbf{p})] = (\mathbf{f}^TV\mathbf{f})^{1/2}$ (Prince, 1994).

3. Standard uncertainty for $S(Q)$

For most diffraction measurements, $I(Q)$ is measured using detectors that directly count quanta. This means that $\sigma[I(Q)] = [I(Q)]^{1/2}$ is true, except when $I(Q)$ approaches zero. For other detection mechanisms, such as image plates, careful study is needed to establish $\sigma[I(Q)]$. For determination of $S(Q)$, usual practice requires the combination of several data sets, which establish instrument flux and background as well as sample scattering. As described previously, propagation of uncertainties to find $\sigma[S(Q)]$ is straightforward (Toby & Egami, 1992). Commonly, smoothing is applied to the background and flux determinations, which introduces minor amounts of correlation between the $S(Q_i)$ values. It was found that this correlation can safely be ignored. Alternatively, this smoothing can be avoided, which eliminates correlation between the $S(Q)$ values.

4. Standard uncertainty for $G(r)$ and $\rho(r)$

The Fourier transform relationship between $S(Q)$ and the PDF is simply a change of coordinates from a vector, $\mathbf{s}$, of values $s_i = S(Q_i) - 1$, to a vector, $\mathbf{g}$, of values $g_j = G(r_j)$ by $\mathbf{g} = T_G^T\mathbf{s}$, where $T_{G,ik} = (2/\pi)Q_i \sin(Q_ir_k)\Delta Q_i$. Likewise, defining vector $\boldsymbol{\rho}$, where $\rho_j = \rho(r_j) - \rho_0$, then $\boldsymbol{\rho} = T_\rho^T\mathbf{s}$, where $T_{\rho,ik} = Q_i \sin(Q_ir_k)\Delta Q_i/(2\pi^2 r_k)$. We note that the uncertainty on $S(Q_i) - 1$ is the same as that of $S(Q_i)$; this means that uncertainties for $G(r)$ and $\rho(r)$ can be obtained from variance–covariance matrix V using $V = T^TV_ST$, where V_S is the variance–covariance matrix for $S(Q)$. If $S(Q)$ is computed without smoothing then $V_{S,ii} = \{\sigma[S(Q_i)]\}^2$ and $V_{S,ij} = 0$ for $i \neq j$. This then allows V_G to be simplified as $V_{G,jk} = \sum_i T_{G,ij}V_{S,ii}T_{G,jk}$ or

$$V_{G,jk} = \frac{4}{\pi^2}\sum_i Q_i^2 \sin(Q_ir_j)\sin(Q_ir_k)\Delta Q_i^2\{\sigma[S(Q_i)]\}^2 \qquad (1)$$

and, likewise,

$$V_{\rho,jk} = \frac{1}{4\pi^4 r_j r_k}\sum_i Q_i^2 \sin(Q_ir_j)\sin(Q_ir_k)\Delta Q_i^2\{\sigma[S(Q_i)]\}^2. \quad (2)$$

The diagonal terms of V_ρ yield $\sigma[\rho_j] = (V_{\rho,jj})^{1/2}$, which reproduces equation (26) of Toby & Egami (1992). The off-diagonal terms are equivalent to equation (A5.3.13) of Egami & Billinge (2003).

If smoothing is used to determine $S(Q)$, then equations (1) and (2) are only approximate, due to the neglect of correlation between $S(Q)$ values. It is possible to include the off-diagonal elements of V_S. However, since this correlation is small, only minor changes would be expected.

It is also worth noting that a sum of form $\sum_i Q_i^2 \sin(Q_ir_j)\sin(Q_ir_k)$ will be largest when $r_j = r_k$ and will decrease as $|r_j - r_k|$ increases. This means that the most significant terms in V_G or V_ρ will be those closest to the diagonal, provided that terms are ordered by increasing or decreasing r. As discussed in Egami & Billinge (2003), this degree of statistical correlation in the PDF decreases as the Q range of the Fourier transform is increased.

5. Uncertainties on results from and fits to $G(r)$ and $\rho(r)$

Commonly, ranges in $G(r)$ and $\rho(r)$ are integrated to determine a coordination number. Integration, differentiation or other results are evaluated from the PDF by multiplying individual $G(r)$ values by constants and then summing them. This means these results are linear functions of the PDF values, which can be expressed as $\mathbf{f}^T\mathbf{g}$, and thus the s.u. for the result is given by $\sigma[\mathbf{f}^T\mathbf{g}] = \mathbf{f}^TV_G\mathbf{f}$.

As was discussed previously, the optimal weighting for least-squares fits to the PDF occurs when the variance–covariance matrix is used in the fit. Thus, the function to be solved is $A^T W_G A\mathbf{p} = A^T\mathbf{g}$, where $W_G = V_G^{-1}$. This is a slightly more complex computation than that customarily done when no off-diagonal terms are present in the weight matrix, but W_G need only be evaluated once when $G(r)$ or $\rho(r)$ is computed. With modern computers, the time needed for the additional $m \times m$ matrix multiplication step, where m is the number of terms in the PDF (typically on the order of 10^3), should be minor. However, if needed, a banded-matrix approximation may be used to increase computational efficiency, since the importance of the V_G terms decreases with distance from the matrix diagonal.

Finally, as was noted previously, the variance–covariance matrix for the p parameters fit in the model to $G(r)$ is given by $V_p = (A^T W_G A)^{-1}$. Thus, the s.u. on parameter p_j is given by $\sigma[p_j] = V_{p,jj}$, provided that model gives a good fit to $G(r)$.

6. Discussion and conclusions

This paper has shown how to compute standard uncertainties on the PDF, on quantities derived from the PDF, and on parameters fit to the PDF. These uncertainties are derived in a manner completely consistent with that used in crystallographic analysis. Where Bragg scattering dominates the contributions to the PDF, the uncertainties for models fit to the PDF will be essentially the same as those fit *via* crystallographic methods, provided the same data range is used in both. However, it is impractical to perform Rietveld refinement at very high Q ranges, so we speculate that PDF fits may in fact offer higher precision. In practice, however, real-space analysis is of greatest value for systems that cannot be modeled well with Rietveld fits.

This paper has estimated statistical uncertainties in real-space structure determination but has not considered the separate issue of systematic errors. PDF computation requires inclusion of many factors that may usually be ignored in crystallographic studies. Properly, these corrections should be considered part of the model and perhaps should be optimized as part of the fitting process. It should also be noted that some instrumental effects, such as a Q-dependent resolution function, can be accurately modeled in a Rietveld fit, but produce aberrations in the PDF.

Our trust of crystallographic methods has been fostered because diffraction data and crystallographic models often agree within the expected statistical uncertainty. Further, the resulting models have been cross-validated with results obtained *via* other techniques. Thus, over its century of development, crystallographic analysis has become one of our most reliable scientific methods.

This validation process is still under way for PDF analysis. The good agreement between crystallographic results and real-space models obtained for selected crystalline materials gives us confidence that systematic errors are not severe. The expressions developed here will now allow researchers to identify when systematic errors are exceeding the expected statistical fluctuations. This in turn may allow for improvements in real-space-analysis techniques, but will also help increase trust in real-space methods.

Finally, it should be noted that many sources of systematic error can be removed using difference-PDF measurements, where the change in the PDF due to variation of an experimental parameter is computed. This is done at the expense of increasing statistical uncertainty. In these cases, accurate uncertainty estimation, as has been developed here, becomes crucial.

This paper arose from discussions while planning the Warren award symposium in honor of Takeshi Egami at the 2003 American Crystallographic Association meeting. We would like to thank Professor Egami for introducing us to real-space analysis. BHT would also like to thank Ted Prince for very instructive lectures and many hours of inspiring discussions. Work at MSU was supported by DOE through DE-FG02-97ER45651 and by NSF through DMR-0304391.

References

Bowron, D. T. & Finney, J. L. (2003). *J. Chem. Phys.* **118**, 8357–8372.
Debye, P. & Menke, H. (1930). *Phys. Z.* **31**, 797.
Dmowski, W., Toby, B. H., Egami, T., Subramanian, M. A., Gopalakrishnan, J. & Sleight, A. W. (1988). *Phys. Rev. Lett.* **61**, 2608–2611.
Egami, T. & Billinge, S. J. L. (2003). *Underneath the Bragg Peaks: Structural Analysis of Complex Materials.* Oxford: Pergamon Press.
Lovesey, S. W. (1984). *Theory of Neutron Scattering from Condensed Matter*, Vol. 1. Oxford: Clarendon Press.
Peterson, P. F., Bozin, E. S., Proffen, T. & Billinge, S. J. L. (2003). *J. Appl. Cryst.* **36**, 53–64.
Petkov, V., Billinge, S. J. L., Larson, P., Mahanti, S. D., Vogt, T., Rangan, K. K. & Kanatzidis, M. G. (2002) *Phys. Rev. B*, **65**, 092105.
Prince, E. (1981). *J. Appl. Cryst.* **14**, 157–159.
Prince, E. (1994). *Mathematical Techniques in Crystallography and Material Science.* New York: Springer-Verlag.
Proffen, T. & Billinge, S. J. L. (1999). *J. Appl. Cryst.* **32**, 572–575.
Proffen, T., Billinge, S. J. L., Egami, T. & Louca, D. (2003). *Z. Kristallogr.* **218**, 132–143.
Toby, B. H. & Egami, T. (1992). *Acta Cryst.* A**48**, 336–346.
Warren, B. E. (1969). *X-ray Diffraction.* Reading: Addison-Wesley. Reprinted (1990). New York: Dover.