PRINCIPAL COMPONENT ANALYSIS OF SUBSTITUENT CONSTANTS

Drahomír HNYK

Institute of Inorganic Chemistry,
Czechoslovak Academy of Sciences, 250 68 Řež near Prague

Received February 14, 1989
Accepted March 17, 1989

Dedicated to Professor Otto Exner on the occasion of his 65th birthday.

The principal component analysis has been applied to a data matrix formed by 7 usual substituent constants for 38 substituents. Three factors are able to explain 99·4% cumulative proportion of total variance. Several rotations have been carried out for the first two factors in order to obtain their physical meaning. The first factor is related to the resonance effect, whereas the second one expresses the inductive effect, and both together describe 97·5% cumulative proportion of total variance. Their mutual orthogonality does not directly follow from the rotations carried out. With the help of these factors the substituents are divided into four main classes, and some of them assume a special position.

Several attempts have already been made to determine the optimum linear free energy relationship (LFER) describing the substituent effects. If we leave out the Taft equation[1] – considered classical at present – and the newer relation by Yukawa and Tsuno[2] both of them stemming in fact from the prototype of all LFERs – Hammett equation[3], then there exist four approaches to solution of this problem, namely those applying the multidimensional statistical methods.

Thus, in the first place, Swain et al.[4] applied the nonlinear least squares treatment to a data matrix comprising 14 reaction series and 43 substituents. The reaction series are represented not only by the original Hammett $\sigma_{p,m}$ constants but also by the values of σ_p^-, σ_p^+, σ_m^+ variables and those describing the inductive effect or derived for various positions of substituents in the naphthalene skeleton. The substituent effects are then expressed by a two-parameter relation where the parameters derived are the nonresonance or field constant and the resonance constant. The drawbacks of this analysis can be summarized in two points: *1*) the data matrix is filled to 38% only, whereby the statistical significance of the whole procedure is decreased, and *2*) the values given for the reactions taking place in the naphthalene skeleton de facto are the $\sigma_{p,m}$ constants which thus appear several times, and also the 3 constants describing the inductive effect are of the same nature. The paper by Nieuwdrop[5] treats 76 selected reactions and equilibria divided into five groups and applies another statistical approach known under the name of factor analysis. The respective log (K_H/K_X)

or log (k_X/k_H) are affected by 17 substituents selected in such way that they might represent, as far as possible, all possible effects on the course of the given reaction. This data matrix is filled to 44%. The results of this procedure show that three constants are necessary to describe the substituent effects on the equilibrium or rate of the set of reactions studied, these constants being obtained by rotation of the factors provided by the factor analysis. The first constant can be compared with the Taft σ_I constants and the other with σ_R^0, the significance of contribution of the third derived constant depending on the reaction type considered. Haldna et al.[6] investigated the effects of 24 substituents on rates or equilibria of 10 reactions using some methods of factor analysis, too, inter alia the principal component analysis (PCA) and spectral-isolation factor analysis. The PCA gave four factors, if a 100% interpretation of the variance is considered. The first one of them, which explains 89% of the variance, gives a good correlation with σ. Applying the second above--mentioned method, the authors of this publication found four factors, but only to two of them can be assigned a physical meaning. One of them can be correlated with the σ or σ^0 constants, the corrrelation coefficient being decreased with increasing number of the factors (2–4) taken into account. This finding is explained by the composite nature of these constants. On the other hand, the other factor correlates with the σ_R^+ constants, hence its meaning can be connected with the resonance effect. The last two factors are insignificant, if we realize that they only explain 2% and 1% of the total variance. The data matrix used in the present paper is filled up to 100% (after supplying some values by multidimensional regression; the individual parameters depend on the reaction type), but there are only 13 substituents proper, because 11 of the substituents are considered twice (for *meta* and for *para* positions), and these are not treated independently in contrast to the procedure given in ref.[5]. Finally, Wold et al.[7] carried out PCA for 28 substituents using not only the Hammett type constants but also others, such as E_s constants, π values, and MR values (for molar refraction) which are used as descriptors here. Since PCA is a considerably flexible method, it is almost always formally successful, even in cases where the descriptors describe different phenomena, e.g. electronic, sterical, polarizability, i.e. those analyzed together in the present paper. However, before the application itself of PCA, the descriptors were transformed into standard scores (then the variables show unit variance), whereby the same statistical significance is alloted to all variables. The result of this analysis is two principal components describing 82% of the overall variance.

RESULTS AND DISCUSSION

The aim of the present paper was to analyze, by means of the PCA method, such a data matrix which would be rid, as far as possible, of all the drawbacks mentioned, although it is, of course, perhaps impossible to avoid some of these drawbacks, particularly if a sufficiently large data set must be treated. For this analysis we chose the logarithms of relative equilibrium constants of 6 reaction series, one of them taking place in the bicyclooctane skeleton and the remaining ones in benzene ring (Fig. 1), i.e. the σ_j substituent constants ($j = I$, m, p°, p, p^+, p_{An}^-, p_{Ph}^-) for 38 substituents covering roughly equally both donors and acceptors; Ph, CH=CHPh, C≡≡CPh, and SOMe can be considered to act simultaneously as donors and acceptors[8]. The same substituents differing only in their position are used only for one reaction (*II*), the same kind of reaction is represented by *V* and *VI* and the respective constants σ_j ($j = p_{\mathrm{An}}^-$, p_{Ph}^-), and the reascn of it was to ensure an at least partial equilibrium between the constants describing the electron-donor and electron-acceptor properties. The values used for the σ_p^+ variable for electron acceptors ($\sigma_p^- > \sigma_p^\circ$) are not a result of measurement but were obtained, for the given data matrix, on the basis of the fact that $\sigma_p^+ = \sigma_p = \sigma_p^\circ$. For electron donors ($\sigma_p^+ < \sigma_p$) it is similarly $\sigma_p^- = \sigma_p^\circ$, which was also utilized in constructing the data matrix which, on the basis of these two presumtions, is then filled to 96·2%. The objectively missing data were completed by the stepwise regression procedure, hence this model data matrix is completely filled.

The procedure used in the present report involves, in its last steps, a solution of secular problem which is generally equivalent to the diagonalization of matrix, in our case the correlation matrix is diagonalized*. This matrix was obtained from a data matrix whose elements were transformed into the standard scores at first. In this way the $\mathbf{R} \equiv [r_{ji}]$ matrix was transformed into $\mathbf{Z} \equiv [z_{ji}]$ matrix, their mutual relation being $\mathbf{R} = \mathbf{Z}\tilde{\mathbf{Z}}/38$. In this case the σ_j variables had the same significance also before the transformation $r_{ji} \rightarrow z_{ji}$ was carried out, because they describe a very narrow region of phenomena. The first skeleton transmits the pure inductive effect, whereas the other one transmits various proportions of the inductive and mesomeric effects depending on the nature of reaction centre and position of X and or Y.

The PCA model $\mathbf{Z} = \mathbf{AF}$ (ref.[9]) was applied to the above-mentioned relation for the čorrelation matrix (**A** means the factor loading matrix and **F** means the matrix of factor scores), which, under the presumption that the factors are non-correlatable, leads to the matrix of reproduced correlations $\mathbf{T} = \mathbf{A}\tilde{\mathbf{A}}$. As already mentioned the results of diagonalization of matrix **Z** are the respective eigenvalues and eigen-

* The first step of PCA involves a selection of the first-factor coefficients a_{j1} (the elements of **A** matrix) carried out in such way as to make the maximum sum of contributions of this factor into the communality. This sum is given by the relation $V_1 = a_{11}^2 + a_{21}^2 + \ldots + a_{n1}^2$, and the a_{j1} coefficients must be chosen in such way that V_1 might be maximum at the conditions $r_{jk} = \sum_{p=1}^{m} a_{jp}a_{jk}$; m means the number of factors. These conditions express that the correlations observed should be replaced by the reproduced ones where $r_{jk} = r_{kj}$, and r_{jj} is the communality h_j^2 of the z_j variable. The whole procedure is similarly repeated for the other factors.

vectors, the eigenvalue expressing the variance explained by each factor $(V_p = \sum_{j=1}^{7} a_{jp}^2)$, whereas the eigenvectors are directly related to the factor loading matrix (after their multiplication by the second square root of the respective eigenvalue of correlation matrix) of the $7 \times m$ magnitude where m means the number of factors taken into account. The overall variance is then defined as a sum of diagonal elements of the correlation matrix $(V = \sum_{p=1}^{m} V_p = 7)$. The squared multiple correlations of the respective vaiiable with the remaining six variables were taken as the initial values of the communalities forming the diagonal of matrix $\boldsymbol{T}$ (in terms of the PCA model only one iteration is carried out for the calculation of communalities). The diagonalization of matrix $\boldsymbol{Z}$ gives the following eigenvalues along with the cumulative proportion of total variance (Table I).

The components obtained represent vectors in a less-dimensional space $(m < 7)$ as compared with the space given by the original variables $(m = 7)$. These are mutually orthogonal and without any physical or chemical meaning. Hence in order to obtain it rotations were carried out which are related to the postulate that the communality of each variable is invariant and its square remains constant, too. These conditions then allow a derivation of the simplicity criterion whose minimization is carried out as the rotation and which can be summarized[9] as follows

$$G = \sum_{p<q=1}^{m} \left[\sum_j a_{jp}^2 a_{jq}^2 - \frac{\Gamma}{7} \left(\sum_j a_{jp}^2\right) \left(\sum_j a_{jq}^2\right) \right], \qquad (1)$$

where the Γ value and the meaning of a_{pq} are decisive for the rotation. If we consider vectors in the rotation always mutually orthogonal, then a_{pq} are elements of factor loading matrix and $\Gamma \in R$ except for $\Gamma = 0, 1$. The first case represents the so-called

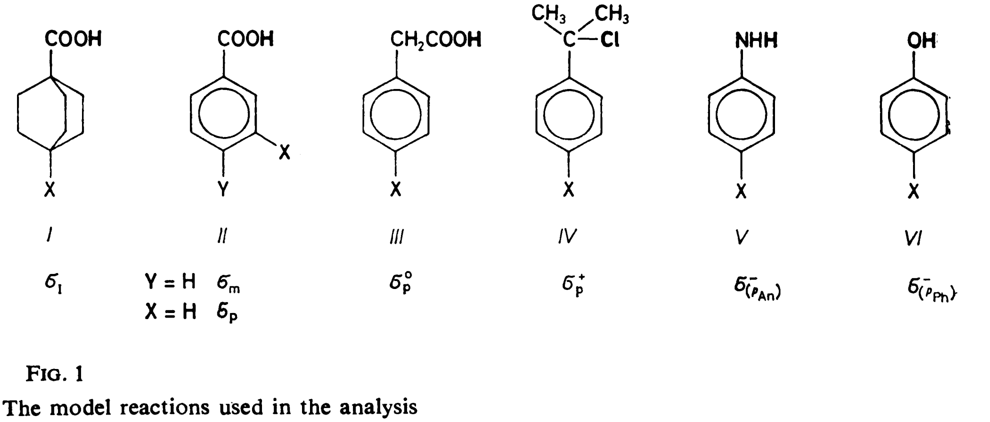

FIG. 1

The model reactions used in the analysis quartimax method and the second case the varimax method. However, if a_{pq} represent elements of the factor structure matrix $\boldsymbol{S}$ ($\boldsymbol{S} = \boldsymbol{AC}$, $\boldsymbol{C}$ is the matrix of correlation coefficients between the factors considered), then the rotated factors cease to be orthogonal after the rotation and represent a result of oblique rotation, $\Gamma \in R$ in the case of the direct oblimin method and $\Gamma \doteq 0$ in the case of direct quartimin method. The convergence criterion chosen for the rotation was 10^{-5}. From Table I it can be seen that the two factors are able to explain 97·5% cumulative proportion of total variance, the addition of the third one will explain 99·4%. (As the rank of the correlation matrix is 5, five factors are formally responsible for 100% interpretation of cumulative proportion of total variance.) The rotations were carried out for the first two factors, and the significance of the third one obviously cannot be rejected either. In the rotation one degree of freedom is consumed by application of the simplicity criterion as such, another one is still left. Inspection of Table II will reveal that the first factor can be considered as a measure of the resonance effect R (the loading of the first factor for σ_I is the least), whereas the meaning of the second factor can be related to the inductive effect I (the largest loading for σ_I). Therefore, the remaining degree of freedom can be consumed by the choice of such Γ in Eq. (*1*) that it might be $a(\sigma_I 1) = 0{\cdot}0$, since this constant by its definition does not reflect any resonance effect. In other words this means an extraction of the inductive effects from the reaction series *II*–*VI*. Hence, if two factors are rotated, then this rotation is mediated by a transformation matrix of 2 × 2 magnitude. As there is no reasonable reason for the above-mentioned factors to be orthogonal, at first we tried an oblique rotation with the presumption that a potential orthogonality should follow from this rotation. The attempt at reaching $a(\sigma_I 1) = 0{\cdot}0$ was unsuccessful in the

TABLE I

The variance explained for each factor together with the respective cumulative proportion of total variance

Factor m	Variance explained	Cumulative proportion of total variance, %
1	6·297	90·0
2	0·527	97·5
3	0·136	99·4
4	0·022	99·7
5	0·018	100·0
6	0·000	100·0
7	0·000	100·0

TABLE II

The matrix **A** along with the rotated factor loadings (Matrix **B**)

j	**A**		**B**	
	Factor 1 (R)	Factor 2 (I)	Factor 1 (R)	Factor 2 (I)
σ_I	0·820	0·562	0·000	0·994
σ_m	0·969	0·210	0·520	0·558
σ_p^o	0·996	−0·044	0·831	0·221
σ_p	0·988	−0·146	0·942	0·080
σ_p^+	0·901	−0·355	1·112	−0·228
$\sigma^-(p_{An})$	0·989	−0·036	0·817	0·229
$\sigma^-(p_{Ph})$	0·964	−0·126	0·900	0·101
V_p	6·297	0·527	4·562	1·468

TABLE III

The factor values for the original variables

i	Factor 1	Factor 2	i	Factor 1	Factor 2
H	−0·406	−1·361	Ph	−0·405	−0·802
Me	−0·797	−1·378	CH_2Cl	−0·297	−0·612
Et	−0·792	−1·408	SMe	−0·786	0·099
i-Pr	−0·780	−1·417	CH_2Ph	−0·693	−1·230
t-Bu	−0·799	−1·445	CH_2SiMe_3	−1·135	−1·447
F	−0·373	0·960	$PO(OEt)_2$	1·005	0·277
Cl	−0·092	0·717	CHO	0·917	−0·005
Br	−0·043	0·704	COPh	0·844	−0·002
I	−0·007	0·446	$CONH_2$	0·549	−0·023
NO_2	1·667	1·803	SO_2CF_3	2·183	2·053
CN	1·266	1·330	SO_2NH_2	1·433	0·752
CF_3	0·731	0·531	SF_5	1·016	1·356
OMe	−1·141	0·076	$SiMe_3$	−0·315	−1·855
NMe_2	−1·991	−0·282	$P(O)Ph_2$	0·890	−0·012
NHAc	−0·847	0·018	OPh	−0·733	0·616
NH_2	−1·719	−0·437	SOMe	0·746	0·948
Ac	0·986	0·145	CH=CHPh	−0·831	−0·620
COOEt	0·771	0·111	N=NPh	0·572	0·041
SO_2Me	1·405	1·495	C≡CPh	−0·072	0·036

case of orthogonal rotation ($a(\sigma_I 1) = 0{\cdot}179$ for $\Gamma > 300$), whereas in the case of oblique rotation $a(\sigma_I 1) = 0{\cdot}0$ for $\Gamma = 0{\cdot}123$, and $\cos \varphi = 0{\cdot}692$ represents the correlation coefficient between these factors (Fig. 2). The results are summarized in Table II. In Fig. 3 the rotated loadings b_{j1} are plotted against b_{j2} in the cartesian coordinate system for each substituent constant. These are divided here roughly into three classes, the first one being represented by σ_I, the second by σ_m (almost the same proportion of inductive and resonance effects), and the third one is formed by the variables predominantly expressing the resonance effect. However, as far as the

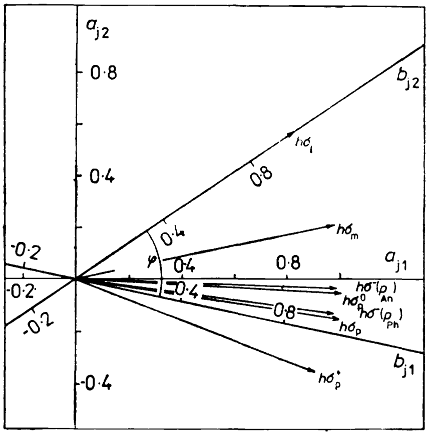

FIG. 2

Representation of the rotation with respect to the original coordinate system along with the communalities of individual variables

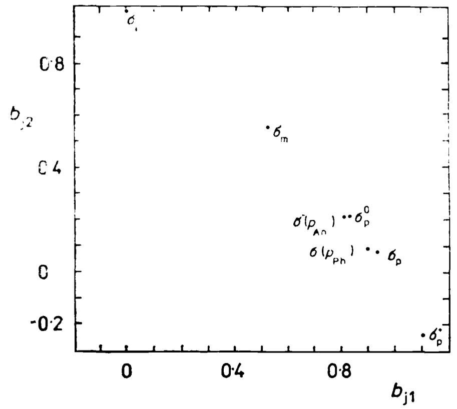

FIG. 3

A plot of b_{j1} against b_{j2} for each variable

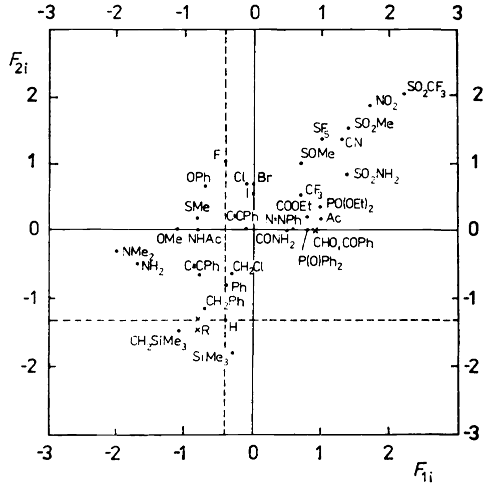

FIG. 4

A plot of the factor 1 (R) against the factor 2 (I) for each substituent last group is concerned, σ_p^+ is somewhat shifted out of it, which obviously has statistical reasons. The coefficient of variance is 322·35 for this variable, whereas for the others it varies within the interval (0·7, 1·9) due to almost zero mean value (0·001 89) and, on the other hand, to the greatest standard deviation of this variable. Table III represents the matrix of factor scores $F = F_{pi}$ (i represents substituents) which is already related to the original variables. From Fig. 4, in which both principal components are plotted against each other for each substituent, it can be seen that the substituents are divided into several classes, which is very close to the conclusions given in ref.[7]. It can be even seen that positions of SMe and NHAc substituents agree with the common chemical practice, which was not the case in the paper mentioned. These substituents form a distinct group of electron donors. The position of OPh is somewhat shifted from that of OMe, and this shift is given, in both its absolute value and direction, by ΔR and ΔI between Me and Ph. The position of Ph shows that this substituent is described by the same model as are alkyl groups, but its incorporation in this group is only formal. Moreover, beside the class of electron acceptors which form another group, we can observe some substituents with unique position. So e.g. $SiMe_3$ has the most negative value of the I component and, therefore, does not directly belong to any group. Due to amphoteric behaviour, C=CPh and C≡CPh also stand outside the groups mentioned. From this standpoint, SOMe is more of an electron acceptor than electron donor.

Addition of a third factor into the rotation would mean a further degree of freedom which, however, cannot be exhausted practically in any way with respect to the composite nature of the constants describing the reactions *II*–*VI*. Obviously, it cannot be stated generally how many factors are necessary for a description of substituent effects, because everything depends on the reaction type considered.

REFERENCES

1. Taft R. W., Lewis I. C.: J. Am. Chem. Soc. *80*, 2436 (1958).
2. Sawada M., Ichihara M., Yukawa Y., Nakachi T., Tsuno Y.: Bull. Chem. Soc. Jpn. *53*, 2055 (1980).
3. Hammett L. P.: *Physical Organic Chemistry*. McGraw-Hill, New York 1940.
4. Swain C. G., Unger S. H., Rosenquist N. R., Swain M. S.: J. Am. Chem. Soc. *105*, 492 (1983).
5. Nieuwdorp G. H. E., de Ligny C. L., van Houwelingen J. C.: J. Chem. Soc., Perkin Trans. 2, *1979*, 537.
6. Haldna U. L., Juga R. J., Tuulmets A. V., Jüriado T. J.: J. Chem. Soc., Perkin Trans. 2, *1987*, 1559.
7. Alunni S., Clementi S., Edlund U., Johnels D., Hellberg S., Sjöström M., Wold S.: Acta Chem. Scand., B *37*, 47 (1983).
8. Exner O.: *Correlation Analysis of Chemical Data*. Plenum Press, New York 1988.
9. Harman H. H.: *Modern Factor Analysis*. Univ. Chicago Press, Chicago 1970.

Translated by J. Panchartek.