Similarity transference of molecular parameters. I. The atomic polar tensors of cyanoacetylene

B. B. Neto, M. N. Ramos, and R. E. Bruns

Citation: The Journal of Chemical Physics **85**, 4515 (1986); doi: 10.1063/1.451772

View online: http://dx.doi.org/10.1063/1.451772

View Table of Contents: http://scitation.aip.org/content/aip/journal/jcp/85/8?ver=pdfcov

Published by the AIP Publishing

Similarity transference of molecular parameters. I. The atomic polar tensors of cyanoacetylene

B. B. Neto[a] and M. N. Ramos
Departamento de Química Fundamental, Universidade Federal de Pernambuco, 50000 Recife, PE, Brasil
R. E. Bruns
Instituto de Química, Universidade Estadual de Campinas, 13100 Campinas, SP, Brasil

(Received 2 June 1986; accepted 11 July 1986)

A similarity transference procedure for the calculation of molecular parameters is proposed. The theoretical relationships between the direct transference method normally used and the proposed procedure are discussed. The importance of adequately defining the similarity models used in the transference calculations is emphasized. As an example, similarity models are constructed using the experimental and the STO-3G and 4-31G molecular orbital values of the atomic polar tensors of the HCN, C_2H_2, CH_3CN, C_4H_2, CH_3CCH, and C_2N_2 molecules. Partial least squares calculations based on these similarity models and using the STO-3G and 4-31G values of the atomic polar tensors of HC_3N result in estimates of the experimental tensors of cyanoacetylene which have about one-half or less the root mean square error of the molecular orbital values. Also the partial least squares detection of potentially unreliable estimates of polar tensor elements is illustrated.

INTRODUCTION

The transference of localized parameters is a common technique used by chemists in attempting to estimate molecular properties when direct experimental observation is difficult. Several examples easily come to mind: isolated atom polarizabilities are often transferred to the molecular environment to calculate molecular polarizabilities[1]; bond energy tables, with average or representative values obtained from a series of molecules for which sufficient thermodynamic data are available, are used to calculate internal energy and enthalpy values[2]; standard bond distances and angles are transferred to predict geometries of molecules for which structural data are incomplete[2]; force constants of molecules with completely analyzed spectra are frequently transferred to molecules for which experimental observations are difficult or band assignments are uncertain.[3] In all such cases it is assumed that the parameters to be transferred are constant and insensitive to the differences between the electronic environments of the original molecules and those of the molecules to which the parameters are transferred. Eventual discrepancies are then explained on the basis of the chemist's knowledge of the differences in the electronic structures of the molecules involved.

For example, the limitations of direct transference procedures are striking for molecules exhibiting delocalization effects, such as conjugated carbon–carbon double bond systems. This is to be expected, since electronic and thermodynamic parameters evaluated using isolated double bonds seldom agree with the values obtained for the double bonds of conjugated systems, and the extent of the difference is not easy to predict.

In this work we apply a multivariate approach to the general transference of molecular parameters in which, instead of assuming a parameter to have the same value for a group of similar molecules, we try to allow for its variation in a natural way, building a model of the similarity holding between the members of the group. The method we employ is based on the well-established principal component analysis,[4] where direct transference corresponds to a zero order principal component model, and variations in parameters for different molecular environments are described by increasing the number of principal components included in the model.

The similarity models used in the calculations are actually determined using the partial least squares (PLS) method introduced by Wold,[5] which is closely related to principal component regression. Linear models are defined which describe variations in atomic polar tensor parameters with differing electronic environments and also provide theoretical estimates of polar tensor elements for molecules where experimental intensity values are lacking.

As an example of the proposed procedure the atomic polar tensors of cyanoacetylene are calculated using experimental and quantum chemical polar tensors of several molecules containing carbon–carbon and carbon–nitrogen triple bonds. These values are then compared with those calculated from the experimental gas phase vibrational intensities and normal coordinates of this molecule.[6]

SIMILARITY TRANSFERENCE OF POLAR TENSORS

The M elements of an atomic polar tensor can be regarded as the coordinates of a point in M-dimensional space. Depending on the situation M can go up to nine (all elements different and not null), but usually it is a much smaller value. To illustrate, seven diagonal ($M = 3$) polar tensors are plotted in Fig. 1(a).

The points representing similar polar tensors naturally form a cluster in M-dimensional space. If the similarity is large, that is, if the points are clustered tightly enough, then

[a] Present address: Instituto de Química, Unicamp, 13100 Campinas, SP, Brasil.

 0021-9606/86/204515-09$02.10

the whole set of polar tensors can be adequately represented by a linear *similarity model*,[7] much in the same way as the truncation of higher order terms in a Taylor series is not prejudicial to dipole moment evaluations for small displacements from the equilibrium geometry. In general, similarity transference of molecular parameters only assumes that the reference parameters are similar to those expected for the molecule whose properties are to be evaluated.

In the simplest of such models each polar tensor element is replaced by the average over the entire set of similar elements. This corresponds to a SIMCA (similarity modeling by class analogy)[8] model with zero principal components:

$$p_{k,\sigma\nu} = \bar{p}_{\sigma\nu} + e_{k,\sigma\nu}, \quad \sigma,\nu = x, y, \text{ or } z, \tag{1}$$

where $p_{k,\sigma\nu} = \partial p_\sigma / \partial \nu$ for the k th polar tensor in the set, and $\bar{p}_{\sigma\nu}$ is the average of all such elements:

$$\bar{p}_{\sigma\nu} = \frac{1}{n} \sum_k \frac{\partial p_\sigma}{\partial \nu}, \tag{2}$$

n being the total number of similar polar tensors.

On using $\bar{p}_{\sigma\nu}$ to represent $p_{k,\sigma\nu}$ the residuals $e_{k,\sigma\nu}$ are left, which contain both experimental and modeling errors, and are a measure of how well the model fits the experimental data. This zero order model corresponds to a single point in M-dimensional space, and the residuals can be used to define a confidence region around it, in which one expects to observe the points for other polar tensors which belong to the same similarity model.[8] In M dimensions this would be a hypersphere [a sphere in Fig. 1(a)].

Direct transference, which corresponds to our zero order model, thus assumes that the polar tensor for the atom being studied is equivalent to an average of polar tensors for atoms having similar electronic environments. Usually, due to scarcity of data, an atomic polar tensor from a selected reference molecule is chosen for transference, instead of an average. For example, the fluorine polar tensor of methyl fluoride has often been chosen to predict intensity values of other fluorine-containing molecules, such as SF_6, UF_6, and UF_5.[9]

The zero order similarity model allows no parameter variation from one molecule to another. Of course it can be improved by the inclusion of principal components. With one component the SIMCA model becomes

$$p_{k,\sigma\nu} = \bar{p}_{\sigma\nu} + t_k b_{\sigma\nu} + e_{k,\sigma\nu},$$
$$k = 1,2,\ldots,n, \quad \sigma,\nu = x, y, \text{ or } z. \tag{3}$$

Here the $\sigma\nu$ element for the k th atom is represented by the average $\bar{p}_{\sigma\nu}$ *plus* a product term, $t_k b_{\sigma\nu}$, calculated so as to minimize the new residuals $e_{k,\sigma\nu}$. In principal component-factor analysis terminology the $b_{\sigma\nu}$ are the loadings of the different variables, the $p_{\sigma\nu}$, and determine the orientation of the principal component (PC) axis with respect to the initial coordinate system. Of all possible axes, the PC axis is the one that contains the maximum value of variance and so best represents the data, according to the least-squares criterion. The t_k, called scores, are the projections of the k polar tensor points on the PC axis. The more similar the actual atomic polar tensors, the more similar their scores. Based on the residuals one can draw a (hyper) cylinder as a region of confidence around the PC axis [Fig. 1(b)].

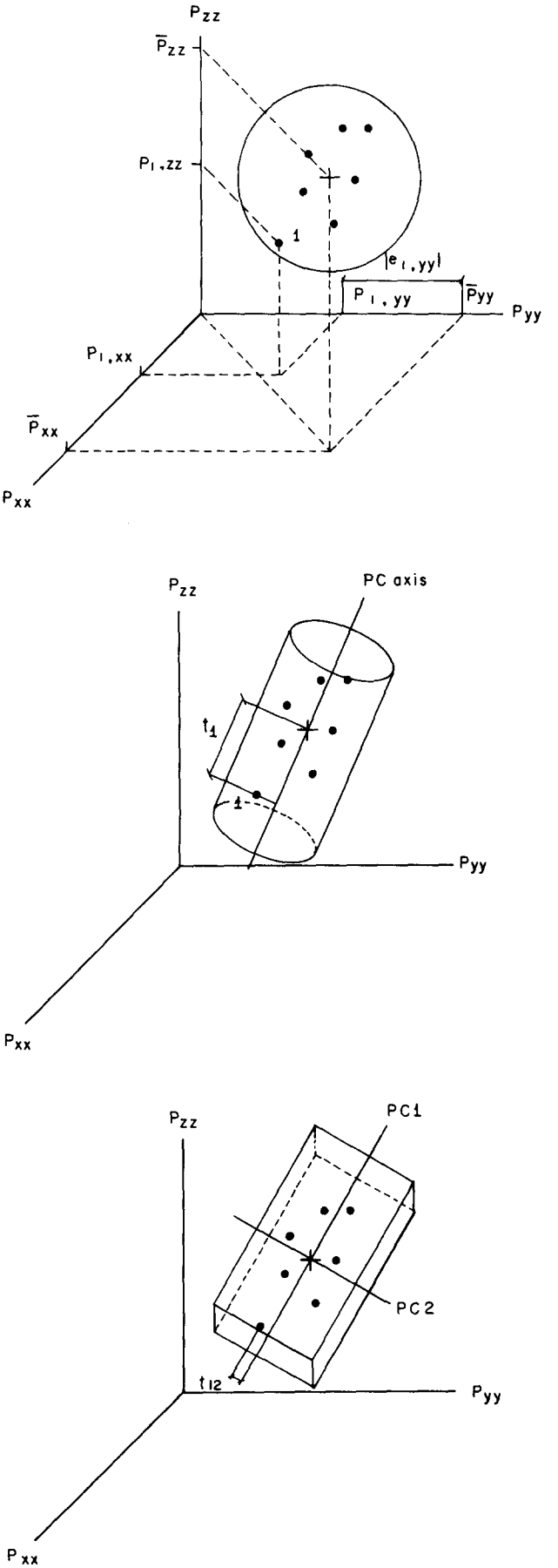

FIG. 1. Principal component models in three dimensions and respective confidence regions. (a) Zero components—a sphere; (b) one component—a cylinder; (c) two components—a box.

If a one component model is not sufficient to give an accurate representation of the data points a two-component model [Fig. 1(c)], in which each point is represented by two scores, can be used:

$$p_{k,\sigma\nu} = \bar{p}_{\sigma\nu} + t_{k1} b_{1,\sigma\nu} + t_{k2} b_{2,\sigma\nu} + e_{k,\sigma\nu},$$
$$k = 1,2,...,n, \quad \sigma,\nu = x, y, \text{ or } z. \tag{4}$$

The second PC axis, whose orientation is given by $b_{2,\sigma\nu}$, contains the maximum amount of remaining variance, considering all possible axes orthogonal to the first one. The score of the k th polar tensor on this axis is given by t_{k2}, and the confidence region in this case is a (hyper)box [see Fig. 1(c)].

Usually the residuals $e_{k,\sigma\nu}$ decrease as more components are added to the model, and additional PC's can be introduced until the data are represented adequately. The more complicated the data structure, the larger the number of PC's required to reduce the residuals to acceptable levels.

For the general case where the expansion is carried through to the A th principal component the equations become

$$p_{k,\sigma\nu} = \bar{p}_{\sigma\nu} + \sum_{a=1}^{A} t_{ka} b_{a,\sigma\nu} + e_{k,\sigma\nu},$$
$$k = 1,2,...,n, \quad \sigma,\nu = x, y, \text{ or } z. \tag{5}$$

The scores t_{ka} of each set of polar tensors are taken as measures of the degree of similarity between the members of the set. The number of components kept in the model A can be determined statistically by means of the cross-validation method.[10] Good fitting models are obtained when the residuals contain little modeling error, i.e., when they reflect essentially experimental error.

The first (and major) step in PC analysis consists of deciding which atomic polar tensors are similar enough to be included in one model. Chemical arguments can be used, e.g., hydrogen polar tensors in one category or model and carbon tensors in another. More objectively, perhaps, one can resort to pattern recognition methods, such as the nearest neighbor rule (KNN),[11] hierarchical cluster techniques,[12] or the SIMCA method itself. The SIMCA method is especially suitable, since it provides a PC formalism for the different levels of approximation involved in the transference procedure. Confidence regions, given by hyperspheres, hypercylinders, or hyperboxes, can be constructed around the PC axes, and an atomic polar tensor is included in one model if its representative point falls within the respective confidence region.

Once similarity models are defined for the various types of atomic polar tensors, they can be used in polar tensor calculations for molecules not included in the model-building phase. PC regression[4] or the related PLS modeling technique[5] can then be used to make predictions from the similarity models.

If theoretical methods are reasonably successful in reflecting the similarities in electronic structures found by experimental observation, as is more or less the case with *ab initio* quantum chemical methods, they can be included in the model calculation. One then arrives at a model that connects experimental values with theoretical results, allowing one to use the latter to make predictions about the former.

Two PC models are then constructed: one for the experimental polar tensors, given by Eq. (5), and the other for the corresponding theoretical values, represented by the analogous equation

$$\pi_{k,\sigma\nu} = \bar{\pi}_{\sigma\nu} + \sum_{a=1}^{A} \theta_{ka} \beta_{a,\sigma\nu} + e_{k,\sigma\nu},$$
$$k = 1,2,...,n, \quad \sigma,\nu = x,y,z. \tag{6}$$

Here the Greek symbols correspond exactly to their Latin counterparts in Eq. (5), the only difference being that they are theoretical, rather than experimental, in origin. Both models—experimental and theoretical—are represented by dotted lines in Fig. 2, for a case in which one component suffices for an adequate description of the data.

The experimental and theoretical scores, t_{ka} and θ_{ka}, respectively, can be related by the linear regressions

$$t_{ka} = \rho_a \theta_{ka} + \Delta_k, \quad a = 1,2,..,A, \tag{7}$$

where ρ_a is the regression coefficient and Δ_k measures the deviation of the k th polar tensor from the regression line. If the theoretical scores mirror reasonably well the similarities in the experimental scores, good regression lines will be obtained in Eq. (7).

This procedure, PC regression, ignores possible correlations between the experimental and theoretical blocks of values, which could improve the regressions in Eqs. (7). This is taken into account by the PLS technique, in which one block is weighted by the other and the regression is performed over latent variables, rather than over principal components. This amounts to relaxing the maximum variance criterion within one single block in order to improve the regression relating one block to the other. The latent variable axes (solid lines in Fig. 2) are tilted with respect to the principal component axes, but lead to better interblock relations.

To predict polar tensor values for atoms absent from the model but nevertheless similar to the ones that were actually employed, one starts by doing quantum chemical calculations to determine the theoretical polar tensor, whose elements are substituted into Eqs. (6) to yield the theoretical scores, θ_{ka}. Experimental scores, t_{ka}, are then obtained from Eqs. (7), and in turn used in Eqs. (5) to predict "experimental" polar tensor elements. Of course, prior to modeling and prediction care must be taken to perform the rotations required to make all polar tensors consistent with one another. The whole procedure is depicted in Fig. 2 by the arrows connecting the three plots.

Misuse of the similarity modeling technique can also be detected with the PLS formalism. If theoretical values are not accurately described by a theoretical similarity model, then inaccurate predictions of experimental results may follow. This can be quantified by means of a statistical F test, as described below.[13]

A PLS prediction for the n'th polar tensor (not included in the model) will yield from Eq. (6) the residuals $\epsilon_{n',\sigma\nu}$ which define for this tensor a residual variance with respect to the model:

$$S^2_{n'} = \sum_{\sigma,\nu} \frac{\epsilon^2_{n',\sigma\nu}}{M - A}. \tag{8}$$

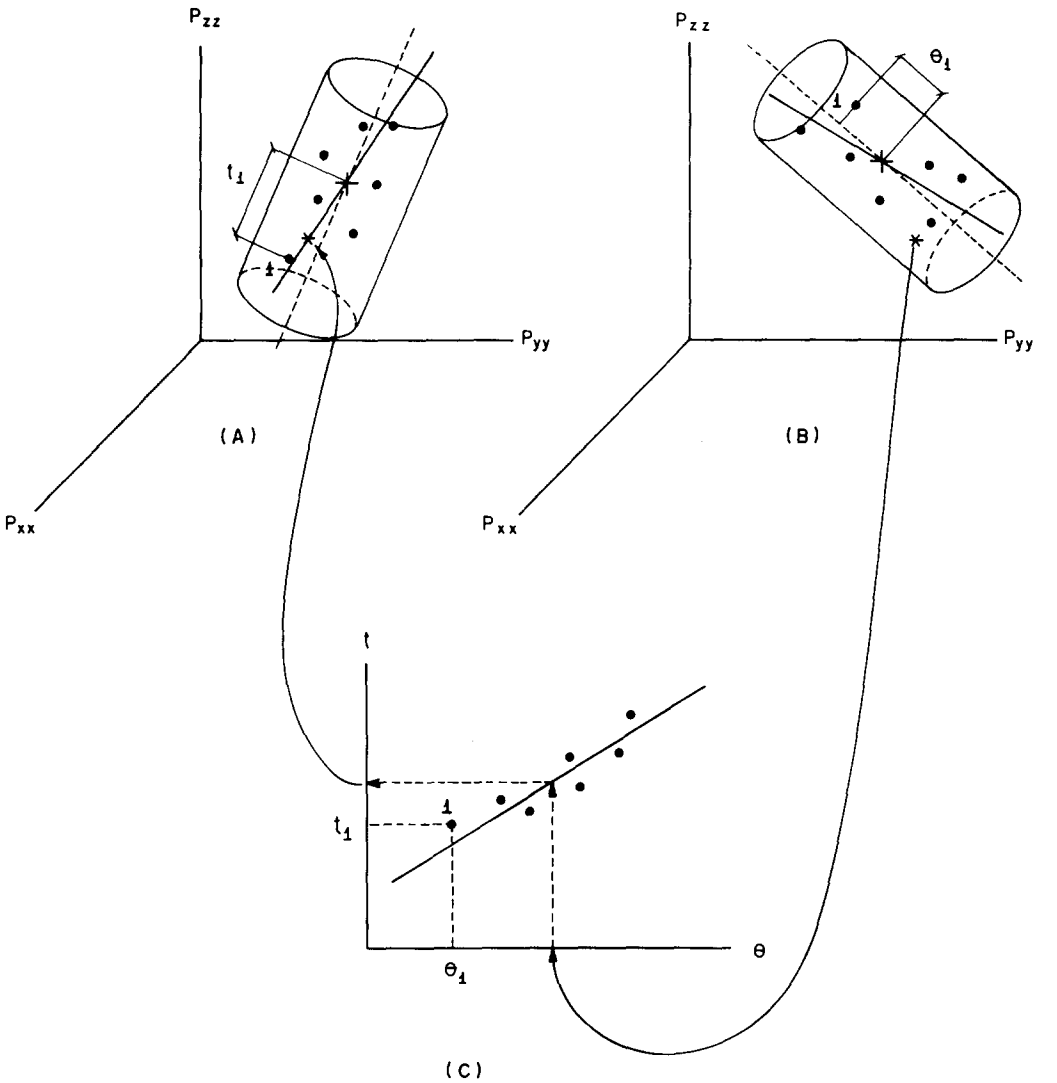

FIG. 2. One component PC models for diagonal polar tensors. (a) Experimental polar tensors; (b) theoretical polar tensors; (c) regression of the experimental scores on the theoretical scores.

Similarly, with the residuals of the tensors used for model building a "typical" variance can be defined

$$S_0^2 = \sum_k \sum_{\sigma,\nu} \frac{\epsilon_{k,\sigma\nu}^2}{(M-A)(n-A-1)}. \tag{9}$$

Assuming a reasonably normal distribution, both variances can be compared in an F test with $(M-A)$ and $(M-A)(n-A-1)$ degrees of freedom,

$$F = S_{n'}^2/S_0^2. \tag{10}$$

An F value showing that the residual variance of tensor n' is significantly larger than the variance of the model set is an indication that the predictions of the experimental polar tensor elements for atom n' are unreliable.

For the applications reported here and, at present, for other attempts to predict IR gas phase intensities using similarity modeling, not only is the F test scheme very approximate, but the calculated model itself is quite unstable and vulnerable to changes in the model-building set of polar tensors. This is the case because, for statistical purposes, IR intensities have been measured for very few molecules, and one is then forced to use half a dozen or less polar tensors to define the similarity model.

Although it is not possible to ascertain the stability of the models calculated here, confidence in the predicted results is improved by the fact that several models result in similar values for the predicted polar tensors. Moreover, the alternative of direct transference, being a special case of PC modeling, suffers more acutely from the same shortcomings.

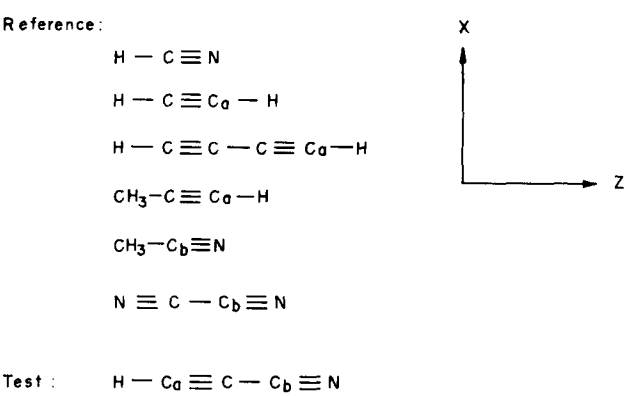

FIG. 3. Molecules used in the PLS prediction of the cyanoacetylene polar tensors.

TABLE I. Experimental and *ab initio* theoretical polar tensor reference values. Units of e; coordinate system as in Fig. 3.

Molecule	Atom	Experimental		STO-3G		4-31G	
		$p_{xx}=p_{yy}$	p_{zz}	$p_{xx}=p_{yy}$	p_{zz}	$p_{xx}=p_{yy}$	p_{zz}
HCN[a]	H	0.237	0.218	0.208	0.393	0.321	0.282
	C	0.084	−0.292	0.042	−0.495	−0.032	−0.103
	N	−0.321	0.074	−0.250	0.102	−0.289	−0.179
C_2H_2[b]	H	0.205	0.183	0.115	0.358	0.253	0.216
	C_a	−0.205	−0.183	−0.115	−0.358	−0.253	−0.216
C_4H_2[c]	H	0.200	0.238	0.122	0.469	0.257	0.281
	C_a	−0.223	−0.394	−0.112	−0.276	−0.262	−0.403
CH_3CCH[d]	H	0.211	0.180	0.119	0.372	0.256	0.203
	C_a	−0.249	−0.465	−0.174	−0.480	−0.301	−0.781
CH_3CN[e]	N	−0.331	−0.171	−0.287	0.018	−0.313	−0.464
	C_b	0.060	0.078	0.118	−0.279	0.025	0.176
C_2N_2[f]	N	−0.245	0.123	−0.231	0.344	...	...
	C_b	0.245	−0.123	0.231	−0.344	...	...

[a] K. Kim and W. T. King, J. Chem. Phys. **71**, 1967 (1979).
[b] K. Kim and W. T. King, J. Mol. Struct. **57**, 201 (1979).
[c] Calculated using the intensities reported in Th. Koops, T. Visser, and W. M. A. Smit, J. Mol. Struct. **125**, 179 (1985).
[d] P. Jona, M. Gussoni, and G. Zerbi, J. Phys. Chem. **85**, 2210 (1981).
[e] Y. Koga, S. Kondo, S. Saëki, and W. B. Person, J. Phys. Chem. **88**, 3152 (1984).
[f] K. Kim and W. T. King, J. Chem. Phys. **80**, 974 (1984).

With increasing amounts of experimental data, similarity modeling of atomic polar tensors will rest on a more solid footing. Indeed, similarity modeling calculations can indicate which experimental measurements should be given highest priority to yield the most secure prediction of polar tensor values.

AN APPLICATION: CYANOACETYLENE

Similarity models were constructed using the linear molecules shown in Fig. 3, the molecular axis being taken as the z axis. The reference (or training) set comprises the molecules employed in the model-building phase. The cyanoacetylene molecule, HC_3N, for which predictions are made, constitutes the test set. The experimental polar tensors of the reference set molecules consistent with the Cartesian system of Fig. 3 are presented in Table I. Where necessary, the experimental polar tensors taken from the literature were rotated to conform to the orientations shown in Fig. 3.

Included in Table I are the *ab initio* STO-3G and 4-31G calculated polar tensor elements for all of the reference set molecules. The coordinate system is the same as the one shown in Fig. 3. The MO calculations were performed running the HONDO package[14] on a DEC PDP-10 mainframe, with standard basis sets as included in the computer program. It should be noted that the 4-31G values for the polar tensors of C_2N_2 are not shown in Table I. The reason is that our 4-31G calculations for this molecule met with convergence difficulties (already experienced by King and coworkers[15]) that led to unacceptably large values for the dipole derivatives. For the STO-3G calculations no such troubles were encountered, and STO-3G values are therefore shown for the entire reference set. Experimental equilibrium geometries[16] were used throughout. The PLS calculations were carried out using the SIMCA-3B program[17] on a CPM 64 kbyte–8-bit microcomputer.

A critical stage in the similarity transference procedure is the definition of the similarity models to be used, which are characterized by the molecules included in the reference sets. Figure 3 contains all molecules having either a carbon–carbon or a carbon–nitrogen triple bond for which complete gas phase intensity data are available. Since the test molecule, cyanoacetylene, contains both types of triple bonds, the other molecules in Fig. 3 are qualified as members of the different categories or classes defining the various similarity models.

A class that suggests itself is that of the acid hydrogen atoms in HCN, C_2H_2, C_4H_2, and CH_3CCH. All these molecules have a classical H–C≡ bond arrangement and one therefore anticipates similar electronic structures for these protons. These structures are not equivalent, of course, because of the different moieties in which the triple bond is embedded. On the other hand, since C–H bonds adjacent to triple bonds have protons with effective charges quite different from those of other protons,[18] other hydrogen polar tensors were not considered for inclusion in this class. Another group of similar atomic polar tensors may be defined by the nitrogen tensors from the HCN, CH_3CN, and C_2N_2 molecules, which display a –C≡N classical bond arrangement.

For the inner atoms the similarity models are a bit more difficult to define. Both the acid and nitrogen carbons have a –C≡ arrangement with differing groups participating in the single and triple bonds. For the carbon atom bonded to nitrogen, HCN, C_2N_2, and CH_3CN provide reference atomic polar tensors. For the acid carbon the C_2H_2, HCN, C_4H_2,

and CH_3CCH molecules could form the reference class. However, separation of the carbon polar tensors into two different similarity models may not be warranted, since the carbon from HCN would be common to both models. This can be confirmed by inspection of Fig. 4, in which 4-31G, STO-3G, and experimental atomic polar tensors are represented graphically. In all three plots the hydrogen polar tensors of the reference class from a compact and well-separated cluster. For the STO-3G results three groupings can be distinguished, corresponding to the hydrogen, nitrogen, and carbon polar tensors. On the other hand, the 4-31G carbon and nitrogen tensors together form a large disperse group and probably only two similarity models should be formed, one for the hydrogen tensors and the other for the rest. In Table II the groups of polar tensors actually used to calculate similarity models for these molecules are summarized. One advantage of using a smaller number of models is evident from this table: more polar tensors can be used to calculate the parameters of each model, leading to statistically more stable solutions.

Once the similarity models are established the atomic polar tensors of cyanoacetylene can be calculated using the 4-31G and STO-3G molecular orbital estimates of the atomic polar tensors in Eqs. (5)–(7). In this way polar tensor values for all atoms but the central carbon of HC_3N are obtained. The remaining polar tensor is obtained from the charge conservation condition, i.e., $\Sigma_\alpha \mathbf{P}_X^{(\alpha)} = \mathbf{0}$. The intensity values are proportional to the squares of the elements of the $\mathbf{P}_Q$ tensor, which contains the dipole moment derivatives with respect to the normal coordinates. This tensor is given by[19]

$$\mathbf{P}_Q = \mathbf{P}_X \mathbf{AUL}', \tag{11}$$

where $\mathbf{P}_X$ is an ordered juxtaposition of the atomic polar tensors and the matrix product $\mathbf{AUL}'$ carries out the transformation from Cartesian to normal coordinates. The normal coordinates for HC_3N were calculated using the general valence force field of Uyemura *et al.*[6]

RESULTS AND DISCUSSION

Table III contains the experimental and molecular orbital polar tensor values for cyanoacetylene along with the partial least squares values using 2, 3, and 4 similarity models, as described in Table II. The root mean square errors for both the A_1 and E symmetry species polar tensor elements, as well as the total root mean square errors, are also listed and can be used as a criterion of accuracy for the different types of calculations.

The total root mean square errors for all the PLS calculations are significantly less than the errors for the 4-31G and STO-3G molecular orbital results. Certainly the minimal basis set STO-3G approximation cannot be expected to provide a wave function which accurately reflects the electronic densities for small geometrical distortions from equilibrium, but the more sophisticated 4-31G basis set gives even poorer results, with an rms error of 0.199 e which is larger than the error of 0.167 e for the STO-3G calculations. King and co-workers[15] have previously commented on the deficiencies of the 4-31G wave functions in calculating electron densities for the C≡N bond. In our case this even led to the exclusion of 4-31G values for C_2N_2, as already mentioned.

On the other hand, the PLS estimates, which have significantly smaller rms errors, are based on crude similarity models, because of the dearth of experimental intensity data for appropriate reference molecules. For example, the 4-31G value of $\partial p_z/\partial z_{C_b}$ for cyanoacetylene fits poorly any of the (4-31G)-based PLS models, and should lead to correspondingly poor predictions. This also occurs for the PLS calculation in which three similarity models were constructed for the STO-3G theoretical polar tensor elements. In spite of these deficiencies, the PLS rms errors are (with the exception of the 4-31G model with four classes) at most one-half of the errors encountered with the direct MO calculations.

More sophisticated wave functions can certainly be expected to provide results with lower errors than the ones obtained for the 4-31G and STO-3G calculations. Such wave functions, however, should also give a more faithful description of the similarities of the molecules involved, thus leading to improved similarity models and lower PLS errors.

The STO-3G theoretical values combined with the reference set experimental values result in better PLS models than do the 4-31G values. The best STO-3G PLS results occur for the calculations in which four similarity models,

TABLE II. Groups of atomic polar tensors used to construct similarity models. Entries in parentheses refer to the absence of C_2N_2 from 4-31G models (see the text).

Number of models	Number of tensors in each model	Description of model	Origin of polar tensors
4	4	H	$HCN,C_2H_2,C_4H_2,CH_3CCH$
	3	C_a	C_2H_2,C_4H_2,CH_3CCH
	3(2)	C_b	$HCN,CH_3CN,(C_2N_2)$
	3(2)	N	$HCN,CH_3CN,(C_2N_2)$
3	4	H	$HCN,C_2H_2,C_4H_2,CH_3CCH$
	6(5)	$C_a + C_b$	$HCN,C_2H_2,CH_3CN,C_4H_2,$ $CH_3CCH,(C_2N_2)$
	3(2)	N	$HCN,CH_3CN,(C_2N_2)$
2	4	H	$HCN,C_2H_2,C_4H_2,CH_3CCH$
	9(7)	$C_a + C_b + N$	$HCN,C_2H_2,CH_3CN,C_4H_2,$ $CH_3CCH,(C_2N_2)$

TABLE III. Experimental, molecular orbital, and partial least squares values for the polar tensors of cyanoacetylene (e).

Symmetry species	Atom	Exptl[a]	4-31G MO	4-31G PLS[b]	4-31G PLS[c]	4-31G PLS[d]	STO-3G MO	STO-3G PLS[b]	STO-3G PLS[c]	STO-3G PLS[d]
($P^{A_1}_{zz}$)	H	0.267	0.306	0.263	0.263	0.263	0.516	0.261	0.261	0.261
	C_a	−0.168	−0.150	−0.157	−0.189	0.076[e]	−0.135	−0.078	−0.090	−0.163
	C_b	0.168	0.583	0.146[e]	0.187[e]	−0.107[e]	−0.050	0.155	0.149[e]	0.152
	N	−0.085	−0.454	−0.240	−0.049	−0.049	0.183	0.017	−0.005	−0.005
	rms error[f]		0.278	0.078	0.023	0.185	0.214	0.068	0.056	0.041
($P^{E}_{xx} = P_{yy}$)	H	0.277	0.235	0.183	0.183	0.183	0.119	0.195	0.195	0.195
	C_a	−0.175	−0.125	−0.075	−0.067	−0.152[e]	−0.055	−0.091	−0.224	−0.184
	C_b	0.217	0.216	0.312[e]	0.298[e]	0.072[e]	0.167	0.178	0.034[e]	0.230
	N	−0.260	−0.307	−0.294	−0.326	−0.326	−0.266	−0.321	−0.303	−0.303
	rms error[f]		0.040	0.085	0.088	0.093	0.102	0.069	0.105	0.047
	Total rms error[f]		0.199	0.082	0.065	0.146	0.167	0.069	0.085	0.044

[a] Data from Ref. 21.
[b] Two similarity models used, as described in Table II.
[c] Three similarity models used.
[d] Four similarity models used.
[e] The PLS residual values indicate that these are unreliable estimates.
[f] $(\text{rms error})^2 = (1/N)\Sigma_i [(P_{ii})_{\text{exptl}} - (P_{ii})_{\text{calcd}}]^2$.

one for each atom of cyanoacetylene (except the central atom), are used. The total rms error in this case, 0.044 e, is about one-fourth of the error resulting from direct use of the STO-3G polar tensor values.

In all the partial least squares calculations the hydrogens formed a class by themselves, as suggested by the plots in Fig. 4. For $\partial p_z/\partial z_H$ the almost identical PLS values of 0.263 and 0.261 e are obtained from the two kinds of similarity models, whereas direct 4-31G and STO-3G calculations yield quite different values, 0.306 and 0.516 e, respectively. A similar situation exists for the $\partial p_x/\partial x_H$ values: the PLS values, 0.183 and 0.195 e, are very close, and the corresponding MO values, 0.235 and 0.119 e, are not. The use of the reference set experimental values thus appears to have a stabilizing effect on the PLS calculated results.

The results for the $\partial p_z/\partial z_N$ element illustrate the ability of the partial least squares method to provide accurate results even when poorly estimated molecular orbital values are occasionally used. In this case the 4-31G and STO-3G values of -0.454 and $+0.183$ e differ by more than 0.6 e and predict opposite signs for the C≡N stretching derivative. Such discrepancies have been observed for this derivative in other molecules, the best known example being HCN, for which only configuration interaction calculations were capable of resolving the $\partial p/\partial r_{CN}$ sign ambiguity.[20] Regardless of whether 4-31G or STO-3G values are used for modeling, however, the PLS results are all close to zero, with the exception of the -0.240 e value obtained when C_a, C_b, and N were all included into a very diffuse 4-31G model. These results are in excellent agreement with the small negative experimental value, which predicts a $\leftarrow \overset{+}{C} \equiv \overset{-}{N} \rightarrow$ change in charge distribution upon stretching. For the bending coordinate, $\partial p_x/\partial x_N$, the PLS, 4-31G and STO-3G values are very similar and in good agreement with the experimental result.

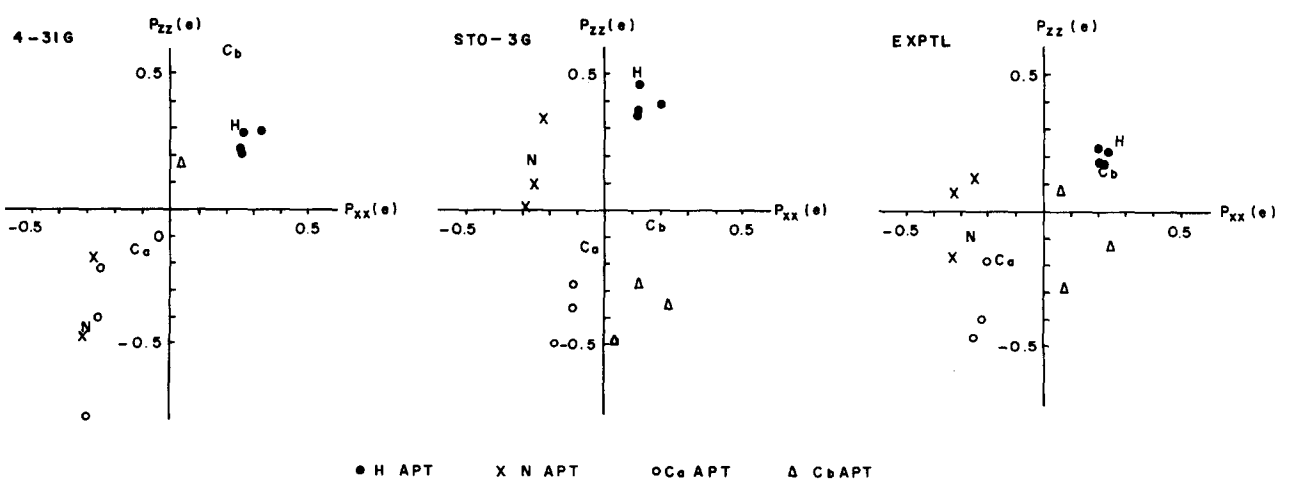

FIG. 4. Graphical representations of polar tensor elements of cyanoacetylene. (a) 4-31G values; (b) STO-3G values; (c) experimental values. The Latin capitals correspond to HC_3N values.

TABLE IV. Experimental and calculated vibrational intensities of cyanoacetylene (km mol^{-1}).

Symmetry species	i	ν_i (cm^{-1})	A_i (exptl)	A_i (4-31G)	A_i (STO-3G)	A_i (9s5p1d/3s2p)[a]	A_i (PLS)[b]	A_i (PLS)[b]
A_1	1	3327	60.4	81.6	225.4	104.8	61.4	55.7
	2	2272	9.92	59.2	11.0	63.8	8.2	12.0
	3	2077	1.93	0.1	50.2	0.2	4.6	10.2
	4	863	0.06	0.0	0.1	0.0	0.6	0.7
E	5	664	68.4	53.4	14.4	...	41.3	36.5
	6	499	8.0	4.1	1.4	...	1.7	2.1
	7	223	0.18	1.0	2.2	...	3.5	1.6
			rms error[c]	21.1	68.2	34.9	10.7	12.8

[a] Value from Ref. 21.
[b] Based on STO-3G results using four and two similarity models, respectively.
[c] $(\text{rms error})^2 = (1/N)\Sigma_i [(A_i)_{\text{exptl}} - (A_i)_{\text{calcd}}]^2$.

For the $\partial p_z/\partial z_{C_a}$ and $\partial p_x/\partial x_{C_a}$ derivatives all the PLS and molecular orbital values have negative signs, with the exception of the 4-31G PLS estimate indicated to be unreliable by its large residual variance. These values are in reasonable agreement with the -0.168 and -0.175 e experimental elements.

Comparisons are more limited for the values of the C_b polar tensor. For $\partial p_z/\partial z_{C_b}$ only two of the PLS results based on the STO-3G MO values are statistically reliable. These values are 0.152 and 0.155 e, in excellent agreement with the 0.168 e experimental value. For $\partial p_x/\partial x_{C_b}$ again only two reliable PLS values are calculated: 0.178 and 0.230 e, using STO-3G values with 2 and 4 similarity models, which compare well with the experimental value of 0.217 e.

In Table IV vibrational intensities calculated for cyanoacetylene using the polar tensors in Table III and Eq. (11) are presented. Intensities corresponding to polar tensor sets for which the PLS method indicated at least one value to be unreliable have not been included in this table.

The root mean square errors for the PLS estimates of the intensities are about half the size of the errors resulting from direct calculation with 4-31G polar tensors. Although the most accurate PLS models were obtained with STO-3G polar tensors, the STO-3G wave function itself leads to poorer intensity values for HC_3N than does the 4-31G wave function. The 4-31G intensity values are in good agreement with the experimental intensities, except for the ν_2 band, for which the 4-31G result, 59.2 km mol^{-1}, is six times the 9.92 km mol^{-1} experimental value. Since this band is assigned to the CN stretching normal mode, this discrepancy is easily attributed to the too large absolute magnitudes calculated for $\partial p_z/\partial z_N$ and $\partial p_z/\partial z_{C_b}$ of cyanoacetylene. All the other polar tensor elements calculated from the 4-31G wave function are in excellent agreement with the experimental values, leading to the accuracy observed for the remaining 4-31G intensities.

Included for comparison are intensity values for the A_1 symmetry species calculated from a 9s5p1d/3s2p wave function.[21] The ν_2 intensity is also badly in error in this case. On the other hand, the PLS estimates of this intensity value are extremely good, differing by only 2 km mol^{-1} ($\sim$20%) from the experimental value. The relative values of the other PLS estimated intensities are also very encouraging, although the ν_5 intensity is calculated to be about one-half as strong as the ν_1 band, whereas in practice A_5 is a little higher than A_1.

FINAL CONSIDERATIONS

Direct transference of atomic polar tensors can be considered as a special case of the transference procedure based on similarity model calculations. This similarity transference procedure can be extended to other molecular parameters such as force constants, bond lengths and angles, etc. Especially intriguing is the possibility of employing different types of parameters (for example, force constants and polar tensors or electro-optical parameters) in the same calculation for simultaneously predicting fundamentally different but related properties (the frequencies and intensities of the fundamental bands of the gas phase infrared spectra).

The PLS calculations are a valuable complement to the molecular orbital calculations of molecular parameters. As a consequence, a library of theoretical and experimental parameters is created and can be continually updated. Indeed the modeling procedure can be used to establish priorities about parameter measurements necessary for more precise predictions. As the molecular orbital estimates of the parameters improve one can also expect more accurate PLS determinations. Since the PLS method is a variable reduction technique, less data are necessary for reliable estimation than in the case of multiple regression, which could also be used to relate experimental and theoretical reference data. With sufficient data, parametric statistical approaches could be included in the similarity transference method.

ACKNOWLEDGMENTS

The authors gratefully acknowledge partial financial support from the Conselho Nacional de Desenvolvimento Científico e Tecnologico (M. N. R. and B. B. N.), the Fundação de Amparo à Pesquisa do Estado de São Paulo (R. E. B.) and the Financiadora de Estudos e Projetos (R. E. B.).

[1] J. Applequist, Acc. Chem. Res. **10**, 79 (1977); J. Appelquist and C. O. Quicksall, J. Chem. Phys. **66**, 3455 (1977).

[2]L. Pauling, *The Nature of the Chemical Bond*, 3rd ed. (Cornell University, Ithaca, 1960).
[3]I. M. Mills, in *Infrared Spectroscopy and Molecular Structure*, edited by M. Davies (Elsevier, Amsterdam, 1963), Chap. 5.
[4]K. V. Mardia, J. T. Kent, and J. M. Bibby, *Multivariate Analysis* (Academic, London, 1979), Chap. 8.
[5]H. Wold, in *Systems Under Indirect Observation*, edited by K. G. Jöreskog and H. Wold (North-Holland, Amsterdam, 1982), Vol. II, Chap. 1.
[6](a) M. Uyemura and S. Maeda, Bull. Chem. Soc. Jpn. **47**, 2930 (1974); (b) M. Uyemura, S. Deguchi, Y. Nakada, and T. Onaka, *ibid.* **55**, 384 (1982).
[7]S. Wold, C. Albano, W. J. Dunn III, U. Eudland, D. Esbensen, P. Geladi, S. Hellberg, E. Johansson, W. Lindberg, and M. Sjöström, in *Chemometrics: Mathematics and Statistics in Chemistry*, edited by B. R. Kowalski (Reidel, Dordrecht, 1984), Chap. 2.
[8]S. Wold, Pattern Recognition **8**, 127 (1976).
[9]W. B. Person and J. Overend, J. Chem. Phys. **66**, 1443 (1977); B. J. Krohn, W. B. Person, and J. Overend, *ibid.* **65**, 969 (1976).
[10]S. Wold, Technometrics **20**, 397 (1978).
[11]T. M. Cover and P. E. Hart, IEEE Trans. Info. Theory **IT-13**, 21 (1967).
[12]W. S. Meisel, *Computer Oriented Approach to Pattern Recognition* (Academic, London, 1972).
[13]M. Sjöström, S. Wold, W. Lindberg, J. A. Persson, and H. Martens, Anal. Chim. Acta **150**, 61 (1983).
[14]M. Dupuis, J. Rys, and H. F. King, Quantum Chemistry Program Exchange **403** (1978).
[15]K. Kim and W. T. King, J. Chem. Phys. **80**, 974 (1984); See also, Y. Koga, S. Kondo, S. Saëki, and W. B. Person, J. Phys. Chem. **88**, 3152 (1984).
[16]I. Suzuki, M. A. Pariseau, and J. Overend, J. Chem. Phys. **44**, 3561 (1966); G. Strey and I. M. Mills, J. Mol. Spectrosc. **59**, 103 (1976); L. Halonen and I. M. Mills, *ibid.* **73**, 494 (1978); M. Tamimoto, K. Kuchitsu, and Y. Morino, Bull. Chem. Soc. Jpn. **44**, 386 (1971); J. L. Duncan, D. C. McKean, P. D. Mallinson, and R. McCullock, J. Mol. Spectrosc. **46**, 232 (1973); M. Tamimoto, K. Kuchitsu, and Y. Morino, Bull. Chem. Soc. Jpn. **42**, 2519 (1969); for C_2N_2, L. E. Sutton, *Supplement to Tables of Interatomic Distances* (Chemical Society, London, 1963); C. C. Constain, J. Chem. Phys. **29**, 864 (1968).
[17]Principal Data Components, 2505 Shepard Blvd., Columbia, MO; Sepanov AB, Ostrand 14, Enskede, Sweden.
[18]W. T. King and G. B. Mast, J. Phys. Chem. **80**, 2521 (1976); R. E. Bruns and B. B. Neto, J. Chem. Phys. **68**, 847 (1978).
[19]W. B. Person and J. H. Newton, J. Chem. Phys. **61**, 1040 (1974).
[20](a) B. Lin, K. M. Sando, C. S. North, H. B. Friedrich, and D. M. Chipman, J. Chem. Phys. **69**, 1425 (1978); (b) J. E. Gready, G. B. Backsay, and N. S. Hush, Chem. Phys. **31**, 467 (1978); (c) A. B. M. S. Bassi, J. Chem. Phys. **68**, 5667 (1978); (d) D. F. Hornig, *ibid.* **68**, 5668 (1978); (e) R. E. Bruns and W. B. Person, *ibid.* **53**, 1413 (1970).
[21]M. N. Ramos, B. B. Neto, R. E. Bruns, and O. M. Herrera, J. Mol. Struct. **142**, 209 (1986).