Journal of Chromatography, 173 (1979) 7–17
© Elsevier Scientific Publishing Company, Amsterdam — Printed in The Netherlands

CHROM. 11,702

REPRODUCIBILITY OF PYROLYSIS–GAS CHROMATOGRAPHIC ANALYSES OF THE MOULD *PENICILLIUM BREVI-COMPACTUM*

GÖRAN BLOMQUIST

National Board of Occupational Safety and Health, Department of Occupational Health, S-901 85 Umeå (Sweden)

ERIK JOHANSSON

Research Group of Chemometrics, Institute of Chemistry, Umeå University, S-901 87 Umeå (Sweden)

BENGT SÖDERSTRÖM

Department of Microbiological Ecology, University of Lund, S-223 62 Lund (Sweden)

and

SVANTE WOLD*

Research Group of Chemometrics, Institute of Chemistry, Umeå University, S-901 87 Umeå (Sweden)

(Received November 20th, 1978)

SUMMARY

The reproducibility of multivariate measurements is discussed. Reproducibility of a second kind is defined, in which part of the variability between samples is described by a principal components model. The use of this generalized reproducibility is shown to give an improved precision in the pyrolysis–gas chromatography of a *Penicillium* species.

INTRODUCTION

The characterization and classification of micro-organisms by chemical means has been under investigation for some time. Reiner[1] showed that different bacterial strains show different "chemical fingerprints" when subjected to pyrolysis followed by gas chromatographic separation of the resulting volatile fragments. This work has been continued by several workers[2,3], including Reiner[4]. More recently, pyrolysis–gas chromatography (Py–GC) has been heavily critisized by Meuzelaar *et al.*[5], mainly because of the difficulty of obtaining reproducibility between different runs on the same sample. Instead, Meuzelaar *et al.*[6] advocate pyrolysis–mass spectrometry as a preferred method for solving the problem.

There is no doubt that Py–GC shows an apparent lack of reproducibility. Fig. 1 shows two consecutive Py–GC runs with the same sample of the mould *Penicillium brevi-compactum*. The situation is not hopeless, however, if the variation between

* To whom correspondence should be addressed.

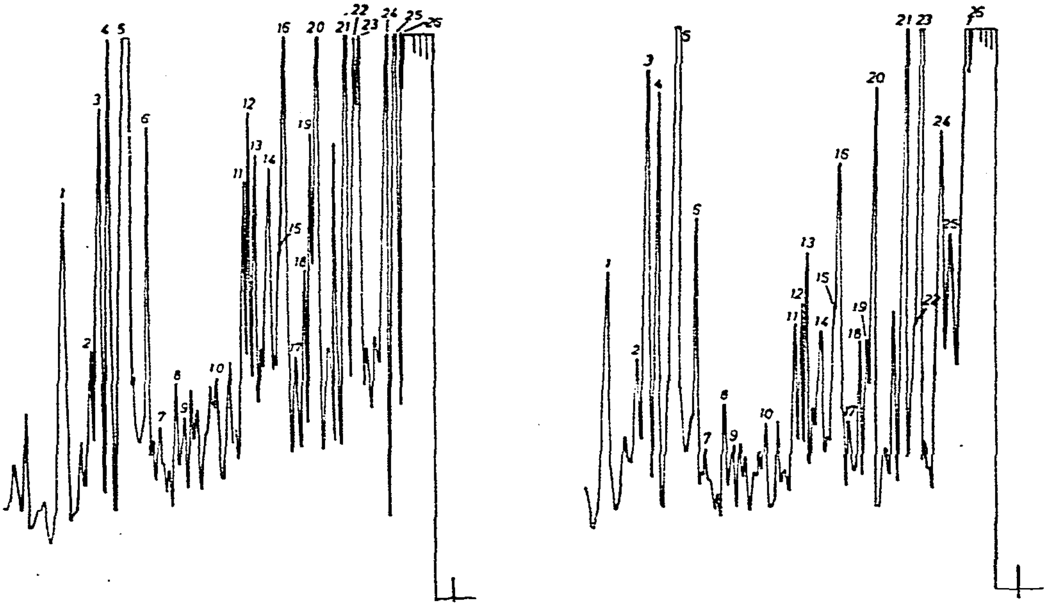

Fig. 1. Two Py–GC runs on samples of *Penicillium brevi-compactum*.

the chromatograms in different runs is to some extent systematic. The variation can then be described by a mathematical model which allows the construction of a second kind of reproducibility.

In this paper we report an investigation of the variability of Py–GC runs on a single fungal strain, *Penicillium brevi-compactum*. We partition this variability into one systematic part and one random part. The use of the systematic part for the identification of new samples is discussed.

REPRODUCIBILITY OF THE FIRST AND SECOND KINDS

The traditional concept of reproducibility (R) involves measurements (y_k) made on a single variable y. The variability of y_k around the mean ($\bar{y}$) or some other measure of the central tendency is defined as the reproducibility of y (see Fig. 2)[7,8]:

$$y_k = \bar{y} + \varepsilon_k \qquad (1)$$

$$R \leftrightarrow \sigma(\varepsilon) \qquad (2)$$

The measurements y_k are described by the mean ($\bar{y}$) and random "noise" (ε). The standard deviation of the noise is related to the reproducibility (R).

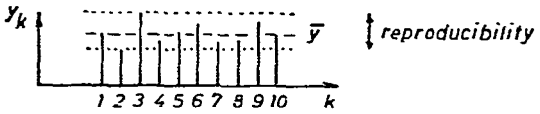

Fig. 2. The reproducibility of a single variable y is related to the variability of measurements y_k around the mean $\bar{y}$.

When multivariate measurements are made, two kinds of reproducibility can be defined. The first relates to the reproducibility of each single variable independently of all the others. This is the same as the traditional concept, and has recently been discussed by Eshuis *et al.*[9]. The second kind relates to the predictability of one multivariate measurement vector from the other multivariate vectors. Thus, we can generalize eqn. (1) to

$$y_k = \Phi + \varepsilon_k \tag{3}$$

Here Φ symbolizes any mathematical model. In eqn. 1 Φ is a simple constant, but we realize that as long as Φ is known to the extent that the deviations ε can be calculated for a single sample measurement y_k, a measure of the reproducibility can be derived from the size of ε.

In the multivariate situation y_k is no more a single number but rather a vector or array of numbers, henceforth denoted by Y_k with the elements y_{ik}, where i is the index of the variables constituting the multivariate measurement; $i = 1,2,\ldots, M$.

In Py–GC, a chromatogram which as those in Fig. 1 can be translated to a vector of numbers Y_k by, for instance, using the areas or the heights of the peaks as variables. Thus, the chromatograms in Fig. 1 are translated to vectors with 26 elements using the peak heights of the numbered peaks. This gives data as shown in Table I.

It can be seen that if the traditional reproducibility related to eqn. 4 is used we have a variation around the mean for each peak of approximately 18% S.D.

We can now ask if it is possible to find a model Φ which is "better" than the ordinary mean, that is, a model which makes the residuals ε have a smaller variability than 18%. We write the multivariate model explicity as

$$y_{ik} = \Phi + \varepsilon_{ik} \tag{4}$$

Model (1) is then

$$y_{ik} = \bar{y}_i + \varepsilon_{ik} \tag{5}$$

It has been shown[10] that it is indeed possible to find a form of Φ which generally gives smaller residuals ε than model (1) and (5). This model, the principal components model[11] or the factor model[12], henceforth called the PCF model, applies to the reproducibility situation discussed here provided that the individual variables i show some kind of correlation with each other.

This corresponds to the model

$$y_{ik} = \bar{y}_i + \sum_{a=1}^{A} \beta_{ia}\,\theta_{ak} + \varepsilon_{ik} \tag{6}$$

Here $\bar{y}_i$ still is the average of variable i and ε the deviations between model and observations (y_{ik}). The products $\beta_{ia}\theta_{ak}$ express the correlation structure between the variables over the group of samples ($k = 1,2,\ldots,N$).

Model (6) has properties which are very desirable in the present context. Firstly, it is generally applicable in the reproducibility situation as the samples, by

TABLE I

PEAK HEIGHTS OF 26 PEAKS (SEE FIG. 1) IN 10 RUNS OF Py–GC ANALYSIS OF *PENICILLIUM BREVI-COMPACTUM*

Each run-vector (row) is normalized to a sum of 1000. Mean values and standard deviations for each variable are shown in the bottom rows.

Run	*Peak*																									
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*	*21*	*22*	*23*	*24*	*25*	*26*
1	47	26	35	46	97	29	30	27	25	33	42	30	26	27	30	26	26	31	37	27	44	30	50	34	57	87
2	30	29	28	38	98	30	18	19	19	23	36	32	28	31	34	26	27	30	37	32	53	32	65	41	64	100
3	29	20	33	38	94	26	15	17	14	18	28	29	25	23	30	16	17	21	25	36	68	25	132	34	37	151
4	23	21	27	32	69	19	22	21	21	29	36	32	29	29	29	27	28	30	38	34	60	34	78	47	72	114
5	36	23	34	42	99	26	19	20	17	23	33	28	25	25	31	19	20	23	30	33	66	26	99	34	47	123
6	40	21	35	38	89	26	17	20	16	20	29	27	24	22	25	17	18	22	30	44	72	23	73	42	44	177
7	32	26	31	40	85	26	31	25	22	33	39	28	25	25	29	24	26	29	33	27	46	26	46	38	69	136
8	25	15	30	37	103	29	12	15	13	15	26	30	28	19	22	16	16	21	29	40	72	35	94	35	52	174
9	29	30	31	34	84	25	17	16	15	18	27	30	28	17	22	17	18	22	28	35	72	30	118	38	53	156
10	37	25	35	39	77	28	26	22	21	27	34	27	28	26	26	24	24	27	36	29	55	29	64	72	55	109
Mean	33	24	32	38	89	26	21	20	18	24	33	29	27	24	28	21	22	26	32	34	61	29	82	42	55	133
S.D.	7.3	4.5	3.0	3.9	11	3.1	6.4	3.8	3.9	6.4	5.4	1.8	1.8	4.3	3.9	4.6	4.6	4.2	4.5	5.5	11	4.0	28	12	11	31

definition, are closely similar[10,13]. Secondly, the number of terms A can be estimated from the data[14] .Thus we can test if model (5) [*i.e.*, model (6) with $A = 0$] or model (6) is better. If we find the latter, it is routine computation to calculate the parameters β and θ which makes the model best approximate the data in the least-squares sense[10,15]. Thirdly, model (6) has an appealing geometric interpretation which will be discussed below in terms of an artificial example.

The S.D. of the residuals ε_{ik}, eqn. 7, gives a measure of the precision, *i.e.*, the variability of the data around the model Φ:

$$S_0 = \left[\sum_{i=1}^{M} \sum_{a=1}^{N} \varepsilon_{ik}^2/(M - A)(N - A - 1)\right]^{1/2} \tag{7}$$

Here M and N are the number of variables and samples, respectively.

IDENTIFICATION OF A NEW SAMPLE

With the traditional reproducibility model (eqn. 5), the identification of a new sample simply involves the comparison of the new sample vector Y^* (having the elements y_i^*) with the mean values $\bar{y}_i$ estimated from earlier samples.

The residuals ε_i^* are calculated as

$$\varepsilon_i^* = y_i^* - \bar{y}_i \tag{8}$$

In other words, each variable is compared with the typical value as in Fig. 2. If all elements fall inside the earlier estimated intervals of variation for the corresponding variables, the new sample is concluded to be identified as being of the present type.

The identification of a new sample according to the reproducibility of the second kind (eqn. 6) is slightly more involved[10]. Firstly, the mean values are subtracted as before giving the initial residuals, now denoted by Z_i^*:

$$Z_i^* = y_i^* - \bar{y}_i \tag{9}$$

However, when it has been found that Z_i have systematic structure, *i.e.*, when it has been found that eqn. 6 describes the earlier data with $A \geqslant 1$, Z_i^* is to be divided into the systematic part and the random part. This involves a regression:

$$Z_i^* = \sum_{a=1}^{A} t_a \beta_{ia} + \varepsilon_i^* \tag{10}$$

Here β_{ia} are the parameters estimated earlier in eqn. 6. The coefficients t_a correspond to the parameters θ in eqn. 6. The t_a values for the new sample are calculated so as to minimize the final residual ε_i^*. Thus the systematic part of Z_i^* is estimated as $\Sigma t_a \beta_{ia}$ and the random part as ε_i^*.

If now ε_{ia} are within the earlier estimated range (as measured by their S.D., eqn. 11), we conclude that the new sample "fits" the same group as the earlier samples; the new sample is identified:

$$S^* = \left[\sum_{i=1}^{M} \varepsilon_i^{*2}/(M - A)\right]^{1/2} \tag{11}$$

Graphically, this identification is simple to understand and is discussed further below with an artificial example.

ARTIFICIAL EXAMPLE

Consider five runs as shown in Fig. 3. The variability according to model (5) is large and no apparent reproducibility is seen. When the mean "chromatogram" (M in Fig. 3) is subtracted from the five runs we obtain instead the picture shown in Fig. 4. Suddenly the five runs look very similar; we immediately realize that the remainders are described by a coefficient θ times "chromatogram" number 1 in Fig. 4.

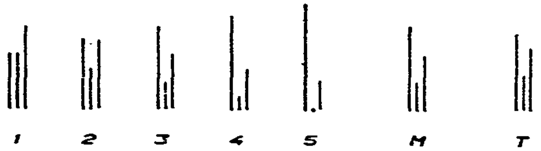

Fig. 3. Five artificial gaschromatograms with three peaks. M is their average. T is the chromatogram of a "test" sample which is to be compared with chromatograms 1–5.

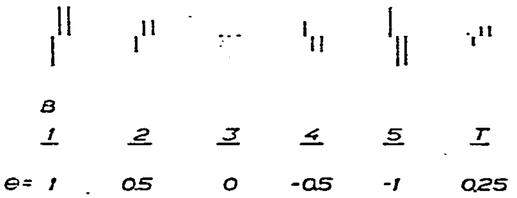

Fig. 4. Same artificial gas chromatograms as in Fig. 3 but with the average chromatogram M subtracted.

We have shown graphically that each chromatogram in Fig. 3 is described as (B is chromatogram 1 in Fig. 4):

$$Y_k = M + B\theta_k \tag{12}$$

A "test" sample which introduced after the chromatograms M and B have been "estimated" from the data (T in Fig. 4) is immediately seen to "fit" the group by first subtracting M.

We see that in this artificial example the traditional reproducibility according to model (5) is very bad. The reproducibility of the second kind according to model (6) is "perfect", a result which is baffling when looking at Fig. 3.

In practice the situation is, of course, less perfect. The residuals never become zero when the optimal number of terms (A) are used in the PCF model (6). However, the residuals ε usually become smaller than those using model (5), that is, some variation between the chromatograms is usually systematic and the information contained in the data is better utilized by using model (6); the precision becomes better by using a more general model.

Whether this decrease in ε is statistically significant, however, must be checked by a statistical procedure. We use a very stringent procedure, cross-validation, details of which will appear elsewhere[14].

Graphical representation

There are other ways of representing the artificial data in Fig. 3. One illustrative way is to construct a space with one orthogonal axis for each variable. We call this space M-space (measurement space). In the artificial example we thus get the three-dimensional space shown in Fig. 5. In this space each sample vector is represented as a point. We see that "samples" all lie along a straight line, corresponding to eqn. 6 with $A = 1$.

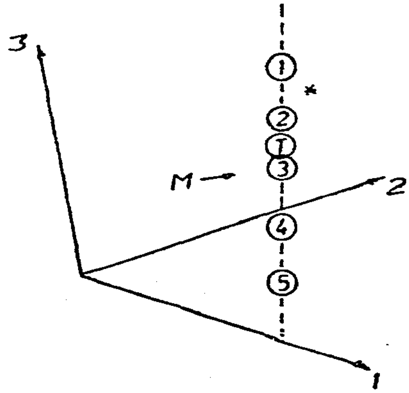

Fig. 5. Three dimensional M-space constructed by using the three "peaks" of samples in Fig. 3 as coordinates x, y and z.

The calculation of the coefficient β in eqn. 6 corresponds to the determination of the direction of the line in the M-space. The coefficient θ_k describe the position of samples k along this line.

The identification of a new sample simply corresponds to seeing if the "sample point" falls close to the line in M-space (the asterisk in Fig. 5). The distance between the sample point and the line is directly measured as the S.D. of the residuals ε_{ik}, eqn. 11.

REAL EXAMPLE: PY–GC OF *PENICILLIUM BREVI-COMPACTUM*

Samples and chemical analysis

Penicillium brevi-compactum Dierckx (CBS 210.28) was grown in Oxoid malt extract broth for 5 days on a rotary shaker (100 rpm) at 22°. Very few conidia were formed during the incubation. The mycelium was harvested by filtration, freeze dried, ground in a mortar and stored in glass tubes in an exsiccator at room temperature.

Pyrolysis investigations were carried out on *ca.* 0.5-mg samples of the fungi. A Pye GCD with a flame-ionization detector and a Pye Pyrolyzer No. 12556 and 12557 was used. As the samples were powdery, the coil method[16] in an 80 mm × 2 mm

TABLE II

RESIDUAL (ε) S.D.s FOR MODEL (6) WITH $A = 2$ AND (1) ALL VARIABLES AND (2) ALL VARIABLES EXCEPT 2, 5, 6, 16 AND 24

The last row shows the values of β_{i1} and β_{i2} in eqn. 6 for case (2). The original data y_{ik} were normalised to $s_i(y) = 1.0$.

Parameter	*Total*	*Variable (i)*											
		1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
$s_i^{(1)}$	0.63	0.50	1.0	0.49	0.51	1.0	1.0	0.47	0.34	0.24	0.38	0.22	0.72
$s_i^{(2)}$	0.48	0.53		0.43	0.64			0.43	0.26	0.21	0.34	0.21	0.62
β_{i1}		−0.10		0.02	−0.11			−0.25	−0.25	−0.28	−0.27	−0.28	−0.06
β_{i2}		0.36		0.41	0.33			0.15	0.19	0.06	0.08	0.06	−0.36

I.D. quartz tube was used. A new ferromagnetic wire and a cleaned quartz tube were used for each sample. A pyrolysis time of 10 sec at 510° was employed throughout. Chromatography was carried out using a 3.0 m × 4 mm I.D. glass column packed with 10% Carbowax 20 M on 100–120-mesh Chromosorb W AW DMCS. The temperature was programmed from 70 to 150° at a rate of 4°/min, with an initial hold for 6 min, a final hold for 30 min and then 10 min at 170°. The carrier gas (nitrogen) flow-rate was 40 ml/min at 25°, with hydrogen and air flow-rates of 50 and 500 ml/min, respectively. The injector temperature was 170° and the detector temperature 200°. Ten samples were analysed over a time period of 30 days.

The 10 resulting gas chromatograms were digitized by using the heights of 26 peaks identifiable in all chromatograms as the values of 26 variables. Each sample data vector was normalized to a sum of 1000 over the 26 peak values. The resulting data are shown in Table I.

Data analysis

First the data were scaled, subtracting from each variable its mean and then dividing each variable by its S.D. (bottom of Table I). This gave each variable a zero mean and the same initial weight. Second, the scaled data matrix was subjected to a principal components analysis (PCA). Cross-validation[14] showed that two product terms $\beta\theta$ were needed to describe the correlation structure ($A = 2$ in eqn. 6).

The standard deviations of the residuals ε_{ik} for each variable i showed that variables 2, 5, 6, 15 and 24 did not participate in the PCF model (see Table II). Hence, these variables were deleted and a new PC analysis was made on the reduced data matrix. The resulting residual S.D.s for each variable are shown in Table II together with the values of the parameters β_{i1} and β_{i2}. These are also plotted against each other in Fig. 6.

Table III shows the values of θ_{1k} and θ_{2k} for the 10 samples ($k = 1,2,\ldots,10$) and the residual S.D. of ε_{ik} for each sample. The total residual S.D. is 0.45 (S_0, eqn. 7).

The plot shown in Fig. 6 gives an indication of the grouping of the variables. Thus, variables having similar β_1 and β_2 values fall close to each other in this plot, and these variables show a similar behaviour over the investigated data set.

We can see a clear clustering of the variables, indicating connections between

13	*14*	*15*	*16*	*17*	*18*	*19*	*20*	*21*	*22*	*23*	*24*	*25*	*26*
0.60	0.69	0.95	0.13	0.17	0.16	0.43	0.66	0.34	0.72	0.67	1.0	0.46	0.57
0.59	0.71		0.17	0.23	0.15	0.43	0.68	0.35	0.63	0.65		0.48	0.59
—0.05	−0.22		−0.27	−0.27	−0.28	−0.25	0.23	0.23	−0.05	0.23		−0.21	0.25
−0.37	−0.07		−0.11	−0.11	−0.08	−0.12	−0.04	−0.06	−0.36	−0.06		−0.24	0.03

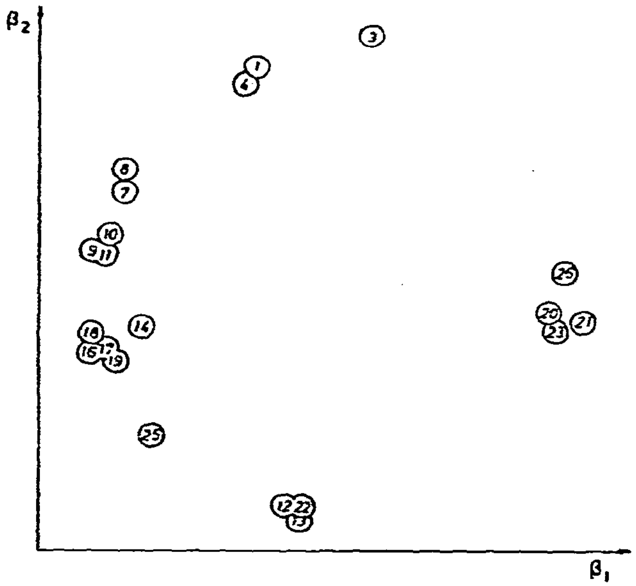

Fig. 6. Plot of β_{i2} against β_{i1} for the variables remaining in analysis (ii) when variables 2, 5, 6, 15 and 24 have been deleted.

the following groups: (1) 1,3,4; (2) 7–11; (3) 14, 16–19; (4) 12, 13, 22; and (5) 20, 21, 23, 26.

A simple representation such as that shown in Fig. 5 for the artificial data set is, of course, more difficult in the present case involving 26 dimensions. We can, however, display different projections of the 26-dimensional *M*-space down on various planes, but this is more interesting when several types of moulds are analysed. Therefore, we show such projections in the following paper.

In summary, the data analysis shows that indeed about 55% of the variation between the 10 chromatograms is systematic. Thus, the reproducibility of the second kind gives about twice as good "precision" as the traditional reproducibility.

TABLE III

VALUES OF θ_{1k} AND θ_{2k} (EQN. 6) FOR RUNS 1–10 FOR THE CASE WHEN VARIABLES 2, 5, 6, 15 AND 24 HAVE BEEN EXCLUDED

The last columns show the data (y) S.D. and residual (ε_{ik}) S.D. for each run.

k	θ_1	θ_2	$S.D. (y)$	$S.D. (\varepsilon)$
1	−5.16	2.19	1.2	0.47
2	−2.44	−2.59	0.87	0.46
3	3.97	0.99	0.88	0.48
4	−2.65	−3.96	1.0	0.34
5	1.03	1.85	0.54	0.33
6	3.02	2.61	1.0	0.58
7	−3.46	1.22	0.98	0.64
8	4.23	−2.00	1.1	0.45
9	3.50	−1.29	0.86	0.43
10	−2.05	0.98	0.66	0.52

DISCUSSION

The fact that multivariate measurements often display correlations between the variables shows that the definition of new kind of reproducibility is needed. The use of the traditional reproducibility relating to the variation of each variable around its mean value results in the unnecessary loss of information and precision. In addition to the primary classification aspects, a PCF analysis as discussed in this paper gives interesting information about the relevance of the variables and the grouping of variables.

In the following paper we discuss the issue of real interest in connection with the Py–GC analysis of fungi, namely the identification of several strains and species. The generalization from the one-group analysis in this paper to several-group analyses is straightforward. Each fungus is described by a separate PCF model (eqn. 6).

New moulds are then classified according to which of the PCF models they fit best. The gain in precision obtained by using a reproducibility based on PCF models instead of the ordinary mean value will be shown to sometimes be of critical importance.

ACKNOWLEDGEMENTS

This project was supported by grants from the Swedish Work Environment Fund and the Swedish Natural Science Research Council.

REFERENCES

1 E. Reiner, *Nature (London)*, 206 (1965) 1272.

2 D. T. Burns, R. J. Stretton and S. D. A. K. Jayatilake, *J. Chromatogr.*, 116 (1976) 107.

3 P. G. Vincent and M. M. Kulik, *Appl. Microbiol.*, 20 (1970) 957.

4 E. Reiner, in C. E. R. Jones and C. A. Cramers (Editors), *Analytical Pyrolysis*, Elsevier, Amsterdam, 1977, pp. 49–56.

5 H. L. C. Meuzelaar, P. G. Kistemaker, W. Eshuis and H. W. B. Engel, in H. H. Johnston and S. W. B. Newsom (Editors), *Rapid Methods and Automation in Microbiology*, Learned Information (Europe) Ltd., Oxford, 1976, pp. 225–230.

6 H. L. C. Meuzelaar, M. A. Posthumus, P. G. Kistemaker and J. Kistemaker, *Anal. Chem.*, 45 (1973) 1546.

7 H. H. Ku, Ch. 2 in American Society of Tool and Manufacturing Engineers, *Handbook of Industrial Metrology*, Prentice Hall, New York, 1967, Ch. 2; Reprinted in *Precision, Measurement and Calibration*, National Bureau of Standards special publication, No. 300, U.S. Department of Commerce, Washington, D.C., 1969, p. 269.

8 W. J. Youden, in I. M. Kolthoff, P. J. Elving and E. B. Sandell (Editors), *Treatise on Analytical Chemistry*, Interscience Encyclopedia, Wiley, New York, 1959, pp. 47–66.

9 W. Eshuis, P. G. Kistemaker and H. L. C. Meuzelaar, in C. E. R. Jones and C. A. Cramers (Editors), *Analytical Pyrolysis*, Elsevier, Amsterdam, 1977, pp. 151–166.

10 S. Wold, *Pattern Recognition*, 8 (1976) 127.

11 R. Gnanadesikan, *Methods for Statistical Data Analysis of Multivariate Observations*, Wiley, New York, 1977.

12 H. H. Harman, *Modern Factor Analysis*, University of Chicago Press, Chicago, Ill., 2nd ed., 1967.

13 A. J. Martens and J. Glas, *Chromatographia*, 5 (1972) 508.

14 S. Wold, *Technometrics*, 20 (1978) 397.

15 H. Wold, in F. N. David (Editor), *Festschrift for J. Neyman*, Wiley, New York, 1966, p. 411.

16 W. C. Thompson, *Lab. Pract.*, 18 (1969) 1074.