Influence of Substituent on Rate and Equilibrium Constants by Means of Spectral-isolation Factor Analysis

Ülo L. Haldna * and Raivo J. Juga
Institute of Chemistry, Academy of Sciences of the Estonian SSR, 15 Akadeemia tee, Tallinn 200108, U.S.S.R.
Ants V. Tuulmets and Toomas J. Jüriado
Department of Chemistry, Tartu State University, Tartu 202400, U.S.S.R.

The key set and spectral isolation methods of factor analysis suggested by Malinowski have been applied to a chemical reactivity data matrix in order to obtain substituent constant scales as orthogonal as possible for the given set of data. The data matrix used consisted of equilibrium and rate constants (24 substituents in the aromatic ring and 10 reaction series). Principal component analysis (p.c.a.) yielded four factors by the indicator function method. The key set method identified the following reaction series: pK values of benzoic acids in water, pK values of phenols in water, pK values of phenols in ethanol–water (23.68 vol.% EtOH), and rate constants for solvolyses of substituted diphenylmethane in ethanol for use as substituent constant scales in correlation analysis of the remaining reaction series. The results of the respective correlation analysis were found to be unsatisfactory: the root mean square error s is 0.23 in logarithmic units. The spectral isolation method enabled us to isolate substituent constant scales whose application to correlation analysis yielded s 0.12, close to the experimental error in the data used, estimated by p.c.a. The first scale is well correlated with the σ° and Hammett σ scales indicating that it reflects the influence of the inductive effect of substituents. The second scale obtained is shown to be related to the resonance effect of substituents because it correlates with σ_R^+ values. It was concluded that the application of the spectral isolation method of factor analysis leads to meaningful substituent constant scales.

The description of the influence of the substituent on rate and equilibrium constants is a classical and important problem of physical organic chemistry. But a sufficiently precise solution to this problem has not yet been obtained. The classical approach to this problem is based on linear free energy relationships (l.f.e.r.).[1] In its modern version the application of l.f.e.r.s is usually accomplished by multiple regression analysis (m.r.a.) of reactivity data. M.r.a. requires that some scales of independent substituent parameters orthogonal among themselves are precisely known and relevant to the reaction studied. Difficulties encountered and the progress in computer technology, on the other hand, have led several investigators to search for an alternative mathematical method for these purposes. As a result attention has been focused on factor analysis (f.a.): during the past decade there has been a remarkable growth in applying f.a. methods to different problems of physical organic chemistry.[2–12]

In most cases the f.a. method employed has been principal component analysis (p.c.a.).[4] P.c.a. is a flexible method so that formal success in its application is guaranteed. But difficulties arise if one tries to find a physical counterpart for the vectors obtained by p.c.a. These vectors are termed abstract because although they have mathematical meaning and are orthogonal to each other, they have no real physical or chemical meaning.[4] Substantial progress in f.a. is due to the target transformation methods suggested by Malinowski.[4] Among these the key set analysis (k.s.a.) and the spectral isolation method (s.i.m.) originally proposed for treating spectroscopic data[13] should probably become quite widely usable. The objective of this paper is to apply k.s.a. and s.i.m. to a classical l.f.e.r. problem in order to obtain new substituent constant scales and compare them with existing ones.

Choice of Data and Methods Used.—We chose 10 reaction series for aromatic compounds because they have been thoroughly studied by traditional m.r.a. methods (see Table 1). The reactions involved are acid–base equilibria and solvolyses. The respective pK and log k values were taken from the compilation of Palm *et al.*[14] The methods used (k.s.a. and s.i.m.) need a completely filled-in data matrix. The pK and log k values for all 24 substituents involved were available for one reaction series only (*i.e.* for the dissociation of phenols in water). For the other nine reaction series we used the correlation analysis equations (1)—(3)[15] for calculation of the missing

$$A = A_0 + \rho^\circ\sigma^\circ + \rho_R^+\sigma_R^+ \quad (1)$$

$$A = A_0 + \rho^\circ\sigma^\circ + \rho_R^-\sigma_R^- \quad (2)$$

$$A = A_0 + \rho^\circ\sigma^\circ + \rho_R^+\sigma_R^+ + \rho_R^-\sigma_R^- \quad (3)$$

points. The values of σ°, σ_R^+, and σ_R^- were taken from ref. 15. The footnote to Table 1 shows which equation has been applied to a given reaction series and how many data points have been involved in the calculations. It should be noted that usually for each series more points were used in the correlation equation than were included in Table 1. The standard error for correlation equations was in the range 0.042—0.086 log units (except for series 1 and 9, see footnote to Table 1). The estimated values for the missing data points have been obtained, as a rule, by interpolation. Therefore the standard error of a given equation can be taken to be the standard error of the estimated points. The correlation coefficient was >0.995 for all series. The number of data points calculated by equation (1)—(3) is 50, *i.e.* 21% of the total number of points in Table 1. The calculated pK and log k values are in parentheses (see Table 1). In each reaction series the data point for the unsubstituted compounds was subtracted from the corresponding value for the substituted compound because in the mathematical model used in s.i.m. the free term is absent.

The data presented in Table 1 were treated by using three f.a. methods: p.c.a., k.s.a., and s.i.m. The use of p.c.a. was

Table 1. Reactivity data used

Substitutents	Reaction series* 1	2	3	4	5	6	7	8	9	10
H	0.800	10.000	11.210	4.210	5.700	6.640	−2.570	4.600	5.180	−4.280
3-Me	1.080	10.100	11.350	4.290	5.880	6.820	−2.760	4.680	5.630	−3.950
4-Me	1.290	10.270	11.490	4.360	5.940	6.830	−2.920	5.080	6.030	−2.910
4-F	(0.510)	9.910	(10.670)	4.150	(5.460)	6.370	−2.230	4.650	(4.150)	−4.030
3-Cl	−0.130	9.120	(10.210)	3.830	5.140	5.900	−1.750	3.520	2.810	−5.920
4-Cl	0.330	9.370	(10.580)	3.980	5.320	6.130	−1.960	4.050	3.830	−4.690
3-Br	−0.160	9.030	11.850	3.810	5.130	5.970	−1.860	3.540	2.840	−5.900
4-Br	(0.170)	9.330	10.570	4.000	5.270	6.100	−1.800	3.860	3.750	−4.790
3-OH	(0.520)	9.330	(10.960)	4.080	5.610	6.740	(−2.320)	4.250	5.100	(−4.690)
4-OH	2.450	9.850	(11.660)	4.500	6.250	7.290	(−3.270)	5.650	6.650	(−0.510)
3-OMe	(0.520)	9.650	(10.960)	3.920	5.590	6.550	−2.400	4.210	4.780	−4.290
4-OMe	2.050	10.200	11.410	4.490	6.030	7.010	−3.220	5.310	6.580	(−0.830)
3-COMe	(−0.190)	9.180	9.990	3.830	5.210	(6.010)	(−1.630)	3.590	3.180	(−6.650)
4-COMe	−0.470	8.050	9.080	3.700	5.100	(5.810)	(−1.360)	2.190	3.510	(−6.260)
3-COOR	0.030	9.100	9.100	(3.840)	5.160	(6.380)	−1.630	3.550	3.090	(−5.740)
4-COOR	(−0.630)	8.500	8.300	(3.740)	5.070	(5.870)	−1.270	2.470	3.490	(−6.140)
3-NO_2	−1.070	8.360	9.350	3.510	4.510	5.440	−0.770	2.520	1.180	−6.780
4-NO_2	−1.730	7.150	7.890	3.440	4.470	5.290	−0.610	1.050	1.390	−7.180
3-NH_2	1.470	9.860	(11.610)	(4.300)	5.790	6.940	−2.780	5.090	6.040	(−3.730)
4-NH_2	3.520	10.440	(12.250)	(5.000)	6.450	7.730	−4.060	6.200	9.120	(0.720)
3-Ph	0.740	9.550	11.100	(4.130)	5.290	6.630	(−2.540)	(4.510)	(5.160)	(−4.330)
4-Ph	0.830	9.450	10.900	(4.260)	5.350	(6.810)	−2.580	4.240	5.360	−3.220
3-Bu^t	0.840	10.050	11.780	4.280	(5.850)	(6.780)	−3.000	4.660	5.820	(−3.920)
4-Bu^t	1.160	10.230	11.550	4.380	(5.920)	(6.850)	−2.990	4.950	5.990	(−2.750)

* The values in parentheses were calculated by the correlation analysis equations (1)—(3).[15] Reaction series (in parentheses are, respectively, the number of the correlation equation, its standard error, and number of points used to obtain the correlation equation): 1, pK_a of pyridine oxides in water, 25 °C (3, 0.172, 26); 2, pK_a of phenols in water, 25 °C; 3, pK_a of phenols in water–ethanol (23.6 mol% EtOH), 25 °C (2, 0.083, 25); 4, pK_a of benzoic acids in water, 25 °C (1, 0.061, 32); 5, pK_a of benzoic acids in water–ethanol (23.6 mol% EtOH), 25 °C (1, 0.042, 40); 6, pK_a of benzoic acids in water–methylcellosolve (48.6 mol% MCS), 25 °C (1, 0.050, 25); 7, log k for the base hydrolysis of ethyl esters of benzoic acids in water–acetone (28.3 mol% acetone), 25 °C (1, 0.059, 30); 8, pK_a of anilines in water, 25 °C (3, 0.086, 31); 9, pK_a of pyridines in water, 25 °C (3, 0.147, 22); 10, log k for the $PhCHClC_6H_4X$ solvolysis reaction in ethanol, 25 °C (1, 0.050, 23).

unavoidable because for k.s.a. and s.i.m. we need the number of factors involved (n.f.). This number was estimated by the indicator function method.[4]

K.s.a. is a target transformation method in f.a. P.c.a. is used to decompose a data matrix, D, into an abstract row matrix, R, and an abstract column matrix, C, such that $D = RC$ within experimental error. In the case of k.s.a. the data matrix is expressed in terms of a key set of data rows, D_k, chosen from the D matrix itself. To accomplish this, a transformation matrix, T, must be found so that $D = RTT^{-1}C = D_k\bar{C}$ where $D_k = RT$ and $\bar{C} = T^{-1}C$. Any arbitrary set of data rows from Table 1 will not necessarily be good for these purposes.[16] It must be stressed that if the key set of rows, D_k, has to consist of reaction series, then instead of D (as given in Table 1) a transformed matrix D^T must be used. Bearing this in mind we can say that k.s.a. is based on the strategy that searches for reaction series unique to each factor. This is achieved mathematically by finding a set of reaction series in the n.f.-dimensional factor space that subtends all other reaction series and whose vector directions are most orthogonal to each other. Caution must be exercised when using this procedure because a key set of series will always be found even when there are actually no such series.[16]

The spectral isolation method (s.i.m.) involves two steps. First, the data matrix D (yet not D^T) is subjected to k.s.a. in order to obtain the D_k matrix consisting of data rows for the unique substituents. In the second step s.i.m. operates with the n.f.-dimensional factor space which subtends all substituent points considered. This space is used to isolate substituent constant values. This is achieved by oblique projection of the substituent points onto each of the n.f. axes passing through the n.f. substituent points previously identified by key set analysis. It is important to note here that the s.i.m. procedure is fully automatic, requiring no preconceived chemical input. But the substituent constant scales are, as a rule, not orthogonal among themselves. However, they are as close to orthogonality as possible for the given set of data. As a result, these scales reproduce the data matrix at a fixed n.f. value with a larger error than the respective orthogonal vectors obtained by p.c.a. Nevertheless, the former should be preferred because they are not devoid of any chemical meaning since they are based on the real reaction series selected by k.s.a.

All programs used were written in FORTRAN IV. Computations were accomplished with an EC-1052 computer (Institute of Cybernetics, Academy of Sciences of the Estonian S.S.R.).

Table 2. Multiregression analysis of 10-(n.f.) reaction series selected by key set analysis as reference scales

N.f.	Key set of reaction series (indexes of columns in Table 1)*	Interval for correlation coefficients†	Root mean errors: Absolute value‡	Relative $100s/\Delta y$
2	4,2	0.950—0.988	0.323	4.3—10.1%
3	4,2,3	0.958—0.988	0.307	4.3—8.2%
4	4,2,3,10	0.963—0.989	0.227	4.1—8.3%

* The remaining 10-(n.f.) reaction series were used as function to be correlated. † For 10-(n.f.) regressions. ‡ Root mean square values for the matrix of errors 24·[10-(n.f.)]

Results and Discussion

Applying p.c.a. to the data matrix (Table 1) we obtained the number of significant factors n.f. 4 (by the indicator function

Table 3. Substituent scales obtained by the spectral isolation method at different n.f. values*

	R(1)			*R*(2)			*R*(3)		*R*(4)
Substituent	N.f. 2	N.f. 3	N.f. 4	N.f. 2	N.f. 3	N.f. 4	N.f. 3	N.f. 4	N.f. 4
H	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3-Me	−0.192	−0.113	−0.102	0.125	0.166	0.112	−0.319	−0.301	−0.194
4-Me	−0.477	−0.477	−0.457	0.687	0.344	0.300	0.027	0.435	−0.285
4-F	0.271	0.000	0.000	0.534	0.000	0.000	1.122	1.987	0.000
3-Cl	1.068	0.795	0.735	0.000	−0.127	0.330	1.104	0.909	0.770
4-Cl	0.582	0.347	0.306	0.534	0.178	0.589	0.953	1.003	0.506
3-Br	0.858	0.238	0.000	−0.958	−1.575	0.000	2.574	0.000	3.384
4-Br	0.661	0.426	0.380	0.538	0.208	0.709	0.949	0.835	0.625
3-OH	0.236	0.351	0.343	0.267	0.498	0.757	−0.492	−0.972	0.080
4-OH	−1.029	−1.205	−1.193	2.667	1.311	1.688	0.803	1.752	−0.247
3-OMe	0.277	0.175	0.155	0.424	0.306	0.596	0.200	0.031	0.237
4-OMe	−0.888	−0.987	−0.948	2.463	1.358	1.523	0.465	1.626	−0.598
3-COMe	1.112	1.096	1.091	−0.494	0.000	0.000	0.000	0.000	0.000
4-COMe	1.467	1.795	1.827	1.429	2.231	2.817	−1.474	−1.932	−0.495
3-COOR	1.116	1.163	1.246	0.860	1.141	0.795	−0.264	1.215	−1.250
4-COOR	1.497	1.946	2.080	1.669	2.649	2.420	−1.990	−0.819	−2.004
3-NO_2	1.909	1.444	1.348	0.506	0.212	1.155	1.848	1.513	1.304
4-NO_2	2.465	2.561	2.558	2.117	2.683	3.751	−0.552	−0.913	−0.048
3-NH_2	−0.367	−0.284	−0.286	0.204	0.174	0.232	−0.333	−0.548	−0.026
4-NH_2	−1.913	−1.624	−1.528	2.969	2.086	2.033	−1.109	−0.086	−1.395
3-Ph	0.105	0.117	0.102	0.314	0.298	0.555	−0.059	−0.446	0.204
4-Ph	0.000	0.000	0.000	1.480	1.120	1.579	0.000	0.000	0.000
3-Bu^t	−0.262	−0.217	−0.253	−0.125	−0.110	0.146	−0.173	−0.921	0.458
4-Bu^t	−0.468	−0.502	−0.498	0.817	0.383	0.495	0.164	0.388	−0.075

* Normalized values, $[R(M)_{sub.} - R(M)_H]/s$, where s is the root mean deviation for the column considered.

values[4]). This n.f. value allows us to reproduce the initial data matrix with a root mean square error s of 0.092 (logarithmic units) that may be interpreted as a mean experimental error of the data in Table 1. The first, second, third, and fourth eigenvalues account for 89, 8, 2, and 1% of their sum, respectively. Consequently, p.c.a. indicates that in the data analysed one factor strongly dominates. However, it should be noted that in p.c.a. the maximum possible amount of variation is assigned to the first factor, then the maximum amount of the remainder is assigned to the second, *etc.* As noted above, there is no reason to search for a physical meaning of row and column vectors obtained by p.c.a. Nevertheless, the first p.c.a. row vector $R(1)$ shows good correlation with Hammett σ values[1] (r 0.987), the relative root mean square error $100s/\Delta y$ being 4.0%. This means that p.c.a. has led us to the well known classical substituent constant scale relevant to the compounds studied.

The application of k.s.a. to the data matrix (Table 1) allowed us to identify a set of reaction series most suitable for the description of data. The identified series were used as substituent constant scales in m.r.a. This set consists of the following reactions series: 4, 2, 3, and 10 (in the sequence of their importance). The results of m.r.a. obtained with the key set of reaction series are shown in Table 2. It should be emphasized that k.s.a. indicated as the first, most important series for the description of the data, reaction series 4, *i.e.* the pK values of substituted benzoic acids in water at 25 °C. This is exactly the reaction series used by Hammett for establishing the σ scale.[1]

The next step was to employ the s.i.m. procedure in order to obtain the values of substituent constants based on the following key substituents identified by k.s.a. (n.f. 4): 3-COMe, 4-Ph, 4-F, and 4-Br. By means of s.i.m. the scales of substituent constants presented in Table 3 were isolated. It is of interest to see how much the selected factor number (n.f.) influences the scales isolated. For these purposes the s.i.m. procedure was also carried out with n.f. 2 and 3. In the former case (n.f. 2) the preceding k.s.a. identified two substituents (3-Cl and 4-Ph) and in the latter case (n.f. 3), three substituents (3-COMe, 4-Ph, and 4-F) as a base for the isolation of substituent constant scales (also presented in Table 3). A comparison of scales $R(M)$, M = 1—4, in Table 3 shows that $R(1)$, $R(2)$, and $R(3)$ depend on the n.f. value selected. The dependence is not large for $R(1)$ but this is not the case for $R(2)$ and $R(3)$. That could be anticipated; the major effect will be isolated more or less uniformly regardless of the n.f. value used. The scales corresponding to minor effects depend strongly on the n.f. value.

Applying the substituent constant scales from Table 3 to the description of data (Table 1) by m.r.a. the following root mean square errors were obtained: n.f. 4, s 0.119; n.f. 3, s 0.178; n.f. 2, s 0.256. As estimated by p.c.a. (n.f. 4), the experimental error of data in Table 1 is probably 0.092 which is close to a value of 0.119 obtained by s.i.m. (n.f. 4). Consequently, s.i.m. affords substituent constant scales that are rather well applicable to describing the data used.

It is of special interest to make a search for a physical meaning of substituent scales isolated by s.i.m. First we correlated the $R(1)$ values at n.f. 2—4 (Table 3) with the Hammett σ values (the latter were used as function values). These correlations yielded n.f. 2, r 0.987, $100s/\Delta y$ 3.9%; n.f. 3, r 0.947, $100s/\Delta y$ 7.9%; n.f. 4, r 0.928, $100s/\Delta y$ 9.2% With increasing n.f. values these correlations were getting worse showing that the Hammett σ values have a composite nature, *i.e.* if more effects are elucidated, the residual $R(1)$ becomes less close to the Hammett σ values. Using the σ° values[1] instead of Hammett σ values very similar correlation parameters were obtained: n.f. 2, r 0.979, $100s/\Delta y$ 4.4%; n.f. 3, r 0.929, $100s/\Delta y$ 8.0%; n.f. 4, r 0.908, $100s/\Delta y$ 9.0%. Therefore the conclusion drawn about the Hammett σ scale is also valid for the σ° scale. The supposed exclusion of resonance effect from the σ° scale does not much influence the relationship between $R(1)$ and σ° when compared with that between $R(1)$ and Hammett σ. The above correlation parameters indicate that to a first approximation the $R(1)$ scale is responsible for the inductive effect of substituents. The next logical step should be

to prove whether the $R(2)$ scale is responsible for the resonance effect. Having this in mind we correlated $R(2)$ values at n.f. 2—4 with σ_R^+ values from Shorter's book.[1,*] The following correlation yielded n.f. 2, r 0.952, $100s/\Delta y$ 12.9%; n.f. 3, r 0.944, $100s/\Delta y$ 13.9%; n.f. 4, r 0.892, $100s/\Delta y$ 19.1%. These results indicate that $R(2)$ may be related to the resonance effect. This satisfactory correlation is astonishing because the $R(2)$ scale includes both electron-donating and -withdrawing substituents.

In terms of l.f.e.r. $\rho_R^+ \neq \rho_R^-$, so two separate scales for describing the resonance effect should be involved.[17] The $R(2)$ values for *meta*-substituents, contrary to the requirements of l.f.e.r., vary greatly and differ considerably from zero. However, the $R(2)$ values for all the substituents studied are in the order $R(2)_{para} > R(2)_{meta}$, demonstrating the occurrence of resonance at the *para*-position.

It should be emphasized that f.a. makes use of formal methods absolutely independent of l.f.e.r. So there is no point in insisting on a correspondence between the scales obtained by s.i.m. and l.f.e.r. Nevertheless, it appeared that s.i.m. produced quantitative scales closely related to the structure effects (I and R effects) introduced long before the discovery of l.f.e.r. operating with the same effects.

The nature of $R(3)$ and $R(4)$ scales is not clear. Their numerical values are rather uncertain as well as depending very much on the n.f. value chosen. The role of $R(3)$ and $R(4)$ scales in reproducing the initial data matrix is rather insignificant as they account for only 2 and 1% of the total variability, respectively.

Acknowledgements

We are indebted to Professor E. R. Malinowski for kindly supplying the listings of k.s.a. and s.i.m. programs, and to Professor V. A. Palm for valuable discussions.

* The correlation σ_R *versus* $R(2)$ was not carried out because only three — R substituents are included in Table 1 (4-COMe, 4-COOR, and 4-NO_2).

References

1 J. Shorter, 'Correlation Analysis of Organic Reactivity,' Research Studies Press, Chichester, 1982, p. 19.
2 M. C. Spanjer, C. L. de Ligny, H. C. Houwelingen, and J. M. Weesie, *J. Chem. Soc., Perkin Trans. 2*, 1985, 1401.
3 D. Johnels, U. Edlund, and S. Wold, *J. Chem. Soc., Perkin Trans. 2*, 1985, 1339.
4 E. R. Malinowski and D. G. Howery, 'Factor Analysis in Chemistry,' Wiley, New York—Chichester, 1980, p. 50—99.
5 R. I. Zalewski and Z. Geltz, *J. Chem. Soc., Perkin Trans. 2*, 1983, 1885.
6 P. H. Weiner, *J. Am. Chem. Soc.*, 1973, **95**, 5845.
7 J. T. Edward, and S. C. Wong, *J. Am. Chem. Soc.*, 1977, **99**, 4229.
8 S. Alunni, S. Clementi, U. Edlung, D. Johnels, S. Hellberg, M. Sjöström, and S. Wold, *Acta Chem. Scand.*, 1973, **B37**, 47.
9 M. Ludwig, O. Pytela, K. Kalfus, and M. Večera, *Collect. Czech. Chem. Commun.*, 1984, **49**, 1182.
10 M. Sjöström and S. Wold, *J. Chem. Soc., Perkin Trans. 2*, 1981, 105.
11 J. T. Edward, M. Sjöström, and S. Wold, *Can. J. Chem.*, 1982, **59**, 2350.
12 Ü. L. Haldna and A. Murshak, *Comput. Chem.*, 1984, **8**, 201.
13 E. R. Malinowski, *Anal. Chim. Acta*, 1982, **134**, 129.
14 'Tables of Rate and Equilibrium Constants of Heterolytic Organic Reactions,' VINITI, Moscow, 1975—1978, vol. 1—5(I).
15 'Tables of Rate and Equilibrium Constants of Heterolytic Organic Reactions,' VINITI, Moscow, 1979, vol. 5(II).
16 E. R. Malinoswki, R. A. Cox, and Ü. L. Haldna, *Anal. Chem.*, 1984, **56**, 778.
17 V. A. Palm, 'Grundlagen der quantitativen Theorie organischer Reaktionen,' Akademie Verlag, Berlin, 1971.

Received 20*th June* 1986; *Paper* 6/1249