component of the potential. On the other hand, it is asymptotic in character and when only a finite number of terms are taken the error is less in magnitude than the first term omitted. Thus, although it can be useful numerically for sufficiently weak anharmonicities its accuracy is essentially limited and, for strong anharmonicities, may not be adequate. This is particularly relevant when dealing with data of such high accuracy that deviations of only a few percent in the temperature factor become significant.

Since for the moment expansion, $T_m(q)$, the inverse FT, $p_m(x)$, always exists, one might have hoped that $p_m(x)$ could be used to represent the p.d.f. even if the perturbation approximation were poor, *i.e.* if $T_m(q)$ differed significantly from $T(q)$. However, the examples considered show that $p_m(x)$ then has regions where it takes significantly negative values despite the fact that it is supposed to be approximating a positive function.

Similar considerations apply to the cumulant expansion. It can be a useful approximation but because of the theorem due to Marcinkiewicz (see Lukacs, 1970) it cannot be the Fourier transform of any p.d.f.; specifically, it must always be positive and therefore cannot follow through a zero of the temperature factor to the negative values that must occur somewhere. Another mathematical difficulty relating to this zero occurs because the cumulant form arises on taking the logarithm of the moment form and so the series for the cumulant form cannot possibly converge for values of q beyond the zero where the logarithm has its singularity.

The numerical results do not allow a clear choice between the moment and the cumulant expansions. Significant differences between the two approximations appear only when perturbation theory is starting to break down. Under these circumstances a procedure that does not rely on perturbation theory is desirable. If one wishes to retain the OPP approach a suitable procedure would be to express the temperature factor directly as the FT of the exact Boltzmann distribution in accordance with equations (1), (2) and (3). Such computations are well within the capabilities of modern large computers.

References

CRAMÉR, H. (1946). *Mathematical Methods of Statistics.* Princeton Univ. Press.

DAWSON, B., HURLEY, A. C. & MASLEN, V. W. (1967). *Proc. R. Soc. London Ser. A*, **298**, 289-306.

KONTIO, A. & STEVENS, E.D. (1982). *Acta Cryst.* A**38**, 623-629.

KURKI-SUONIO, K., MERISALO, M. & PELTONEN, H. (1979). *Phys. Scr.* **19**, 57-63.

LUKACS, E. (1970). *Characteristic Functions*, 2nd. ed, p. 213. London: Griffin.

MAGNUS, W. & OBERHETTINGER, F. (1949). *Formulas and Theorems for the Special Functions of Mathematical Physics* (Engl. transl.), p. 80. New York: Chelsea Publishing Co.

MAIR, S. L. (1980). *J. Phys. C*, **13**, 1419-1425.

SCHERINGER, C. (1984*a*). *Acta Cryst.* A**41**, 73-79.

SCHERINGER, C. (1984*b*). *Acta Cryst.* A**41**, 79-81.

TANAKA, K. & MARUMO, F. (1983). *Acta Cryst.* A**39**, 631-641.

WILLIS, B. T. M. (1969). *Acta Cryst.* A**25**, 277-300.

Acta Cryst. (1985). A**41**, 85-89

The Estimation of Average Molecular Dimensions. 2.* Hypothesis Testing with Weighted and Unweighted Means

BY ROBIN TAYLOR AND OLGA KENNARD

Crystallographic Data Centre, University Chemical Laboratory, Lensfield Road, Cambridge CB2 1EW, England

(*Received* 8 *June* 1984; *accepted* 6 *September* 1984)

Abstract

The average value (μ) of a molecular dimension may be estimated by a weighted ($\bar{x}_w$) or unweighted ($\bar{x}_u$) mean. Computer simulations show that $\bar{x}_u$ can be used in hypothesis tests, since the distribution of $(\bar{x}_u - \mu)/\sigma(\bar{x}_u)$ is closely approximated by Student's t distribution. In contrast, hypothesis tests based on the weighted mean are inexact and potentially misleading.

* Part 1: Taylor & Kennard (1983).

I. Introduction

In a previous paper (Taylor & Kennard, 1983) we discussed some of the problems involved in estimating average molecular dimensions from crystallographic data. The average value of k observations of a molecular dimension (x_i, $i = 1, 2, \ldots, k$) is usually estimated in one of two ways. The simplest procedure is to calculate the unweighted mean, $\bar{x}_u$:

$$x_u = \sum_{i=1}^{k} x_i/k, \qquad (1)$$

0108-7673/85/010085-05$01.50

the standard error of which is given by

$$\sigma(x_u)=\left[\sum_{i=1}^{k}(x_i-\bar{x}_u)^2/k(k-1)\right]^{1/2}. \quad (2)$$

Alternatively, a weighted mean ($\bar{x}_w$) can be used:

$$\bar{x}_w=\sum_{i=1}^{k}w_ix_i\Big/\sum_{i=1}^{k}w_i, \quad (3)$$

where

$$w_i=1/\sigma^2(x_i). \quad (4)$$

Here, $\sigma(x_i)$ is the least-squares e.s.d. of x_i, multiplied by an empirical 'correction factor' to allow for systematic errors in the diffraction experiment. A correction factor of 1·5 is probably adequate for most purposes (Hamilton & Abrahams, 1970; Taylor & Kennard, 1985). The standard error of $\bar{x}_w$ may be estimated from the formula

$$\sigma(\bar{x}_w)=\left[1\Big/\sum_{i=1}^{k}w_i\right]^{1/2}. \quad (5)$$

In general, the ith observation of a molecular dimension can be expressed as

$$x_i=\mu_i+\varepsilon_i=\mu+(\mu_i-\mu)+\varepsilon_i, \quad (6)$$

where μ_i is the true value of the dimension in the ith crystal structure and ε_i is the experimental error in its measurement. As a first approximation, ε_i may be regarded as a random variable from a normal distribution with zero mean and standard deviation $\sigma(x_i)$. If the molecular dimension is not significantly affected by changes in its chemical and crystallographic environment, then

$$\mu_i\simeq\mu_j,\ i\neq j \quad (7)$$

and (6) becomes

$$x_i=\mu+\varepsilon_i. \quad (8)$$

However, if the dimension is sensitive to changes in its environment, then

$$\mu_i\neq\mu_j,\ i\neq j, \quad (9)$$

i.e. the true value of the dimension in the ith crystal structure will differ from the true value in the jth structure. In this case, we are effectively trying to determine the population mean, μ, of the μ_i, *i.e.* the average value of the dimension over all possible environments.

If we were to estimate the average value of the dimension from the ith observation alone, the standard error of the estimate would be $[\sigma^2(\mu)+\sigma^2(x_i)]^{1/2}$; this quantity may be called the *precision of x_i as an estimate of the mean.* The first term, $\sigma^2(\mu)$, accommodates the uncertainty in the estimate due to environmental effects, *i.e.* $\sigma^2(\mu)$ represents the real *physical* variation in the dimension as the crystal-field environment is changed. The second term, $\sigma^2(x_i)$, accommodates the uncertainty in the estimate due to experimental errors in the measurement of x_i. If $\sigma^2(\mu)>\sigma^2(x_i)$ for all i (*i.e.* the dimension is very sensitive to changes in its environment), the various observations will not differ much in precision. Consequently, the average value of the dimension should be estimated by $\bar{x}_u$ rather than $\bar{x}_w$. Conversely, when $\sigma^2(\mu)$ is small, the observations will differ in precision and the weighted mean may be preferable to the unweighted mean. However, if there are uncertainties in the $\sigma(x_i)$ [*i.e.* if $\sigma(x_i)$ is not an exact estimate of the standard deviation of ε_i in (8)], the weights used in calculating $\bar{x}_w$ will be subject to sampling errors. The true standard error of $\bar{x}_w$ is then likely to exceed the value given by (5). This will also be the case if environmental effects are not insignificant, *i.e.* $\sigma^2(\mu)$ is not negligibly small.

In this paper, we consider the use of weighted and unweighted means in hypothesis tests. Specifically, we discuss how to test the null hypothesis

$$H_0\colon \bar{x}=\mu_0$$

against the alternative hypothesis

$$H_1\colon \bar{x}\neq\mu_0,$$

where $\bar{x}$ is the estimated average value of a molecular dimension and μ_0 is some postulated value. Depending on whether $\bar{x}$ is an unweighted or weighted mean, it is natural to base the test on the statistics

$$d_u=(\bar{x}_u-\mu_0)/\sigma(\bar{x}_u) \quad (10)$$

or

$$d_w=(\bar{x}_w-\mu_0)/\sigma(\bar{x}_w). \quad (11)$$

H_0 will be rejected in favour of H_1 if $|d_u|$ (or $|d_w|$) exceeds some critical value, d_α. In order to compute d_α, we must know the significance level of the test ($=\alpha$) and the distribution of the test statistic. The former may be chosen arbitrarily by the analyst and is not considered here. However, the latter depends on several factors and is investigated below by computer simulation.

II. Methodology: generation of d_u and d_w values by simulation

Several computer simulations were performed, some to investigate the distribution of d_u and others to investigate the distribution of d_w. The procedure used to generate simulated values of d_u and d_w was as follows. Four parameters were set at the beginning of the simulation: sample size k; minimum e.s.d. $\sigma_{\min}$; maximum e.s.d. $\sigma_{\max}$; variance due to environmental effects $\sigma^2(\mu)$ [in some simulations, a fifth parameter, $\sigma(f)$, was used: see III-3]. Pseudo-random number generators (*NAG Fortran Library Manual*, 1983) were then used to generate an artificial sample of observations $[x_i, \sigma(x_i), i=1,2,\ldots,k]$. Each $\sigma(x_i)$ was chosen

at random from a uniform distribution in the range σ_{min}–σ_{max}. Each x_i was drawn from a normal distribution* with mean = 1, standard deviation = $[\sigma^2(\mu)+\sigma^2(x_i)]^{1/2}$. In simulations designed to investigate d_u, the k observations thus generated were used to calculate $\bar{x}_u$ and $\sigma(\bar{x}_u)$, from (1) and (2) respectively. The statistic d_u was then computed from (10), using $\mu_0 = 1$ (this corresponds to the null hypothesis being true, since each x_i was drawn from a distribution with unit mean). In simulations designed to investigate d_w, the quantities $\bar{x}_w$ and $\sigma(\bar{x}_w)$ were calculated from (3) and (5), and d_w from (11) (again with $\mu_0 = 1$). Further values of d_u or d_w were generated by repetition of the complete procedure.

III. Results

III-1. *Unweighted mean, environmental effects large*

This is the simplest case. When $\sigma^2(\mu) \gg \sigma^2(x_i)$ for all i, the observations do not vary in precision. If the null hypothesis is true, standard statistical theory shows that d_u follows Student's t distribution with $(k-1)$ degrees of freedom (Cruickshank & Robertson, 1953). Thus, H_0 can be rejected in favour of H_1 if $|d_u|$ exceeds the tabulated value of $t_{k-1,\alpha}$. The test will be at the $100(1-\alpha)\%$ confidence level, *i.e.* the probability of falsely rejecting the null hypothesis will be α. For large samples ($k > \sim 20$) the t distribution is well approximated by the standard normal distribution.

III-2. *Unweighted mean, environmental effects moderate or small*

Although $\bar{x}_u$ is not the best (*i.e.* most precise) estimate of the mean when environmental effects are small, it may nevertheless be used, *e.g.* if the observational e.s.d.'s are not available. Since the $\sigma(x_i)$ are now comparable with, or larger than, $\sigma(\mu)$, the various x_i differ in precision. This violates a fundamental assumption made in the derivation of the t distribution (Student, 1908). Consequently, d_u is no longer distributed as Student's t with $(k-1)$ degrees of freedom (Cochran, 1954). A computer simulation was performed to illustrate this point. Ten thousand simulated values of d_u were generated, using the simulation parameters $k = 5$, $\sigma_{min} = 0{\cdot}001$, $\sigma_{max} = 0{\cdot}030$, $\sigma(\mu) = 0$. Fig. 1(*a*) shows a probability plot of the resulting d_u distribution against a t distribution with $(k-1) = 4$ degrees of freedom. Small but systematic deviations of the observed points from the ideal straight line of unit slope are evident. For comparison, Fig. 1(*b*) shows the probability plot produced by a second simulation in which $\sigma(\mu)$ was set to 0·3 [*i.e.* $\gg \sigma(x_i)$]. Here, the plotted points lie close to the ideal straight line.

When $\sigma(\mu)$ is small, Cochran (1954) suggested that the distribution of d_u may be approximated by a t distribution, *provided that the number of degrees of freedom is reduced from the 'ideal' value of* $(k-1)$. He gave formulae for estimating the 'effective' number of degrees of freedom. However, the deviations of the plotted points from the ideal straight line in Fig. 1(*a*) are small enough to suggest that this adjustment may not be necessary in practice. In particular, the fit between the observed d_u distribution and the t distribution with $(k-1)$ degrees of freedom is good at the extremities of the plot, *i.e.* in those regions of the distribution that are important in hypothesis testing. In order to investigate the consequences of assuming that d_u follows a t distribution with $(k-1)$

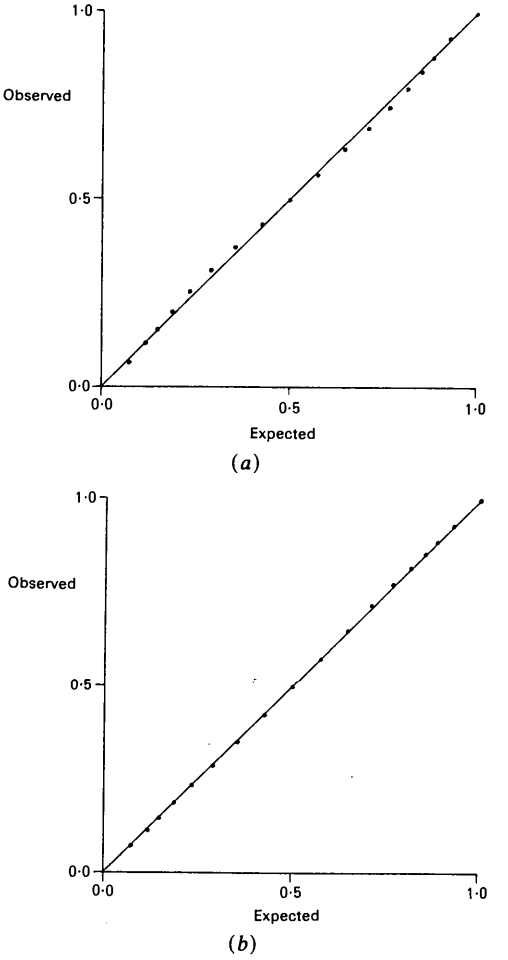

Fig. 1. (*a*) Plot of observed cumulative probability distribution of 10 000 simulated d_u values against cumulative distribution expected for a t distribution with $(k-1)$ degrees of freedom. Simulated d_u values were generated with parameters $k = 5$, $\sigma_{min} = 0{\cdot}001$, $\sigma_{max} = 0{\cdot}030$, $\sigma(\mu) = 0$. (*b*) As for (*a*), except with $\sigma(\mu) = 0{\cdot}3$.

* The results described in this paper are therefore based on the assumption that x_i is normally distributed. This is probably a good approximation for many molecular parameters. Some of the situations in which it is not a good approximation will be discussed in a subsequent paper.

Table 1. *Results of simulations investigating distribution of d_u as a function of $\sigma(\mu)$*

See text for explanation of symbols. Table gives percentage of simulated $|d_u|$ values that exceeded $t_{k-1,0.05}$.

		σ_{min}-σ_{max}		
$\sigma(\mu)$	Sample size	0·001-0·005	0·001-0·010	0·001-0·030
0·0	$k=5$	3·6	3·4	3·0
	$k=10$	4·9	4·3	4·5
	$k=20$	4·6	4·7	4·5
$\frac{\sigma_{min}+\sigma_{max}}{2}$	$k=5$	4·7	4·7	4·5
	$k=10$	4·4	5·0	4·8
	$k=20$	5·3	5·2	5·1
0·3	$k=5$	4·8	5·0	5·0
	$k=10$	5·0	4·6	4·9
	$k=20$	5·0	5·3	4·8

degrees of freedom, even when $\sigma(\mu)$ is small, we performed the following simulation. Six thousand values of d_u were generated, using the simulation parameters $k=5$, $\sigma_{min}=0{\cdot}001$, $\sigma_{max}=0{\cdot}005$, $\sigma(\mu)=0$. A count was made of the number of occasions on which $|d_u|$ exceeded $t_{k-1,0.05}$ (*i.e.* $t_{4,0.05}=2{\cdot}776$). At the end of the simulation, there were 216 such occasions, *i.e.* about 3·6% of the d_u values were formally significant at the 95% confidence level. Since this is less than the ideal proportion of 5%, the assumption that d_u follows a t distribution with $(k-1)$ degrees of freedom seems to be safe. This conclusion was supported by several other simulations with various values of k, σ_{min}, σ_{max} and $\sigma(\mu)$; results are summarized in Table 1.

III-3. *Weighted mean, environmental effects negligible*

If $\sigma(\mu)$ is negligibly small and the $\sigma(x_i)$ are accurate (*i.e.* exact estimates of the experimental standard deviations of the x_i), the distribution of d_w will be normal. Thus, H_0 can be rejected at the 95% confidence level if $|d_w|$ exceeds 1·96 (since 95% of observations in a normal distribution lie within ±1·96 standard deviations of the mean). However, if the $\sigma(x_i)$ are inaccurate there will be errors in the weights used to calculate $\bar{x}_w$ and $\sigma(\bar{x}_w)$, and d_w will not be normally distributed. This was investigated as follows. Six thousand simulated values of d_w were generated, using the simulation parameters $k=5$, $\sigma_{min}=0{\cdot}001$, $\sigma_{max}=0{\cdot}005$, $\sigma(\mu)=0$. The procedure used in the simulation was exactly as described in § II, except that the weights used in computing $\bar{x}_w$ and $\sigma(\bar{x}_w)$ were not calculated from (4), but from

$$w_i = 1/s^2(x_i), \tag{12}$$

where

$$s(x_i) = f_i\sigma(x_i). \tag{13}$$

Here, f_i is a number chosen at random from a normal distribution with unit mean and standard deviation = $\sigma(f)=0{\cdot}1$.* The purpose of the procedure was to introduce random errors into the weights used in calculating $\bar{x}_w$ and $\sigma(\bar{x}_w)$.

Table 2. *Results of simulations investigating distribution of d_w as a function of $\sigma(f)$*

See text for explanation of symbols. Table gives percentage of simulated $|d_w|$ values that exceeded $z_{0.05}$ (= 1·96). All simulations were performed with $\sigma(\mu)=0$.

		σ_{min}-σ_{max}		
$\sigma(f)$	Sample size	0·001-0·005	0·001-0·010	0·001-0·030
0·0	$k=5$	5·4	5·0	5·8
	$k=10$	5·3	5·1	5·5
	$k=20$	4·9	4·9	5·2
0·1	$k=5$	6·1	5·1	5·8
	$k=10$	5·3	5·9	6·1
	$k=20$	6·0	6·2	5·5
0·2	$k=5$	8·7	8·6	7·9
	$k=10$	9·5	9·3	8·9
	$k=20$	9·1	9·7	9·7
0·3	$k=5$	15·3	14·8	14·3
	$k=10$	19·7	17·3	16·0
	$k=20$	22·6	21·7	18·7

Of the six thousand values of $|d_w|$ thus generated, 368 exceeded 1·96, *i.e.* about 6·1% of the d_w values were formally significant at the 95% confidence level. Thus, the probability of falsely rejecting the null hypothesis was very slightly increased by the random errors in the w_i. Several other simulations were performed with different values of k, σ_{min}, σ_{max} and $\sigma(f)$; results are summarized in Table 2. They show that the distribution of d_w is relatively insensitive to small errors in the weights (*c.f.* Cochran & Carroll, 1953).

III-4. *Weighted mean, environmental effects not negligible*

Ideally, the weighted mean should only be used when $\sigma(\mu)$ is negligibly small. However, it may easily be used by mistake when $\sigma(\mu)$ is appreciably greater than zero. In this case, we may anticipate that d_w will not be normally distributed even if the $\sigma(x_i)$ are accurate. This was confirmed as follows. Six thousand values of d_w were generated, using the simulation parameters $k=5$, $\sigma_{min}=0{\cdot}001$, $\sigma_{max}=0{\cdot}005$, $\sigma(\mu)=(\sigma_{min}+\sigma_{max})/2=0{\cdot}003$ [weights were calculated from (4), not from (12)]. A count was made of the number of $|d_w|$ values that exceeded 1·96. The result is given in Table 3, together with those obtained from other simulations with different values of k, σ_{min}, σ_{max} and $\sigma(\mu)$. The table shows that the distribution of d_w is

* On very rare occasions, the number drawn from the normal distribution was less than 0·2. In order to avoid generating very small values of $s(x_i)$, we equated f_i to 0·2 on these occasions.

Table 3. *Results of simulations investigating distribution of d_w as a function of $\sigma(\mu)$*

See text for explanation of symbols. Table gives percentage of simulated $|d_w|$ values that exceeded $z_{0.05}$ (= 1·96).

		σ_{min}–σ_{max}		
$\sigma(\mu)$	Sample size	0·001–0·005	0·001–0·010	0·001–0·030
0·0	$k=5$	4·4	5·4	4·6
	$k=10$	5·2	5·3	5·1
	$k=20$	4·9	5·0	4·9
$\frac{\sigma_{min}+\sigma_{max}}{8}$	$k=5$	6·9	10·4	18·3
	$k=10$	7·1	11·4	24·9
	$k=20$	7·3	12·8	30·3
$\frac{\sigma_{min}+\sigma_{max}}{4}$	$k=5$	13·4	20·9	31·6
	$k=10$	14·1	24·3	42·4
	$k=20$	15·6	28·9	51·6
$\frac{\sigma_{min}+\sigma_{max}}{2}$	$k=5$	30·1	41·5	52·8
	$k=10$	33·0	50·0	61·9
	$k=20$	35·6	52·5	71·8

extremely sensitive to the presence of environmental effects. For example, with the simulation parameters given above, some 30% of the d_w values were formally significant at the 95% confidence level (*i.e.* the *true* confidence level of the test was about 70%). We conclude that the use of weighted means in hypothesis tests is extremely dangerous when $\sigma(\mu)$ is not negligibly small.

IV. Conclusions

In our previous study (Taylor & Kennard, 1983) we concluded that the unweighted mean is probably satisfactory for most samples of crystallographic data. The present work reinforces this conclusion since it shows that unweighted means may be used in hypothesis tests with little difficulty. Some approximations are necessary when environmental effects are small, but they are unlikely to cause problems in practice. In contrast, the use of d_w in hypothesis tests cannot be recommended. This is mainly because the value of $\sigma(\bar{x}_w)$ obtained from (5) may be a gross underestimate of the true standard error of the weighted mean if environmental effects are not negligible.

Olga Kennard is a member of the external staff of the Medical Research Council.

References

COCHRAN, W. G. (1954). *Biometrics*, **10**, 101–129.
COCHRAN, W. G. & CARROLL, S. P. (1953). *Biometrics*, **9**, 447–459.
CRUICKSHANK, D. W. J. & ROBERTSON, A. P. (1953). *Acta Cryst.* **6**, 698–705.
HAMILTON, W. C. & ABRAHAMS, S. C. (1970). *Acta Cryst.* A**26**, 18–24.
NAG Fortran Library Manual (1983). Mark 10, Vol. 5.
STUDENT (1908). *Biometrika*, **6**, 1–25.
TAYLOR, R. & KENNARD, O. (1983). *Acta Cryst.* B**39**, 517–525.
TAYLOR, R. & KENNARD, O. (1985). In preparation.

SHORT COMMUNICATIONS

Contributions intended for publication under this heading should be expressly so marked; they should not exceed about 1000 *words; they should be forwarded in the usual way to the appropriate Co-editor; they will be published as speedily as possible.*

Acta Cryst. (1985). A**41**, 89–90

Tensorial classification of non-magnetic ferroic crystals: tensors of rank $N \leq 4$.

By D. B. LITVIN, *Department of Physics, The Pennsylvania State University, The Berks Campus, PO Box* 2150, *Reading, PA* 19608, *USA*

(*Received* 2 *March* 1984; *accepted* 12 *July* 1984)

Abstract

The tensorial classification of all non-magnetic ferroic crystals is given for all possible macroscopic tensorial properties of rank $N \leq 4$.

A ferroic crystal can be classified according to the point-group symmetry of the non-ferroic or prototypic phase and the point-group symmetries of the ferroic crystal's domains. Each class of ferroic crystals is denoted by a symbol **GFH**, where F is a symbol denoting 'ferroic', **G** the point group of the non-ferroic phase, and **H** the point group of one of the ferroic crystal's domains (Aizu, 1970, 1976*a*,*b*). The domains of a ferroic crystal can possibly be distinguished by the values of the components of a macroscopic tensorial property associated with each of the domains. Whether or not one can distinguish some or all of the domains has led to the additional tensorial classification of ferroic crystals (Aizu, 1969, 1970). Aizu has tabulated the magnetic classes of ferroic crystals and the tensorial classification of ferroelectric, ferromagnetic and ferroelastic crystals (Aizu, 1970; see also Cracknell, 1972). A method has been presented (Litvin, 1984) to determine the tensorial classification of non-magnetic ferroic crystals for an arbitrary tensorial property.

0108-7673/85/010089-02$01.50