PRINCIPAL COMPONENT REGRESSION WITH CHEMICAL SHIFT INCREMENTS. I. *p*-DISUBSTITUTED BENZENES AND 2-NAPHTHYL DERIVATIVES*

Miroslav HOLIK

Department of Theoretical and Physical Chemistry, Masaryk University, 611 37 Brno, Czech Republic; e-mail: holik@chemi.muni.cz

Received November 6, 1995
Accepted January 19, 1996

Prediction of ^{13}C substituent chemical shifts in 14 series of *para*-disubstituted benzenes and in 2-substituted naphthalenes was based on principal component regression with chemical shift increments for the *ipso*, *ortho*, *meta* and *para* position of monosubstituted benzenes. Mean-centered matrix of shift increments was submitted to singular value decomposition and principal component regression was used for the projection of the investigated substituent chemical shifts and for the calculation of regression coefficients. Residual standard deviation between experimental and fitted values in *para*-disubstituted benzenes was in agreement with absolute values of "an electron demand" of substituents. Inspection of the regression parameters revealed that for the prediction of chemical shifts in 2-substituted naphthalenes the combination of chemical shift increments was better than the use of single increments. It is believed that the presented procedure is general and can be used for other aromatic or heteroaromatic systems.

Key words: Linear transformation; Multivariable regression analysis; Principal component regression; Chemical shift increments; ^{13}C NMR substituent chemical shifts.

For the assignment of NMR signals of an investigated molecule various experimental procedures have been used; a very popular among them is to compare the chemical shifts in the molecule under study with those of properly selected model compounds. For instance, chemical shifts of disubstituted benzenes, can be regarded as composed from chemical shift increments from monosubstituted derivatives. Generally, this effect is not exactly additive, but rather proporcional[1,2]. Thus, in the series of disubstituted benzenes with one variable (X) and one fixed (Y) substituent the regression (Eq. (*1*)) of the substituent chemical shifts, SCS, (i.e., $d = \delta(X \neq H) - \delta(X = H)$) with chemical shift increments**, CSI, represents very useful method of the chemical shift assignment[2].

* Presented in part at the *VIth International Conference on the Correlation Analysis in Chemistry, Prague, September 1994.*

**Substituent chemical shifts of monosubstituted benzenes[3], $\mathbf{z}_j$, are called here chemical shift increments

$$\mathbf{d}_j = b_j \cdot \mathbf{z}_j \qquad (1)$$

This procedure is sometimes called "the Lynch plot" according to the author[1] who introduced it into the ^{13}C NMR spectroscopy.

For ^{13}C SCS of C-1, C-2, C-3, and C-4 in *para*-disubstituted benzenes and corresponding CSI, e.i., $\mathbf{z}_i$, $\mathbf{z}_o$, $\mathbf{z}_m$, and $\mathbf{z}_p$, these four equations can be summarized into the matrix equation (*2*), where *m* is the number of compounds in the series (the number of variable substituents).

$$\underset{m}{\left[\mathbf{d}_1\ \mathbf{d}_2\ \mathbf{d}_3\ \mathbf{d}_4\right]} = \underset{m}{\left[\mathbf{z}_i\ \mathbf{z}_o\ \mathbf{z}_m\ \mathbf{z}_p\right]} \cdot \begin{bmatrix} b_{i1} & 0 & 0 & 0 \\ 0 & b_{o2} & 0 & 0 \\ 0 & 0 & b_{m3} & 0 \\ 0 & 0 & 0 & b_{p4} \end{bmatrix} \qquad (2a)$$

$$\mathbf{D}_{m,4} = \mathbf{Z}_{m,4} \cdot \mathbf{B}_{4,4} \qquad (2b)$$

We can easy imagine that the variation in, e.g., $\mathbf{d}_1$ can be explained not only by the variation in $\mathbf{z}_i$ but rather by some proper combination of all four CSI, i.e., $\mathbf{z}_i$, $\mathbf{z}_o$, $\mathbf{z}_m$, and $\mathbf{z}_p$, Eq.(*3*).

$$\mathbf{d}_1 = b_{i1} \cdot \mathbf{z}_i + b_{o1} \cdot \mathbf{z}_o + b_{m1} \cdot \mathbf{z}_m + b_{p1} \cdot \mathbf{z}_p \qquad (3)$$

In this case, the matrix of regression coefficients, **B**, will be no more diagonal but can possess some off-diagonal elements not equal to zero, Eq. (*4*). Generally, the matrix **D** can differ from the matrix **Z** in the number of columns, and this regression can be now used also for other aromatic system with more or less ^{13}C chemical shifts.

$$\underset{m}{\left[\mathbf{d}_1\ \mathbf{d}_2\ \ldots\ \mathbf{d}_q\right]} = \underset{m}{\left[\mathbf{z}_i\ \mathbf{z}_o\ \mathbf{z}_m\ \mathbf{z}_p\right]} \cdot \begin{bmatrix} b_{i1} & b_{i2} & \ldots & b_{iq} \\ b_{o1} & b_{o2} & \ldots & b_{oq} \\ b_{m1} & b_{m2} & \ldots & b_{mq} \\ b_{p1} & b_{p2} & \ldots & b_{pq} \end{bmatrix} \qquad (4a)$$

(CSI) in order to emphasize their use as substituent constants. In the following text, boldface upper case letters are used for matrices, boldface lower case letters for vectors, primes for the transposition of matrices or vectors, and lower case italics for scalars.

$$\mathbf{D}_{m,q} = \mathbf{Z}_{m,4} \cdot \mathbf{B}_{4,q} \qquad (4b)$$

This equation can be regarded as a linear transformation of the substituent effect from the coordinate systen of monosubstituted benzenes to the other aromatic system; the **B** matrix being so called transformation matrix. In order to get unbiased regression parameters in the matrix **B** it is necessary to take both substituent chemical shift vectors ($\mathbf{d}_k$) and chemical shift increments vectors ($\mathbf{z}_j$) mean centered, therefore, no constant (location) paramater is used in the presented equations. For the explanation of one matrix (e.g.,**Y**) with another one (e.g., **X**) the partial least-squares (PLS) method[4,5] was used. Although well suited for calibration purposes, for the explanation and prediction of SCS of any aromatic compounds by CSI constants the abovementioned linear transformation seems to be more proper.

For the evaluation of suitability of model matrix **Z** to explain ^{13}C SCS of some aromatic system (matrix **D**) it is necessary to project the data matrix **D** into the data $\hat{\mathbf{D}}$ calculated by least square method and to check the difference between **D** and $\hat{\mathbf{D}}$ matrices.

Multiparameter linear regression is generally written in matrix form as in the Eq. (*5*).

$$\mathbf{D}_{mn} = \mathbf{Z}_{mp} \cdot \mathbf{B}_{pn} + \mathbf{E}_{mn} \ , \qquad (5)$$

where *m* is the number of measurements (the number of compounds in the series), *p* is the number of independent variables (number of substituent constants z_j), and *n* is the number of dependent variables (i.e., number of SCS investigated). The **E** is the matrix of residuals, the elements of which are minimized by the least-squares method. This process gives the optimal estimation of parameters *b* in matrix **B** according to the Eq. (*6*) and fitted data $\hat{\mathbf{D}}$ are calculated as in Eq. (*7*).

$$\mathbf{B} = (\mathbf{Z}' \cdot \mathbf{Z})^{-1} \cdot \mathbf{Z}' \cdot \mathbf{D} \qquad (6)$$

$$\hat{\mathbf{D}} = \mathbf{Z} \cdot (\mathbf{Z}' \cdot \mathbf{Z})^{-1} \cdot \mathbf{Z}' \cdot \boldsymbol{D} = \mathbf{H} \cdot \mathbf{D} \qquad (7)$$

Each data matrix can be partitioned into three characteristic matrices by Singular Value Decomposition[6] (SVD); for the **Z** matrix the result is represented by Eq. (*8*).

$$\mathbf{Z}_{mp} = \mathbf{U}_{mp} \cdot \mathbf{S}_{pp} \cdot (\mathbf{V}_{pp})' \ , \qquad (8)$$

where **U** and **V** are orthonormal matrices, i.e., $\mathbf{V} \,.\, \mathbf{V}' = \mathbf{V}' \,.\, \mathbf{V} = \mathbf{U}' \,.\, \mathbf{U} = \mathbf{I}_p$ and **S** is the diagonal matrix of singular values ordered in such a way that $s_{11} > s_{22} > \ldots > s_{pp}$. Pseudoinverse of **Z** is then given by Eq. (*9*).

$$(\mathbf{Z}_{mp})^{+} = \boldsymbol{V}_{pp} \cdot (\boldsymbol{S}_{pp})^{-1} \cdot (\boldsymbol{U}_{mp})' \qquad (9)$$

In the case of correlation of the column vectors in matrix **Z**, the very small *s* values are omitted and the size of the **S** matrix is decreased from *p* to *q* before the matrix inverse. The corresponding decrease in the size is performed also on the matrices **U** and **V** and the pseudoinverse is then calculated according to Eq. (*10*).

$$(\mathbf{Z}^{\neq}_{mp})^{+} = \mathbf{V}_{pq} \cdot (\mathbf{S}_{qq})^{-1} \cdot (\mathbf{U}_{mq})' \qquad q < p \qquad (10)$$

After substituting from Eqs (*8*) and (*9*) or Eq. (*10*) into the Eq. (*7*) we get Eq. (*11*) which is an essential of the so called principal component (PC) regression[6].

$$\hat{\mathbf{D}} = \mathbf{U} \cdot \mathbf{U}' \cdot \mathbf{D} \qquad (11)$$

Thus, the theoretical values $\hat{\mathbf{D}}$ can be calculated by projection of the experimental values **D** with the help of the projection matrix constructed from the orthogonal vectors obtained by SVD of the model matrix **Z**. Either is the matrix $\mathbf{U}_{mp}$ used as the whole or reduced in size *p* to *q* as in Eq. (*10*).

Success in the explanation of the data **D** by the model matrix **Z** can be assessed with the sum of squares of residuals (*ssr*),

$$ssr = (\mathbf{D} - \mathbf{U} \cdot \mathbf{U}' \cdot \mathbf{D})' \cdot (\mathbf{D} - \mathbf{U} \cdot \mathbf{U}' \cdot \mathbf{D}) \qquad (12)$$

or by the residual standard deviation of the whole matrix (*RSD*) or of the individual columns *j* (rsd_j); see Eq. (*13*) or (*15*), respectively.

$$RSD = [ssr/(n(m - q))]^{1/2} \qquad q \leq p \qquad (13)$$

$$rsd_j = [\sum (d_j - \hat{d}_j)^2/(m - q)]^{1/2} \quad q \le p \tag{14}$$

Pseudoinverse can also be used for the estimation of parameters *b*, Eq. (*15*).

$$\mathbf{B}^{\neq}_{pn} = (\mathbf{Z}^{\neq}_{mp})^{+} \cdot \mathbf{D}_{mn} \tag{15}$$

If the explanation of data **D** by the model matrix **Z** is successful, the parameter matrix $\mathbf{B}^{\neq}$ obtained from data matrices **D** and **Z** (so called training set), can be used for the prediction of data $\mathbf{D}^{*}$ from the new data $\mathbf{Z}^{*}$ not included in **Z** (so called test set), e.g., Eq.(*16*).

$$\mathbf{D}^{*}_{1n} = \mathbf{Z}^{*}_{1p} \cdot \mathbf{B}^{\neq}_{pn} \tag{16}$$

If all PCs are used for the calculation of pseudoinverse (Eq. (*9*)) the parameter matrix $\mathbf{B}^{\neq}$ is equivalent to the matrix **B** (Eq. (*6*)).

CALCULATIONS AND DATA

All calculations were carried out with double precission on a PC with 486 processor. Standard subroutines were used for linear regression and singular value decomposition. For the PLS calculations the published algorithm was applied[5].

p-Disubstituted Benzenes

The ^{13}C NMR substituent chemical shifts (SCS) calculated for C-1, C-2, C-3, and C-4 atoms of 1,4-disubstituted benzenes, from data presented in Tables I–III of ref.[7a] were used as data matrix $\mathbf{D}_{14,4}$. Both variable (X on C-1) and fixed (Y on C-4) substituents comprised NMe_2, NH_2, OMe, F, Cl, Br, Me, H, CF_3,CN, COOEt, COMe, NO_2, and CHO groups or atoms forming 14 series each with 14 compounds (**D1**–**D14**). For each substituent Y, these data were mean-centered to corresponding matrices **DC1**–**DC14**. As the independent variables model, the matrix $\mathbf{Z}_{14,4}$ of Ewing chemical shift increments[3], z_j, was used rather than the matrix **D8** with Y = H; the **C8** and **Z** matrices differ slightly due to not exactly same measurement conditions. Column centered matrix **Z** gave matrix **ZC** and corresponding standardized matrix was **ZS**.

For the purpose of the prediction the matrices **D1** (Y = NMe_2) and **Z** were separated to the training sets, **Dm** and **Zm** containing minimal set of substituents according to Taft[8], i.e., X = NMe_2, OMe, F, Cl, H, Me, COOMe, and NO_2 and to the test sets, **Dt** and **Zt** containing remaining substituents. All matrices were also column-mean-centered to **DmC**, **ZmC**, **DtC**, and **ZtC**, respectively. In order to check the effect of standardization, sometimes recommended in the literature (e.g., ref.[4b]), all matrices were standardized to **DmS**, **ZmS**, **DtS**, and **ZtS**, respectively.

2-Naphthyl Derivatives

The ^{13}C SCS data of 21 2-substituted naphthalenes were taken from the literature cited in ref.[4b] forming the data matrix $\mathbf{D}_{21,10}$. The substituents were arranged in the order: H, Me, Et, *t*-Bu, CH_2Br, F, Cl, Br, I, COMe, CHO, COOMe, COOH, CN, NO_2, NH_2, NMe_2, OMe, OH, NHCOMe, and OCOMe. Matrix of Ewing increments z_j, $\mathbf{Z}_{21,4}$, was constructed for the projection and both matrices were column-mean-centered to **DC** and **ZC** and standardized to **DS** and **ZS** matrices, respectively.

Data Pretreatment And Number Of Principal Component Used

Standardization of the data to the unit variance is not recommendable since it gives unnaturally high weight to the ^{13}C chemical shift variations in the position *meta* to the variable substituent X. This fact can be inferred from the loadings, **V**, obtained after the SVD of corresponding data matrices **ZC** and **ZS**, i.e., column-mean-centered and scaled to the unit variance, respectively (see Table I); for the sake of simplicity the loadings in the upper part of the Table I were multiplied by ten and rounded to whole numbers.

From the left part of the Table I (**ZC)** it is clear that the 1st principal component (PC) explains mainly variations in the *ipso* SCS (value 9 in the first column), the second PC is responsible for variations in *para* position (value 8) and the third reflects the variations at the *ortho*-C atom (value 8). These three vectors can be orthogonally rotated by the angles –25°, 25°, and 15°, respectively, to put value 10 into the corresponding position and nearly zero into the others. Necessity of the 4th PC for the explanation of variations in the *meta* position is clear on the first sight (value 10 in the fourth column). Real contribution of the 4th PC to SCS values depends not only on the loading but also on

Table I

Loadings from the SVD of centered ($\mathbf{ZC}_{14,4}$) and standardized ($\mathbf{ZS}_{14,4}$) chemical shift increments data for the projection of SCS of 1,4-disubstituted benzenes

Position	ZC				ZS			
	1	2	3	4	1	2	3	4
				V . 10				
1	9	4	2	0	5	–3	6	5
2	–4	4	8	0	–6	2	–1	8
3	0	0	0	10	3	9	2	0
4	–2	8	–5	0	–5	0	8	–3
				S . V				
1	49.8	7.7	2.0	0.0	3.1	–1.0	1.4	0.6
2	–23.5	7.0	6.6	0.0	–3.4	0.8	–0.1	0.8
3	0.8	–0.5	0.2	2.0	2.1	2.9	0.5	0.0
4	–14.0	15.5	–4.0	0.1	–3.0	0.1	1.9	–0.4

the magnitude of the corresponding singular value. The lower part of the Table I shows the results of the multiplication of the **S** and **V** matrices: the value 2.0 confirms that the contribution of the 4th PC is substantial. Therefore, for the most calculations all four PC were used.

The picture is not so clear in the case of standardized data **ZS**, i.e., column-mean-centered and scaled to the unit variance, which is believed (ref.[4b]) not to hide systematic variations in the positions with the small initial variance like *meta*-SCS. However, these small variations are unnaturally exaggerated (see value 9 for the second PC). Moreover, loadings for the first PC show extensive mixing and exclusion of the 4th PC from the calculation as it was done in ref.[4b] could seriously affect the reproducibility of the chemical shifts in the *ipso*, *ortho* and *para* positions; see columns 4 in the right-low part of the Table I.

RESULTS AND DISCUSSION

Projection

The experimental ^{13}C SCS of 1,4-disubstituted benzenes, $\mathbf{DC}_{14,4}$, were projected with the help of Eq. (*11*) and the quality of this projection was tested by the residual standard deviation, *RSD*, Eq. (*13*). This assessment shows, how well the ^{13}C SCS in *p*-disubstituted benzenes could be explained by the chemical shift increments, $\mathbf{z}_j$, calculated from the chemical shifts of monosubstituted benzenes[3]. In this particular case, the quality of the projection can be also regarded as a measure of the additivity of chemical shift increments. The *RSD* values in Table II show that halogens and alkyl as Y substituent suit better the additivity scheme than the electron acceptor or electron donor groups. This finding is in good agreement with so called "electron demand", η, of Y substituent which was calculated by "Dual Substituent Parameter–Nonlinear Regresion" DSP–NLR method[7a] and confirmed by "The Second Order Regression Analysis" approach[7b] with Taft σ_I and σ_R^0 constants using the ^{13}C SCS of *para*-C atom to the variable substituents X. The electron demand of substituent Y has positive value for all π donor Y groups, essentially zero for Y = H, Me and halogens, and negative value for all π acceptor Y groups. An inspection of the Table II reveals that *RSD* values correspond very well to the absolute values of the electron demands of Y groups: correlation of η with (sign(η)**RSD*) gives the correlation coefficient $r = 0.9813$. The dissimilarity of ^{13}C SCS in mono and disubstituted benzenes caused by electron demand of Y substituent can, in this way, be classified without help of any empirical σ constants.

Naturally, there are some differences between *RSD* and η absolute values. They are caused mainly by the fact, that the PC regression approach utilizes an information hidden in all the ^{13}C chemical shifts of the molecule while for the calculation of electron demand[7] only ^{13}C SCS of C-4 were used. E.g., the electron demands of NMe_2 and NH_2 groups differ significantly, while the similarity of both groups revealed by PC regression (Eq. (*11*)) is comparable (*RSD* = 0.497 and 0.576, respectively). The same Table II shows also the residual standard deviations (rsd_j) for individual positions in the benzene ring calculated according to Eq. (*14*). It can be seen that the main difference

between NMe_2 and NH_2 series consists in rsd_4 values for SCS of C-4 atom (*ipso* to the Y substituent); the enhanced dissimilarity of C-4 SCS in NH_2 series, which can be caused by higher electronegativity of NH_2 group with respect to NMe_2, is responsible for higher electron demand value. Probably, the same effect is the reason for high rsd_4 and η values in the methyl series. For π acceptor Y group the electron demand is saturated mainly by the resonance effect, since the largest dissimilarity is observed in *ipso* (*para* to Y) position. On the contrary, the electron donor groups NMe_2 and NH_2 affect mainly the chemical shifts in the *ortho* position (*meta* to Y) probably by π-polarization mechanism. In this sense, the PC regression analysis can be of great help in investigation of changes of electron density in the molecules caused by substitution.

As the next, the PC projection was applied on the ^{13}C SCS data of 21 2-substituted naphthalenes which were analyzed before by Johnels at al. by PLS method[4]. Standard deviations (*sd*) in Table III, column 1, confirm that we use exactly same data as the

TABLE II
Assessment of the projection of SCS for *p*-disubstituted benzenes from the CSI data[a]

No.	Y (at C-4)	rsd_1	rsd_2	rsd_3	rsd_4	*RSD*	η[b]
1	NMe_2	0.557	0.742	0.223	0.283	0.497	0.257
2	NH_2	0.612	0.770	0.247	0.548	0.576	0.516
3	OMe	0.513	0.489	0.148	0.454	0.427	0.432
4	F	0.398	0.281	0.124	0.298	0.292	0.098
5	Cl	0.178	0.152	0.096	0.190	0.158	0.028
6	Br	0.229	0.221	0.122	0.217	0.202	0.019
7	Me	0.263	0.171	0.125	0.384	0.256	0.307
8	H	0.177	0.104	0.123	0.142	0.139	0.058
9	CF_3	0.544	0.416	0.143	0.468	0.421	–0.422
10	CN	0.845	0.636	0.175	0.816	0.674	–0.591
11	COOEt	0.630	0.442	0.177	0.598	0.495	–0.480
12	COMe	0.728	0.480	0.170	0.543	0.520	–0.494
13	NO_2	0.928	0.640	0.246	0.610	0.653	–0.712
14	CHO	0.856	0.596	0.218	0.604	0.612	–0.603
–[c]	rand (1)	1.075	1.060	0.718	1.435	1.102	–
–[c]	rand (2)	0.877	1.061	0.936	0.722	0.907	–
–[c]	rand (3)	1.215	0.894	1.365	0.927	1.118	–

[a] Projection, Eq. (*11*); RSD, Eq. (*13*); rsd_j, Eq. (*14*). [b] Electron demand values from ref.[7]. [c] Data in SCS matrix substituted by normally distributed random numbers $N(0,1)$.

cited authors (cf. Table I in ref.[4a]). Individual residual standard deviations, *rsd*, were assessed against the total residual standard deviation, *RSD* by Fisher *F* test: the null hypothesis states that variance of residuals (i.e., differences between calculated and found SCS) for an individual position in molecule (rsd_j) does not significantly differ from that one for all positions (*RSD*).

The significant differences between experimental data (SCS) and data reproduced from CSI ($\mathbf{z}_j$) were found only for the both *ortho* positions, i.e. C-1 and C-3 (cf. column 2 in Table III). This result is in agreement with former finding that tables of *ortho*-SCS values for the monosubstituted benzenes cannot be used to predict ^{13}C chemical shifts in 2-naphthalenes[4a,9]. This is attributed to the fact that C-2,C-1 and C-2,C-3 π bond orders in naphthalene[9] are distinctly non-equivalent while the corresponding π bond order in benzene is near to their mean value.

TABLE III
Assessment of the projection of SCS for 2-substituted naphthalenes from the CSI data[a]

Method		PCR			PLS		
Data matrix		**DC**	**DC**	**DS**	**DS**	**DS**	**DS**[b]
No. of PC		4	3	3	3	3	3[b]
RSD		0.49	0.55	0.43	0.42	0.48	1.45
Position	*sd*[c]	*rsd*	*rsd*	*rsd*	*rsd*	*rsd*[d]	*rsd*[b]
1	9.74	1.04[e]	1.04[e]	0.17	0.17	0.20	1.9[e]
2	16.19	0.31	0.32	0.17	0.19	0.22	3.5[e]
3	4.80	0.84[e]	0.82[e]	0.42	0.40	0.46	2.1[e]
4	0.80	0.32	0.71[e]	0.42	0.40	0.45	0.4
5	0.19	0.17	0.17	0.89[e]	0.89[e]	1.03[e]	0.2
6	2.12	0.26	0.34	0.16	0.14	0.16	0.3
7	0.57	0.28	0.39	0.60[e]	0.57[e]	0.66[e]	0.4
8	1.30	0.24	0.32	0.28	0.25	0.29	0.4
9	0.96	0.32	0.34	0.38	0.36	0.42	0.4
10	2.73	0.21	0.39	0.10	0.09	0.10	0.3

[a] Projection, Eq. (*11*); RSD, Eq. (*13*); rsd_j (j = 1, 2...10), Eq. (*14*). [b] Data from ref.[4a]. [c] Standard deviation of 21 SCSs from their mean value. [d] Prediction, Eq. (*16*), predicted data not included in training set. [e] Null hypothesis $\mathbf{H}_0$: $RSD = rsd_j$ is rejected according to *F* test on the 0.05 confidence level.

If only three PCs were taken for the projection then the significant difference between experimental and reproduced data were found also for the SCS on the C-4 atom (column 3 in Table III). This is not surprising since the fourth PC is responsible for the projection of the SCS in *meta* position (see above).

Totally different results were obtained when the data were standardized before calculations (column 4 in Table III). Significant disagreement was found for the experimental and calculated SCS not only on the C-5 but also on C-7, which seems to be not accessible to any reasonable explanation. Practically the same results were calculated also by the PLS method (column 5 in Table III) and so this peculiarity cannot be accounted for by different computational approach (PCA instead of PLS) but rather unproper preprocessing of the data (see above).

For the prediction of unknown SCS (see next section), it is necessary to evaluate the matrix of parameters **B** by Eq. (*6*) or Eq. (*15*). This calculation can be quite useful also for the better explanation of the findings from the projection. Table IV shows the parameters of multiparameter regression of ^{13}C SCS in 2-substituted naphthalenes on the CSI, $\mathbf{z}_j$ (j = *ipso*, *ortho*, *meta*, and *para*). It is clear, that above-mentioned SCS in both *ortho* positions (C-1 and C-3) can be predicted by proper combination of $\mathbf{z}_o$ and $\mathbf{z}_p$; addition and subtraction of $\mathbf{z}_p$ simulates the increase and decrease in the bond order between C-2, C-1 and C-2, C-3 atoms, respectively.

The sign at b_m plays probably the same role in the case of SCS on the C-6 and C-8 atoms. The C-8 and C-10 atoms could, according to the number of bonds, correspond

TABLE IV
Multiparameter regression[a] of SCS in 2-naphthyl derivatives with CSI

d	$\mathbf{b}_i$	$\mathbf{b}_o$	$\mathbf{b}_m$	$\mathbf{b}_p$
1	–	1.01	–	0.52
2	1.01	–	–	–
3	–	0.96	–	–0.49
4	–	–0.07	1.07	0.07
5	–	–	–	–
6	–	–	0.36	0.44
7	–	–0.08	0.47	0.14
8	–	–0.08	–0.36	0.30
9	–	–	–	–0.22
10	–	–	–0.55	0.49

[a] Equation (*5*), m= 21, n = 10, p = 4; parameters b_j calculated according to Eq. (*6*), only those significant by Student's t -test at the 0.05 level are given ($t_{0.05,17}$ = 2.110).

to C_{para} to substituent, however, their SCSs depend clearly not only on $\mathbf{z}_p$ but also on $\mathbf{z}_m$. Only SCS that cannot be explained by any combination of CSIs belongs to C-5; its variation with the change of 2-substituent is very small (cf. *sd* = 0.19 in Table III) and therefore, this SCS can be regarded as constant.

Prediction

For the prediction of ^{13}C SCS the matrix **B** has to be calculated from some well selected training sets **Dm** and **Zm**. The training set substituents should comprise at least two from each group: electron donors, electron acceptors, halogen atoms and alkyl groups and hydrogen[8]. Then, for another compound, not included in the training set, prediction of SCS ($\mathbf{D}^*$) is gained from the CSI ($\mathbf{Z}^*$) and $\mathbf{B}^{\neq}$ using the Eq. (*16*).

TABLE V
Prediction of SCS for new *p*-dimethylaminobenzene derivative using training sets and CSI test set[a]

CSI for		Residuals[b]				*ssr*[c]
No.	X	*i*	*o*	*m*	*p*	
		A: Training sets **DmC**, **ZmC** (data centered)				
2	NH_2	1.46	–0.60	0.01	–0.13	2.50
6	Br	0.47	–0.86	0.38	–0.39	1.26
9	CF_3	1.75	0.68	0.18	0.43	3.74
10	CN	1.29	0.11	0.59	0.70	2.52
12	COMe	0.38	–1.39	0.62	–0.26	2.53
14	CHO	–0.68	–2.00	0.32	–0.65	4.97
	std[d]	0.90	0.98	0.24	0.51	
		B: Training sets **DmS**, **ZmS** (data standardized)				
2	NH_2	–3.12	–1.53	4.41	4.40	51.00
6	Br	2.57	0.35	1.58	–0.47	9.47
9	CF_3	1.30	0.47	–3.14	1.91	15.52
10	CN	4.77	1.51	–2.37	3.42	42.28
12	COMe	–0.80	–2.23	–2.48	0.97	12.94
14	CHO	–1.23	–2.57	–3.76	1.25	24.12
	std[d]	2.86	1.67	3.23	1.76	

[a] Equations (*15*) and (*16*). [b] Residual $e = (d_t - d^*)$; SCS(exp) = d_t, SCS(calc) = d^*. [c] $ssr = \Sigma e^2$. [d] Standard deviation from the mean value.

The series 1 from the Taft collection of data[7a], i.e., 4-substituted *N*,*N*-dimethylanilines, **D1**, was selected for this calculation. The Table V shows the results of the prediction for the SCS of the compounds not included in the training set. When the data are only centered before calculations than residuals (i.e. SCS(predicted) – SCS(experimental)) lay in the range about ±2 (Table V, part A). However, if the standardized data are used for calculation of matrix **B**, then, since $\mathbf{z}_j$ vector of individual compound (m = 1) cannot be standardized, the residuals are 2–3 times larger (Table V, part B). In all the calculations four principal components were used since with three PC the test values *ssr* were larger.

Based on the PLS prediction[5], the result of projection of 2-naphthyl derivatives to the space of CSI was evaluated (Table III, column 6). In this case, both matrices $\mathbf{D}_{21,1}$ and $\mathbf{Z}_{21,4}$ were standardized and 21 projections were made always excluding the predicted SCS vector from the training sets. The *RSD* and *rsd* for experimental and predicted values were slightly worse than from PLS without excluding predicted vector (Table III, column 5) but no resemblance with the published results[4a] (Table III, column 7) was found.

CONCLUSIONS

Recommended procedure for the projection and the prediction of ^{13}C SCS of any series of aromatic substances from the CSI corresponding to monosubstituted benzenes could be:

1. to calculate projection of SCS ($\hat{\mathbf{D}}$) from experimental data (**D**) and the **H** (hat) matrix or corresponding PC (**U** . **U**′) from CSI (**Z**) by Eq. (*7*) and (*11*), respectively,

2. to assess the quality of projection by *RSD*, Eq. (*13*), and/or *rsd*, Eq. (*14*),

3. *a*) if $RSD \leq 0.3$, the quality of the projection is good (see Table II, Y = H, CH_3, and halogen atoms) and CSI can be used for prediction of SCS by simple regression, Eq. (*2*),

b) for $0.3 \leq RSD \leq 0.9$, the multiparameter prediction, Eq. (*15*) and Eq. (*16*), is the choice,

c) if $RSD \geq 0.9$ (see *RSD* values in Table II, calculated for randomly selected data rand (1), rand (2), and rand (3)), matrix of parameters **B** should be calculated by the principal component regression, Eqs (*6*) or (*15*), and checked by Student's *t* test if at all any combination of the CSI is suitable for the prediction of the data or not (cf. Table IV, **d** = 5).

REFERENCES

1. Lynch B. M.: Can. J. Chem. *55*, 541 (1977); Membrey F., Ancian B., Doucet J. P.: Org. Magn. Reson. *11*, 580 (1978); Membrey F., Boutin L., Doucet J. P.: Tetrahedron Lett. *21*, 823 (1980); Membrey F., Ancian B., Doucet J. P.: J. Chem. Soc., Perkin Trans. 2 *1980*, 1399;

Chandrasekaran R., Perumal S., Wilson D. A.: Magn. Reson. Chem. *25*, 1001 (1987); Holík M.: Chemom. Int. Lab. Systems *19*, 225 (1993).
2. Holik M.: Org. Magn. Reson. *9*, 491 (1977); Holik M., Belusa J., Brichacek J.: Collect. Czech. Chem. Commun. *43*, 610 (1978); Holik M., Potacek M., Svaricek J.: Collect. Czech. Chem. Commun. *43*, 734 (1978); Holik M., Mistr A., Laznicka V.: Collect. Czech. Chem. Commun. *43*, 739 (1978); Solcaniova E., Toma S.: Org. Magn. Reson. *14*, 138 (1980); Solcaniova E., Toma S., Fiedlerova A.: Org. Magn. Reson. *14*, 181 (1980); Perez C., Schleinitz K. D., Grundemann E.: Z. Chem. *22*, 260 (1982); Holik M., Paveska P., Mlynarik V.: J. Mol. Struct. *114*, 15 (1984); Holik M., Matejkova B.: Collect. Czech. Chem. Commun. *55*, 261 (1990).
3. Ewing D. F.: Org. Magn. Reson. *12*, 499 (1979).
4. a) Johnels D., Edlund U., Johansson E., Wold S.: J. Magn. Reson. *55*, 316 (1983); b) Johnels D., Edlund U., Wold S.: J. Chem. Soc., Perkin Trans. 2 *1985*, 1339; c) Johnels D., Edlund U., Johansson E., Cocchi M.: J. Chem. Soc., Perkin Trans. 2, *1989*, 1773.
5. Geladi P., Kowalski B. R.: Anal. Chim. Acta *185*, 1 (1986); Lindberg W., Persson J.-A., Wold S.: Anal. Chem. *55*, 643 (1983).
6. a) Mandel J.: Am. Statistician *36*, 15 (1982); b) Lorber A.: Anal. Chem. *56*, 1004 (1984).
7. a) Bromilow J., Brownlee R. T. C., Craik D. J., Sadek M., Taft R. W: J. Org. Chem. *45*, 2429 (1980); b) Holik M.: Magn. Reson. Chem. *30*, 189 (1992).
8. Ehrenson S., Brownlee R. T. C., Taft R. W.: Prog. Phys. Org. Chem. *10*, 1 (1973).
9. Craik D. J., Ternai B.: Org. Magn. Reson. *15*, 268 (1981).