Use of the Hammett Equation in Substituted Thiophenes

Sergio Alunni and Sergio Clementi *
Dipartimento di Chimica, Università di Perugia, Via Elce di Sotto 10, 06100 Perugia, Italy
Cynthia Ebert and Paolo Linda
Facoltà di Farmacia, Università di Trieste, Trieste, Italy
Giuseppe Musumarra
Istituto dipartimentale di Chimica e Chimica Industriale, Università di Catania, Catania, Italy
Michael Sjöström and Svante Wold
Institute of Chemistry, University of Umeå, Umeå, Sweden

The application of multivariate statistics to linear free-energy relationships in the thiophene series increases the understanding of the Hammett equation. Principal components analysis shows that α-substituted thiophenes require the σ constants used for *para*-substituted benzene derivatives, owing to a strict proportionality of the substituent effects in the heterocyclic and homocyclic systems. Partial least-squares analysis indicates that two independent effects are linearly transferred from benzenes to thiophenes.

Although a complete understanding of substituent effects is still far from being achieved, the use of linear free-energy relationships for describing the systematic variation of properties within a series of compounds is of great value in physical organic chemistry. The great success of the Hammett equation is due to the possibility of deriving a mathematical model thus enabling both the prediction of properties of substrates not yet investigated and speculation on the reaction mechanism on the basis of the statistical results.

However, the Hammett equation is a simple method of bivariate statistical analysis (simple linear regression) and applies to homogeneous series only. Moreover, it was derived for benzene derivatives and its use for modelling properties of compounds different from benzenes is not obvious. As part of our interest in the chemistry of heteroaromatic molecules we have pointed out the duality of the Hammett approach.[1]

In applying the Hammett equation to a series of monosubstituted benzenes the substituent parameters are kept constant, and the reaction parameters result from the statistical analysis. But, if one wishes to compare the substituent effects in the same reaction for two series of *para*-monosubstituted benzenes and α-monosubstituted thiophenes, the bivariate statistical method is not sufficient to solve the problem and an additional assumption must be made.

Either one assumes that the substituent effect scale remains constant and therefore the same σ values can be used in both series, or one assumes that the reaction constant is indeed constant and fits the derivatives in both series to the same equation.

In the former case the variation due to the change of the aromatic ring between the varying substituent and the reaction centre is reflected by the variation of the ρ values between the two series (sensitivity to substituent effects). However, the implicit assumption that the interaction between the substituent and the aromatic moiety remains constant is not supported on theoretical grounds.

In the latter case, the assumption that the reaction constant should be the same under the same experimental conditions seems sounder, but this approach requires the definition of an individual σ constant for each moiety linked to the reaction centre (*e.g.* 5-methyl-2-thienyl), and no resulting statistical parameter can simply be related to the variation induced by the change of the aromatic ring. In previous work our group[2] and many other authors[3] preferred to keep the σ values constant, whereas other authors, who investigated side-chain reactivities in detail, used the second approach.[4] As pointed out by Wold and his co-workers in analysing the success of the Hammett equation in terms of the shell model, the two assumptions differ in what is considered to be the 'inner' shell.[5]

The application of recently developed computer packages of multivariate statistics such as SIMCA–MACUP[6,7] can solve this problem. The methods allow the comparison of matrices of data with each other. Thus we report here the results of the comparison between thiophenes and benzenes using a data set which includes the pK_a values of carboxylic acids as well as electrophilic and side-chain reactivities taken from the relevant literature.[2,4,8–20]

The data set was chosen with the statistical requirements of the methods used in mind. In particular the number of substituents and reactions is as large as possible, provided that at least four measurements are available for each variable (reaction) and object (substituted aromatic compounds). A total of 15 reactions was used (see Table 1). Five of the series were measured under exactly the same experimental conditions.

The data set so constructed was analysed in two ways, applying two different statistical methods.

(A) A matrix was built in which each reaction was taken as a variable (15 reactions) and each substrate was considered as an independent object (12 compounds). This matrix was analysed by the SIMCA method, based on the philosophy of fitting disjointed principal components models to each separate class of objects.

(B) A matrix was built considering six substituents (objects) only, common to both series, for a number of variables (nine reactions for benzenes: X-block; eight reactions for thiophenes: Y-block). Where a reaction centre is already present in the molecules, substitution refers to the *para*-position in benzenes and to the second α-position in thiophenes. The matrix was analysed by the PLS2 method, a partial least-squares analysis where the latent variables t and u are extracted as components from each block X and Y, respectively, while simultaneously a linear relationship between the latent variables t and u of the same dimensionality is optimised.

Principal Components Analysis (*PCA*).—The SIMCA method is described in detail in refs. 6 and 7. Therefore only a brief summary is given here. The data matrix X contains elements $x(ki)$ (reactivity or equilibrium data) where index i is used for the reactions (variables) and k is used for the substrates (objects).

The statistical parameters $\bar{x}(i)$, $b(ia)$, and $t(ak)$ in equation (1)

Table 1. Relative reactivities data set

		Benzenes								Thiophenes							
PCA Number	Reaction	PLS Number	Ref.	OMe	Me	H	Cl	Br	NO_2	PLS Number	Ref.	OMe	Me	H	Cl	Br	NO_2
1	Hydrogen exchange	1	8	6.83	3.05	0		−0.58			8			8.61			
2	Detritiation									10	16		2.32	0	−0.77	−0.89	
3	Acetylation	2	9	5.47	2.11	0	−1.68	−1.85									
4	Trifluoroacetylation									11	*2b*	6.25	2.58	0	−0.24	−0.34	
5	Bromination									12	*2a*		2.59	0	−0.28	−0.42	
6	Bromination	3	10	5.20	2.73	0	−0.84	−1.21									
7	Chlorination									13	*2a*			0	−0.38	−0.48	−5.96
8	Chlorination	4	10	7.66	2.92	0	−0.39	−0.51									
9	σ^+	5	11	−0.78	−0.31	0	0.11	0.15	0.79		*2a*			−0.79			
10	Solvolysis PNB									14	*4b*	5.21	1.91	0		−0.86	
11	$ArSO_2Cl + PhNH_2$	6	12	−0.40	−0.16	0			0.78	15	17	−1.16	−1.14	−1.09	−0.80	−0.77	0.38
12	Solvolysis AEA		*4a*	4.39	1.52	0					*4a*		6.60	4.73		3.99	
13	$ArSO_2NHPh + OH$	7	13	0.27	0.19	0	−0.43		−1.31	16	18		−0.33	−0.57	−1.14	−1.14	−2.30
14	ArCHO + OH	8	14	1.06	0.49	0	−0.46		−2.11		19			0.31		−0.26	
15	Hammett σ	9	15	−0.28	−0.16	0	0.22	0.22	0.78	17	20	0.41	0.45	0.68	0.89	0.92	1.39

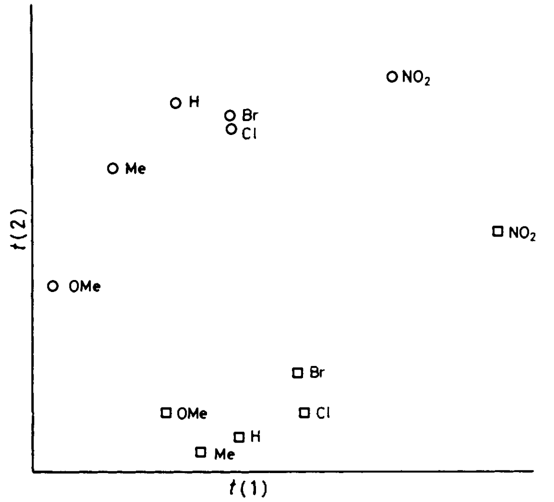

Figure 1. Principal components plot for the A matrix. The circles represent benzenes and the squares represent thiophenes

$$x(ki) = \bar{x}(i) + \sum_{a} b(ia)\, t(ak) + e(ki) \qquad (1)$$

are estimated by minimizing the squared residuals $e(ki)$, together with the number of significant cross-terms A, by means of the NIPALS algorithm. The data are first autoscaled in order to give all variables the same initial importance. The analysis then proceeds by model expansions. Initially a model with $A = 0$ is fitted to the data, which means that each variable is given as its mean value. Then these averages are subtracted from the matrix elements $x(ki)$ thus giving residuals of dimension zero. If these residuals contain systematic information (according to cross-validation as in ref. 21) the $b(i1)t(1k)$ term is estimated. Then new residuals are calculated by subtracting this term from each element. If the new residuals (dimension one) still contain systematic information, additional $b(i)t(k)$ terms are then estimated one after the other until the residuals just contain noise.

The analysis permits an illustration of the data structure to be obtained. Plots of one component against another can be considered as windows on the data set showing inhomogeneities, if present. The plot of the first against the second component, containing most of the information, usually gives the best idea on the presence of subgroupings.

The SIMCA analysis of the A matrix shows that two components only are significant according to the cross-validation procedure on autoscaled data, reweighted in order to give the same initial importance to each block of reactions proceeding by the same mechanism. Consequently, the mathematical model is represented by a plane in the 15-dimensional space. The projection of each object on the model is illustrated in Figure 1, and the resulting statistical parameters are listed in Tables 2 (objects parameters: scores) and 3 (variables parameters: weights, averages, loadings, and modelling powers). The data do contain structure as almost 85% of the total variance is explained by the two-component model.

Figure 1 clearly indicates that the objects (substrates) cluster into two groups: benzenes and thiophenes. Therefore the multivariate approach positively states that the change in the aromatic system is far more dramatic than the change in substituents, otherwise we should have found in the plot pairs of equally substituted compounds. Interestingly, the first dimension $t(1)$, which contains most of the information (48%), separates the substituents according to their known electronic effect from electron-donors to electron-withdrawers. However, the span covered by substituted thiophenes is larger than that of substituted benzenes.

Table 2. Objects parameters (scores: t) and distances from the model [$s(k)$] for the PCA of the A matrix. The reference standard deviation $s(0)$ is 0.24

Substituent (k)	[Ring]	$t(1)$	$t(2)$	$s(k)$
Methoxy	Benzene	−3.11	−0.07	0.27
Methyl	Benzene	−1.99	0.84	0.10
Hydrogen	Benzene	−0.96	1.35	0.13
Chlorine	Benzene	−0.06	1.14	0.08
Bromine	Benzene	0.01	1.24	0.08
Nitro	Benzene	2.77	1.52	0.14
Methoxy	Thiophene	−1.22	−1.06	0.25
Methyl	Thiophene	−0.59	−1.35	0.23
Hydrogen	Thiophene	0.07	−1.24	0.25
Chlorine	Thiophene	1.14	−1.03	0.05
Bromine	Thiophene	1.09	−0.74	0.16
Nitro	Thiophene	4.53	0.35	0.08

Table 3. Variables parameters (weights, averages, loadings, and modeling powers) for the PCA of the A matrix.

Reaction	$w(i)$	ave	$b(1)$	$b(2)$	MPOW
1	0.087	0.311	−0.032	−0.277	0.54
2	0.237	0.039	−0.292	−0.121	0.31
3	0.116	0.094	−0.129	−0.206	0.65
4	0.125	0.206	−0.305	−0.013	0.38
5	0.232	0.122	−0.271	−0.111	0.08
6	0.130	0.152	−0.128	−0.207	0.65
7	0.124	−0.212	−0.088	−0.236	0.67
8	0.101	0.196	−0.124	−0.213	0.63
9	1.290	−0.153	0.227	0.359	0.66
10	0.215	0.336	−0.525	0.099	0.41
11	0.855	−0.373	0.117	0.439	0.69
12	0.243	0.858	0.053	−0.446	0.42
13	0.722	−0.488	−0.248	0.073	0.76
14	0.575	−0.081	−0.262	−0.206	0.81
15	2.028	0.993	0.438	−0.363	0.87

Accordingly, there is no simple linear relationship capable of modelling benzenes and thiophenes at the same time. Therefore this result shows that it is incorrect to use a unique parameter (such as the reaction constant) for the whole set.

The SIMCA method also permits an evaluation of the statistical significance of the groupings by computing the interclass distance.[6] This is accomplished by fitting the objects of each class (B and T) to its own model and to the model of the other class, and evaluating the standard deviations obtained. The resulting distance of 3.4 confirms that monosubstituted benzenes and thiophenes are indeed well separated groups, since the value is significant on the $p = 0.01$ level of the F-distribution.

Partial Least-squares Analysis (*PLS*).—PCA is not aimed at finding out the relationships existing between one or more 'dependent' variables and a group of explanatory variables. In chemistry this is usually dealt with by multiple regression analysis (MRA) techniques.[15] However, MRA assumes that all the 'descriptor' variables are independent of each other, error-free, and 100% relevant to the problem. The partial least-squares (PLS) method[22] was recently developed to handle these problems in an alternative way, where the relevance of each independent variable results from the statistical analysis.

When the problem under investigation does not involve a single dependent variable, there are in fact two blocks of variables, and it becomes possible to define a dependent matrix

Table 4. Model expansions results for the independent PC analyses of the benzene and thiophene blocks of the B matrix. The figures indicate the fraction of variance explained by the model after including each A dimension

	%V(B)	%V(T)
$A = 1$	97	93
$A = 2$	99	97

Table 5. Scores from the PLS analysis: the t parameters are the latent variables for the X block and the u parameters are the latent variables of the Y block

Substituent	t(1)	u(1)	t(2)	u(2)
Methoxy	2.52	2.36	0.35	0.59
Methyl	1.32	2.40	−0.16	−0.18
Hydrogen	0.33	0.39	−0.34	−0.43
Chlorine	−0.35	−0.39	−0.01	−0.20
Bromine	−0.49	−0.46	−0.01	−0.17
Nitro	−3.33	−3.30	0.17	0.38

Table 6. Weights, averages, and loadings from the PLS analysis

Reaction	w(i)	ave	b(1)	b(2)
X block (benzenes)				
1	0.105	0.030	0.250	0.228
2	0.108	−0.078	0.252	0.068
3	0.120	−0.024	0.252	0.083
4	0.093	0.014	0.250	0.237
5	1.915	−0.013	−0.499	−0.599
6	1.459	0.080	−0.287	0.079
7	0.975	−0.297	0.284	−0.499
8	0.530	−0.140	0.291	−0.101
9	2.658	0.346	−0.501	0.507
Y block (thiophenes)				
10	0.143	−0.046	0.256	−0.034
11	0.125	0.048	0.253	0.216
12	0.146	0.008	0.255	−0.004
13	0.198	−0.179	0.236	−0.344
14	0.315	0.058	0.500	0.564
15	1.208	−0.922	−0.338	0.608
16	0.837	−0.732	0.361	−0.116
17	2.757	2.176	−0.505	0.363

Y and an 'independent' matrix X.[7] The question is whether or not the members of the Y matrix can be described as a function of the members of the X matrix.

In general, this problem is handled by computing principal components models for each of the two matrices, followed by establishment of a linear relationship between the principal components of these two blocks, respectively. Instead of this two-step procedure, it is now possible to make a single analysis in which the two steps are achieved simultaneously.

This method is called PLS2, and current experience shows that it is computationally much faster than PCA followed by MRA, and that it leads to a better prediction of the members of the Y matrix. The PLS2 method gives a description of the X matrix by one principal component-like model [equation (1)], a description of the Y matrix by an analogous model [equation (2)], and predictive relations between the latent variables t and u [equation (3)], where $d(a)$ is a proportionality coefficient for each dimension.

$$y(kj) = \bar{y}(j) + \sum_a c(ja)\,u(ak) + f(kj) \qquad (2)$$

$$u(ak) = d(a)\,t(ak) + h(ak) \qquad (3)$$

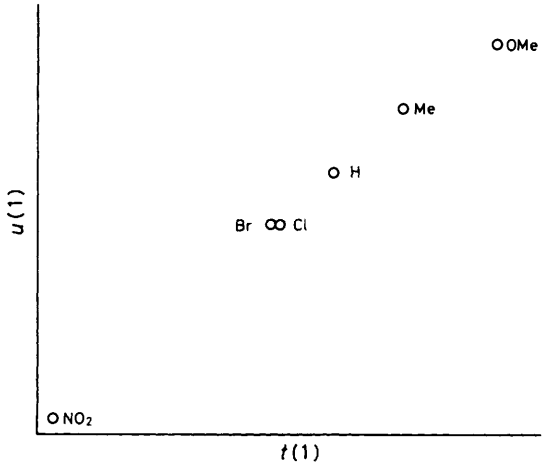

Figure 2. Plot of the first latent variable of the thiophene block against the first latent variable of the benzene block

The algorithm used in this MACUP method, presented in detail elsewhere,[7] is iterative for each dimension as in PCA. It consists of finding the latent variables of the X matrix $t(ak)$ from starting values of $u(ak)$ and the X elements, and then recomputing the latent variables of the Y matrix $u(ak)$ from the Y elements and the $t(ak)$ values, until the process converges. The meaning of b and t is the same as in PCA and can therefore be used in the same way in *e.g.* classification.

The **B** matrix is therefore formed by a block of y variables (eight thiophene reactions) and a block of x reactions (nine benzene reactions) for the same objects (the six substituents available).

It is possible to: (a) detect, by PCA, the structure of each block; (b) find out, by PLS2, how well the elements of the Y block can be predicted from the elements of the X block, or, in other words, how much of the systematic variation of the X block can be transferred to the Y block, and how many independent effects this relationship involves; (c) estimate, by comparing the two methods, whether the thiophene block has any systematic variation that cannot be explained in terms of the corresponding benzenes, spotting out peculiar intrinsic behaviour of thiophenes, if any.

The independent PCA of the two blocks shows that in both cases a two-component model describes almost completely the data structures. The numerical results for the model expansions are reported in Table 4.

The partial least-squares analysis (PLS2) is aimed at finding out the inner relationship between the latent variables (principal components) of each block under the constraint of maximizing the correlation between them. The results, obtained after pretreatment of missing data on the basis of the whole principal components model, are reported in Tables 5 and 6.

Again two components are significant according to cross-validation. The first component explains almost 93% of the total variance of the Y block, while the second one explains a further 4%, up to 97%. The results are illustrated in Figures 2 and 3.

Figure 2 shows the relationship between the first component (latent variable) of the benzene block, t(1), against the first component (latent variable) of the thiophene block, u(1). The points lie on a straight line, thus indicating that a large degree (93%) of simple linear relationship exists between the two sets of data.

To a first approximation, since the order of the substituents along the latent variable plot is the usual one from electron-withdrawers to electron-donors, this large degree of pro-

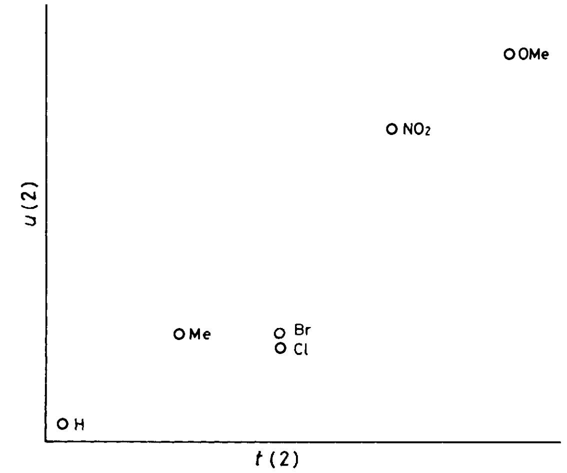

Figure 3. Plot of the second latent variable of the thiophene block against the second latent variable of the benzene block

portionality can be ascribed to the 'intrinsic' electronic effect typical of each substituent.

However, the second latent variable is also significant, indicating that a second effect, independent of the first, is also transferable from the benzene to the thiophene block. This orthogonal effect is much less important as it is responsible for a much smaller fraction of variance. The high value (1.5) of the coefficient of the inner relationship, *d*(2), indicates that this effect is much more important in thiophenes than in benzenes.

The order (Figure 3) in which the substituents line up along the straight line (methoxy, nitro, methyl, and hydrogen, with halogens behaving in a somewhat peculiar way) might be tentatively ascribed to an interaction of the substituents with the aromatic ring, enhanced in the thiophene series because of the presence of sulphur.

Since the fit of both the PCA and PLS models goes up to 97% these two effects apparently explain completely the data structure. Therefore the thiophene block does not exhibit any peculiar behaviour of its own.

The present result must also take into account our previous findings on the nature of substituent effects, showing the existence of well separated subgroups such as alkyls, halogens, acceptors, and donors.[23,24] We have stated that this discrete nature of substituents forces any mathematical model to describe almost exclusively the intergroup structure.[24]

Consequently, the relationship between thiophenes and benzenes might just be due to the presence of subgroupings. It is possible, however, to test how much of the variance of the *Y* matrix explained by PLS should be related to the intergroup variance, by replacing the *X* matrix with an appropriate simulation of groups. Thus an *X* matrix with three variables was constructued. For donors all three variables were zero, for alkyls and hydrogen (100), for halogens (010), and for acceptors (001). The PLS calculation with this matrix shows that the first latent variable explains 60% of the total variance, and the second a further 27%. This should be compared with 97% on using the benzene *X* matrix.

Accordingly, most of the structure of the data set should be ascribed to the variation between groups, but it is possible to observe that the large degree of proportionality found between thiophenes and benzenes also includes a significant contribution of the intergroup variance. In other words, this 'unique' effect pointed out by the first latent variable indicates that the relationship between thiophenes and benzenes involves mainly the relative effects of each group of substituents, but also some degree of systematic behaviour of substituents within each group.

Conclusions.—The combination of the results obtained by PCA and PLS clearly points out a number of features. (1) The variation of the aromatic ring is the most important structural change between the two series, even within the same reactions. Accordingly a unique reaction constant for all the substrates is not warranted.

(2) There is an approximate linear relationship between the major substituent effects in the two series. This explains why the application of separate Hammett models to substituted benzenes and thiophenes using the same σ values works fairly well. The variation of the ρ values from benzenes to thiophenes for the same reaction is indeed due to the different, but proportional, effect of the substituents on the ring, as previously suggested.[1,25]

(3) The data analysis shows that a single model can adequately describe substituent effects in both benzene and thiophene series, *i.e.* the reactivities of thiophenes can be predicted from those of benzenes.

(4) The model indicates that two independent effects are linearly transferred from benzenes to thiophenes. Tentatively, the predominant effect can be ascribed to the 'intrinsic' electronic characteristics of substituents, which are transmitted proportionally to the reaction centre, whereas the second effect, perhaps due to a different interaction between the substituents and the aromatic moiety, is much smaller.

(5) The fact that two significant dimensions (according to cross-validation) explain 97% of the variance, both in the PCA of the two separate benzene and thiophene matrices and in the PLS of their reaction, shows that the data are unusually precise and that the conclusions drawn are statistically highly significant.

Acknowledgements

Thanks are due to C.N.R. for a grant to S. C., to M.P.I. (Italy) for grants to S. C., P. L., and G. M., and to N.F.R., F.R.N., and S.T.U. (Sweden) for grants to M. S. and S. W. Helpful discussions with E. Johansson (Umeå) and D. Spinelli and G. Consiglio (Bologna) are also acknowledged.

References

1 S. Clementi and G. Marino, *Chem. Scr.*, 1977, **11**, 87.
2 (*a*) G. Marino, *Adv. Heterocycl. Chem.*, 1971, **13**, 235; (*b*) S. Clementi and G. Marino, *J. Chem. Soc., Perkin Trans. 2*, 1972, 71; (*c*) P. Linda, A. Lucarelli, G. Marino, and G. Savelli, *ibid.*, 1974, 1610.
3 J. Ridd, *Phys. Methods Heterocycl. Chem.*, 1971, **4**, 55.
4 (*a*) E. A. Hill, M. L. Gross, M. Stasiewicz, and M. J. Manion, *J. Am. Chem. Soc.*, 1969, **91**, 7381; (*b*) D. S. Noyce, C. A. Lipinski, and R. W. Nichols, *J. Org. Chem.*, 1972, **37**, 2615.
5 M. Sjostrom and S. Wold, *Acta Chem. Scand.*, 1981, **B35**, 537.
6 S. Wold and M. Sjostrom, in 'Chemometrics: Theory and Application,' ed. B. R. Kowalski, ACS Symposium Series no. 52, 1977, ch. 7.
7 S. Wold, C. Albano, W. J. Dunn, K. Esbensen, S. Hellberg, E. Johansson, and M. Sjostrom, in 'Food Research and Data Analysis', eds. H. Martens and H. Russwurm, Applied Science Publishers, London, 1983, p. 147.
8 S. Clementi and A. R. Katritzky, *J. Chem. Soc., Perkin Trans. 2*, 1973, 1077.
9 H. C. Brown and G. Marino, *J. Am. Chem. Soc.*, 1962, **84**, 1658.
10 L. M. Stock and H. C. Brown, *Adv. Phys. Org. Chem.*, 1963, **1**, 96.
11 H. C. Brown and Y. Okamoto, *J. Am. Chem. Soc.*, 1958, **80**, 4979.
12 O. Rogne, *J. Chem. Soc. B*, 1971, 1855.
13 G. Dauphin, A. Kergomard, and H. Verschambre, *Bull. Soc. Chim. Fr.*, 1967, **9**, 3395.

14 W. J. Bover and P. Zuman, *J. Chem. Soc., Perkin Trans. 2*, 1973, 786.

15 J. Shorter, 'Correlation Analysis of Organic Reactivity,' Research Study Press, New York, 1982.

16 A. R. Butler and C. Eaborn, *J. Chem. Soc. B*, 1968, 370.

17 E. Maccarone, G. Musumarra, and G. A. Tomaselli, *Ann. Chim. (Rome)*, 1973, **63**, 861, and references quoted therein.

18 A. Ballistreri, E. Maccarone, and G. Musumarra, *J. Chem. Soc., Perkin Trans. 2*, 1977, 984.

19 W. J. Scott, W. J. Bower, K. Bratin, and P. Zuman, *J. Org. Chem.*, 1976, **41**, 1952.

20 D. Spinelli, R. Noto, and G. Consiglio, *J. Chem. Soc., Perkin Trans. 2*, 1976, 747.

21 S. Wold, *Technometrics*, 1978, **20**, 397.

22 H. Wold, 'Systems under indirect observation,' eds. K. G. Joreskog and H. Wold, North Holland, Amsterdam, 1982, vol. 2.

23 D. Johnels, S. Clementi, W. J. Dunn, U. Edlund, H. Grahn, S. Hellberg, M. Sjostrom, and S. Wold, *J. Chem. Soc., Perkin Trans. 2*, 1983, 863.

24 S. Alunni, S. Clementi, U. Edlund, D. Johnels, S. Hellberg, M. Sjostrom, and S. Wold, *Acta Chem. Scand.*, 1983, **B37**, 47.

25 M. Fiorenza, A. Ricci, G. Sbrana, G. Pirazzini, C. Eaborn, and J. G. Stamper, *J. Chem. Soc., Perkin Trans. 2*, 1978, 1232.

Received 19*th March* 1984; *Paper* 4/436