Acta Cryst. (1977). A**33**, 441–444

Test of the Behaviour of the MDKS Formula

By B. Busetta

Laboratoire de Cristallographie et de Physique Cristalline associé au CNRS, Université de Bordeaux I,
351 *cours de la Libération*, 33405 *Talence, France*

(*Received* 2 *October* 1976; *accepted* 14 *January* 1977)

A study of the efficiency of functions related to the MDKS formula in sorting the triplet invariants was performed on eight solved structures. Both D and S terms were checked separately. A new mixed function $D+S$ is described. The K scale factor departs from the commonly used values and depends on the nature of the structure.

Introduction

In crystal structure determination, the number of unknown parameters is much less than the number of available data. Thus we have an overdetermined system and relations between the data may be computed.

For instance, the triplet cosine invariants may be estimated from the moduli of structure factors and their use in direct methods is a valuable improvement.

If agreement is reached on the dependence of the triplet cosine invariants on the moduli of the (normalized) structure factors, problems begin with the mathematical formulation of this dependence. Indeed, the mathematical approach is complicated and approximations must always be made.

In our view, the accuracy of the experimental data is not sufficient to obtain triplet cosine invariants from a mathematical equation. The different cosine invariant computations merely provide a means of sorting the cosine invariants rather than the exact value of the cosines (Busetta & Comberton, 1974).

The triplet cosine invariants are considered as depending on statistical moments of increasing order: moment of order $1/\sqrt{N}$, *i.e.*

$$|E_{H_1}E_{H_2}E_{H_1+H_2}|$$

(N = number of atoms in the cell);
moment of order $1/N$, *i.e.*

$$\sum_K (|E_K|^2-1)(|E_{H_1+K}|^2-1) + (|E_K|^2-1)(|E_{H_1+H_2+K}|^2-1) + (|E_{H_1+K}|^2-1)(|E_{H_1+H_2+K}|^2-1);$$

moment of order $1/N\sqrt{N}$, *i.e.*

$$\sum_K (|E_K|^2-1)(|E_{H_1+K}|^2-1)(|E_{H_1+H_2+K}|^2-1).$$

In a recent paper (Giacovazzo, 1976) the posititivity of the triplet cosine invariants in $P\bar{1}$ is related to the values of the moments of order $1/N$ (called Q) and $1/(N\sqrt{N})$ (called A) by a function $G=k[1+A/(1+Q)]$, where k is positive, and $1+Q$ corresponds to the variance associated with the value of A.

Another suitable function to sort the triplet cosine invariants (Hauptman, 1972) is the constrained MDKS formula where D is a constrained (or conditional) form of A:

$$D=\langle \varepsilon_{H_1+H_2+K} \| E_K|>t, |E_{H_1+K}|>t\rangle_K$$

with $\varepsilon_H=|E_H|^p-\langle |E_K|^p\rangle_K$ (p is generally 2), and S a constrained form of Q:

$$S=\langle \varepsilon_{H_1+K} \| E_K|>t\rangle_K+\langle \varepsilon_{H_2+K} \| E_K|>t\rangle_K + \langle \varepsilon_{H_3+K} \| E_K|>t\rangle_K .$$

A fast computation of D was described in a previous paper (Busetta, 1976). With Giacovazzo's (1976) notation, the sorting function involved by the MDKS formula is $H=m(A-kQ)$ where k and m are positive. The triplet cosine invariant and Q vary in opposite senses. Conversely, a recent approximation [Giacovazzo (1976), equation (12)] forecasts variations in the same sense:

$$B=1+Q\simeq \frac{2}{P}(|E_{H_1}|^2+|E_{H_2}|^2+|E_{H_1+H_2}|^2) - 3\left(\frac{1}{N}+\frac{1}{N\sqrt{N}}|E_{H_1}E_{H_2}E_{H_1+H_2}|\cos\varphi\right).$$

To clear up this problem we checked these expressions on different solved structures.

Sorting factor

We have pointed out that the largest errors in the estimated triplet cosine invariants are due to positive cosines computed negative or near zero (Busetta & Comberton, 1974; Busetta, 1976). As the actual number of positive cosines is much greater than the number of negative ones, the number of cosines wrongly estimated as negative is sufficiently large to give a mean error between the estimated and actual cosine invariants greater than the corresponding error obtained if all the cosines were set equal to +1·0.

This is why we introduced the weight $A'_{H,K}$ (Busetta

& Comberton, 1974) using only the positive estimated triplet cosines. Because of these spurious negative cosines, in order to compare easily the efficiency of different formulae for a number of solved structures, it is meaningless to use the mean error in the estimated triplet cosine invariants. We define therefore a sorting factor as follows. The triplet cosine invariants are grouped in sets of 100 elements in which $A_{H,K}(2\sigma_3/\sigma_2^{3/2}|E_HE_KE_{H-K}|)$ may be considered as constant. In a set, the mean cosine invariant is $\cos_{\text{moy}}$, roughly equal to the theoretical value $I_1(A)/I_0(A)$. We denote $\cos_{\text{true}(+)}$ and $\cos_{\text{true}(-)}$ as the averages respectively of the actual 25 greatest and 25 least cosine invariants. Then, the triplet invariants of this set are ranked according to decreasing values of the sorting F function. We denote $\cos_{\text{est}(+)}$ and $\cos_{\text{est}(-)}$ as the mean values of the actual cosine invariants of the 25 triplet invariants respectively at the top and bottom of this sorted list.

The efficiency of the sorting may be defined by the ratios:

$$\alpha_+=\frac{\cos_{\text{est}(+)}-\cos_{\text{moy}}}{\cos_{\text{true}(+)}-\cos_{\text{moy}}},\quad \alpha_-=\frac{\cos_{\text{est}(-)}-\cos_{\text{moy}}}{\cos_{\text{true}(-)}-\cos_{\text{moy}}}.$$

α_+ is the sorting factor corresponding to the top of the ranked list and α_- to the bottom. The spurious negative estimation of positive invariants involves $\alpha_+>\alpha_-$.

The maximum value of α is $+1$, where the 25 greatest (or least) cosine invariants are correctly selected (perhaps not in the right order, but as all the proposed formulae involve α values far from $+1$, this drawback is no problem). An α value near zero proves no efficiency of the sorting function while a negative value means that the sorting function must work in the reverse sense.

In this paper, we report for each test a single value which is the average of the α values of the different sets of 100 cosine invariants.

Mixed function

In each set of 100 cosine invariants, associated with a given mean A value, the mean triplet cosine invariant is $I_1(A)/I_0(A)$ associated with a r.m.s.d. $\sigma_{\cos}(A)$ (Hauptman, 1972).

The actual values of the cosine invariants ($\sigma=0$) are supposed to follow the theoretical distribution f_A (Hauptman, Fisher, Hancock & Norton, 1969). If, with a sorting function D, the rank r_D is attributed to a triplet cosine invariant its most probable value is:

$$\alpha_Df_A^{-1}(r_D)+(1-\alpha_D)\,\frac{I_1(A)}{I_0(A)}$$

associated with a r.m.s.d. $\sigma=\sqrt{(1-\alpha_D)}\sigma_{\cos}(A)$, where $0<\alpha_D<1$ is the estimated efficiency of the sorting function D (Busetta & Comberton, 1974).

Another sorting function S will give for the same cosine invariant

$$\alpha_Sf_A^{-1}(r_S)+(1-\alpha_S)\,\frac{I_1(A)}{I_0(A)}$$

associated with a r.m.s.d. $\sigma=\sqrt{(1-\alpha_S)}\sigma_{\cos}(A)$.

Combination of these two values, weighted by $1/\sigma^2$, provides a new most probable cosine:

$$\frac{1}{\dfrac{1}{1-\alpha_S}+\dfrac{1}{1-\alpha_D}}\left\{\frac{1}{1-\alpha_D}\left[\alpha_Df_A^{-1}(r_D)+(1-\alpha_D)\frac{I_1(A)}{I_0(A)}\right]\right.$$
$$\left.+\frac{1}{1-\alpha_S}\left[\alpha_Sf_A^{-1}(r_S)+(1-\alpha_S)\frac{I_1(A)}{I_0(A)}\right]\right\},$$

that is to say

$$\frac{(1-\alpha_S)\alpha_D}{2-\alpha_S-\alpha_D}\left\{f_A^{-1}(r_D)+\frac{\alpha_S}{\alpha_D}\,\frac{1-\alpha_D}{1-\alpha_S}\,f_A^{-1}(r_S)\right.$$
$$\left.+2\,\frac{1-\alpha_D}{\alpha_D}\,\frac{I_1(A)}{I_0(A)}\right\}. \qquad (2)$$

For each function S (or D) two sorting factors α_{S+} and α_{S-} (or α_{D+} and α_{D-}) are defined according to whether the rank r_S (or r_D) corresponds to the top or the bottom of the ranked list. We use α_+ if $f^{-1}(r)>I_1(A)/I_0(A)$ and α_- if not. If S and D have their MDKS meaning, (2) may be considered as a modified MDKS formula, where K may have four different values corresponding to $(\alpha_{D+},\alpha_{S+})$, $(\alpha_{D+},\alpha_{S-})$, ...

Dependence of the structure

Theoretically, direct methods deal with randomly distributed atoms. This is not the case in most actual structures where some dominant features (polycyclic framework, full extended chain, *etc.*) involve a large number of Patterson overlaps. The presence of coincident interatomic vectors results in an increase in the average values $\langle(|E_K|^2-1)^2\rangle_K$ and $\langle(|E_K|^2-1)^3\rangle_K$ (Hauptman, 1964).

It was pointed out that, in this case, the MDKS formula is more efficient for computing the triplet cosine invariants than the usual triple-products formula or D (Duax, Weeks & Hauptman, 1972). We intend to check on solved structures the efficiency of the S term in sorting the triplet cosine invariants.

Table 1 gives the sorting factors obtained with S, D, and $S+D$ as sorting functions for different solved structures, according to increasing values of $\langle(|E_K|^2-1)^2\rangle_K$. All the computations were done with $p=0{\cdot}50$ in the expressions for S and D to avoid discrepancies produced by large values of E. When both α_{S+} and α_{S-} were computed negative, a new sorting was done according to the increasing values of S, *i.e.* in the sense used in the MDKS formula.

The use of the D function in sorting the triplet cosine invariants is always worthwhile and should be standard in direct methods. Except in the case of CAF (see the first column in Table 1 for the abbreviations) the α_{D+}

values are always $>0{\cdot}20$. The $\alpha_{D+}=0{\cdot}90$ observed in the case of the centrosymmetric structure FUR is not significant. When sorting with the S function, the sorting factor decreases as the mean value $m=\langle(|E_K|^2-1)^2\rangle_K$ increases. For small values of m (say $m<1{\cdot}8$), α_S is positive. S must work in the sense forecast in Giacovazzo's approximation. For large values of m ($m>1{\cdot}8$) α_S is negative. Then S must work in the sense forecast in MDKS.

When $m<1{\cdot}8$ the mixed $S+D$ function does not give a more efficient sorting than D. In this case S and D influence the triplet cosines in the same way, *i.e.* S involves roughly the same errors as D. Conversely, when m is large, we have generally $\alpha_{S+D}>\alpha_S+\alpha_D$ and S seems to correct the errors involved in D.

Dependence on the mathematical expressions used in S and D

Table 2 gives the sorting factors observed in the estimation of the triplet cosine invariants for different values of p in the expressions of S and D. To save core during the computations and also to avoid discrepancies which may be involved by large values of $|E|^p$, we fixed the upper value of $|E|^p$ at 2·56.

For D, it is difficult to find a general law and the results are roughly the same whatever the value of p, except in the case of CAF for which $p=2$ seems more efficient.

For S, the required expression is different according to the nature of the structure. When there are few

Table 1. *Dependence of the sorting factors on the nature of the structure*

$\overset{\bullet}{\alpha}_S$ corresponds to sorting according to increasing values of S (MDKS sense).

		$\langle(\lvert E_K\rvert^2-1)^2\rangle_K$	$\langle(\lvert E_K\rvert^2-1)^3\rangle_K$	α_{S+}	α_{S-}	$\overset{\bullet}{\alpha}_{S+}$	$\overset{\bullet}{\alpha}_{S-}$	α_{D+}	α_{D-}	α_{S+D+}	α_{S+D-}
Theoretical		1·00	2·00								
AZE	3-chloro-1,2,4-triphenylazetidin-2-one, $Pca2_1$, 2($ClN_8O_{21}H_{16}$) (Colens, Declercq, Germain, Putzeys & van Meerssche, 1974)	$1{\cdot}20_4$	$3{\cdot}22_7$	0·168	0·111			0·491	0·323	0·507	0·310
EST	2,3-dimethoxy-13,8-diazaestron $P2_12_12_1$, $C_{17}O_3N_2H_{22}$ (unpublished results)	$1{\cdot}41_8$	$4{\cdot}79_5$	−0·016	0·013			0·211	0·204	0·199	0·216
DEA	deaza-1-isotubercidin picrate $P2_1$, $C_{12}H_{16}N_3O_4.C_6H_2N_3O_7$ (Ducruix, Riche & Pascard, 1976)	$1{\cdot}52_9$	$5{\cdot}66_7$	0·066	0·063			0·354	0·270	0·356	0·276
TIM	timolol maleate $P1$, 2($SO_3N_4C_{13}H_{24}.C_4O_4H_4$) (Gadret & Leger, to be published)	$1{\cdot}57_5$	$6{\cdot}03_6$	0·061	0·028			0·499	0·336	0·499	0·333
VAL	valinomycin, $P1$ 2($C_{54}H_{90}N_6O_{18}$) (Karle, 1975)	$1{\cdot}91_6$	$9{\cdot}57_5$	−0·066	−0·027	0·079	0·034	0·413	0·258	0·418	0·281
CAF	complex caffein–sulfanylamide $P2_1/c$, $C_8N_4O_2H_{10}.SO_3N_2C_8H_{10}$ (Gadret, Carpy & Leger, to be published)	$2{\cdot}00_1$	$9{\cdot}21_6$	−0·219	−0·082	0·245	0·073	0·065	0·220	0·351	0·298
DAU	daunomycin hydrochloride $P2_1$ $C_{27}H_{29}NO_{10}{\cdot}HCl$ (Courseille, to be published)	$2{\cdot}00_8$	$12{\cdot}64_2$	−0·196	−0·111	0·226	0·096	0·283	0·117	0·341	0·227
FUR	ethyl furocoumarylate $P2_1/n$ $C_{14}O_5H_{10}$ (Bravic, to be published)	$3{\cdot}34_6$	$25{\cdot}65_0$	−0·333	−0·085	0·254	0·111	0·900	0·687	0·900	0·697

Table 2. *Dependence of the sorting factors on the mathematical expression used for S and D*

For each structure the α_+ and α_- factors are reported respectively in the right and the left columns. For structures with $\langle(|E|^2-1)^2\rangle>1{\cdot}8$ the sorting was according to increasing values of S (*i.e.* MDKS sense).

		AZE		EST		VAL		CAF		DAU		FUR	
S	$p=0{\cdot}25$									0·183	0·105		
	0·50	0·178	0·109	−0·016	0·013	0·079	0·034	0·245	0·073	0·226	0·093	0·254	0·111
	1·00	0·272	0·153	0·001	−0·014	0·108	0·032	0·167	0·041	0·153	0·057	0·225	0·093
	2·00	0·324	0·181	0·038	0·017	0·065	0·030	−0·044	0·041	0·131	0·059	0·083	0·067
	3·00	0·319	0·188										
D	$p=0{\cdot}25$									0·267	0·109		
	0·50	0·491	0·323	0·211	0·204	0·413	0·258	0·065	0·220	0·280	0·121	0·900	0·687
	1·00	0·489	0·320	0·209	0·192	0·451	0·264	0·081	0·180	0·311	0·115	0·860	0·657
	2·00	0·491	0·330	0·229	0·197	0·491	0·305	0·149	0·206	0·263	0·129	0·860	0·708

Table 3. *Influence of wrong thermal parameters on the sorting factors*

All the computations were performed with $p=0{\cdot}50$. The actual mean thermal parameter is underlined.

	AZE			TIM			CAF			DAU		
S	$\underline{B=4{\cdot}0}$	$\underline{0{\cdot}168}$	$\underline{0{\cdot}111}$	$B=4{\cdot}50$	0·014	0·024	$B=0{\cdot}5$	0·214	0·064	$B=0{\cdot}8$	0·204	0·077
	$B=8{\cdot}0$	0·201	0·049	$B=6{\cdot}00$	0·024	0·025	$\underline{B=4{\cdot}0}$	$\underline{0{\cdot}245}$	$\underline{0{\cdot}073}$	$B=3{\cdot}50$	0·211	0·089
				$\underline{B=7{\cdot}60}$	$\underline{0{\cdot}061}$	$\underline{0{\cdot}028}$				$\underline{B=4{\cdot}90}$	$\underline{0{\cdot}216}$	$\underline{0{\cdot}091}$
				$B=9{\cdot}10$	0·097	0·029				$B=6{\cdot}50$	0·207	0·088
D	$\underline{B=4{\cdot}0}$	$\underline{0{\cdot}491}$	$\underline{0{\cdot}324}$	$B=4{\cdot}50$	0·456	0·315	$B=0{\cdot}5$	0·296	0·225	$B=0{\cdot}8$	0·174	0·082
	$B=8{\cdot}0$	0·295	0·192	$B=6{\cdot}00$	0·481	0·323	$\underline{B=4{\cdot}0}$	$\underline{0{\cdot}065}$	$\underline{0{\cdot}220}$	$B=3{\cdot}50$	0·276	0·120
				$\underline{B=7{\cdot}60}$	$\underline{0{\cdot}499}$	$\underline{0{\cdot}336}$				$\underline{B=4{\cdot}90}$	$\underline{0{\cdot}297}$	$\underline{0{\cdot}123}$
				$B=9{\cdot}10$	0·474	0·268				$B=6{\cdot}50$	0·288	0·106

Patterson overlaps ($m<1{\cdot}8$) the best value of p is 2. Conversely, when a large number of Patterson overlaps occur ($m>1{\cdot}8$) a p value near 0·5 is best.

Error involved by errors in E factors

The theory deals with ideal values, but this is not the case in practice. First, the observed structure factors are affected by unavoidable experimental errors, but the main errors in the E factors are due to a wrong determination of the mean thermal parameter. For instance, when the resolution is too low, and this is often the case with large molecules, the computed thermal parameter is underestimated.

We have studied how the efficiency of both S and D sorting functions is affected by incorrect thermal parameters. Table 3 gives the sorting factors observed in the estimation of triplet cosine invariants for E values computed with different thermal parameters.

The efficiency of the D function remains unchanged for a large range of B values and therefore errors in the estimation of B are not troublesome. The results for the S function are more surprising and it seems that for small values of m, an unreasonable increase of B is efficient in improving the sorting of positive triplet cosine invariants. An examination of the reflexions directly involved in the triplet invariants clearly shows that the ability of these reflexions to give positive cosine invariants is dependent on the space vector ($\sin\theta$) of the reflexion and increases with increasing $\sin\theta$. This observation may be an explanation of the improvement of the S sorting when B increases abnormally.

Conclusion

It was pointed out (Duax *et al.*, 1972) that MDKS 'affords a better evaluation of the cosine invariants for structures with great overlap in the Patterson synthesis'. Conversely, the modified triple product evaluation is 'adequate' when there is no extensive overlap in the Patterson synthesis. Our study allows us to define the border between these two cases [a mean value $\langle(|E_H|^2-1)^2\rangle_H$ around 1·8]. It also shows that the use of the MDKS formula when no overlap occurs may be unwise, because K is positive in that case.

We showed that the value of p in $\varepsilon_H=|E_H|^p-\langle|E|^p\rangle$ may have values between 0·5 and 2 in the evaluation of D, but must have fixed values (0·5 or 2) in the evaluation of S. In all the cases the use of D (or $D+KS$) affords worthwhile improvements in the determination of the structure by direct methods.

References

Busetta, B. (1976). *Acta Cryst.* A**32**, 359–362.

Busetta, B. & Comberton, G. (1974). *Acta Cryst.* A**30**, 564–568.

Colens, A., Declercq, J. P., Germain, G., Putzeys, J. P. & van Meerssche, M. (1974). *Cryst. Struct. Commun.* **3**, 119–122.

Duax, W. L., Weeks, C. M. & Hauptman, H. (1972). *Acta Cryst.* B**28**, 1857–1863.

Ducruix, A., Riche, C. & Pascard, C. (1976). *Acta Cryst.* B**32**, 2467–2471.

Giacovazzo, C. (1976). *Acta Cryst.* A**32**, 927–1038.

Hauptman, H. (1964). *Acta Cryst.* B**17**, 1421–1433.

Hauptman, H. (1972). *Crystal Structure Determination, The Role of the Cosine Seminvariants.* New York: Plenum.

Hauptman, H., Fisher, J., Hancock, H. & Norton, D. A. (1969). *Acta Cryst.* B**25**, 811–814.

Karle, I. L. (1975). *J. Amer. Chem. Soc.* **97**, 4379–4385.