Acta Cryst. (1953). **6**, 698

The Comparison of Theoretical and Experimental Determinations of Molecular Structures, with Applications to Naphthalene and Anthracene

By D. W. J. Cruickshank

Chemistry Department, The University, Leeds 2, England

and A. P. Robertson

Mathematics Department, The University, Glasgow W. 2, Scotland

(*Received* 10 *February* 1953)

An earlier discussion of the use of statistical significance tests in the comparison of experimental and theoretical determinations of molecular structures is extended, and the application of multivariate tests is shown. The accuracy of atomic coordinates determined by Fourier or least-squares methods, irrespective of whether the peaks overlap or not, is discussed in detail. Formulae are given for the errors of molecular parameters in terms of the errors of the atomic parameters of the crystal structure. The methods are applied in a discussion of the latest results on naphthalene and anthracene, and it is shown that while it is hardly necessary to postulate any errors in the molecular-orbital theory for anthracene, there are very significant discrepancies for naphthalene.

1. Introduction

In X-ray crystallography, when a molecular* structure has been determined from a set of experimental data, it is often of great interest to decide whether or not it is in agreement with some theoretical structure. In general the two sets of molecular parameters will be different and the crystallographer will have to decide whether the differences are small enough to have arisen quite easily from errors in the experimental determination, or whether they are so large as to rule out this possibility. Similar problems arise in the comparison of different experimental determinations of the same molecular structure or of similar chemical groupings. The purpose of this paper is to indicate what sort of an answer can be given to these questions.

It is contended that comparisons of two structures should be made by statistical significance tests based on proper estimates of the accuracy of the results. This has already been advocated by Cox & Cruickshank (1948) and Cruickshank (1949*a*); but these discussions presented only an incomplete account of significance tests for the comparison of single parameters (e.g. the comparison of an experimentally determined bond length with a theoretical value). A fuller discussion of one-parameter significance tests is given in § 2 of this paper, and in § 3 the methods are extended to cover the simultaneous comparison of any number of parameters, thus making possible the comparison of structures as wholes. These two sections are largely the formulation of standard statistical procedures (see e.g. Kendall 1943, 1946) in the crystallographic situation. Naturally the difference between experimental and theoretical results is often so large as to render unnecessary the use of the formal apparatus of significance tests, but in those cases where, roughly speaking, the differences are of the order of magnitude of the errors significance tests are the only objective method of comparison.

Correction of systematic errors (particularly the finite-series effect) and estimation of the accuracy of the results are a necessary preliminary to the application of significance tests. The two papers above, following Booth (1945, 1946), give a discussion of this problem which is valid for atomic coordinates determined by the Fourier method when the peaks do not overlap. Cruickshank (1952) has introduced the modified differential Fourier method as a general means of determining atomic coordinates, which is valid whether the peaks overlap or not. When there is no overlapping and certain other conditions are satisfied, the usual Fourier method with back corrections by Booth's method (1945, 1946) approximates to the modified differential Fourier method. The estimation of the accuracy of results found by this general method is discussed in § 4. The method has formal similarities with Hughes's (1941) application of the least-squares method, for which the estimated errors can be derived by standard formulae. A comparison of the accuracies of the two methods is made in § 4. This part of the discussion is a more general form of that given earlier by Cruickshank (1949*b*), which was valid only for non-overlapping peaks.

A distinction must be made between the crystallographic structure parameters and the molecular parameters. First, the molecular parameters are bond lengths and angles, and not the positions of atoms in the unit cell. Secondly, a molecular parameter may

* The word molecular is used as a convenient way of designating the part of the crystallographic structure which is of interest, irrespective of the fact that the word may not always be correct in the chemical sense.

be estimated from more than one set of crystallographic parameters, because the assumed symmetry of the free molecule may be higher than its crystallographic symmetry. § 4 is concerned with discussing the errors of the crystallographic structure parameters, and § 5 is concerned with expressing the errors of the molecular parameters in terms of the former.

As examples, in § 6 the preceding theory is applied to recent investigations on naphthalene and anthracene.

2. One-parameter significance tests

This section summarises the tests for comparison with respect to one parameter. By introducing the statistical notions, it lays the foundation for subsequent generalization to more than one parameter. For definiteness we suppose that the one parameter is the length of a particular bond.

The experimental estimate l of the (unknown) true bond length is obtained by interpretation of the observed reflexion intensities. l differs by 'errors' from the unknown λ. The fundamental supposition (common to all metrical science) about the 'errors' is that for a given experimental procedure (understanding by this both the experimental arrangement and the interpretive process) l is a random variable with a probability distinction, i.e. there is a probability density function $f(l)$ such that $f(l)dl$ is the probability that the experimental estimate of λ lies in the range l to $l+dl$. (Since l must have some value $\int_0^\infty f(l)dl = 1$.) Like λ, $f(l)$ is unknown, but, from general considerations of the experimental procedure, reasons will be given in § 4 for supposing, when certain systematic errors have been corrected, that l is distributed normally about λ with unknown standard deviation σ; thus

$$f(l) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left\{\frac{-(l-\lambda)^2}{2\sigma^2}\right\}. \qquad (2\cdot1)$$

By the methods discussed in §§ 4 and 5 an estimate s of σ can be made. When the number ν of degrees of freedom on which this estimate is based is large (say $\nu > 30$) s may be treated as a moderately accurate estimate of σ, and so to a high degree of approximation l is distributed normally about mean λ with standard deviation s, i.e. $t = (l-\lambda)/s$ is a normal random variable with zero mean and unit standard deviation. (ν is the difference between the number of crystallographically independent planes observed and the number of independent parameters derived from the data.) In fact for all values of ν the random variable t has 'Student's' t distribution with ν degrees of freedom:

$$f(t) = \frac{1}{(\pi\nu)^{\frac{1}{2}}} \frac{\Gamma(\frac{1}{2}(\nu+1))}{\Gamma(\frac{1}{2}\nu)} \frac{1}{(1+t^2/\nu)^{\frac{1}{2}(\nu+1)}}, \qquad (2\cdot2)$$

$\Gamma(x)$ being the Gamma-function. This distribution tends to normality as $\nu \to \infty$. Because t involves only the unknown λ, and not the unknown σ, this distribution forms the basis of the statistical significance test used to compare an experimental bond length with a theoretical value.

Suppose the experimental data yield an estimated bond length l_o, with estimated standard deviation s_o based on ν degrees of freedom, and we wish to compare this with a theoretical value λ_o. On the tentative hypothesis that λ_o is the true value,

$$t_o = (l_o-\lambda_o)/s_o \qquad (2\cdot3)$$

is a value of a random variable t having a Student distribution with ν degrees of freedom. From the tables of this distribution, the probability P that $|t| \geqslant |t_o|$ can be found. If this is very small, it indicates that the occurrence of the results l_o and s_o on the tentative hypothesis is a rare and surprising event, and we are led to suspect, or even to reject, the hypothesis that the experimental data come from a structure with true bond length λ_o. On the other hand if P is not small, we conclude that the experimental data are not inconsistent with a value λ_o for the bond length, though they do not provide evidence to *prove* that λ_o is the *correct* value. When P is so small as to cast doubt upon the hypothesis, we say that l_o is significantly different from λ_o; just how small P has to be for this is arbitrary, and is a compromise between the danger of making a false judgment, which increases as we allow larger values of P to be significant, and the possibility of being unable to make any judgment, which increases as we restrict significance to smaller values of P. Most purposes are served by the following table:

$P > 0\cdot05$	$= 5\%$	not significant,
$0\cdot05 > P \geqslant 0\cdot01$	$= 1\%$	possibly significant,
$0\cdot01 > P \geqslant 0\cdot001$	$= 0\cdot1\%$	significant,
$0\cdot001 > P$		highly significant.

It is to be noted that P is not the chance of some specified observation being made, e.g. if the experimental and theoretical values of a bond length are 1·390 Å and 1·410 Å, and give $P = 0\cdot01$, it is not implied that if the theory is correct the chance of observing 1·390 Å is 1 in 100. Strictly speaking, the chance of making some exactly specified observation is zero; though it is true that if we are rounding off to the nearest 0·01 Å, we may speak of the probability of observing 1·39 Å, but the P values found by significance tests have nothing to do with this.

The situation is that we have to set up arbitrary and necessarily imperfect conventions for deciding whether to reject or retain a theory; the above significance tests provide us with a consistent and objective method of doing this.

In many problems $\nu > 30$, and so to a high degree of approximation the t distribution may be treated as normal. In this case

$$P = 1-\left(\frac{2}{\pi}\right)^{\frac{1}{2}}\int_0^{t_o}\exp\{-\tfrac{1}{2}t^2\}dt = 1-\mathrm{erf}\,(t_o/\sqrt{2})\,. \quad (2{\cdot}4)$$

The values of t_o at the various significance points are then

$$P = 0{\cdot}05 \qquad t_o = 1{\cdot}960\,,$$
$$P = 0{\cdot}01 \qquad t_o = 2{\cdot}576\,,$$
$$P = 0{\cdot}001 \qquad t_o = 3{\cdot}291\,.$$

It is possible to compare two experimentally determined bond lengths and to test the hypothesis that the true lengths are the same. Let the determinations be l_1 and l_2, with estimated standard deviations s_1 on ν_1 degrees of freedom and s_2 on ν_2 degrees of freedom. When both ν_1 and ν_2 are large the hypothesis may be tested on the normal law by taking

$$t_o = (l_1-l_2)/(s_1^2+s_2^2)^{\frac{1}{2}}, \quad (2{\cdot}5)$$

supposing the determinations to be independent. When either or both of ν_1 and ν_2 are small the distribution is more complicated and is not tabulated; a discussion of the problem will be found in Kendall (1946, § 21).

There is a difference between the above definition of P and that given by Cruickshank (1949*a*). In the former definition P was defined as the probability that a bond length A could be observed as greater than another bond length B by at least δl by chance, although really equal to B. Corresponding to this on the present definition, P is the probability that the difference between A and B *irrespective of sign* could be observed as greater than δl. The present definition is to be preferred, as it is in line with the usual statistical practice, and as the many-parameter definition given in the next section reduces to it for one parameter. An accidental fault of the wording of the previous definition was that the words 'by at least δl' were unfortunately omitted.

3. Many-parameter significance tests

When it is desired to compare two structures with respect to several parameters (e.g. atomic coordinates, bond lengths, angles between bonds) the one-parameter tests described in the last section can be applied to each parameter separately, but it may then be difficult to interpret the results if some parameters show significant differences and some not. As this method takes no account of possible correlations between parameters, it is preferable to use a test which considers all parameters simultaneously.

Suppose, then, we wish to consider the parameters $x_1, x_2, \ldots, x_n$, having (unknown) true values $\xi_1, \xi_2, \ldots, \xi_n$. As in the one-parameter case, we suppose that the x_i are random variables normally distributed with means ξ_i. In general they will be correlated, and their joint probability distribution will be the multivariate normal form

$$f = (2\pi)^{-\frac{1}{2}n}(\det\alpha)^{-\frac{1}{2}}\exp\{-\tfrac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha^{ij}(x_i-\xi_i)(x_j-\xi_j)\}\,, \quad (3{\cdot}1)$$

where $\alpha = (\alpha_{ij})$ is the variance matrix of the x_i, and $\alpha^{-1} = (\alpha^{ij})$ is its inverse. The variance matrix α of the x's is a generalization of the variance σ^2 of l in the one-parameter case; it is symmetric and its diagonal elements give the variances of the x's, and the other elements their covariances. Thus

$$\alpha_{ii} = \mathrm{var}\,(x_i) = \sigma_{xi}^2;\ \alpha_{ij} = \mathrm{cov}\,(x_i, x_j) = \varrho_{ij}\sigma_{xi}\sigma_{xj}\,,$$

where ϱ_{ij} is the correlation coefficient of x_i and x_j, or

$$\alpha_{ij} = \int\int\ldots\int(x_i-\xi_i)(x_j-\xi_j)f\,dx_1dx_2\ldots dx_n\,.$$

As with σ^2 in the one-parameter case, the variance matrix α is unknown, but we can make an estimate $a = (a_{ij})$ of it by the methods discussed in §§ 4 and 5. Assuming this has been done on ν degrees of freedom, we take as a generalization of the statistic t of § 2, the statistic T given by

$$T^2 = \sum_{i=1}^{n}\sum_{j=1}^{n}a^{ij}(x_i-\xi_i)(x_j-\xi_j)\,, \quad (3{\cdot}2)$$

where $a^{-1} = (a^{ij})$ is the inverse matrix to $a = (a_{ij})$. When $n = 1$, T reduces to t, for in this case the estimated variance matrix has a single term s^2, with inverse $1/s^2$, so $T^2 = (x-\xi)^2/s^2 = t^2$.

To test the hypothesis that the values $x_{o1}, x_{o2}, \ldots, x_{on}$, with estimated variance matrix $a_o = (a_{o,ij})$ based on ν degrees of freedom, found in a particular experimental determination, come from a structure in which the true values of these parameters are $\xi_{o1}, \xi_{o2}, \ldots, \xi_{on}$ we calculate

$$T_o^2 = \sum_{i=1}^{n}\sum_{j=1}^{n}a_o^{ij}(x_{oi}-\xi_{oi})(x_{oj}-\xi_{oj})\,. \quad (3{\cdot}3)$$

From the known distribution of T^2, the probability P that $T^2 \geqslant T_0^2$ can be found. If this is small we reject the hypothesis. More precisely, the significance levels given in § 2 can be used, e.g. if $0{\cdot}01 > P \geqslant 0{\cdot}001$, we say that the set of values $(x_{o1}, x_{o2}, \ldots, x_{on})$ differs from $(\xi_{o1}, \xi_{o2}, \ldots, \xi_{on})$ by an amount which is significant.

Just as the distribution of t tends to normality as $\nu \to \infty$, so the distribution of T^2 tends to the χ^2 distribution with n degrees of freedom (for the χ^2 distribution see e.g. Kendall (1943) or Weatherburn (1947)). When ν is sufficiently large, it is a sufficient approximation to use the simpler χ^2 distribution in place of the exact distribution. In rough terms, when $n = 1$ the χ^2 approximation is useful for $\nu > 30$ (as in § 2), and for $n = 6$ it is useful for $\nu > 60$.

4. The accuracy of atomic coordinates found by the least-squares or modified differential Fourier methods

4·1. *Exact formulae*

Both the modified differential Fourier method and the least-squares method lead to a set of simultaneous

equations linear in the small parameter refinements ε_{sj} (Cruickshank, 1952, equations (3·9) and (3·10)), ε_{sj} denoting a small variation of the jth coordinate of atom s. These equations may be written

$$\sum_{s,j} c_{ri,sj}\varepsilon_{sj} = b_{ri} , \qquad (4·1)$$

where, in the modified differential Fourier method,

$$c_{ri,sj} = \frac{1}{V}\frac{2\pi}{a_i}\sum_3 h_i \frac{\partial|F_c|}{\partial x_{sj}} \sin(\theta_r - \alpha) ,$$

and

$$b_{ri} = \frac{1}{V}\frac{2\pi}{a_i}\sum_3 h_i(|F_o|-|F_c|)\sin(\theta_r-\alpha) ;$$

and in the least-squares method

$$c_{ri,sj} = \sum_u w \frac{\partial|F_c|}{\partial x_{ri}}\frac{\partial|F_c|}{\partial x_{sj}} ,$$

and

$$b_{ri} = \sum_u w(|F_o|-|F_c|)\frac{\partial|F_c|}{\partial x_{ri}} .$$

In these equations F_o and F_c are the observed and calculated structure factors, h_i is the plane index for the direction x_i, a_i a cell side, V the unit cell volume, α the phase angle, $\theta_r = 2\pi(\sum_i h_i x_{ri}/a_i)$, $\sum_3$ is a summation over all observed planes, $\sum_u$ a summation only over symmetry independent planes, $|F_c|$ is treated in the derivatives as a function of the parameters x_{sj}, and w is the weight of each independent plane.

For either method we may write b_{ri} in the form

$$b_{ri} = \sum_u k_{sj,u}(|F_{o,u}|-|F_{c,u}|) , \qquad (4·2)$$

thus taking the summation only over symmetry independent F's.

The solutions of (4·1) are

$$\varepsilon_{ri} = \sum_{s,j} c^{-1}_{ri,sj} b_{sj} ,$$

where $c^{-1}_{ri,sj}$ is an element of the matrix inverse to $c_{ri,sj}$. Using (4·2), this can be written

$$\begin{aligned}\varepsilon_{ri} &= \sum_{s,j}\sum_u c^{-1}_{ri,sj}k_{sj,u}(|F_{o,u}|-|F_{c,u}|) \\ &= \sum_u d_{ri,u}(|F_{o,u}|-|F_{c,u}|) , \qquad (4·3)\end{aligned}$$

where

$$d_{ri,u} = \sum_{s,j} c^{-1}_{ri\,sj}k_{sj,u} .$$

Equation (4·3) gives ε_{ri} as a linear sum over independent ΔF_u's.

Accordingly, the variance of ε_{ri} is

$$\sigma^2(\varepsilon_{ri}) = \sum_u (d_{ri,u})^2\sigma^2(F_u) . \qquad (4·4)$$

A simplification of (4·4) is possible for the least-squares method since by definition $w_u = 1/\sigma^2(F_u)$. To achieve this, we rewrite (4·2) as

$$b_{ri} = \sum_u \varkappa_{ri,u} w_u^{1/2}(|F_{o,u}|-|F_{c,u}|) ,$$

and (4·3) as

$$\varepsilon_{ri} = \sum_u \delta_{ri,u} w_u^{1/2}(|F_{o,u}|-|F_{c,u}|) ,$$

where

$$\delta_{ri,u} = \sum_{s,j} c^{-1}_{ri,sj}\varkappa_{sj,u} .$$

Then

$$\sigma^2(\varepsilon_{ri}) = \sum_u (\delta_{ri,u})^2 w_u\sigma^2(F_u) \qquad (4·5)$$

$$= \sum_u (\delta_{ri,u})^2 \qquad (4·6)$$

$$= \sum_u\sum_{s,j}\sum_{t,k}(c^{-1}_{ri,sj}\varkappa_{sj,u})(c^{-1}_{ri,tk}\varkappa_{tk,u}) .$$

But

$$\sum_u \varkappa_{sj,u}\varkappa_{tk,u} = \sum_u w\frac{\partial|F_c|}{\partial x_{sj}}\frac{\partial|F_c|}{\partial x_{tk}} = c_{sj,tk} ;$$

hence

$$\sigma^2(\varepsilon_{ri}) = \sum_{s,j}\sum_{t,k} c^{-1}_{ri,sj}c_{sj,tk}c^{-1}_{ri,tk} = c^{-1}_{ri,ri} . \qquad (4·7)$$

The covariance of ε_{ri} and ε_{sj} is

$$\mathrm{cov}(\varepsilon_{ri},\varepsilon_{sj}) = \sum_u d_{ri,u}d_{sj,u}\sigma^2(F_u) . \qquad (4·8)$$

For least-squares this simplifies to

$$\mathrm{cov}(\varepsilon_{ri},\varepsilon_{sj}) = c^{-1}_{ri,sj} . \qquad (4·9)$$

The essence of the simplification for least squares is the step from (4·5) to (4·6). With any but the weighting appropriate for the standard deviations, we do not have $w_u\sigma^2(F_u) = 1$. The modified differential Fourier method is equivalent to taking an artificial weighting $w_u \propto p_u/f_r$, where p_u is the number of planes related by symmetry in each crystallographic form, but, since $\sigma^2(F_u) \propto f_r/p_u$ is not a necessary relation, no corresponding simplification is possible. For the same reason the correctly weighted least-squares equations, or the Fourier method with equivalently weighted coefficients, lead to the parameter estimates having the lowest variances.

The accuracy of coordinates found by least squares, using the function

$$R_2 = \Sigma w''(|F_o|^2-|F_c|^2)^2 ,$$

or from the corresponding Patterson function, may be derived in a similar manner.

By the central limit theorem (see e.g. Kendall, 1943) the probability distribution of errors for an atomic coordinate is approximately normal, because of the large number of independent F's, irrespective, within rather wide limits, of the distribution laws for the individual F's.

4·2. *Approximate formulae*

Just as the equations for the ε_{ri} may often be approximately simplified (see Cruickshank, 1952), so may the formulae for the accuracy of the parameters. The first simplification is to take an approximate

form for the matrix of coefficients $c_{ri,sj}$. This is discussed in the paper just referred to, and its consequence is that the coefficients $d_{ri,u}$ of (4·4) are simplified. In particular, if all the peaks are spherical and are well resolved without overlapping in an orthogonal cell of a centrosymmetric structure, the matrix is diagonal and we have approximately in the Fourier method

$$\sigma^2(\varepsilon_{ri}) = \left\{\frac{1}{V^2}\frac{4\pi^2}{a_i^2}\sum_u \lambda_{r,u}^2\sigma^2(F_u)\right\}\Big/\left(\frac{\partial^2\varrho}{\partial x_i^2}\right)^2, \quad (4\cdot10)$$

where $(\partial^2\varrho/\partial x_i^2)$ is the second derivative of the electron density evaluated at the position of atom r, and $\lambda_{r,u}$ is a sum of trigonometric terms depending on the plane u and the position of atom r (Cruickshank, 1949*a*), the numerator being the variance of the slope of the density.

The second simplification is that it is then often sufficient to replace $\lambda_{r,u}$ by a term which does not depend on the exact position of atom r, though it may depend on whether r is in a general or a special position in the cell. The various approximations which can be made in this way have been discussed by Cruickshank & Rollett (1953); some less general remarks about this were made earlier by Cruickshank (1949*a*, *b*). In the simplest case of an atom in a general position in many centrosymmetric space groups, (4·10) has the approximate form

$$\sigma^2(\varepsilon_{ri}) = \left\{\frac{1}{V}\frac{4\pi^2}{a_i^2}\sum_3 h_i^2\sigma^2(F)\right\}\Big/\left(\frac{\partial^2\varrho}{\partial x_i^2}\right)^2. \quad (4\cdot11)$$

4·3. *Estimation of* $\sigma(F_u)$

To estimate coordinate standard deviations by the preceding formulae, it is first necessary to estimate the $\sigma(F_u)$. Coordinates found by the least-squares or modified differential Fourier methods (or their appropriate approximations) may be in error owing to:

(*a*) experimental errors in the $|F_o|$'s;
(*b*) imperfections of the molecular model used for the F_c's;
(*c*) errors in the unit cell size;
(*d*) computational approximation errors.

We will not discuss the last two sources of error, which in any case do not enter through the preceding formulae.

Comparison of (4·3) and (4·4) suggests that a simple way of estimating (*a*) and (*b*) together is to take $||F_{o,u}|-|F_{c,u}||$ as an estimate of $\sigma(F_u)$, $F_{c,u}$ being the final value of the calculated structure factor. This procedure certainly allows for the errors in the $|F_o|$'s, and makes some sort of allowance for the imperfections of the F_c's, and has the incidental merit of being easy to apply. It is open to criticism that it treats the errors (*b*) as random, whereas they are systematic, the errors in the different $F_{c,u}$ being correlated. The objection is valid in principle, but the method may be used because of the necessity of making some estimate of (*b*). The check against its improper application is that the difference map should show no strong features attributable to the calculated model. Often this is so, and the method then gives a fairly satisfactory estimate of the errors.

Sometimes it is of interest to find the errors produced by the errors in the $|F_o|$ values only. For this purpose the random errors of each $|F_o|$ may be estimated by the agreement between several independent observations, or, if Geiger-counter techniques are being used, from the expected fluctuations in the number of counts. Unless crystals of different sizes are used, this will not take account of absorption and extinction errors.

These remarks have a bearing on the choice of the weights w in the least-squares method. The weights are sometimes estimated from a study of the experimental errors in the $|F_o|$'s, which, of course, can be made before the refinement starts. With this choice of w's, (4·7) will estimate only the parameter errors due to the experimental errors in the $|F_o|$'s. To include the effects of the imperfections of the calculated model, either (4·4) must be used with estimated $\sigma(F_u)$'s not related to the w's, or another set of w's is necessary. This new set cannot be properly estimated until the refinement is nearly complete, as it must come from a study of the residuals $|F_{o,u}-F_{c,u}|$. Alternatively, sometimes only the relative weights are determined from a study of the experimental errors in the $|F_o|$'s, and at the end of the analysis the estimated variance of the observation of unit weight is taken as

$$s^2 = (\sum_u w|F_{o,u}-F_{c,u}|^2)/(m-n), \quad (4\cdot12)$$

where m is the number of independent observations and n is the number of parameters determined. This method, again, may be open to the objection that the relative weighting is not appropriate to the total errors.

The factor $(m-n) = \nu$ in place of m in (4·12) is statistically correct, and allows for the reduction of the residuals as the number of parameters is increased. Strictly, an effect of this sort ought to be allowed for in the estimation of the errors of the Fourier method, but provided $((m-n)/m)^{\frac{1}{2}}$ is close to unity, the previous discussion will be satisfactory.

5. The accuracy of molecular parameters

The molecular parameters are functions of the atomic parameters. Their standard deviations may be obtained either by expressing them directly in terms of the $\sigma(F)$'s, or in terms of the variances and covariances of the atomic parameters. Thus if a molecular parameter $m = \sum_i l_i x_i$ is a linear function of a number of atomic parameters x_i,

$$\sigma^2(m) = \sum_i \sum_j l_i l_j \operatorname{cov}(x_i, x_j). \quad (5\cdot1)$$

In three-dimensional work the latter often takes a very simple form, but for unresolved bond lengths in two-dimensional projections it is sometimes more convenient to express the bond-length errors directly in terms of the $\sigma(F)$'s.

The simplification for three-dimensional resolved structures is that the covariances between different atoms are approximately zero (except possibly for non-centrosymmetric structures with heavy atoms). On this assumption we now give a short list of the variances and covariances of some molecular parameters in terms of the variances of the atomic parameters.

If l is the bond length between two independent atoms, having variances $\sigma^2(a)$ and $\sigma^2(b)$ in the direction of the bond

$$\sigma^2(l) = \sigma^2(a)+\sigma^2(b) . \quad (5\cdot2)$$

If the bond is across a centre of symmetry

$$\sigma^2(l) = 4\sigma^2(a) . \quad (5\cdot3)$$

If β is the angle formed at B between two bonds AB and BC

$$\sigma^2(\beta) = \frac{\sigma^2(A)}{AB^2}+\sigma^2(B)\left(\frac{1}{AB^2}-\frac{2\cos\beta}{AB.BC}+\frac{1}{BC^2}\right)+\frac{\sigma^2(c)}{BC^2} , \quad (5\cdot4)$$

where $\sigma^2(A)$ and $\sigma^2(C)$ are the variances of A and C in the directions at right angles to AB and BC respectively, and $\sigma^2(B)$ is the variance of B in the direction of the centre of the circle passing through A, B and C. If A and B are related by a centre of symmetry, $\sigma^2(\beta)$ is given by replacing A by the centre of symmetry O, which has no error, and using the half length OA in place of AB.

If two bonds have no common atoms their covariance, α_{ij}, will be zero.

If l_i and l_j are the lengths of two bonds AB and BC, with common atom B,

$$\alpha_{ij} = \mathrm{cov}\ (l_i, l_j) = \sigma^2(B)\cos\beta , \quad (5\cdot5)$$

where β is the angle, and $\sigma^2(B)$ is the variance of B in the direction of the tangent to the circle through A, B and C. If A and B are related by a centre of symmetry

$$\mathrm{cov}\ (l_i, l_j) = 2\sigma^2(B)\cos\beta . \quad (5\cdot6)$$

The covariance between the length AB and the angle β is

$$\mathrm{cov}\ (AB, \beta) = -\frac{\sigma^2(B)\sin\beta}{BC} , \quad (5\cdot7)$$

where $\sigma^2(B)$ is the variance of B in the direction of AB. If either A and B or B and C are related by a centre of symmetry

$$\mathrm{cov}\ (AB, \beta) = -\frac{2\sigma^2(B)\sin\beta}{BC} . \quad (5\cdot8)$$

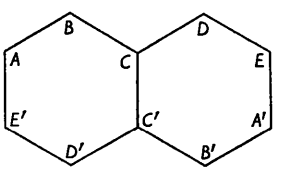
Sometimes the assumed symmetry of a molecule may be higher than its crystallographic symmetry; the different crystallographic values may then be averaged to give the molecular parameters. As an example of the variance of a molecular parameter found in this way, suppose that there is no crystallographic relation between two bonds AB and BC, having a common atom B, but that the molecular symmetry assumes the two bonds equal. The variance of the averaged bond length is then

$$\sigma^2(l) = \tfrac{1}{4}(\sigma^2(A)+4\sigma^2(B)\cos^2\tfrac{1}{2}\beta+\sigma^2(C)) , \quad (5\cdot9)$$

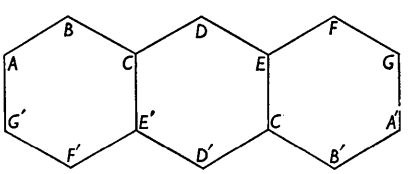
where $\sigma^2(A)$ is the variance of A in the direction AB, $\sigma^2(B)$ is the variance of B in the direction bisecting the angle β, and $\sigma^2(C)$ is the variance of C in the direction CB, and no correlation has been assumed in the errors of A, B and C.

6. Applications to naphthalene and anthracene

Very complete experimental redeterminations of the crystal and molecular structures of naphthalene (Abrahams, Robertson & White, 1949*a*, *b*)



and anthracene (Mathieson, Robertson & Sinclair, 1950; Sinclair, Robertson & Mathieson, 1950)



have been described recently. Ahmed & Cruickshank (1952) have made corrections to these results for finite-series effects, and have estimated the bond-length standard deviations. They found that the chemically equivalent but crystallographically non-equivalent bond lengths all agreed within the estimated standard deviations (e.s.d's), and hence derived weighted mean estimates of the chemically equivalent bonds. These are given in Tables 1 and 2, together with their e.s.d's.

Theoretical bond lengths have been determined by Coulson, Daudel & Robertson (1951) by the method of molecular orbitals. These also are given in Tables 1 and 2, together with the differences between the

Table 1. *Averaged bond lengths in naphthalene*

Bond	Experimental	E.s.d.	Theoretical	Difference	$\lvert t_o\rvert$
AB	1·365 Å	0·006 Å	1·384 Å	−0·019 Å	3·17
BC	1·425	0·005	1·416	+0·009	1·80
AE'	1·404	0·009	1·406	−0·002	0·22
CC'	1·393	0·010	1·424	−0·031	3·10

Table 2. *Averaged bond lengths in anthracene*

Bond	Experimental	E.s.d.	Theoretical	Difference	$\|t_o\|$
AB	1·371 Å	0·006 Å	1·382 Å	−0·011 Å	1·83
BC	1·424	0·005	1·420	+0·004	0·80
CD	1·396	0·004	1·406	−0·010	2·50
AG'	1·408	0·010	1·410	−0·002	0·20
CE'	1·436	0·007	1·430	+0·006	0·86

experimental and theoretical results, and the modulii of the ratios, $|t_o|$, of these differences to the e.s.d's.

The e.s.d's may be regarded as moderately accurate, since there were several hundred independent experimental observations for each structure and $||F_o|-|F_c||$ was used as estimate of $\sigma(F)$. Accordingly, significance tests for the comparison of experiment and theory may be based on the χ^2 distribution, or, for one parameter, on the normal distribution.

The comparisons between theory and experiment will be made first bond by bond, and then for all the bond lengths of each molecule taken together. The latter comparisons show clearly that the theory is considerably better for anthracene than for naphthalene.

For naphthalene the differences between experiment and theory on the bonds *BC* and *AE'* are not significant. For the bond *AB*, $|t_o| = 3{\cdot}17$, and for *CC'*, $|t_o| = 3{\cdot}1$; for both bonds $0{\cdot}01 > P > 0{\cdot}001$, and so the differences are significant. For anthracene the differences for *AB*, *BC*, *AG'* and *CE'* are not significant. For *CD*, $|t_o| = 2{\cdot}5$, with $0{\cdot}05 > P > 0{\cdot}01$, so this difference is possibly significant. These comparisons show that the theory is more satisfactory for anthracene than for naphthalene, but give no overall figure of merit for either molecule. To obtain this, we now apply the multivariate significance tests of § 3.

The e.s.d's of the three coordinates of a given atom vary slightly with direction, and the bond length e.s.d's of Tables 1 and 2 were calculated allowing for this. To simplify the calculation of the covariances between the averaged bond lengths, it was assumed that the errors for a given atom were independent of direction. On this assumption the elements of the variance matrix of the averaged bond lengths of naphthalene in terms of the variances of the atomic coordinates are, taking $\beta = 120°$,

$$\sigma^2(AB) = \tfrac{1}{4}(\sigma^2(A)+\sigma^2(B)+\sigma^2(D)+\sigma^2(E)),$$
$$\sigma^2(BC) = \tfrac{1}{4}(\sigma^2(B)+\sigma^2(D)+4\sigma^2(C)\cos^2\tfrac{1}{2}\beta),$$
$$\sigma^2(AE') = \sigma^2(A)+\sigma^2(E),$$
$$\sigma^2(CC') = 4\sigma^2(C),$$
$$\mathrm{cov}\,(AB, BC) = \tfrac{1}{4}(\sigma^2(B)+\sigma^2(D))\cos\beta,$$
$$\mathrm{cov}\,(AB, AE') = \tfrac{1}{2}(\sigma^2(A)+\sigma^2(E))\cos\beta,$$
$$\mathrm{cov}\,(BC, CC') = -\tfrac{1}{2}(4\sigma^2(C)\cos\tfrac{1}{2}\beta) = -2\sigma^2(C)\cos\tfrac{1}{2}\beta,$$
$$\mathrm{cov}\,(AB, CC') = \mathrm{cov}\,(BC, AE') = \mathrm{cov}\,(AE', CC') = 0.$$

The variance matrix, a_{ij}, was calculated taking the atomic variances as

$$\sigma^2(A) = \sigma^2(E) = 38{\cdot}8\times10^{-6}\ \text{Å}^2,$$
$$\sigma^2(B) = \sigma^2(D) = 32{\cdot}6\times10^{-6}\ \text{Å}^2, \quad \text{and}$$
$$\sigma^2(C) = 24{\cdot}1\times10^{-6}\ \text{Å}^2.$$

This matrix was inverted and

$$T_o^2 = \sum_{i=1}^{4}\sum_{j=1}^{4}\delta_i a^{ij}\delta_j$$

was calculated, the δ's being the differences between the experimental and theoretical bond lengths.

The elements for the variance matrix for anthracene are similar, though

$$\sigma^2(CE') = \sigma^2(C)+\sigma^2(E),$$
$$\mathrm{cov}\,(BC, CE') = \mathrm{cov}\,(CD, CE') = \tfrac{1}{2}(\sigma^2(C)+\sigma^2(E))\cos\beta.$$

The atomic variances used were

$$\sigma^2(A) = \sigma^2(G) = 39{\cdot}7\times10^{-6}\ \text{Å}^2,$$
$$\sigma^2(B) = \sigma^2(F) = 28{\cdot}0\times10^{-6}\ \text{Å}^2,$$
$$\sigma^2(C) = \sigma^2(E) = 20{\cdot}6\times10^{-6}\ \text{Å}^2, \quad \text{and}$$
$$\sigma^2(D) = 21{\cdot}6\times10^{-6}\ \text{Å}^2.$$

For naphthalene $T_o^2 = 23{\cdot}15$, corresponding, for four degrees of freedom, approximately to $P = 0{\cdot}0001$. For anthracene $T_o^2 = 11{\cdot}39$, corresponding, for five degrees of freedom, approximately to $P = 0{\cdot}05$. Thus treating the molecules as wholes the difference between theory and experiment is highly significant for naphthalene, but only possibly significant for anthracene. We see that the multivariate significance tests have clarified the comparison between theory and experiment, and have demonstrated that the theory is much more satisfactory for anthracene than for naphthalene.

It is true that these conclusions are indicated by the r.m.s. differences between experiment and theory, 0·019 Å for naphthalene and 0·007 Å for anthracene, and by the ratio of these to the bond length e.s.d's. But it is only by evaluating T_o^2 that the correlation of errors between the different bond lengths can be properly taken into account, and the comparisons made on a uniform basis.

These comparisons have been made to test the hypothesis that the theoretical values are the true values, and have ignored the admitted imperfections of the theory (Coulson *et al.*, 1951), which are estimated to produce corrections up to 0·015 Å per bond. The value of the comparisons is that they show that it is hardly necessary to postulate any errors in the theory for anthracene but that important errors occur in the theory for naphthalene. One noticeable feature is that the mean theoretical bond lengths are too long for both molecules. For naphthalene the mean experimental bond length (weighted according to the number of bonds of each kind) is 1·396 Å, and the mean theoretical is 1·403 Å. For anthracene the means are 1·403 Å and 1·407 Å. Using the previous coordinate variances,

and allowing for the correlations between bond lengths, the e.s.d's of the mean bond lengths are 0·0028 Å for naphthalene and 0·0021 Å for anthracene. The difference of the means is possibly significant for naphthalene, but, even if the theoretical mean is readjusted, it is still necessary to postulate theoretical errors for this molecule up to about 0·020 Å to obtain satisfactory agreement between theory and experiment.

References

ABRAHAMS, S. C., ROBERTSON, J. M. & WHITE, J. G. (1949*a*). *Acta Cryst.* **2**, 233.

ABRAHAMS, S. C., ROBERTSON, J. M. & WHITE, J. G. (1949*b*). *Acta Cryst.* **2**, 238.

AHMED, F. R. & CRUICKSHANK, D. W. J. (1952). *Acta Cryst.* **5**, 852.

BOOTH, A. D. (1945). *Nature, Lond.* **156**, 51.

BOOTH, A. D. (1946). *Proc. Roy. Soc.* A, **188**, 7.

COULSON, C. A., DAUDEL, R. & ROBERTSON, J. M. (1951). *Proc. Roy. Soc.* A, **207**, 306.

COX, E. G. & CRUICKSHANK, D. W. J. (1948). *Acta Cryst.* **1**, 92.

CRUICKSHANK, D. W. J. (1949*a*). *Acta Cryst.* **2**, 65.

CRUICKSHANK, D. W. J. (1949*b*). *Acta Cryst.* **2**, 154.

CRUICKSHANK, D. W. J. (1952). *Acta Cryst.* **5**, 511.

CRUICKSHANK, D. W. J. & ROLLETT, J. S. (1953). *Acta Cryst.* **6**, 705.

HUGHES, E. W. (1941). *J. Amer. Chem. Soc.* **63**, 1737.

KENDALL, M. G. (1943). *The Advanced Theory of Statistics*, vol. 1. London: Griffin.

KENDALL, M. G. (1946). *The Advanced Theory of Statistics*, vol. 2. London: Griffin.

MATHIESON, A. McL., ROBERTSON, J. M. & SINCLAIR, V. C. (1950). *Acta Cryst.* **3**, 245.

SINCLAIR, V. C., ROBERTSON, J. M. & MATHIESON, A. McL. (1950). *Acta Cryst.* **3**, 251.

WEATHERBURN, C. E. (1947). *Mathematical Statistics*. Cambridge: University Press.

Acta Cryst. (1953). **6**, 705

Electron-Density Errors at Special Positions

BY D. W. J. CRUICKSHANK AND J. S. ROLLETT

Chemistry Department, The University, Leeds 2, England

(*Received* 10 *February* 1953)

Approximate formulae are given for the electron-density and slope errors at special positions in any space group. An example shows that the errors at special positions can be several times those at general positions.

Density errors

The electron density $\varrho(x, y, z)$ is given by the triple Fourier series of structure factors, $F(hkl)$, as

$$\varrho(x, y, z) = \frac{1}{V}\sum_{3}|F|\cos\{2\pi(hx/a+ky/b+lz/c)-\alpha\}$$

$$= \frac{1}{V}\sum_{3}|F|\cos(\theta-\alpha)\,, \quad \text{say}\,. \qquad (1)$$

The F's may be related by symmetry, and in terms of the independent F's we may write (1) as

$$\varrho = \frac{1}{V}\sum_{\text{indep.}}|F|\sum_{\text{form}}\cos(\theta-\alpha)\,, \qquad (2)$$

where the inner summation is over all planes of the same crystallographic form.

If each independent $|F|$ has a standard deviation $\sigma(F)$, the standard deviation of the error in the electron density, by the law for the combination of errors, is

$$\sigma(\varrho) = \frac{1}{V}\left\{\sum_{\text{indep.}}\sigma^2(F)\left[\sum_{\text{form}}\cos(\theta-\alpha)\right]^2\right\}^{\frac{1}{2}}. \qquad (3)$$

Equation (3) shows that the error varies from point to point in the unit cell. However, if there are a large number of terms in the summation, the error is nearly constant over large regions of the cell, as (3) is the sum of squares of cosine terms. These approximate values of the error depend on whether the position considered is a general one (x, y, z), or a special one, such as $(0, 0, z)$ or $(x, \bar{x}, z)$, and on the space group.

For the type of position considered let those planes in each form with the same $|\cos(\theta-\alpha)|$ be said to constitute a sub-form. Let

$$m = \sum_{\text{sub-form}}\cos(\theta-\alpha)/|\cos(\theta-\alpha)|\,, \qquad (4)$$

so that if all the planes in a sub-form have the same $\cos(\theta-\alpha)$, m is just the number of planes in that sub-form. Let η be the r.m.s. value of $\cos(\theta-\alpha)$ in a sub-form for positions of the given type; then approximately

$$\sigma(\varrho) = \frac{1}{V}\left\{\sum_{\text{all sub-forms}}[\eta m\sigma(F)]^2\right\}^{\frac{1}{2}}, \qquad (5)$$

since, on squaring the summation for each form in (3),