PRINCIPAL COMPONENT ANALYSIS AND TARGET TESTING OF SUBSTITUENT EFFECT USING CARBONYL STRETCHING FREQUENCY AND ^{13}C NMR CHEMICAL SHIFT DATA MATRICES*

Ghazwan F. FADHIL

Department of Chemistry,
College of Science, University of Basrah, Basrah, Iraq

Received August 6, 1991
Accepted April 30, 1992

Principal component analysis technique has been applied to analyse the substituent effect on carbonyl stretching frequency and ^{13}C NMR chemical shifts. The general formula for the investigated molecules is X–G–Y, where X represents the set of substituent (OMe, Me, F, Cl, Br, CN and NO_2), Y is the probe site and G is benzene ring. According to the indicator function two significant components are responsible for the substituent effect. The validity of several substituent parameters have been investigated by target testing technique. Invariably substituent parameters derived by iterative multiple linear regression analysis viz. σ_R (Reynolds), σ_F (Reynolds) and σ_R (NMR) have lower SPOIL values when compared with other substituent parameters. Model designing of IR and ^{13}C NMR data matrices separately have shown that models which incorporate σ_R (Reynolds) and σ_F (Reynolds) or σ_R (NMR) and a substituent field parameters have the lowest root mean square error RMSE. Substituent effect on several properties are better correlated with Reynolds' σ_R and σ_F than with other commonly used substituent parameter(s). The orthogonality of substituent parameters used in the model can be achieved by including the methyl group in the substituent set.

The effect of a substituent transmitted through *p*- or *m*- substituted benzene ring on spectroscopic properties e.g. IR stretching frequency[1-4] and ^{13}C NMR chemical shift[5-9] has been investigated by several workers. To rationalize the substituent effect two different statistical models are used: Hammett type model (cf. Eq. (*1*)) and dual substituent parameters DSP (cf. Eq. (*2*))

$$\delta P = \rho \sigma \,, \tag{1}$$

$$\delta P = \rho_R \sigma_R + \rho_F \sigma_F \,, \tag{2}$$

* Presented as a poster at the 2nd Czechoslovak Chemometrics Conference, Brno 1990.

where δP is $(P_X - P_H)$, which is the substituent effect on the probe site whose spectroscopic property P being measured, ρ is a transmission factor for the property being investigated, σ an appropriate classical reactivity Hammett parameter (substituent parameter) such as σ_p, σ_p^+, σ_m etc., ρ_R and ρ_F are transmission factors for the resonance and field effects, respectively, and σ_R and σ_F are substituent parameters which measure the effect of resonance and field, respectively. Several substituent parameters have been suggested for the resonance and field effect viz.: *R* (SLH) and *F* (SLH, ref.[10]), *M* (Dewar) and *F* (Dewar, refs[11,12]), σ_R (general data, ref.[6]), σ_R (NMR, ref.[13]), σ_R (IR, ref.[14]), σ_R and σ_F (Reynolds, ref.[15]), σ_R (theor., ref.[16]), σ_F (Taft, ref.[13]), σ_F (Grob, ref.[17]), σ_F (Charton, ref.[18]), σ_F (Adcock, ref.[19]).

Usually simple or multiple linear regression analysis is performed to model ^{13}C NMR substituent chemical shift (SCS). However, several researchers have applied principal component analysis technique successfully to model ^{13}C NMR SCS[22 - 26]. The advantage of principal component analysis is that requires no prior knowledge of the number of component needed to model the data matrix[20].

In order to solve the controversy whether Hammett type model or DSP model is better in rationalizing the substituent effect, we have applied principal component analysis to two data matrices. The first data matrix consists of the substituent effect on the IR stretching frequency of the carbonyl group. The second data matrix consistes of the substituent effect on the remote ^{13}C NMR chemical shift. Our aim, inter alia, is to investigate the validity of various substituent parameters by utilising target testing technique and model designing. Below we give brief description of principal component and target testing techniques.

Malinowski and Howery[20] have defined a function to detect the significant number of components in a data matrix they called factor (component) indicator function *IND* (Eq. (3)),

$$IND = RE / (c - n)^2 , \tag{3}$$

$$RE = \left[\frac{\sum_{j=n+1}^{c} \lambda_j}{c(c-n)} \right]^{1/2} , \tag{4}$$

where *RE* is the real error, λ_j is the secondary eigenvalue (eigenvalue due to error), *j*, *c* and *n* are the number of rows, columns and components in the data matrix, respectively. *IND* function is a function of secondary eigenvalues, the number of rows and columns in the data matrix and the significant number of components. Hence the behaviour of the *IND* function varies with the number of components. We increased the number of components gradually and calculated each time *IND* function. As the

number of components increased the *IND* function decreased in value and reached a minimum when the significant number of components was achieved. To test the validity of a component Malinowski and Howery[23] have designed a target test to examine the validity of a test component in reproducing the data matrix. They defined a function called *SPOIL* as in Eq. (5),

$$SPOIL = \frac{RET}{EDM} \cong \frac{RET}{REP}, \tag{5}$$

where *RET* is the real error in target, *REP* is the real error in predicted target and *EDM* is the real error from the data matrix. According to Malinowski and Howery a *SPOIL* value between 0 and 3 is indication of an acceptable component. A component which gives a *SPOIL* value between 3 and 6 is moderately acceptable while a component with *SPOIL* value greater that 6 is not acceptable.

CALCULATIONS

Each datum in the IR stretching frequency matrix and in the ^{13}C NMR matrix was constructed as $P_X - P_H$, thus the location parameter was not used. Principal component analysis was performed for the covariance matrix, since the absolute error of each data column in each matrix were similar. For the same reason standerization was not applied[20]. Calculation were carried out on NEC ACOS 800 computer using FACTANAL computer programme*.

RESULTS AND DISCUSSION

Molecular systems used in this study have the general formula X–G–Y. X is the substituent attached *para* or *meta* to G, which is the main body of the molecular system and it is benzene ring by choice. Y is the rest of the molecular system where the probe lies in.

Principal Component Analysis of Substituent Effect

We have analysed two data matrices for the substituent effect. The IR data matrix which contains the substituent effect on the carbonyl IR stretching frequency for molecular systems presented in Table I. The latter matrix is the ^{13}C NMR SCS for carbon atoms with their molecular systems shown in Table II. Results for principal component analysis are presented in Table III. For both matrices the *IND* function reaches a minimum

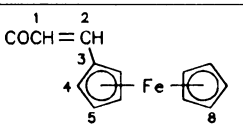
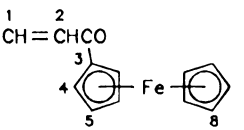
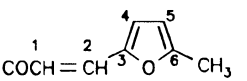
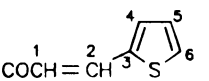
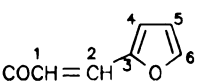
* The programme is available from Professor E. R. Malinowski, Department of Chemistry and Chemical Engineering, Stevens Institute of Technology, Hoboken, New Jersey 07030, U.S.A.

TABLE I
Molecular system X-C_6H_4-Y, for IR data (ν(CO) stretching, in CCl_4) matrix construction

Entry	Y	Position of X[a]	Ref.
1	$OCON(CH_3)_2$	*para*	10
2	$SCON(CH_3)_2$	*para*	10
3	$CON(CH_3)_2$	*para*	9
4	$CH{=}CHCON(CH_3)_2$	*para*	9
5	$COCH_3$	*para*	8
6	COOH	*para*	8
7	COOH	*meta*	8
8	$COCH_3$	*meta*	8
9	$CH{=}CHCON(CH_3)_2$	*meta*	9
10	$CON(CH_3)_2$	*meta*	9

[a] X = OMe, Me, F, Cl, Br, CN, NO_2.

TABLE II
Molecular systems *p*-X-C_6H_4-Y[a] for ^{13}C NMR ($CDCl_3$) chemical shift data matrix construction

Entry	Y	Carbon atoms used in the data matrix	Ref.
1		1, 2, 3, 4, 5, 8	7
2		1, 2, 3, 4, 5, 8	7
3		1, 2, 3, 4, 5, 6	8
4		1, 2, 3, 4, 5, 6	8
5		1, 2, 3, 4, 5, 6	8

[a] X = OMe, Me, F, Cl, CN, NO_2.

at two abstract factors, which indicates that two components are sufficient to reproduce each data matrix i.e. the dual substituent parameter DSP model (cf. Eq. (*2*)) is valid for both data matrices.

Target Testing of Substituent Parameters

The validity of several substituent parameters mentioned in the introduction has been tested by utilising target testing technique, each data matrix was tested separately. Table IV gives *SPOIL* values at two components, since results from principal component analysis have indicated that two components are sufficient to reproduce the substituent effect data matrices. Results for testing substituent resonance parameters mentioned in the introduction, for both data matrices, have shown that σ_R (Reynolds) has the lowest *SPOIL* value. σ_R (NMR) has the next lowest *SPOIL* value, while other substituent resonance parameters, viz. *R* (SLH), σ_R (GD), *M* (Dewar), σ_R (IR) and σ_R (theor.), have shown acceptable *SPOIL* values, their *SPOIL* values were significantly higher than σ_R (Reynolds) and σ_R (NMR). We attribute this to the effectiveness of the iterative multiple linear regression analysis of substituent effect on ^{13}C NMR chemical shift for 3- and 4-substituted styrene[15] and 4-substituted benzene[13], respectively. Similarly target testing for substituent field parameters using both data matrices separately

TABLE III
Variation of *IND* function with increasing number of components

No. of factors	Real error	*IND* function
	IR data matrix	
1	1.2541	$3.48 \cdot 10^{-2}$
2	0.4301	$1.72 \cdot 10^{-2}$
3	0.3092	$1.93 \cdot 10^{-2}$
4	0.2764	$3.07 \cdot 10^{-2}$
5	0.2170	$5.42 \cdot 10^{-2}$
6	0.1568	0.1568
	^{13}C NMR data matrix	
1	0.2518	$1.007 \cdot 10^{-2}$
2	0.1148	$7.177 \cdot 10^{-3}$
3	$7.2099 \cdot 10^{-2}$	$8.011 \cdot 10^{-3}$
4	$5.1608 \cdot 10^{-2}$	$1.290 \cdot 10^{-2}$
5	$4.7854 \cdot 10^{-2}$	$4.785 \cdot 10^{-2}$

have shown that σ_F (Reynolds) has the lowest *SPOIL* value when compared with other types of σ_F (cf. Table IV).

This result is in good agreement with the above finding regarding the effectiveness of the iterative multiple linear regression analysis method for deriving substituent parameters. Other substituent parameters for the field effect viz. σ_F (Grob), σ_F (Adcock), σ_F (Taft) and σ_F (Charton) have more congested SPOIL values. *F* (Dewar) and *F* (SLH) have relatively higher *SPOIL* values than other substituent filed parameters.

Model Designing

We have constructed DSP models by using substituent parameters with acceptable *SPOIL* values. Modeling IR data matrix gave only three models with root mean square error *RMSE* value closer to the calculated real error *RE* at two components (cf. Table V). These models are (σ_R (NMR), σ_F (Taft)), (σ_R (NMR), σ_F (Charton)) and (σ_R (Reynolds), σ_F (Reynolds)). Their *RMSE* values are very low when compared with other models' *RMSE* values. Modeling ^{13}C NMR data matrix disclosed several models with *RMSE*

TABLE IV
SPOIL values at two components for substituent parameters

Tested substituent parameter	Data matrix used to test substituent parameter		Ref.
	IR	^{13}C NMR	
σ_R (Reynolds)	1.85	0.28	15
σ_R (NMR)	2.07	0.71	13
σ_R (SLH)	2.55	2.07	10
σ_R (general data)	2.56	1.31	20
M (Dewar)	2.61	1.91	11, 12
σ_R (theor.)	2.67	1.62	16
σ_R (IR)	2.92	1.37	14
σ_R^n	3.99	3.03	29
σ_F (Reynolds)	0.80	0.43	15
σ_F (Grob)	1.13	1.49	17
σ_F (Adcock)	1.20	0.79	19
σ_F (Taft)	1.18	1.85	13
σ_F (Charton)	1.24	1.57	18
F (Dewar)	1.68	2.43	11, 12
F (SLH)	1.75	2.43	10

values closer to *RE* (cf. Table V) at two components. The lowest RMSE value was given by σ_R (Reynolds), σ_F (Reynolds)) model. This result is in accordance with that of target testing i.e. substituent parameters derived by iterative multiple linear regression analysis methods, are the best in reproducing the substituent effect.

Condition for Non-Fortuitous Correlation

The quality of a correlation for a property with dual substituent parameter could be fortuitous if the substituent parameters in Eq. (*2*) were not orthogonal i.e. the correlation coefficient between the substituent parameters used in the equation was high e.g. 0.9. Figure 1 shows a plot between σ_R (Reynolds) and σ_F (Reynolds) with correlation coefficient 0.6. The methyl group is an outlier and if it is deleted from the correlation the correlation coefficient will be 0.9. Hence if a dual substituent correlation was performed this would give very high correlation coefficient for Eq. (*2*). Thus for chemically meaningful correlation the substituent set must include the methyl group as a necessary condition.

TABLE V
DSP models constructed from IR and ^{13}C NMR data matrices with root mean square error (*RMSE*) values

DSP model	*RMSE* values	
	IR	^{13}C NMR
σ_R (general data), σ_F (Taft)	0.709	0.112
σ_R (general data), σ_F (Charton)	0.735	0.110
σ_R (general data), σ_F (Grob)	0.935	0.132
σ_R (general data), σ_F (Adcock)	0.931	0.126
σ_R (NMR), σ_F (Taft)	0.658	0.119
σ_R (NMR), σ_F (Charton)	0.664	0.112
σ_R (NMR), σ_F (Grob)	0.846	0.127
σ_R (NMR), σ_F (Adcock)	0.849	0.118
σ_R (IR), σ_F (Taft)	0.834	0.117
σ_R (IR), σ_F (Charton)	0.863	0.116
σ_R (IR), σ_F (Grob)	1.043	0.140
σ_R (IR), σ_F (Adcock)	1.015	0.132
σ_R (Reynolds), σ_F (Reynolds)	0.689	0.104
R (SLH), *F* (SLH)	0.730	0.122
M (Dewar), *F* (Dewar)	0.719	0.121

TABLE VI
Comparison of correlation quality for Reynolds' model and other substituent parameter(s) models

Molecular system	Correlated property		Statistical results[a]		Other[b] models		Statistical results		Ref.
			r	*f*			*r*	*f*	
p-X-C_6H_4-NHCOCH$_3$	IR	ν(CO)[c]	0.981	0.205	σ_I,	σ_R^{BA}	0.979	0.21	4
p-X-C_6H_4-N=C=S	^{13}C	SCS[d]	0.998	0.064	–	–		–	27
p-X-C_6H_4-N=$\overset{\alpha}{C}H$-$\overset{\beta}{C}H$=$\overset{\gamma}{C}H$-C_6H_5	^{13}C	SCS							
	α-C		0.983	0.197		σ_p^+	0.999	0.024	9
	β-C		0.996	0.087		σ_p	0.976	0.211	9
	γ-C		0.999	0.016		σ_p	0.993	0.118	9
p-X-C_6H_4-CH(C_6H_5)-C_6H_4-F	^{19}F	SCS	0.995	0.085	σ_I,	σ_R^0	–	0.144	28
p-X-C_6H_4-CH=N-C_6H_4-F (*p*)	^{19}F	SCS	0.999	0.024	σ_I,	σ_R^0	–	0.062	28
p-X-C_6H_4-CH=CH-C_6H_4-F (*p*)	^{19}F	SCS	0.999	0.038	σ_I,	σ_R^{BA}	–	0.060	28
p-X-C_6H_4-N=C=S	π-electron density[e]		0.985	0.182	–	–	–	–	27

[a] Reynolds' DSP model. [b] Using different substituent parameters, see refs. [c] Stretching. [d] Remote C-atom. [e] Relative, CNDO.

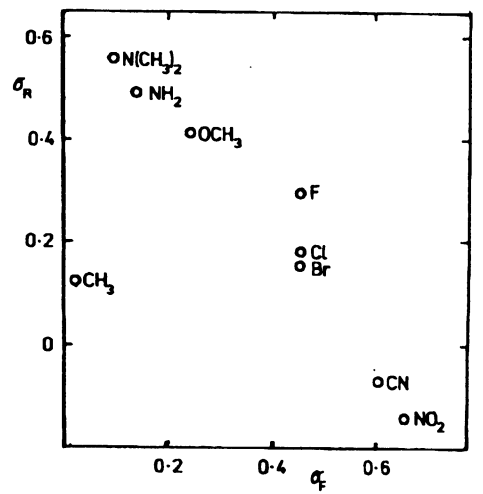


FIG. 1
Plot of σ_R (Reynolds) against σ_F (Reynolds)

Correlation of Reynolds' Substituent Parameters with the Substituent Effect of Different Properties

Target testing of both data matrices have shown that Reynolds' substituent parameters have the lowest *SPOIL* value when compared with other substituent parameters mentioned in the introduction. Modelling ^{13}C NMR data matrix demonstrated that the lowest *RMSE* value was given by a model which uses Reynolds' substituent parameters. In order to cast more light on Reynolds' substituent parameters, we have correlated Reynolds' substituent parameters with the substituent effect on IR stretching frequency, ^{13}C NMR and ^{19}F NMR chemical shifts and π-electron atomic charges, for several molecules. Results of such correlations are presented in Table VI. Other workers[4,9,27,28] have correlated the same substituent effect on the mentioned properties with different substituent parameter(s) on the same set of molecules we have used. For comparison their correlation coefficients (*r*) and (*f*) values are presented in Table VI. Generally the (*r*) and (*f*) values for DSP model using Reynolds' substituent parameters are better than (*r*) and (*f*) values derived from other models using other type of substituent parameters.

We are grateful to Professor O. Exner for helpful discussion.

REFERENCES

1. Laurence C., Berthelot M.: J. Chem. Soc., Perkin Trans. 2 *1972*, 98.
2. Spaargaren K., Kruk C., Molenarr-Langeveld T. A., Korver P. K., Van Der Haak P. J., Deboer T.: Spectrochim Acta, A *28*, 965 (1972).
3. Perjéssy A., Jones R. G., McClair S. L., Wilkins J. M.: J. Org. Chem. *48*, 1266 (1983).
4. O'Conner C. J., McLennan D. J., Calver D. J., Lomax T. D., Porter A. J., Rogers D. A.: Aust. J. Chem. *37*, 497 (1984).
5. Nelson G. L., Levy G. C., Crigioli J. D.: J. Am. Chem. Soc. *94*, 3089 (1972).
6. Ehrenson S., Brownlee R. T. C., Taft R. W.: Prog. Phys. Org. Chem. *10*, 1 (1973).
7. Solcaniova E., Toma S., Fiederova A.: Org. Magn. Reson. *14*, 181 (1980).
8. Musumarra G., Ballistreri F.: Org. Magn. Reson. *14*, 385 (1980).
9. Radeglia R., Kim D. G., Bodeker J.: J. Prakt. Chem. *326*, 505 (1984).
10. Hansch C., Lee A., Unger S. H., Kim K. H., Nikaitani D., Lien B. J.: J. Med. Chem. *16*, 1207 (1973).
11. Dewar M. J. S., Grisdale P. J.: J. Am. Chem. Soc. *84*, 3539 (1962).
12. Dewar M. J. S., Grisdale P. J.: J. Am. Chem. Soc. *84*, 3548 (1962).
13. Bromilow J., Brownlee R. T. C., Lopez V. C., Taft R. W.: J. Org. Chem. *44*, 4766 (1979).
14. Katritzky A. R., Topsom R. D.: Chem. Rev. *77*, 639 (1977).
15. Reynolds W. F., Gomes A., Maron A., MacIntyre W., Tanin A., Hamer G. K., Peat I. R.: Can. J. Chem. *61*, 2376 (1983).
16. Marriott S., Topsom R. D.: J. Chem. Soc., Perkin Trans. 2 *1985*, 1045.
17. Grob G. A., Schaub B., Schlageter M. G.: Helv. Chim. Acta *63*, 57 (1980).
18. Charton M.: Prog. Phys. Org. Chem. *13*, 199 (1981).
19. Adcock W., Abeywickrema A. N.: J. Org. Chem. *47*, 2957 (1982).
20. Malinowski E. R., Howery D. G.: *Factor Analysis in Chemistry*. Wiley – Interscience, New York 1980.

21. Wold S., Albano C., Dunn W. J. III, Edlund U., Esbensen K., Geladi P., Helleberg S., Johansson E., Lindberg W., Sjostrom M. in: *Chemometrics. Mathematics and Statistics in Chemistry* (B. R. Kowalski, Ed.). Reidel, Dordrecht 1983.
22. Musumarra G., Wold S., Gronowitz S.: Org. Magn. Reson. *17*, 118 (1981).
23. Reynolds W. F., Gomes A., MacIntyre D. W., Munder R. G., Tanin A., Wong H. E., Hamer G. K., Peat I. R.: Can. J. Chem. *61*, 2367 (1983).
24. Edlund U., Grahn H., Helleberg S., Sjostrom M., Wold S., Clementi S., Dunn W. J. III: J. Chem. Soc., Perkin Trans. 2 *1983*, 863.
25. Exner O., Buděšínský M.: Magn. Reson. Chem. *27*, 276 (1989).
26. Buděšínský M., Exner O.: Magn. Reson. Chem. *27*, 585 (1989).
27. Danihel I., Kristian P., Böhm S., Kuthan J.: Chem. Zvesti *38*, 39 (1984).
28. Dayal S. K., Taft R. W.: J. Am. Chem. Soc. *95*, 5595 (1973).
29. Hoefnagel A. J., Wepster B. M.: J. Am. Chem. Soc. *95*, 5357 (1973).