References

Bateman, A. (1953). *Higher Transcendental Functions*, Bateman Manuscript Project, parts I and II. New York: McGraw Hill.

Giacovazzo, C. (1980). *Direct Methods in Crystallography*. London: Academic Press.

Hauptman, H. & Karle, J. (1953). *Acta Cryst.* **6**, 136–141.

Karle, J. & Hauptman, H. (1953). *Acta Cryst.* **6**, 131–135.

Klug, A. (1958). *Acta Cryst.* **11**, 515–543.

Lenstra, A. T. H. (1974). *Acta Cryst.* A**30**, 363–369.

Lenstra, A. T. H. (1979). *Bull. Soc. Chim. Belg.* **88**, 359–368.

Lenstra, A. T. H., Petit, G. H. & Geise, H. J. (1979). *Cryst. Struct. Commun.* **8**, 1023–1029.

Lindgren, B. W. (1976). *Statistical Theory*, 3rd ed. New York: Macmillan.

Neuts, M. F. (1973). *Probability*. Boston: Allyn & Bacon.

Parthasarathi, V. & Parthasarathy, S. (1975). *Acta Cryst.* A**31**, 38–41.

Petit, G. H., Lenstra, A. T. H. & Van Loock, J. F. (1981). *Acta Cryst.* A**37**, 353–360.

Rohatgi, V. K. (1976). *An Introduction to Probability Theory and Mathematical Statistics*. New York: Wiley.

Shmueli, U (1982). *Acta Cryst.* A**38**, 362–371.

Shmueli, U. & Kaldor, U. (1981). *Acta Cryst.* A**37**, 76–80.

Shmueli, U. & Wilson, A. J. C. (1981). *Acta Cryst.* A**37**, 342–353.

Srinivasan, R. & Parthasarathy, S. (1976). *Some Statistical Applications in X-ray Crystallography*. Oxford: Pergamon.

Van de Mieroop, W. (1979). PhD thesis (in Dutch), Univ. of Antwerp.

Wilson, A. J. C. (1949). *Acta Cryst.* **2**, 318–321.

Wilson, A. J. C. (1950*a*). *Acta Cryst.* **3**, 397–398.

Wilson, A. J. C. (1950*b*). *Research*, **3**, 48.

Wilson, A. J. C. (1969). *Acta Cryst.* B**25**, 1288–1293.

Wilson, A. J. C. (1978). *Acta Cryst.* A**34**, 986–994.

Acta Cryst. (1983). A**39**, 562–565

Moments of the Probability Density Function of R_2 Approached *Via* Conditional Probabilities. II. Completely Correct and Completely Incorrect Models in Space Group $P\bar{1}$

By W. K. L. Van Havere and A. T. H. Lenstra

University of Antwerp (UIA), Department of Chemistry, Universiteitsplein 1, B-2610 Wilrijk, Belgium

(*Received* 15 *April* 1982; *accepted* 14 *February* 1983)

Abstract

With the help of conditional probabilities formulas are derived for the first and second moment of R_2 as a function of the size of the model. The formulas are valid in the space group $P\bar{1}$ for two extreme cases, *viz* completely correct and completely incorrect models. Incorporation of the observed intensities enables one to obtain accurate *a priori* estimates of $\langle R_2\rangle$ and $\sigma(R_2)$. The theory agrees very well with simulated experiments.

1. Introduction

In automated structure determinations of single crystals, one may use the mathematical residual function R_2 to discriminate between correct and incorrect models. The applicability of R_2 as a discriminator function increases sharply if one has at one's disposal an *a priori* evaluation of its average value and spread. That is to say, in order to be able to use statistical decision methods in an automated analysis one needs to know for the crystallographic situation at hand either the probability distribution of the residual R_2 or the moments of this distribution.

Until recently, the assumption of an infinite data set allowed only the prediction of the first moment (mean value) but precluded the evaluation of the higher moments. The break-through came with the introduction of the calculus of conditional probability. In part I (Van Havere & Lenstra, 1983) we laid down the general principles of the new theory and derived expressions for the first and second moments of the probability density function of the residual R_2 for completely correct and completely incorrect structure models in space group $P1$. The results for $P1$ may serve as a model for all primitive non-centrosymmetric space groups. In this paper we will derive similar expressions tor space group $P\bar{1}$, which may serve as a parent for all primitive centrosymmetric space groups.

2. Moments of R_2

Throughout this work E_o will refer to the observed magnitude of the normalized structure factor belonging to a structure containing N atoms in the asymmetric unit. Likewise E_c will refer to the calculated magnitude of an E value of a model containing n atoms in the asymmetric unit. The definition of R_2 is

$$R_2 \equiv \sum_H (E_o^2 - \eta^2 E_c^2)^2 / \sum_H E_o^4 \qquad (2.1)$$

0108-7673/83/040562-04$01.50

with η^2 describing the fraction of the scattering power of the model *versus* the total structure:

$$\eta^2 \equiv \eta_c^2/\eta_o^2. \qquad (2.2)$$

Taking point atoms of equal scattering power, η_c^2 becomes n, the number of atoms in the model, and η_o^2 becomes N, the number of atoms in the asymmetric part of the unit cell. The reciprocal vector $H \equiv (h,k,l)$ can span any subset of the total space. The normalized structure factors are defined as

$$\bar{E}_o \equiv \bar{E}_o(H) = (2/N)^{1/2} \sum_{j=1}^{N} \cos(2\pi H r_j). \qquad (2.3)$$

We take over the approximations and apply the mathematical machinery already developed (Van Havere & Lenstra, 1983). Since full details on the derivation can be found in our previous article we will here only sketch briefly the argument.

For the space group $P\bar{1}$ one can write

$$\langle R_2;\mathscr{E}_o\rangle = 1 + \eta^4 \frac{\sum_H \langle E_c^4;E_o\rangle}{\sum_H E_o^4} - 2\eta^2 \frac{\sum_H E_o^2\langle E_c^2;E_o\rangle}{\sum_H E_o^4} \qquad (2.4)$$

and

$$\begin{aligned}\sigma^2(R_2;\mathscr{E}_o) = \Big\{&\sum_H \eta^8(\langle E_c^8;E_o\rangle - \langle E_c^4;E_o\rangle^2)\\ &-\sum_H 4\eta^6 E_o^2(\langle E_c^6;E_o\rangle\\ &-\langle E_c^4;E_o\rangle\langle E_c^2;E_o\rangle)\\ &+\sum_H 4\eta^4 E_o^4(\langle E_c^4;E_o\rangle - \langle E_c^2;E_o\rangle^2)\Big\}\\ &\times\Big\{\sum_H E_o^4\Big\}^{-2}, \qquad (2.5)\end{aligned}$$

where $\mathscr{E}_o$ is the set of observed structure factors as used in the calculations. The notation $\langle E_c^n;E_o\rangle$ means the value of E_c^n averaged over the coordinates of the model in direct space under the constraint of E_o. With (2.4) and (2.5) the problem of finding the moments of $P(R_2)$ is shifted to finding the moments of the conditional intensity distribution $P(E_c;E_o)$. We will do so for the limiting cases of a completely incorrect and a completely correct model.

2.1. *Incorrect models*

An incorrect model is characterized by the absence of any correlation between the set of observed structure factors and the set of calculated structure factors belonging to the (partial) model. Therefore

$$\langle E_c^n;E_o\rangle = \langle E_c^n\rangle. \qquad (2.1.1)$$

The moments of E_c can be obtained either directly by averaging the structure-factor equations over direct space assuming equal probability for all points in this space, or by using intensity distributions derived by Wilson (1949). Wilson, assuming a large number of atoms evenly distributed in the asymmetric part of the unit cell, has shown that for space group $P\bar{1}$ one gets

$$P(E_c) = (2/\pi)^{1/2} \exp(-E_c^2/2). \qquad (2.1.2)$$

The moments are given by (Shmueli, 1982)

$$\langle E^{2n}\rangle = 2^{1-n}\frac{(2n-1)!}{(n-1)!}, \quad n = 1, 2, 3, \ldots. \qquad (2.1.3)$$

Substitution of (2.1.3) into (2.4) and (2.5) yields

$$\langle R_2;\mathscr{E}_o\rangle = \Big\{\sum_H (E_o^4 - 2\eta^2 E_o^2 + 3\eta^4)\Big\}\Big/\sum_H E_o^4 \qquad (2.1.4)$$

and

$$\sigma^2(R_2;\mathscr{E}_o) = \Big\{\sum_H (8\eta^4 E_o^4 - 48\eta^6 E_o^2 + 96\eta^8)\Big\}\Big\{\sum_H E_o^4\Big\}^{-2}. \qquad (2.1.5)$$

2.2 *Correct models*

In a completely correct model observed and calculated structure factors are correlated. Now, starting from the conditional probability function $P(E_o;E_c)$ derived by Srinivasan & Parthasarathy (1976), employing Bayes theorem and using marginal distribution functions $P(E_c)$ and $P(E_o)$ of the form of (2.1.2), one finds

$$P(E_c;E_o) = \left(\frac{2\eta_o^2}{\pi(\eta_o^2 - \eta_c^2)}\right)^{1/2} \exp\left\{-\frac{\eta_c^2 E_o^2 + \eta_o^2 E_c^2}{2(\eta_o^2 - \eta_c^2)}\right\} \times \cosh\left\{\frac{\eta_o\,\eta_c\,E_o\,E_c}{\eta_o^{\wedge} - \eta_c^2}\right\}. \qquad (2.2.1)$$

Since we are using asymptotic conditions in the number of atoms, the following formulas are strictly speaking only valid in situations in which total structure, partial model and difference structure are all large. Following arguments presented in Appendix *A*, we derive for the moments of (2.2.1)

$$\langle E_c^{2n};E_o\rangle = \frac{(2n-1)!}{(n-1)!} 2^{1-n}\left(\frac{\eta_o^2 - \eta_c^2}{\eta_o^2}\right)^n \times {}_1F_1\left(-n; \frac{1}{2}; -\frac{1}{2}\frac{\eta_c^2 E_o^2}{\eta_o^2 - \eta_c^2}\right), \qquad (2.2.2)$$

for $n = 1, 2, 3, \ldots$,

in which ${}_1F_1$ is a confluent hypergeometric function.

Substitution of the relevant moments of (2.2.2) into (2.4) and (2.5) gives

$$\langle R_2;\mathscr{E}_o\rangle = \left\{\sum_H E_o^4(\eta^8 - 2\eta^4 + 1) + \sum_H E_o^2(6\eta^6 - 2\eta^2)(1-\eta^2) + \sum_H 3\eta^4(1-\eta^2)^2\right\}\Big/\sum_H E_o^4 \qquad (2.2.3)$$

and

$$\sigma^2(R_2;\mathscr{E}_o) = \left\{\sum_H E_o^6(16\eta^{14} - 32\eta^{10} + 16\eta^6)(1-\eta^2) + \sum_H E_o^4(168\eta^{12} - 144\eta^8 + 8\eta^4)(1-\eta^2)^2 + \sum_H E_o^2(384\eta^{10} - 48\eta^6)(1-\eta^2)^3 + \sum_H 96\eta^8(1-\eta^2)^4\right\}\left\{\sum_H E_o^4\right\}^{-2}. \qquad (2.2.4)$$

The path of $\langle R_2\rangle$ and $\sigma(R_2)$ in a generalized case can be seen from Figs. 1 and 2.

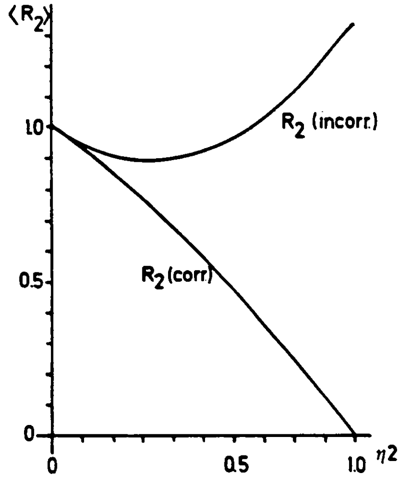

Fig. 1. $\langle R_2\rangle$ as a function of the model size. The summations $\sum_H E_o^n$ are replaced by $\langle E_o^n\rangle$ as found from equation (2.1.3).

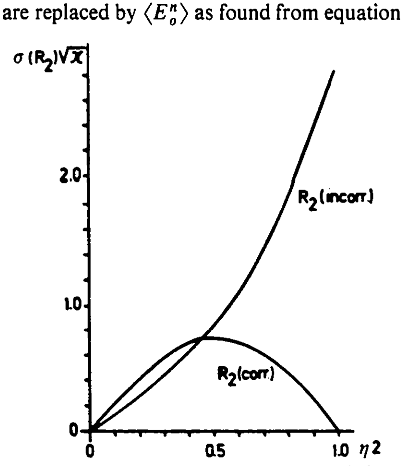

Fig. 2. $\sigma(R_2)$ as a function of the model size.

3. Verification

Using simulated *observed* structures one avoids any systematic pecularities of a real structure and thus disagreements between theory and *experiment* can be uniquely attributed to flaws in the theoretical argumentation. The applicability of the theory depends, of course, on the robustness of its results towards the very pecularities of real structures which are disregarded in the theory. In the preceding paper (Van Havere & Lenstra, 1983) we have demonstrated that for space group $P1$ the theory is in excellent agreement with calculations based on simulated structures as well as on a real structure. Here, we confine ourselves to test the theory against simulated structures.

In our example the *observed* structure is a set of 100 atomic positions randomly placed in the asymmetric part of the unit cell, with a corresponding set of 1530 *observed* reflections. The atoms are taken as point atoms with equal scattering power. To test the theory for completely incorrect models, 10 000 independent and thus completely incorrect models were generated with n ($n \le 100$) randomly placed atoms in the asymmetric part of the unit cell. Substitution of their E_c values into (2.1) gives R_2(exp). $\langle R_2(\text{exp})\rangle$ was obtained by averaging over the 10 000 trials. Furthermore, $\sigma^2[R_2(\text{exp})]$ can be calculated as

$$\left\{\sum_{\text{exp}} [R_2(\text{exp})]^2 - [\sum_{\text{exp}} R_2(\text{exp})]^2\right\} \times 10^{-4}. \qquad (3.1)$$

Table 1 gives the comparison of $\langle R_2(\text{exp})\rangle$ and $\sigma^2[R_2(\text{exp})]$ with $\langle R_2(\text{th})\rangle$ and $\sigma^2[R_2(\text{th})]$, calculated from (2.1.4, 5).

The theory for completely correct models was tested using the same simulated structure as before, but restricting the set ε_o to 70 reflections. Random samples of n ($n \le 100$) correct atomic positions were selected to represent the models and to compute the E_c values. As in the previous case, substitution into (2.1) gives R_2(exp), while $\langle R_2(\text{exp})\rangle$ and $\sigma^2[R_2(\text{exp})]$ were obtained after 100 000 of such trials for each n. Table 2 gives the comparison of $\langle R_2(\text{exp})\rangle$ and $\sigma^2[R_2(\text{exp})]$ with $\langle R_2(\text{th})\rangle$ and $\sigma^2[R_2(\text{th})]$, calculated from (2.2.3, 4). The choice $N = 100$ and the number of reflections used (1530 or 70) is purely arbitrary. Experience has shown that the number of trials (10 000 for incorrect and 100 000 for correct models) is large enough to achieve convergence for both the average and the spread.

The values given in Tables 1 and 2 show a very satisfactory agreement between theory and simulated experiments. They allow the conclusion that the formalism and the derived formulas are correct. The small discrepancies that are left can, as was demonstrated earlier (Van Havere & Lenstra, 1983), be attributed to the use of asymptotic intensity distri-

Table 1. *Comparison for incorrect models of* $\langle R_2(\mathrm{exp})\rangle$ *and* $\sigma^2[R_2(\mathrm{exp})]$ *with theoretical values*

n	$\langle R_2(\mathrm{th})\rangle$	$\langle R_2(\mathrm{exp})\rangle$	$\sigma^2[R_2(\mathrm{th})]$	$\sigma^2[R_2(\mathrm{exp})]$
0	1·0000	1·0000	0·00000	0·00000
1	0·9414	0·9408	0·00001	0·00001
20	0·9045	0·9033	0·00006	0·00006
30	0·8892	0·8875	0·00013	0·00013
40	0·8955	0·8934	0·00026	0·00026
50	0·9235	0·9208	0·00049	0·00050
60	0·9731	0·9702	0·00089	0·00092
70	1·0448	1·0412	0·00154	0·00162
80	1·1374	1·1335	0·00254	0·00272
90	1·2520	1·2469	0·00402	0·00430
100	1·3883	1·3827	0·00613	0·00668

Table 2. *Comparison for correct models of* $\langle R_2(\mathrm{exp})\rangle$ *and* $\sigma^2[R_2(\mathrm{exp})]$ *with theoretical values*

n	$\langle R_2(\mathrm{th})\rangle$	$\langle R_2(\mathrm{exp})\rangle$	$\sigma^2[R_2(\mathrm{th})]$	$\sigma^2[R_2(\mathrm{exp})]$
0	1·0000	1·0000	0·00000	0·00000
25	0·8221	0·8225	0·00472	0·00428
50	0·5855	0·5876	0·01300	0·01232
75	0·3021	0·3042	0·00793	0·00763
100	0·0000	0·0000	0·00000	0·00000

butions for a finite number of atoms and to a lesser extent to the elimination of double summations in (2.5).

WVH thanks the Belgian organization IWONL for financial support. The help of Professor H. J. Geise in the preparation of this manuscript is gratefully acknowledged. We also wish to thank G. H. Petit and J. F. Van Loock for stimulating discussions.

APPENDIX *A*

The moments of the intensity distribution given in (2.2.1) are defined as

$$\langle E_c^\mu;E_o\rangle = \left(\frac{2\eta_o^2}{\pi(\eta_o^2-\eta_c^2)}\right)^{1/2} \times \int\limits_0^\infty E_c^\mu \exp\left[-\frac{\eta_c^2E_o^2+\eta_o^2E_c^2}{2(\eta_o^2-\eta_c^2)}\right] \times \cosh\left(\frac{\eta_o\,\eta_c\,E_o\,E_c}{\eta_o^2-\eta_c^2}\right)\,\mathrm{d}E_c. \qquad (A.1)$$

Using the identity (Bateman, 1953, II)

$$\cosh(z) = (\pi z/2)^{1/2}\,I_{-1/2}(z), \qquad (A.2)$$

equation (A.1) can be transformed to

$$\langle E_c^\mu;E_o\rangle = \left[\frac{\eta_o^3\,\eta_c\,E_o}{(\eta_o^2-\eta_c^2)^2}\right]^{1/2} \times \int\limits_0^\infty E^{\mu+1/2} \exp\left[-\frac{\eta_c^2E_o^2+\eta_o^2E_c^2}{2(\eta_o^2-\eta_c^2)}\right] \times I_{-1/2}\left(\frac{\eta_o\,\eta_c\,E_o\,E_c}{\eta_o^2-\eta_c^2}\right)\,\mathrm{d}E_c. \qquad (A.3)$$

Using a generalization of Weber's first exponential integral (Bateman, 1953, II),

$$\int\limits_0^\infty J_\nu(at)\exp(-p^2t^2)\,t^{\mu-1}\,\mathrm{d}t = \frac{\Gamma(\tfrac{1}{2}\nu+\tfrac{1}{2}\mu)\left(\dfrac{1}{2}\dfrac{a}{p}\right)^\nu}{2p^\mu\,\Gamma(\nu+1)}\;{}_1F_1\left(\tfrac{1}{2}\nu+\tfrac{1}{2}\mu;\,\nu+1;\,-\frac{a^2}{4p^2}\right),$$

$$\mathrm{Re}(\nu+\mu)>0,\quad a\in C,\quad \mathrm{Re}(p^2)>0 \qquad (A.4)$$

together with the identity (Bateman, 1953, II)

$$I_\nu(z) = \exp(-i\tfrac{1}{2}\nu\pi)\,J_\nu[z\exp(i\pi/2)]$$
$$-\pi < \arg(z) \le \pi/2 \qquad (A.5)$$

and Kummer's first transformation (Bateman, 1953, I)

$${}_1F_1(a;b;x) = e^x\,{}_1F_1(b-a;\,b;\,-x), \qquad (A.6)$$

we can write equation (A.3) as

$$\langle E_c^\mu;E_o\rangle = \frac{\Gamma\left(\dfrac{\mu+1}{2}\right)}{\Gamma\left(\dfrac{1}{2}\right)}\,2^{\mu/2}\left(\frac{\eta_o^2-\eta_c^2}{\eta_o^2}\right)^{\mu/2} \times {}_1F_1\left(-\frac{\mu}{2};\frac{1}{2};-\frac{1}{2}\frac{\eta_c^2E_o^2}{\eta_o^2-\eta_c^2}\right), \qquad (A.7)$$

which, using the identity (Bateman, 1953, I)

$$\Gamma(\mu+\tfrac{1}{2}) = 2^{(1-2\mu)}\sqrt{\pi}\,\frac{\Gamma(2\mu)}{\Gamma(\mu)} \qquad (A.8)$$

and taking $\mu=2n$, $n=1, 2, 3, \ldots$, reduces to

$$\langle E_c^{2n};E_o\rangle = \frac{(2n-1)!}{(n-1)!}\,2^{(1-n)}\left(\frac{\eta_o^2-\eta_c^2}{\eta_o^2}\right)^n \times {}_1F_1\left(-n;\frac{1}{2};-\frac{1}{2}\frac{\eta_c^2E_o^2}{\eta_o^2-\eta_c^2}\right). \qquad (A.9)$$

As can be seen from the definition of ${}_1F_1$ (Bateman, 1953, I),

$${}_1F_1(a;b;x) = 1 + \frac{a}{b}\frac{x}{1!} + \frac{a(a+1)}{b(b+1)}\frac{x^2}{2!} + \ldots, \qquad (A.10)$$

the moments reduce in our case to an nth-degree polynomial in x, because a is a negative integer. For instance, the fourth moment can be written as

$$\langle E_c^4;E_o\rangle = \eta^4E_o^4 + 6\eta^2(1-\eta^2)\,E_o^2 + 3(1-\eta^2)^2. \qquad (A.11)$$

References

Bateman, H. (1953). *Higher Transcendental Functions*, Bateman Manuscript Project, parts I and II. New York: McGraw Hill.

Shmueli, U. (1982). *Acta Cryst.* A**38**, 362–371.

Srinivasan, R. & Parthasarathy, S. (1976). *Some Statistical Applications in X-ray Crystallography*. Oxford: Pergamon.

Van Havere, W. K. L. & Lenstra, A. T. H. (1983). *Acta Cryst.* A**39**, 553–562.

Wilson, A. J. C. (1949). *Acta Cryst.* **2**, 318–321.