A Comparison of the Methodologies of Principal Components Analysis and Linear Solvation Energy Relationships for the Correlation and Rationalization of Solvent Effects on ^{13}C -NMR Shifts of Indenyllithium

Mortimer J. Kamlet^a and Robert W. Taft^b

^aNaval Surface Weapons Center, White Oak Laboratory, Silver Spring, MD 20903-5000, and ^bDepartment of Chemistry, University of California, Irvine, CA 92717, USA

Kamlet, Mortimer J. and Taft, Robert W., 1986. A Comparison of the Methodologies of Principal Components Analysis and Linear Solvation Energy Relationships for the Correlation and Rationalization of Solvent Effects on ^{13}C -NMR Shifts of Indenyllithium. – Acta Chem. Scand. B 40: 619–624.

Solvent effects on a planar carbanion, indenyllithium, were analyzed by the solvatochromic comparison method, and the conclusions compared with the results of an earlier analysis of the same data using the method of principal components analysis. It was shown that the results for the single hydrogen bond donor in the solvent set "poison" the data matrix so that the principal components analysis loses important information and generates misinformation. It is suggested that the PCA method is most successful when the data have been prescreened in earlier correlations by physical organic chemists.

For several years, we have been involved in a dialogue with the chemometricians, S. Wold and M. Sjöström (W and S) regarding the comparative efficacy of their method of principal components (PC) analysis versus our method of linear solvation energy relationships (LSERs) for the correlation and prediction of properties depending on solute-solvent interactions. The disputation began when W and S suggested in no uncertain terms¹ that, rather than reflecting fundamental and rather simple physicochemical relationships, as the present authors had proposed, LSERs should be regarded merely as *local empirical models of similarity or locally valid linearizations of complicated relationships*. In response, we pointed out² that a single set of *solvatochromic parameters* and a general LSER of the form of eqn. (1) have been used to correlate and predict hundreds of diverse properties of solutes and solvents.

$$XYZ = XYZ_0 + A(\delta_{\text{H}}^2)_1 \bar{V}_2 + B\pi^*_1 \pi^*_2 + C\alpha_1 \beta_2 + D\beta_1 \alpha_2 \quad (1)$$

We suggested that, in the light of the number and diversity of the properties correlated by the same equation and parameters, W and S would be correct in their appraisal of LSERs only if their definition of the term "*local*" were expanded to include every area of chemistry and some of biology where physicochemical, biochemical, pharmacological, and toxicological properties depend on interactions between solutes and solvents.

Using the convention that subscript 1 applies to the solvent and subscript 2 to the solute, $(\delta_{\text{H}}^2)_1$ and $\bar{V}_2$ in eqn. (1), the square of the solvent Hildebrand solubility parameter and the solute molar volume are the terms that measure the endoergic cavity forming process. The solvatochromic parameters, π^*_1 and π^*_2 , measure exoergic solute-solvent dipolarity/polarizability interactions, and $\alpha_1 \beta_2$ and $\beta_1 \alpha_2$ measure exoergic solute-solvent hydrogen bonding interactions, α being the solvatochromic measure of solute or solvent HBD acidity, and β the solvatochromic measure

of HBA basicity (HBD = hydrogen bond donor, HBA = hydrogen bond acceptor).

When the XYZ term in eqn. (1) involves properties of a single solute in a series of solvents, the parameters relating to the solutes are subsumed into the constants. Further, when the XYZ studied does not involve the formation of a cavity or a change in cavity volume, as with the property discussed in the present paper, the term in $(\delta_{ii})_1$ drops out, and the general LSER takes the form of eqn. (2):

$$XYZ = XYZ_0 + s\pi^* + a\alpha + b\beta. \quad (2)$$

In the light of the pejorative assessment by W and S of LSERs, it is appropriate to point out some of the shortcomings of their method of PC analysis when applied to the correlation of solute-solvent interactions. Thus, in another paper,³ we present evidence that PC analysis has most often been effective in the field of solubilities and solute-solvent interactions when the input data used by the chemometricians have been prescreened for appropriateness and accuracy by physical organic chemists. For example, after demonstrating the strong effects of solute-solvent hydrogen bonding interactions on UV/visible spectra,^{4,5} we carefully assembled a subset of spectral data wherein such interactions were excluded.⁶ We then used this subset for 47 indicator solutes to construct our π^* scale of solvent dipolarity/polarizabilities.⁶ W and S later⁷ used this same subset to derive two principal components, and made much of the fact that correlations of the same and other spectral data with their θ_1 and θ_2 principal components were marginally better than corresponding correlations with our single π^* parameter. They did not point out that the marginal superiority of their PCs would have all but disappeared if the comparisons had been with our correlations using the two-parameter $(\pi^* + d\delta)$ formalism.⁸ Further, our correlations involved the same parameters as had been used in hundreds of other correlations, so that it has been possible to compare the solvent dipolarity effects on the spectra with corresponding effects on other physicochemical properties. For comparison, in the five years since publication, their UV/visible θ_1 and θ_2 principal components have never, to our knowledge, been used to correlate any other spectral or other type property.

As abundantly demonstrated elsewhere,³ the

main shortcoming of the practitioners of PC analysis is their routine inclusion in the computational matrixes of properties that do not relate to the effect being studied, while often failing to include intrinsic properties that are necessary for the correlation. Also, since these analysts have usually been more conversant with statistics than with chemistry, they have often failed to identify and exclude experimental results which are obviously inaccurate or which, as in the example discussed here, reflect an additional unconsidered effect. Thus, a larger amount of inappropriate data in the PC analysis procedure has often overwhelmed a smaller amount of accurate and appropriate data. The bad input has poisoned the good, and the result has been what computer scientists refer to as GIGO, "garbage in, garbage out". As an example, which we analyze in detail elsewhere,³ PC analysis of factors influencing solubilities in water, S_w , of 83 organic nonelectrolytes⁹ gave, with the best four of eight principal components, a correlation coefficient, r , of 0.77 and a standard deviation, sd , of 0.59 log unit. By comparison, analysis of S_w of a 105 solvent data set of comparable complexity by eqn. (1) and the LSER method gave $r = 0.9954$ and $sd = 0.14$ log unit.^{10,11}

W and S have not been as guilty in this regard as other chemometricians because, as has been mentioned earlier, their practice in the field of solute-solvent interactions (as well as with substituent effects)¹ has been to treat data which had been prescreened in earlier correlations by physical organic chemists. There has been one exception and, in the present paper, we point out the important information that has been lost and the serious misinformation that has been generated in a PC analysis by Wold and coworkers where the input data had not been prescreened.¹²

Results and discussion

The above-mentioned study involved the effects of thirteen solvents on ¹³C-NMR chemical shifts

Table 1. Data used in correlations. Chemical shifts in ppm.

No.	Solvent C ₄ -C ₅	π^*	β	α	θ_1	θ_2	C ₅ -C ₆ H ₁₂	C ₁ -C ₅	C ₂ -C ₅	C ₂ -C ₅	C _{3a} -C ₅
1.	TEA	0.14	0.71	0	9.80	-0.80	117.53	-24.79	-2.96	9.06	3.69
2.	i-Pr ₂ O	0.27	0.49	0	9.63	-0.82	117.48	-24.76	-2.79	9.09	3.89
3.	Et ₂ O	0.27	0.47	0	6.08	0.22	116.13	-24.00	-0.71	12.01	4.53
4.	THP	0.51	0.54	0	4.30	0.83	115.24	-23.55	0.14	13.54	5.06
5.	DME	0.53	0.41	0	2.08	0.62	114.36	-22.27	1.35	14.77	5.75
6.	THF	0.58	0.55	0	1.47	1.27	114.25	-22.38	1.53	15.70	5.60
7.	CH ₃ CN	0.75	0.37	0.19	-4.69	-0.28	112.49	-18.89	6.01	18.37	6.51
8.	Pyridine	0.87	0.64	0	-6.15	-0.25	112.82	-17.99	6.61	19.34	7.05
9.	DMSO	1.00	0.76	0	-6.55	-0.85	111.71	-17.59	7.22	19.10	7.08
10.	DMF	0.88	0.69	0	-7.33	-0.06	111.19	-17.59	7.29	20.21	7.29
11.	HMPA	0.87	1.05	0	-8.64	0.13	110.19	-16.95	7.79	21.14	7.96
12.	Dioxane	0.55	0.37	0	1.7		114.52	-22.51	1.07	15.92	5.34

of a planar lithium carbanion, indenyllithium (*I*). The stated purpose of the work was to derive a solvent scale useful as a guide for the generation of delocalized carbanions, but before discussing their PC analysis, we present here an analysis of the solute-solvent interactions by our *solvatochromic comparison method*.

The ¹³C-NMR shifts of interest are assembled in Table 1 together with the solvent π^* , α , and β values. The low solubilities of dioxane and anisole led Wold and coworkers to exclude these from their initial PC analysis; we included the shifts in dioxane in ours. Their initial analysis was performed on the data matrix containing the ¹³C chemical shifts at five positions referenced to cyclohexane, i.e., five variables measured in eleven solvents. A two-component model was said to be adequate by CV (cross-validation), accounting for 79 % of the original data standard deviation. A somewhat better two-component model, describing 94 % of the standard deviation of the four variables, was obtained when the shifts at the other four positions were referenced internally to the corresponding shifts at the 5-position (the 1- and 3-positions are equivalent, as are 3a and 7a, 4 and 7, and 5 and 6). Wold and coworkers' θ_1 and θ_2 principal components from the latter analysis¹² are included in Table 1. In order to compare our results with their best results, we carried out analyses of the shifts at the C₅ position referenced to cyclohexane, and at the C₁, C₂, C_{3a}, and C₄ positions referenced to C₅.

Before undertaking the correlations by eqn.

(1), it was necessary to consider what solvent effects should be expected with the ¹³C shifts, and hence what parameters to include in the correlations. (a) Since the NMR measurement entailed no change in solute volume, we excluded the cavity term in going from eqn. (1) to eqn. (2). (b) Because we had earlier found NMR shifts to be strongly influenced by dipole/dipole or charge/dipole interactions,¹³ we expected important $s\pi^*$ terms. (c) Because association of HBA bases with Li⁺ should lessen ion pairing with indenide ion, we also expected dependences on HBA basicity; but since these are only second order effects, $b\beta$ terms were expected to be smaller than $s\pi^*$ or aa terms. (d) Hydrogen bonding by HBD solvents to indenide ion was expected to lead to large aa terms, applicable only to acetonitrile, the single solvent in the study for which $\alpha \neq 0.00$. (e) Goodness of fit of two-parameter correlations with π^* and β should be adequate for 11 of the solvents, but less so with the inclusion of the results in acetonitrile. (f) The importance of canonical structures with charge at the 1-, 2-, and 3-positions to the total indenide ion resonance hybrid should cause effects to be strongest for the chemical shifts of the carbons on the five-membered ring.

The correlations agreed fully with the expectations. The 12-solvent multiple linear regressions in π^* , β , and α are given by eqn. (3)–(7). The *r* and *sd* values in parentheses are the goodness of fit measures if the term in aa is excluded. The two-parameter correlations for the 11-solvent

data set excluding acetonitrile have the same coefficients of π^* and β and similar r and sd values.

$$\delta(C_5-C_6H_{12}) = (119.8 \pm 0.8) - (7.2 \pm 1.0)\pi^* - (2.4 \pm 1.4)\beta - (6.2 \pm 5.0)\alpha$$

$$r = 0.963 \text{ (0.955), } s.d. = 0.8 \text{ (0.8)} \quad (3)$$

$$\delta(C_1-C_5) = (-29.0 \pm 0.8) + (9.0 \pm 0.9)\pi^* - (3.9 \pm 1.4)\beta + (9.9 \pm 4.6)\alpha$$

$$r = 0.980 \text{ (0.967), } s.d. = 0.7 \text{ (0.9)} \quad (4)$$

$$\delta(C_2-C_5) = (-7.2 \pm 0.9) + (12.6 \pm 1.1)\pi^* - (3.6 \pm 1.6)\beta + (12.7 \pm 5.4)\alpha$$

$$r = 0.984 \text{ (0.971), } s.d. = 0.9 \text{ (1.1)} \quad (5)$$

$$\delta(C_{3a}-C_5) = (6.7 \pm 1.4) + (13.8 \pm 1.7)\pi^* - (1.1 \pm 2.5)\beta + (4.9 \pm 8.4)\alpha$$

$$r = 0.962 \text{ (0.960), } s.d. = 1.3 \text{ (1.1)} \quad (6)$$

$$\delta(C_3-C_5) = (2.5 \pm 0.4) + (4.4 \pm 0.5)\pi^* - (1.1 \pm 0.8)\beta + (1.6 \pm 2.5)\alpha$$

$$r = 0.970 \text{ (0.968), } s.d. = 0.4 \text{ (0.4)} \quad (7)$$

The italicized terms in eqns. (3)–(7), although almost certainly reflecting real solute-solvent interaction effects, are not statistically significant at the 95 % confidence level. Correspondingly, in most of these instances, the θ_2 terms in Wold and coworkers' analyses¹² were also not statistically significant. The large uncertainties in these equations are not due to calculational shortcomings, but rather reflect an intrinsic source of imprecision in the measurements. The most closely analogous investigation to which Eliasson *et al.*¹² could relate their findings was an UV/visible spectroscopy study of alkali metal salts of 1,3-diphenylbut-1-ene by Burley and Young.¹⁴ In the time frame of the UV/visible measurements, these workers were able to discern that their lithium salts were contact ion pairs (CIP) in the less dipolar solvents, solvent-separated ion pairs (SSIP) in more dipolar solvents, and mixtures of CIP and SSIP in solvents of intermediate dipolarity.

A similar situation existed for indenyllithium in the solvents of Table 1, but the differences between the CIPs and SSIPs were not discernible (but led to averaged signals) in the time frame of the NMR measurements. Thus, for example, the markedly different shifts in diethyl ether and diisopropyl ether could be rationalized and, hence, Wold and coworkers' markedly different θ_1 and

θ_2 values for these solvents (Table 1). In all other solvent property scales, including π^* , β , $E_T(30)$, X_R , DN , B , Py , etc., these solvents have very similar values.¹⁵ Here, however, for an SSIP, the size of the solvent separating the anion from the cation would have a strong effect on its properties.

The important conclusions which we have drawn from the above correlations are as follows. Reflecting strong dependences on solute HBD acidity, the two-parameter correlations for the $C_5-C_6H_{12}$, C_1-C_5 and C_2-C_5 chemical shifts deteriorate in quality when the CH_3CN data are included, but goodness of fit is restored when a term in α is also allowed. The dependence on solvent π^* values, and hence charge density in the carbanion, is highest at C_{3a} and C_{7a} . However, probably because of steric effects, ion pairing and hydrogen bonding at $C_{3a,7a}$ are minimal. Hydrogen bonding by the HBD solvent to *I* exerts its strongest effects at the positions on the five-membered ring. Although statistically significant at the 95 % confidence level in only one instance, the small effect of solvent HBA basicity is probably real and most evident at C_1 and C_2 .

Analysis of conclusions of Wold and coworkers.

With the above as prologue, we now consider the conclusions arrived at in Wold and coworkers' analyses of the same data. To avoid misquoting them inadvertently, the following text in italics is verbatim. (a) *For the PC analysis ... using four chemical shift variables, i.e., having the $C_{5,6}$ as an internal reference ... a two component model was obtained which described 94 % of the standard deviation of the four variables.* As in their case, we found that a two-parameter model best correlated the results in eleven of the twelve solvents. However, in our LSER procedure, the fact that an additional solute-solvent interaction influenced the shift in acetonitrile was signalled by the deterioration in statistical fit when this solvent was included in the two-parameter correlation. The PC analysis, on the other hand, gives no such signal when an additional interaction occurs for only one solvent in the matrix, but rather the additional effect is apportioned between the θ_1 and θ_2 values for that solvent. As a result, the PC analysis loses important information regarding absolute and relative effects on the ^{13}C shifts by hydrogen bonding involving the solvent as donor

and the carbanion as acceptor and, although experimentally probably correct, the acetonitrile results poison the total θ_1 – θ_2 parameter set.

(b) *One would intuitively expect some correlation between solvent basicity scales ... and the principal component(s) of the shift matrix. However, for the complete solvent matrix this is not observed, instead θ_1 which is the dominating "effect", has a moderate correlation to the "acidity" or "polarity" scales... The best fit is found for π^* , $r = 0.86$, $n = 13$. If anisole is excluded, the correlation is improved, $r = 0.96$, $n = 12$.*

This would be a reasonable conclusion if the principal components were "pure" parameters, i.e., reflecting the effect of only a single type solute-solvent interaction (as is the case for the π^* , α , and β solvatochromic parameters¹⁶). However, θ_1 and θ_2 are not pure parameters. A correlation of θ_1 with π^* and β is given by eqn. (8a) for 12 solvents and by eqn. (8b) for 11 solvents excluding acetonitrile. To be noted particularly is how,

$$\theta_1 = (15.2 \pm 1.7) - (22.3 \pm 2.0)\pi^* - (2.8 \pm 2.9)\beta$$

$$n = 12, r = 0.973, s.d. = 1.8 \quad (8a)$$

$$\theta_1 = (16.0 \pm 1.7) - (21.2 \pm 2.0)\pi^* - (4.8 \pm 3.0)\beta$$

$$n = 11, r = 0.978, s.d. = 1.6 \quad (8b)$$

in our analysis, the dependence of the first principal component on solvent HBA basicity increases when the "poisoned datum for acetonitrile" is excluded.

(c) *The second component, θ_2 , does not show any acceptable correlation to any of the empirical solvent scales, and can be considered as specific for the complex system at hand. Another conclusion than can be reached from this study is that ion-pair solvation cannot successfully be explained by a single solvent effect ... nor a solvent scale derived in structurally different systems, at least not for such a variety of solvents.*

No wonder they did not see the intuitively expected relationship of the second component with any of the scales of solvent HBA basicity – the dependence on HBA basicity, which is demonstrated without any statistical uncertainty in eqn. (4), is already included in their first component! Furthermore, considering the intrinsic imprecision of the measurements, the β scale accommodates the effects of ion-pairing quite well.

(c) *To conclude, the present study confirms the complexity of the ion-pair system when it concerns*

the solvation of ion-pair(s). Reasonable correlations to single solvent scales are obtained only on subsets of the original solvent matrix, which would indicate that, apart from cation solvation, solvent effects exercised on the anion also reflect the anion-cation distance and thus the reflected π -polarization. These latter interactions are probably governed by other solvent characteristics than those valid for the cation solvation process and are more important for the non-ethereal subsets.

Finally, we note that modern data analytic methods, such as PC-CV and PLS give direct possibilities to find the systematic behavior in multivariate data sets and to relate data sets to each other. The direct approach gives results which are easier to interpret than those using indirectly derived solvent scales (and other empirical "substituent" scales) of dubious relevance for a given problem.

Based on the above, it is no wonder that Wold considered the fundamental relationships governing solvent effects to be complicated, while we consider them to be quite simple. From their PC analysis, they concluded that different effects apply to different subsets of the data (ethereal vs. nonethereal), and that the ion-pair system is complex and unique. As with the UV/visible principal components, they did not relate their indenyllithium principal components to any other solvent effect studies in the paper under discussion nor, to our knowledge, in any more recent papers. To relate their indenyllithium principal components to their UV/visible principal components would go against their definition of "local" and the core of their reasoning (although in both instances their θ_1 parameter correlates reasonably well with π^*). If the above represent typical conclusions of the chemometric approach, then PC analysis provides no more than a vague and murky understanding of solute-solvent interactions and, indeed, often a misunderstanding.

On the other hand, using the LSER approach without distinction between ethereal and nonethereal subsets, we observe all the effects intuitively expected by Wold and coworkers and by us. We find the relative effects of solvent dipolarity/polarizability, HBA basicity, and HBD acidity to be chemically quite reasonable (for example, the coefficients of π^* for the shifts of the 1-, 2-, and 3a-carbons of the carbanion and the coefficients of α for the 1- and 2-carbons are, as expected higher than the coefficients for ^{13}C shifts

of the carbonyl carbon of acetone).¹³ We also find that there is nothing particularly unusual about the solvent effects on ¹³C chemical shifts of indenyllithium, but rather that they fit well into a mutually consistent and mutually supportive framework of over 600 solute-solvent interactions which have been correlated by the same set of solvatochromic parameters.

Relative predictive ability. An important claim of the proponents of principal components analysis has been that PCA leads to better predictability of new experimental data than LSER or LFER. It has frequently been unclear to us how this is the case. In the present instance, for example, it is possible to predict with reasonable accuracy effects of other solvents on ¹³C shifts at all positions from published solvatochromic parameters of hundreds of solvents, or estimated parameters for hundreds more. If a new solvent has unusual substituents which make parameter estimation unreliable, simple UV/visible measurements allow facile determination of π^* and β , which in turn allow prediction of ¹³C-NMR shifts. It is not obvious, however, how one might use Wold and coworkers' principal components for predictions of new experimental data. Granted, it is possible to estimate θ_1 for a new solvent from their correlation with our π^* parameter or through our somewhat more precise correlation with π^* and β in eqn. (8b), but how does one estimate θ_2 , which they claim is system-specific and which we believe contains elements of HBA basicity and possibly molar volume, but which is mainly a scavenger of experimental residuals?

The above comparison might seem trivial in the sense that, once the chemical factors influencing solvent effects on NMR shifts of indenyllithium have been elucidated, it is unlikely that anyone needs to predict shifts in additional solvents. It is not trivial, however. In *exactly analogous situations*, PC analyses and LSERs have been used¹⁷ to correlate toxicological properties of industrial pollutants used to the golden orfe fish and the fathead minnow. The data bases mainly comprise compounds that act by a *non-specific nonreactive* toxicity mechanism governed mainly by transport properties, but also include a number of toxicants that also act by *specific reactive* toxicity mechanisms (reactions of RCH=O toxicants with enzyme amine groups, reactions of Michael addend toxicants with en-

zyme sulfhydryl groups, *in vivo* hydrolysis of carboxylic esters, etc.). For these data bases, the reactive toxicants "poison" the nonreactive PCA matrixes in exactly the same way as acetonitrile "poisoned" the indenyllithium PCA matrix. In these instances, we have shown that the ability to identify and exclude reactive "outliers" and the consequently improved predictive capability of LSERs over PC analyses is critical to any meaningful analysis of the problem, and to the ability to predict fish-kill properties of suspected pollutants. This is possible with LSER but not with PC analyses as currently practiced.

Acknowledgement. The work by MJK was done under Naval Surface Weapons Center Independent Research Task IR-210.

References

1. Sjöström, M. and Wold, S. *Acta Chem. Scand. B* 35 (1981) 537.
2. Kamlet, M. J. and Taft, R. W. *Acta Chem. Scand. B* 39 (1985) 611.
3. Kamlet, M. J. and Taft, R. W. *J. Chemometrics*. To be submitted.
4. Kamlet, M. J. and Taft, R. W. *J. Am. Chem. Soc.* 98 (1976) 377.
5. Taft, R. W. and Kamlet, M. J. *J. Am. Chem. Soc.* 98 (1976) 2886.
6. Kamlet, M. J., Abboud, J.-L. M. and Taft, R. W. *J. Am. Chem. Soc.* 99 (1977) 6027.
7. Sjöström, M. and Wold, S. *J. Chem. Soc. Perkin Trans. 2* (1981) 104.
8. Taft, R. W., Abboud, J.-L. M. and Kamlet, M. J. *J. Am. Chem. Soc.* 103 (1981) 108.
9. Chastrette, M., Rajzmann, M., Chanon, M. and Purcell, K. F. *J. Am. Chem. Soc.* 107 (1985) 1.
10. Taft, R. W., Abraham, M. H., Doherty, R. M. and Kamlet, M. J. *Nature* 313 (1985) 384.
11. Kamlet, M. J., Doherty, R. M., Abboud, J.-L., Abraham, M. H. and Taft, R. W. *J. Pharm. Sci.* 75 (1986) 338.
12. Eliasson, B., Johnels, D., Wold, S. and Edlund, U. *Acta Chem. Scand. B* 36 (1982) 155.
13. Taft, R. W. and Kamlet, M. J. *Org. Magn. Reson.* 14 (1980) 485.
14. Burley, J. W. and Young, R. N. *J. Chem. Soc. (B)* (1971) 1018.
15. Reichardt, C. *Solvent Effects in Organic Chemistry*. Verlag Chemie, Weinheim, 1979, Chapter 7.
16. Kamlet, M. J., Abboud, J.-L. M., Abraham, M. H. and Taft, R. W. *J. Org. Chem.* 48 (1983) 2877.
17. Kamlet, M. J., Doherty, R. M., Taft, R. W., Abraham, M. H. and Abraham, D. J. *Environ. Sci. Technol.* Submitted.

Received January 2, 1986.