Studies of ^{13}C N.M.R. Substituent Chemical Shifts of Disubstituted Benzenes Using Multivariate Data Analysis

Dan Johnels,[a] Ulf Edlund,*[,a] Erik Johansson,[b] and Marina Cocchi[c]
[a] *Department of Organic Chemistry, University of Umeå, S-901 87 Umeå, Sweden*
[b] *AB Hässle, S-431 83 Mölndal, Sweden*
[c] *Dipartimento di Chimica, Via Campi 183, 41100 Modena, Italy*

The non-additivity in the ^{13}C substituent chemical shifts (SCS) of 1,3-and 1,4-disubstituted benzenes has been studied using multivariate data analytical methods. Most of the non-additivity is systematic and can be predicted with high accuracy (0.2 ppm) from carbon shifts of monosubstituted benzenes. This means that no additional substituent effects are needed and that the non-additivity contributions of the SCS are clustered in the same way as previously noticed for ^{13}C SCS in monosubstituted benzenes.

A number of studies have been conducted on ^{13}C n.m.r. substituent induced chemical shifts (SCS) in aromatic systems. Commonly, the SCS are analysed by means of a dual substituent parameter (DSP) equation[1] (1) where the substituent

$$\delta_{SCS} = \rho_I\sigma_I + \rho_R\sigma_R \quad (1)$$

parameters, σ_I and σ_R, are claimed to reflect the substituent field and resonance effects, respectively. The σ_R values are chosen out of four different resonance scales, mostly based on measure of best fit. The ρ_I and ρ_R terms are susceptibility constants which are characteristic of the measured position. Some attempts have been made to analyse the ^{13}C SCS in disubstituted systems,[2] where both substituents are varied. The non-additivity of substituent effects observed in these systems[2] have initiated the present investigation.

A number of different models have been proposed to account for these non-additivities, originating from interactions between the substituents. The models range from simple one parameter equations[3] to different extensions of the DSP equation.[4] The extended models represent different approaches to describe changes in 'electron demand' placed on the substituents in disubstituted systems.

The most popular model so far is the DSP-non linear resonance (DSP-NLR) model equation (2)[4c–e,5] where ε is

$$\delta_{SCS} = \rho_I\sigma_I + \rho_R\sigma_R/(1 - \varepsilon\sigma_R) \quad (2)$$

another regression parameter expressing the electron demand placed on the substituent in the studied position.

In our previous studies of SCS in monosubstituted aromatic systems, we used the Principal Component (PC) and Partial Least Squares (PLS) data analysis.[6] There are several advantages using these methods. The parameterization of the resulting model is determined during the analysis, and all positions in a given system are modelled simultaneously. Furthermore, if the goal is to predict the chemical shifts of unknown or new structures in an optimal fashion, the PC/PLS approach is also a method of preference. Firstly, the shift values of all positions are modelled, and hence possible to predict, contrary to the traditional approach where only remote positions are analysed, one at a time. Secondly, the conventional studies use only one single global model, thereby overlooking systematic information present in local clusters, if present. Thirdly, the substituents included in the DSP-type studies are limited to substituents where ρ_I and σ_R values have been determined.

In our earlier multivariate studies it was found that the majority of the substituents belonging to one out of four groups, alkyls, acceptors, donors, or halogens.[6a] A limited number of substituents were found in between these clusters but these functional groups are only rarely included in multiparameter correlation studies. This type of clustering was also revealed by plotting σ_I *vs.* σ_R. If the continuity concept for a global model, like the DSP model, should be fulfilled, a larger number of the substituents should be chosen outside these groups,[6d] in order to get a fair representation of the total substituent domain.

Using only substituents similar to those recommended in the basis set, we found that the use of local models for each class of substituents significantly improved the description of SCS compared with DSP models.[6d] So if emphasis is on the predictive ability of the SCS model, local models for each substituent type (alkyls, acceptors *etc.*) are recommended. A global, unifying view of substituent effects, as inherited from the classical structure–reactivity studies, will reduce the predictive ability. A thorough comparison between DSP and PC approaches in aromatic substrates, where the probe position is located in rigid conjugated indicator groups has recently appeared.[7]

In this context, one should mention an alternative way to the prediction of chemical shifts, *i.e.* the use of 'knowledge-based' systems founded on shift databases or libraries.[8] One advantage of this 'hard model concept' is its generality, provided that the shift library is adequate both in size and information content. A severe limitation is, however, that no chemical information is gained as in PC or regression methods, *i.e.* in terms of components or 'effects.'

We have now extended our studies to *m*- and *p*-disubstituted benzenes, in order to investigate the behaviour of the non-additives of the ^{13}C SCS in these systems, and to study to what extent the non-additivities relate to the SCS in monosubstituted benzenes.

Methods

^{13}C N.m.r. chemical shifts of 1,3- and 1,4-disubstituted benzenes with 15 common substituents were collected from the literature.[4c–e] In total, 240 disubstituted systems were analysed. Since the same type of substituents appear in both positions, a symmetry in the data matrices is obtained. Hence, we get six data matrices, four matrices for the different positions in the 1,3 disubstituted system, and two matrices for the different positions in the 1,4 system, Figure 1.

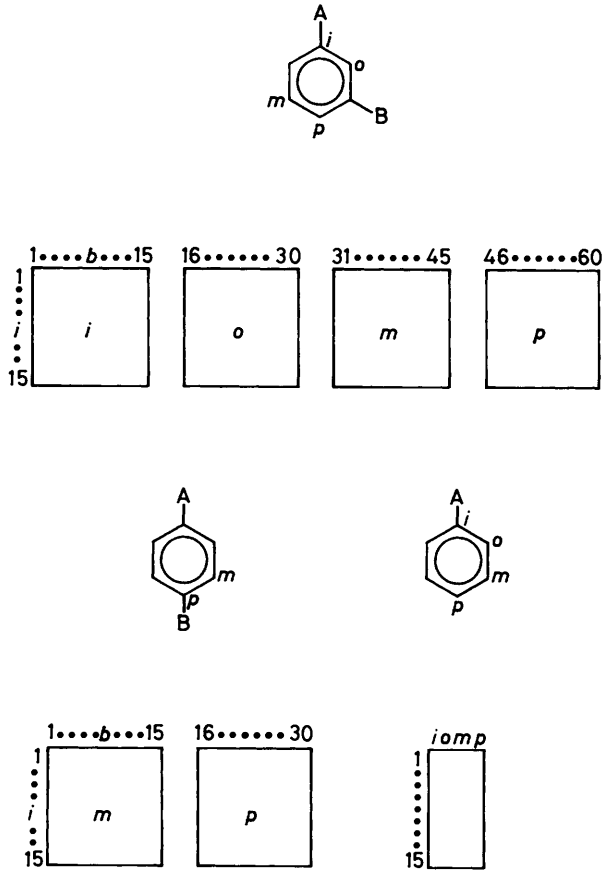

Figure 1. Construction of data matrices used in the calculations. *Substituents* (A *and* B) *in number order:* NMe_2, NH_2, OMe, F, Cl, Br, Me, H, CF_3, CN, COOEt, COMe, NO_2, H, CHO, OCOMe.

Table 1. ^{13}C N.m.r. chemical shifts for *p*-chlorobenzaldehyde.

	Position			
	ipso	*ortho*	*meta*	*para*
Reported shifts[a]	137.93	130.84	129.38	140.98
Predicted shifts	134.81	131.17	129.45	140.88
Our measured shifts	134.80	130.87	129.45	140.95

[a] Reference 4(*c*).

The PC and PLS methods can handle missing data in the matrices, provided that they are randomly distributed. However, this is not the case for the 1,3-disubstituted series. Hence the missing values were filled in by the shift values predicted by separate PC-analyses of the four shift matrices of the 1,3 system.

In order to extract the non-additivity in these matrices, the data in each matrix were transformed according to equation (3)

$$y'_{ik} = y_{ik} - \bar{y}_i - \bar{y}_k + \bar{\bar{y}} \quad (3)$$

where y'_{ik} is the non-additivity of substituent i and k in a given position. The original shift value is given by y_{ik}, $\bar{y}_i$ is the mean value for substituent i in a row in the data matrix, $\bar{y}_k$ is the mean value of substituent k in the column and $\bar{\bar{y}}$ is the grand mean value.

The non-additivity matrices, obtained by this transformation were merged systemwise, i.e. the 1,3 system gives rise to a 15 × 60 matrix and the 1,4 system to a 15 × 30 matrix, according to Figure 1. These data matrices were then subjected to PC analysis to extract the systematic behaviour of the data. This analysis was followed by a PLS analysis, in order to determine to what extent the systematic variation can be accounted for by the use of the chemical shifts of monosubstituted benzenes.

In short, the data is modelled in the following way in a PC analysis:[9]

$$y'_{ik} = \sum_{a=1}^{A} u_{ia} b_{ak} + e_{ik} \quad (4)$$

where the u_{ia} is the component value for substituent i, b_{ak} is the loading of substituent k. These terms are derived in a least squares fashion, *i.e.* by minimizing the square of the residuals e_{ik}.

The number of components, A, are determined by a crossvalidation method to secure that the components are statistically significant, *i.e.* to avoid over-parametrization. The resulting components are orthogonal to each other.

In the PLS method,[9a,b,10] the matrix containing the dependent variables is modelled as equation (4) while the matrix containing the independent variables is modelled simultaneously as:

$$x_{im} = \bar{x}_m + \sum_{a=1}^{A} f_{ia} p_{am} + g_{im} \quad (5)$$

where x_{im} is the scaled chemical shift in the monosubstituted benzene having substituent i in position m. The mean value in position m is given by $\bar{x}_m$, t_{ia} correspond to the component value of substituent i and p_{am} is the loading of position m. Finally the model error is given as g_{im}.

The two matrices are related as shown in equation (6) where

$$u_{ia} = c_a t_{ia} + h_{ia} \quad (6)$$

c_a is the least-squares inner regression coefficient and h_{ia} is the residual. Again, the number of components are determined by the cross-validation method.

In the PLS analysis, the X-matrix, containing the SCS in monosubstituted benzenes was scaled to give a total variance of one for each variable *i.e.* position in the monosubstituted benzene. Moreover, in the PLS analysis the data for the monosubstituted benzenes were deleted in the Y-block.

Results and Discussion

In the first PC-analysis, we observed that the reported shift value for the *ipso*-carbon in *p*-chlorobenzaldehyde deviated considerably from the value predicted by the model. In order to investigate the cause for this unexpected discrepancy, we remeasured the ^{13}C n.m.r. spectrum of this compound, Table 1. The shift values thus obtained are in accordance with the predicted values. Hence, the corrected shift values were used in the reported analyses.

The number of components and the percentage of variance explained by the different PC and PLS models are reported in Table 2.

A two-component PC model of the non-additivities in the 1,4 system accounts for 94% of the variance. The corresponding figure for the 1,3 system is 76% using a three-component model. Thus, the major part of the variation of the non-additivities is systematic.

In order to investigate if the systematic part can be modelled by the SCS of the corresponding monosubstituted benzenes, separate PLS analysis was carried out for the two different systems.

As evident from Table 2, 88% (83/94 × 100) of the systematic

variation is explained by the shift data of the mono-substituted benzenes in the 1,4 series, and 80% (61/70 × 100) in the 1,3 system.

In Table 3, the total averages in each position in the different systems are reported, together with the standard deviations at various stages of the analyses.

The standard deviation around the grand mean indicates that the largest substituent effects occur, as expected, in *ipso*, *ortho*, and *para* positions, while the *meta* positions are relatively unaffected.

From the standard deviation after the transformation according to equation (3), which removes the additivity from the data, it is evident that the non-additivity is small, *i.e.* in the order of 0.5 ppm except for the *para* position in the 1,4 series, where the standard deviation is 1.17 ppm.

After the PC modelling only 0.2 ppm, approximately, of the shift variation remains unexplained. The PLS analysis gives rise to standard deviations of the same magnitude, the only exception being the *para* position in the 1,4 system. In this context, it should be mentioned here that standard deviations of the order of 0.2 ppm were found when the SCS in 2-substituted naphthalenes were described by SCS in mono-substituted benzenes. Since the instrumental resolution is 0.05 ppm an unaccountable value of 0.2 ppm is quite acceptable using a global model.

The largest single contribution to the standard deviation of the *para* position in the 1,4 series is the NMe_2 substituent. This is an indication that the behaviour of this substituent is slightly altered in this system, relative to what is observed in the corresponding mono-substituted benzene. It is also noted that the electron demand of this substituent in the DSP/NLR treatment was considerably different from what was expected.[4c] It was proposed that the reason for this discrepancy is an alteration of the geometry of the substituent towards greater coplanarity to the benzene ring in the 1,4 system. If one accepts this explanation to the deviation, then it is understandable that the shifts can not be modelled by the SCS of the mono-substituted compound. When this substituent is deleted from the *Y*-block, the standard deviations in the *para* position are 1.04, 0.19, and 0.39, respectively (values in parenthessis in Table 3). The standard deviations in the *meta* position are not altered to any significant extent.

A qualitative view of the ability to model ^{13}C n.m.r. chemical shifts in these disubstituted systems from the shifts in mono-substituted benzenes is given in Figure 2, where the predicted shifts are plotted against the experimental shifts for 50 randomly selected entries in the data matrices. As is evident, the modelling of the shift data in disubstituted systems from shift data in mono-substituted systems is satisfactory.

In Figures 3(*a*), (*b*) and 4(*a*), (*b*) the two first components in the PC and PLS models are plotted for the 1,4 and 1,3 system, respectively. The same clustering of substituents according to substituent type, *i.e.* acceptor, alkyl, donor, or halogen, in the PC and PLS models is evident in all figures. The deviation between the PC and PLS model observed in the 1,4 system corresponds to a rotation of the model plane around the first component. Based on these observations we conclude that the models have approximately the same direction in the multidimensional space. However, as the number of substituents are limited, the derivation of local models for each substituent class is not recommended.

The observation of substituent clusters deserves some additional comment. An overwhelming majority of the reported multiparameter SCS studies uses at the maximum 10–15 substituents. If the substituents are non-uniformly distributed in the substituent domain, this number of substituents is not sufficient to manifest grouping into 3–4 groups. Hence, in each individual data analysis, such grouping can not be demonstrated. As mentioned, substituents are normally chosen according to some recommended basis set,[11] where substituents with different characteristics are proposed, *i.e.* they can be rated either as alkyls, acceptors, donors or halogens. We have shown that such a choice creates clustered data, a condition which can be proven if the number of substituents in each cluster is sufficient.[6a] Moreover, if one wishes to predict properties, like chemical shifts etc., local models are more suitable than one single global model.[6d] We have earlier shown that simple local models for each substituent type together has a significantly improved predictive ability relative to a more complicated global model. The σ_I and σ_R constants bear no information about the intra-cluster behaviour, but are merely labels of the substituent type.[6d] In fact, these parameters are correlated within each cluster.

The proliferation of new σ-scales is contrary to the original unifying concept. At present there exists in the literature more scales than substituents in any recommended basis set. This is in fact an unintended approach to the use of local models.

Table 2. Number of components (*A*) and explained variance (%) by the different multivariate models.

	System			
	1,3-disubstituted benezenes		1,4-disubstituted benzenes	
Model	A^a	%Variance	A^a	%Variance
PC	3	76	2	94
PLS	3	61	2	83

[a] Number of significant components according to cross-validation.

Table 3. Total average and standard deviations in the various positions at different stages of the statistical analyses.

	System					
	1,3-disubstituted benzenes				11,4-disubstituted benzenes	
	ipso	*ortho*	*meta*	*para*	*meta*	*para*
Total avg.	139.85	120.33	129.70	123.40	125.00	138.20
SD[a]	13.66	6.88	0.51	5.65	0.89	5.34
SD[b]	0.23	0.62	0.18	0.40	0.56	1.17(1.04)[e]
SD[c]	0.19	0.28	0.12	0.17	0.25	0.16(0.19)
SD[d]	0.23	0.39	0.12	0.27	0.19	0.51(0.39)

[a] Standard deviation around the grand mean value. [b] Standard deviation after transformation according to equation (3). [c] Standard deviation of the residuals from the PC model. [d] Standard deviation of the residuals from the PLS model. [e] Standard deviations in the different models when the NMe_2 is deleted from the modelling.

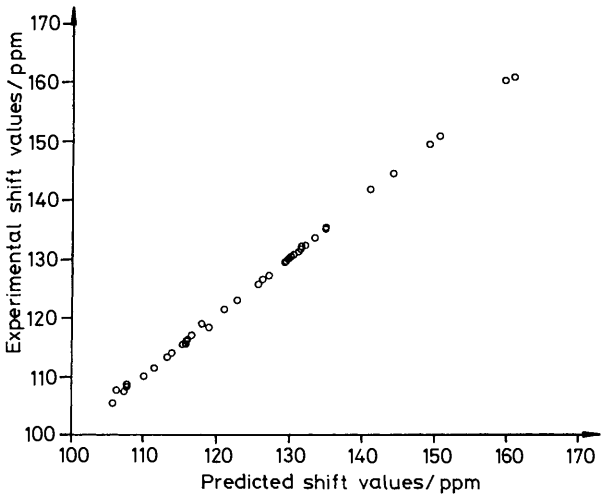

Figure 2. Plot of the experimental ^{13}C n.m.r. chemical shift values against the predicted shift values for 50 randomly selected entries from the data matrices.

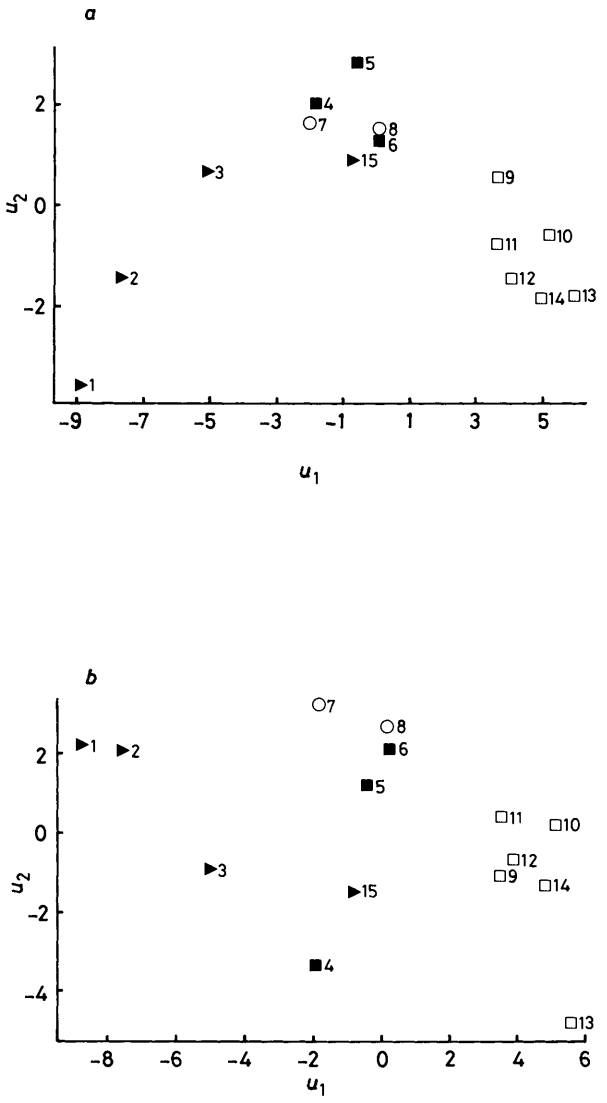

Figure 3. Plot of the two first components of the PC (*a*) and PLS (*b*) analysis of the 1,4-disubstituted benzenes.

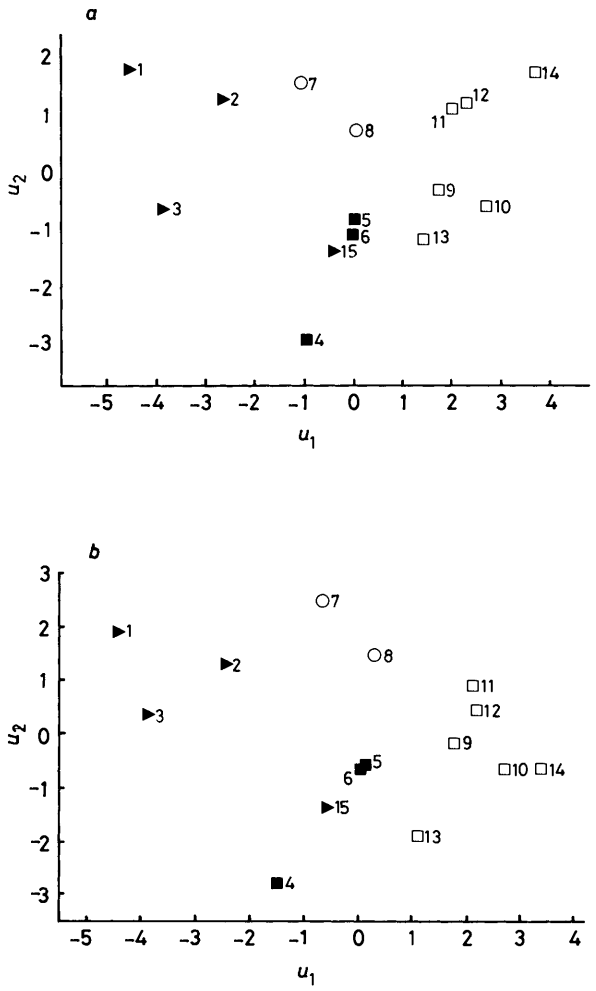

Figure 4. Plot of the two first components of the PC (*a*) and PLS (*b*) analysis of the 1,3-disubstituted benzenes.

Conclusions

It is shown that the major part of the non-additivity in ^{13}C n.m.r. shifts in 1,3- and 1,4-disubstituted benzenes is systematic and can be modelled from the shifts of mono-substituted benzenes using multivariate data analytical methods. This means, in general, that no new additional parameters or effects are necessary to describe the non-additivity. Furthermore, the same clustering of the substituents is observed, as previously found in monosubstituted aromatic systems. Although a single global statistical model gives acceptable shift predictions, *i.e.* in the order of 0.2 ppm, the observed clustering suggests that the predictive strength of this approach can be further enhanced by the use of local models.

Hence, instead of using a DSP/NLR approach, where each position demands as many regression calculations as the number of substituents, we present an easier and more efficient method, where all positions are adequately described using a single two or three component PCA/PLS model.

Acknowledgements

Financial support by the Swedish Natural Science Research Council is gratefully received.

References

1 S. Ehrenson, R. T. C. Brownlee, and R. W. Taft, *Prog. Phys. Org. Chem.*, 1973, **10**, 1.

2 J. Bromilow, R. T. C. Brownlee, R. D. Topsom, and R. W. Taft, *J. Am. Chem. Soc.*, 1976, **98**, 2020.
3 (*a*) B. M. Lynch, *Can. J. Chem.*, 1977, **55**, 541; (*b*) N. Inamoto, S. Masuda, K. Tori, and Y. Yoshimura, *Tetrahedron Lett.*, 1978, 4547; (*c*) F. Membrey, L. Boutin, and J.-P. Doucet, *ibid.*, 1980, 823; (*d*) F. Yamada, Y. Nakai, T. Nishiyama, and S. Nishi, *Technol. Rep. Kansai Univ.*, 1984, **25**, 103.
4 (*a*) D. A. Happer and G. J. Wright, *J. Chem. Soc., Perkin Trans. 2*, 1979, 694; (*b*) M. Mishima, M. Fujio, and Y. Tsuno, *Mem. Fac. Sci., Kyushu Univ., Ser. C*, 1980, **12**, 227; (*c*) J. Bromilow, R. T. C. Brownlee, D. J. Craik, M. Sadek, and R. W. Taft, *J. Org. Chem.*, 1980, **45**, 2429; (*d*) R. T. C. Brownlee and M. Sadek, *Aust. J. Chem.*, 1981, **34**, 1593; (*e*) J. Bromilow, R. T. C. Brownlee, D. J. Craik, and M. Sadek, *Magn. Reson. Chem.*, 1986, **24**, 862.
5 (*a*) R. G. Jones and G. Allen, *Org. Magn. Reson.*, 1982, **19**, 196; (*b*) N. K. Wilson and R. D. Zehr, *J. Org. Chem.*, 1982, **47**, 1184.
6 (*a*) D. Johnels, U. Edlund, H. Grahn, S. Hellberg, M. Sjöström, S. Wold, S. Clementi, and W. J. Dunn III, *J. Chem. Soc., Perkin Trans. 2*, 1983, 863; (*b*) D. Johnels, U. Edlund, E. Johansson, and S. Wold, *J. Magn. Reson.*, 1983, **55**, 316; (*c*) D. Johnels and T. A. Holak, *Chem. Scr.*, 1984, **24**, 203; (*d*) D. Johnels, U. Edlund, and S. Wold, *J. Chem. Soc., Perkin Trans. 2*, 1985, 1339; (*e*) G. Musumarra, D. Carbone, D. Johnels, and U. Edlund, *Magn. Reson. Chem.*, 1986, **24**, 209.
7 O. Exner and M. Budesinsky, *Magn. Reson. Chem.*, 1989, **27**, 27.
8 H. Kalchhauser and W. Robien, *J. Chem. Inf. Comput. Sci.*, 1985, **25**, 103, and references therein.
9 (*a*) S. Wold, M. Sjöström, and S. Hellberg, *Bull. ISI.*, 1987, **195**, 1; (*b*) S. Wold, C. Albano, W. J. Dunn III, U. Edlund, K. Esbensen, P. Geladi, S. Hellberg, E. Johansson, W. Lindberg, and M. Sjöström, in 'Chemometrics, Mathematics and Statistics in Chemistry,' ed. B. R. Kowalski, Riedel, Dordrecht, Holland, 1984, p. 17; (*c*) S. Wold, K. Esbensen, and P. Geladi, *Chemometrics and Intelligent Lab. Syst.*, 1987, **2**, 37.
10 P. Geladi and B. R. Kowalski, *Anal. Chim. Acta*, 1986, **185**, 1.
11 J. Bromilow, R. T. C. Brownlee, V. O. Lopez, and R. W. Taft, *J. Org. Chem.*, 1979, **44**, 4766.

Received 29*th November* 1988; *Paper* 8/04718I