Journal of Mathematical Chemistry Vol. 29, No. 1, 2001

On the optimal selection of principal components in QSPR studies

Emili Besalú [a] and Leonel Vera [b]

[a] *Department of Chemistry and Institute of Computational Chemistry, University of Girona, Campus de Montilivi, 17071 Girona, Catalonia, Spain*
E-mail: emili@iqc.udg.es

[b] *Department of Chemistry, Universidad Católica del Norte, Avenida Angamos 0610, Casilla 1280, Antofagasta, Chile*
E-mail: lvera@ucn.cl

Received 31 October 2000

A heuristic method to sort principal components is analysed. The obtained arrangements are property dependent and it is demonstrated how the procedure is equivalent to the called Most Predictive Variable Method. As an application of the new algorithm, a Quantitative Structure–Property Relationships (QSPR) study is performed over the set of the 18 structural isomers of the octane molecule. The original molecular descriptors are obtained from a quantum similarity matrix related to the molecular family. The analysis is based on the use of linear models where distinct sets of principal components act as optimal descriptors for 6 physicochemical molecular properties. The proposed algorithm allows to determine sequences of the first Principal Components which are identified as forming the optimal descriptors set for each of the 6 studied properties. The benefits of the new approach are revealed when comparing the obtained results with classical ones arising from a standard principal component analysis study.

KEY WORDS: QSAR, sorting principal components, most predictive variable method, cross-validation, quantum similarity

1. Introduction

An important field of investigation in contemporaneous chemistry is based on the prediction of molecular properties, either physicochemical or biological. Within this paradigm, the QSAR/QSPR (Quantitative Structure–Activity or Structure–Property Relationships) fields are one of the most known, see for example the reviews [1–7] and the references cited there. Many times, the related methods are based on the use of molecular descriptors and it is very frequent to manipulate the parameters using linear techniques as Multilinear Regression [8], Principal Component Analysis (PCA) [9–11], Partial Least Squares (PLS) [12] or the computation of the so-called $q^{(2)}$ coefficient [13].

Here is presented a simple and useful algorithm to sort the principal components (PCs) in a property dependent fashion. In the first part of this work, it is demonstrated

0259-9791/01/0100-0021$19.50/0

how the Most Predictive Variable Method (MPVM) of Cuadras and co-workers [14,15] degenerates in a simple and heuristic criterion: to sort the PCs according to the respective squared correlation coefficient with respect to the molecular properties vectors. In this way, it will be shown how PCs that are not related to a maximum reservoir of variance can be, instead, well conditioned to enter in a linear scheme capable to grasp a notable amount of "property variance". Of course, if different properties or molecular families are studied, the sorting of the PCs is different too.

As an application example, the family of 18 structural isomers of the octane molecule will be studied and a set of 6 physicochemical properties will be correlated. The theoretical molecular parameters come from the manipulation of a Molecular Quantum Similarity Matrix (MQSM, see [16–21] and appendix). The final linear equations depend on different sets of PCs. Comparisons will be done between the classical PCA results and the ones coming from the proposed algorithm.

2. The heuristic method

In a previous article [22], a methodology for the prediction of molecular properties was described. The molecular descriptors, previously obtained from a molecular quantum similarity matrix, were pre-processed with the PCA technique [9–11]. The final treatment was carried out using neural networks [23].

In the present work, the same molecular quantum similarity matrix is used. The goal is to manipulate the PCs in order to enhance its performance in QSPR studies. At the end, it will be shown how using only linear methods, the results are substantially improved respect to the original article. The new proposed algorithm establishes a new PCs selection criterion. The PCs sorting is oriented to the molecular property under study.

It is not our intention to affirm that a linear method will be always better than another one based on neural networks. Instead, we promote the idea that a method such as the PCA, which has nothing to do with the molecular properties, can be oriented to them in a similar way as it is done when using the PLS technique. That is, PCs related to a small or intermediate amount of data variance are best candidates to correlate (and predict) molecular properties and enter into linear QSPR models.

Given a set of n molecules and m descriptors forming an $n \times m$ matrix $\boldsymbol{A}$, the related m PCs (loadings) generates new molecular coordinates (scores). Within the classical and standard approach in order to reduce the problem dimension, $p < m$ first components attached to the p biggest eigenvalues are selected. Then, the molecular family is understood to be described for a new $n \times p$ matrix $\boldsymbol{C}$. Usually, this new data matrix enter into the QSPR treatment and the same coordinates are used to classify different molecular properties of a given molecular family. Thus, the selection of molecular descriptors is performed irrespective to the properties.

Here, it is proposed to use a distinct matrix $\boldsymbol{C}$ for every molecular property vector. The matrix $\boldsymbol{C}$ will contain, for every property, a different set of PCs. It follows the

algorithm that allows, given an n-dimensional molecular property vector $\boldsymbol{y}$, the selection of the optimal and property-oriented principal descriptors:

1. Define the original $n \times m$ descriptors matrix $\boldsymbol{A}$.
2. Obtain the m PCs: $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m)$.
3. Give a molecular property vector $\boldsymbol{y}$.
4. Compute the squares of the m linear correlation coefficients between every principal component, $\boldsymbol{x}_i$, and the vector $\boldsymbol{y}$: r_i^2, $i = 1, 2, \ldots, m$.
5. Sort the PCs according to the decreasing value of the r_i^2 coefficients.
6. Give the number of PCs to be considered: p.
7. Choose the first p vectors obtained in step 5.

In this algorithm, the selected descriptors are orthogonal because they are PCs but this is not compulsory because the method can be applied to any kind of column vector descriptors. In the first case, which appears to be the most common choice, the computed squared correlation coefficients $\{r_i^2\}$ are additive. That is, if r_i^2 and r_j^2 are coefficients associated to two different and orthogonal vectors ($i \neq j$), a linear regression model involving both vectors will give a squared correlation coefficient equal to $R^2 = r_i^2 + r_j^2$. This is a general result that can be extended to any set of more than two original orthogonal column descriptors.

3. Equivalence with Cuadras' method

The sorting method proposed here, despite to be initially derived heuristically, has a solid mathematical foundation. In this section will be demonstrated how the presented method always generates the same eigenvectors sorting as the method opposed by Cuadras and co-workers [14,15]. The original formulation of Cuadras defines the following *predictavility coefficient* within the Most Predictive Variable Method (MPVM) approach:

$$\chi^2(\boldsymbol{y}, \boldsymbol{x}_i) = \frac{(\boldsymbol{y}^{\mathrm{T}}\boldsymbol{x}_i)^2}{\sum_j (y_j - \overline{y})^2 \lambda_j}. \tag{1}$$

Here, the descriptors column vector, $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots)^{\mathrm{T}}$ is a transformed principal component normalised and with a null mean value. The column vector $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^{\mathrm{T}}$ contains the molecular properties. The scalar $\overline{y}$ is the mean value of the $\boldsymbol{y}$ vector components and the λ_j terms are the eigenvalues attached to the original eigenvectors.

If the scale and origin of the $\boldsymbol{y}$ vector are changed, the new obtained one, $\boldsymbol{y}'$, is related to the former by a simple expression:

$$\boldsymbol{y}' = a\big(\boldsymbol{y} - \overline{y}\mathbf{1}\big),$$

where **1** represents the column vector having all the elements equal to the unity and the parameter a stands for a normalisation constant. As the mean value of the elements of $\boldsymbol{y}'$ vector is null, $\overline{y}' = 0$, the new expression for χ^2 in terms of the vector $\boldsymbol{y}'$ would be

$$\chi^2(\boldsymbol{y}', \boldsymbol{x}_i) = \frac{(\boldsymbol{y}'^{\mathrm{T}}\boldsymbol{x}_i)^2}{\sum_j (y'_j)^2\lambda_j}. \tag{2}$$

This allows to demonstrate that the parameter χ^2 is invariant to the change of scale due to the multiplication by an arbitrary constant a:

$$\chi^2\big(\boldsymbol{y}', \boldsymbol{x}_i\big) = \frac{(a(\boldsymbol{y} - \overline{y}\,\mathbf{1})^{\mathrm{T}}\boldsymbol{x}_i)^2}{\sum_j (ay_j - a\overline{y})^2\lambda_j} = \frac{((\boldsymbol{y} - \overline{y}\,\mathbf{1})^{\mathrm{T}}\boldsymbol{x}_i)^2}{\sum_j (y_j - \overline{y})^2\lambda_j} = \chi^2\left(\frac{\boldsymbol{y}'}{a}, \boldsymbol{x}_i\right).$$

Also, from the previous expression,

$$\chi^2\big(\boldsymbol{y}', \boldsymbol{x}_i\big) = \chi^2\left(\frac{\boldsymbol{y}'}{a}, \boldsymbol{x}_i\right) = \frac{(\boldsymbol{y}^{\mathrm{T}}\boldsymbol{x}_i - \overline{y}\mathbf{1}^{\mathrm{T}}\boldsymbol{x}_i)^2}{\sum_j (y_j - \overline{y})^2\lambda_j}$$

and due to the fact that the elements of the vector x_i have null mean value, $0 = \sum_j \boldsymbol{x}_{ij}$, the scalar product $\mathbf{1}^{\mathrm{T}}\boldsymbol{x}_i$ is zero:

$$\mathbf{1}^{\mathrm{T}}\boldsymbol{x}_i = (1\ 1\ \ldots\ 1)\begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{in} \end{pmatrix} = \sum_j x_{ij} = 0.$$

Then,

$$\chi^2\big(\boldsymbol{y}', \boldsymbol{x}_i\big) = \chi^2\left(\frac{\boldsymbol{y}'}{a}, \boldsymbol{x}_i\right) = \frac{(\boldsymbol{y}^{\mathrm{T}}\boldsymbol{x}_i)^2}{\sum_j (y_j - \overline{y})^2\lambda_j}$$

and

$$\chi^2\big(\boldsymbol{y}', \boldsymbol{x}_i\big) = \chi^2\left(\frac{\boldsymbol{y}'}{a}, \boldsymbol{x}_i\right) = \chi^2(\boldsymbol{y}, \boldsymbol{x}_i).$$

In consequence, the predictavility coefficient χ^2 is invariant to both, the shift and scaling of the vector $\boldsymbol{y}$.

If the constant a is chosen to be a normalisation factor, one can interpret the formula (2) as a *canonical form* to express the predictavility coefficient. Its denominator is a constant related to every molecular family and also to every property vector. In other words, given a molecular family and a property vector, the coefficient χ^2 is *proportional* to the square of the scalar product $\boldsymbol{y}'^{\mathrm{T}}\boldsymbol{x}_j$. At the same time, and due to the fact that the vectors $\boldsymbol{y}'$ and $\boldsymbol{x}_i$ are normalised, this scalar product coincides with the linear correlation coefficient r_i between the respective elements. So, finally one is able to write the following relations:

$$\chi^2(\boldsymbol{y}, \boldsymbol{x}_i) = \chi^2\big(\boldsymbol{y}', \boldsymbol{x}_i\big) \propto \big(\boldsymbol{y}'^{\mathrm{T}}\boldsymbol{x}_i\big)^2 = r_i^2,$$

and conclude that Cuadras' coefficient is merely proportional to the correlation coefficient r_i^2. The PCs sorting imposed by the use of the $\chi^2(\boldsymbol{y}, \boldsymbol{x}_i)$ term is exactly the same as the obtained according to the values of r_i^2. Our proposal is to use this squared linear correlation coefficient as a parameter to measure the predictive power of a descriptor. Some of the advantages of this procedure are:

- It gives the same descriptor ordering than the coefficient χ^2.
- It is more intuitive and directly interpretable.
- It is easier to compute than χ^2 because there is no need of diagonalising any matrix. It is not necessary to know any eigenvalue.
- It is not forced to use it over a set of descriptors having the property to be orthogonal or to be eigenvectors of a matrix.
- If the descriptors are orthogonal (independently if they come from a process of diagonalisation or not), the numerical values r_i^2 are additive. The sum coincides with the squared coefficient R^2 that will eventually be obtained if a multilinear regression computation is performed.

4. Results and discusion

The PCs sorting method has been tested in the study of the 18 octane structural isomers. Six physicochemical properties have been considered: Gibbs free energy of formation, enthalpy of formation, standard entropy, normal boiling point, density and refraction index. The molecules are identified in table 1, together with the experimental data properties [24,25].

Figures 1 and 2 present the collection of representations showing as the independent variable the number of PCs employed in the linear correlation study. The dependent parameters are: R^2 (squared correlation coefficient coming from a multilinear fitting), r^2_{cv} (squared correlation coefficient), $s_{n,\mathrm{cv}}$ (relative standard deviation) and f_{cv} (statistical significance according to the Fisher test [26]). All these parameters, except the first one, refer to cross-validation calculations.

Figure 1 presents the graphs obtained when the PCs addition is performed following the classical criteria of retention of maximal data variance. This classical ordering defines what we call the *canonical order* (CO) for the PCs. Figure 2 reflects the same kind of results but this time the PCs are entered following the sequence established by the proposed algorithm. As this new PCs sorting is property dependent, we call it a *specific order* (SO). To denote a SO, the sequence of vectors is indicated using the numbering they have in the CO.

The statistical parameter $s_{n,\mathrm{cv}}$ has been defined for each property as the standard deviation for a given number of PCs (this applies for both the CO and the SO) divided by the maximum value obtained relative to the same property and during the cross-validation process. As the definition of $s_{n,\mathrm{cv}}$ is relative, this allows to perform rapid visual inspections along the graphs: the parameter always achieves its maximum value in one of the graphs in figure 1 and never in figure 2.

Table 1
Experimental data for the 18 structural octane isomers. The symbols, from left to right, are: free energy of formation, enthalpy of formation, standard entropy, normal boiling point, density at 20°C and refractive index at 20°C. The three first data values come from [24] and the last three ones were taken from [25].

	Molecule	ΔG_f° (Kcal/mol)	ΔH_f° (Kcal/mol)	S° (Kcal/mol K)	T_b (°C)	ρ (g/ml)	n
1	n-octane	4.14	−49.82	110.82	126	0.703	1.3974
2	2,2,3,3-tetramethylbutane	4.88	−53.99	94.34	106	0.824	1.4695
3	2-methyl-3-etylpentane	5.08	−50.48	105.43	116	0.719	1.4040
4	3-methyl-3-etylpentane	4.76	−51.38	103.48	118	0.727	1.4078
5	2,2,3-trimethylpentane	4.09	−52.61	101.62	110	0.716	1.4030
6	2,2,4-trimethylpentane	3.13	−53.57	101.62	99	0.692	1.3915
7	2,3,3-trimethylpentane	4.52	−51.73	103.14	115	0.726	1.4075
8	2,3,4-trimethylpentane	4.32	−51.97	102.99	113	0.719	1.4042
9	2,3-dimethylhexane	4.23	−51.13	106.11	116	0.712	1.4011
10	2,4-dimethylhexane	2.80	−52.44	106.51	109	0.700	1.3929
11	2,5-dimethylhexane	2.50	−53.21	104.93	109	0.694	1.3925
12	3,3-dimethylhexane	3.17	−52.61	104.70	112	0.710	1.4001
13	3,4-dimethylhexane	4.14	−50.91	107.15	118	0.719	1.4041
14	2,2-dimethylhexane	2.56	−53.71	103.06	107	0.695	1.3935
15	3-ethylhexane	3.95	−50.40	109.51	119	0.714	1.4016
16	2-methylheptane	3.06	−51.50	108.81	118	0.698	1.3949
17	4-methylheptane	4.00	−50.69	108.35	118	0.705	1.3979
18	3-methylheptane	3.29	−50.82	110.32	119	0.706	1.3985

Respect to the results evaluation, the following characteristics and general trends can be mentioned:

- When considering the process of multilinear fitting, in order to reproduce a value of R^2 greater than 0.8 using the CO, at least the first 12 PCs must be chosen. At the contrary, if the SO approach is considered, only 4, 5 or 6 PCs must be used. Of course, this result is a consequence of the additive properties of the r_i^2 parameters mentioned above and to the algorithm definition: the method is designed to choose the vectors that contribute most to the value of R^2. But this is a first warning: this shows that the CO is not optimal to carry out linear fittings and, by extension, it is expected to be a bad choice to start a cross-validation process.
- Using the SO and less than 16 PCs, the relative standard deviation ($s_{n,\mathrm{cv}}$) never overcomes the value of 0.1, irrespective to the studied property.
- If an optimal number of PCs is chosen, the statistical significance test parameter (f_{cv}) is less that 0.001 (significance level units) when the SO is considered. With respect to the CO, the related values are substantially greater, in most cases greater than 0.3 irrespective to the number of PCs. This behaviour is reproduced in all the properties, except for the entropy.
- Perhaps one of the best parameters available to measure the predictive capability is the $q^{(2)}$ coefficient [13]. It constitutes an estimation of the value of the squared

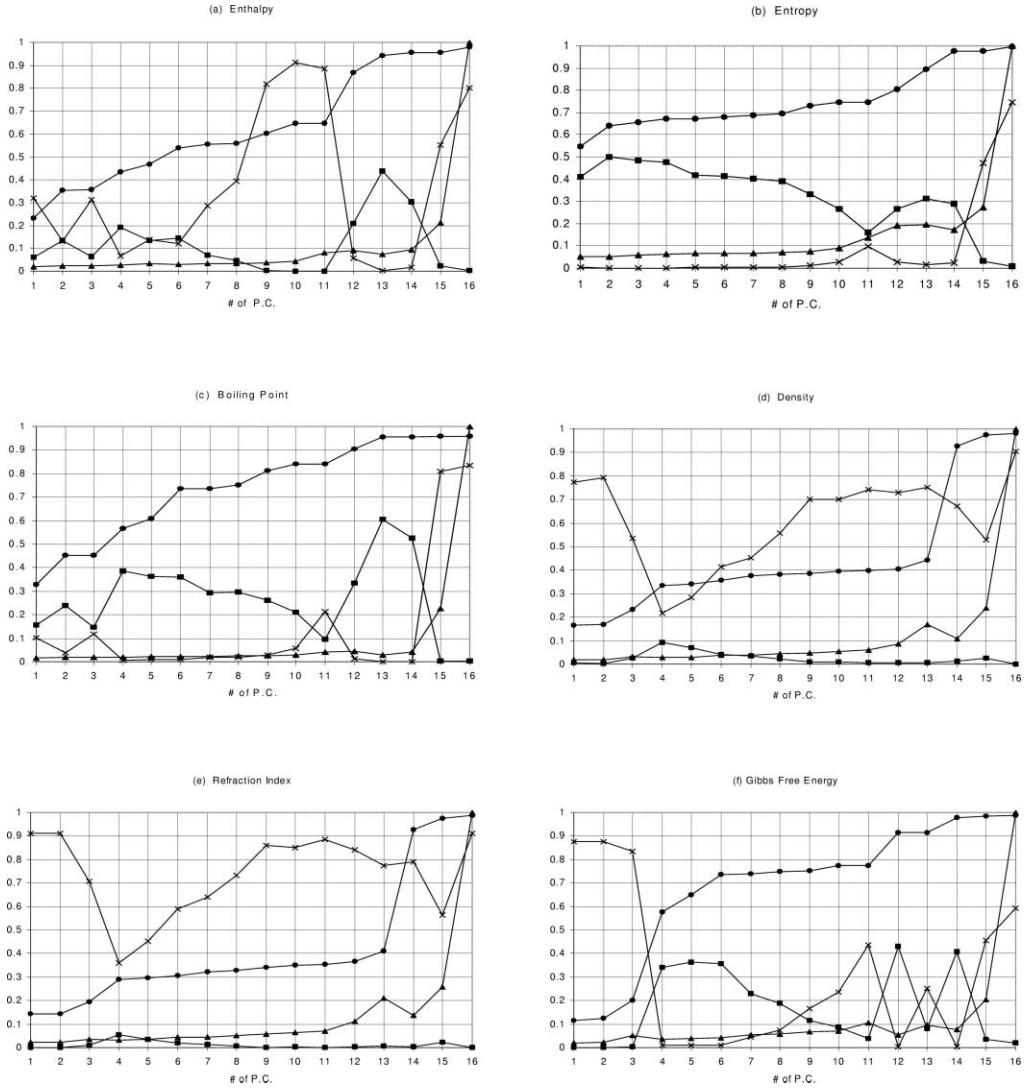

Figure 1. Linear fitting and cross-validation results obtained for the studied family. The independent variable is the number of principal components sorted by the canonical order. The dependent variables are: ● R^2 (square of the correlation coefficient for the multilinear fitting), ■ r^2_{cv} (square of the correlation coefficient resulting from the process of cross-validation), × f_{cv} (level of statistical significance for the cross-validation results, according to Fisher test) and ▲ $s_{n,\mathrm{cv}}$ (relative standard deviation for the process of cross-validation).

correlation coefficient related to the cross-validated data, r^2_{cv}. This last parameter has been computed in this work instead of the approximate one. When using the CO, r^2_{cv} achieves small values. The best value was 0.6 when the normal boiling points were studied. Moreover, the result for 13 PCs is unstable. On the other hand, when the SO is considered, the variable is especially stable for 3 properties (entropy, boiling point and free Gibbs energy). Also, the achieved numerical

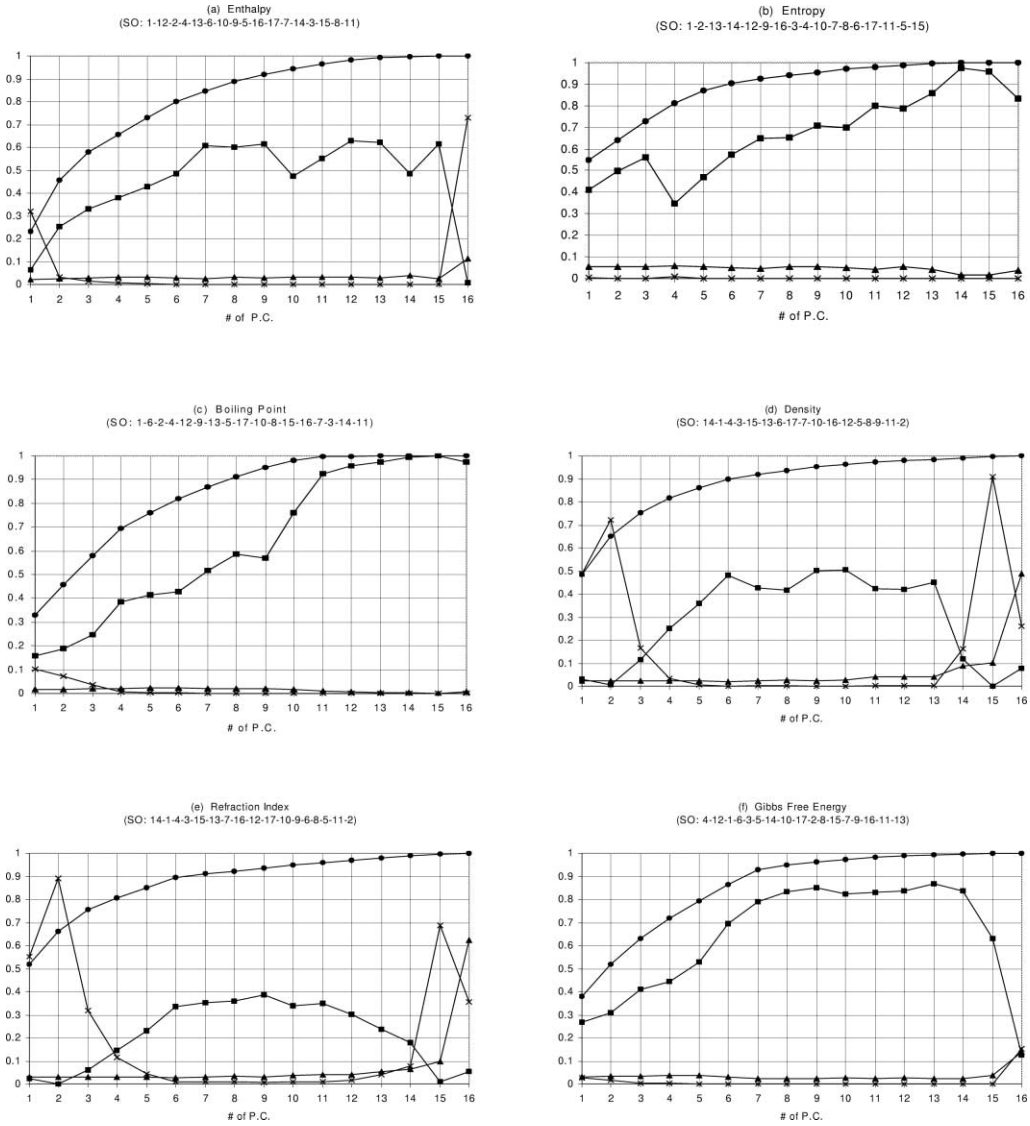

Figure 2. Linear fitting and cross-validation results obtained for the studied family. The independent variable is the number of principal components sorted by the specific order determined by de text algorithm and the respective property. For each graph, the eigenvectors are denoted by means of the numbering of the canonical order. The dependent variables are the same of figure 1.

values are quite good: 0.8 or more. For the other three properties, the obtained values are not optimal for a linear treatment, but, in any case, they are much better than the ones coming when the CO is considered.

From the previous considerations, it can be understood that the SO is most suitable for data preprocessing and selection techniques when generating predictive models for

Table 2
Predicted property values for the octane isomers using the technique of linear cross-validation and the canonical ordering for the first 9 eigenvectors. The mean percentual errors were obtained following the formula $e = \frac{100}{18}\sum_{i=1}^{18}\left|\frac{\text{Experimental value}(i)-\text{Computed value}(i)}{\text{Experimental value}(i)}\right|$.

Molecule	ΔG_f° (Kcal/mol)	ΔH_f° (Kcal/mol)	S° (Kcal/mol·K)	T_b (°C)	ρ (g/ml)	n
1	1.53	−52.76	109.79	116.61	0.68	1.38
2	4.13	−52.55	101.71	107.73	0.71	1.40
3	3.54	−54.18	98.19	105.15	0.73	1.41
4	5.62	−50.49	103.61	119.64	0.75	1.42
5	5.16	−52.30	99.07	111.45	0.77	1.44
6	4.14	−53.08	99.87	107.66	0.76	1.43
7	4.72	−51.59	102.93	116.77	0.74	1.42
8	4.03	−51.35	106.02	113.91	0.70	1.39
9	3.68	−53.08	101.43	106.24	0.74	1.42
10	1.99	−54.44	102.56	102.99	0.70	1.40
11	3.68	−51.54	106.56	110.31	0.71	1.40
12	4.00	−51.74	104.87	113.10	0.77	1.43
13	3.18	−51.96	106.86	119.14	0.71	1.40
14	3.50	−50.78	109.76	118.53	0.69	1.39
15	4.34	−50.79	106.87	123.21	0.74	1.42
16	3.35	−51.66	107.30	116.15	0.69	1.39
17	3.34	−50.75	110.39	119.03	0.69	1.39
18	3.17	−50.98	110.15	113.95	0.66	1.37
Mean percentage error	22.83	2.34	2.53	3.91	4.00	1.26

a training family. Either if the data will enter *a posteriori* in a linear or a non-linear process, the criteria promoted here is better than the classical ones described by Kaiser [10] or Cattell [11], which does not consider the molecular property. As an example, one can consider the results concerning the enthalpy and the free energy: in reference [22] using the first 9 PCs in the CO a small value of R^2 is obtained, $s_{n,\text{cv}} = 0.04$ and very bad values of $f_{\text{cv}} = 0.82$ and $r^2_{\text{cv}} = 0.003$ are reproduced (see figure 1(a)). On the contrary, as it is shown in figure 2(a), if 9 PCs are selected in the SO 1-12-2-4-13-6-10-9-5, then all the parameters improve considerably ($R^2 = 0.92$, $s_{n,\text{cv}} = 0.027$, $f_{\text{cv}} = 0.00012$ and $r^2_{\text{cv}} = 0.61$). Also, comparing the Gibbs free energies and looking to figures 1(f) and 2(f) similar enhancements are obtained using the SO 4-12-1-6-3-5-14-10-17 instead to the canonical one: R^2 changes from 0.75 to 0.96, $s_{n,\text{cv}}$ from 0.06 to 0.02, f_{cv} from 0.17 to less than 0.00001 and r^2_{cv} from 0.12 up to 0.85.

When comparing figures 1 and 2, if the CO is considered the insertion of a new PC in the processes of cross-validation introduces instability among the dependent parameters. On the contrary, when the SO is used, the tendency to the uniformity is achieved. A paradigmatic example of this behaviour is found for the r^2_{cv} coefficient associated to the boiling point property.

Table 3
Predicted property values for the octane isomers using the technique of linear cross-validation and the specific ordering of 9 eigenvectors for every property. The mean percentual errors were obtained using the same formula appearing in table 2.

Molecule	ΔG_f° (Kcal/mol)	ΔH_f° (Kcal/mol)	S° (Kcal/mol·K)	T_b (°C)	ρ (g/ml)	n
1	3.89	−48.39	116.53	124.56	0.71	1.40
2	4.83	−53.23	96.46	105.32	0.76	1.43
3	5.22	−52.13	102.75	108.66	0.73	1.40
4	4.77	−51.38	103.38	118.31	0.73	1.41
5	3.93	−52.82	102.27	109.74	0.69	1.38
6	2.74	−53.72	103.23	105.49	0.70	1.40
7	4.46	−51.29	104.83	119.10	0.75	1.42
8	4.34	−52.67	100.09	111.67	0.75	1.41
9	4.57	−51.70	106.61	122.26	0.70	1.39
10	2.89	−52.44	108.32	110.63	0.70	1.39
11	2.44	−52.73	102.85	108.00	0.67	1.38
12	3.67	−52.13	104.28	109.13	0.71	1.41
13	4.04	−50.88	106.47	109.20	0.72	1.41
14	2.88	−53.74	106.35	113.55	0.70	1.40
15	3.63	−52.27	107.10	120.84	0.72	1.39
16	3.63	−51.27	105.65	113.72	0.70	1.39
17	3.74	−49.87	108.68	120.30	0.70	1.40
18	2.55	−51.94	110.13	117.36	0.71	1.40
Mean percentage error	7.12	1.19	1.70	2.91	1.81	0.70

Tables 2 and 3 contain the estimated property values for every property and when 9 PCs are chosen in both, the specific and canonical orders. The quality of the results has to be extracted form the comparison with the data contained in table 1: the rows of percentage errors in tables 2 and 3 claim that the SO is, in this sense, better than the CO.

Finally, some words of caution must be mentioned. For every studied property in this work, the number of optimal descriptors has to be considered high. The authors believe that this characteristic arises from the fact that an isomeric family was studied. Results concerning other families and the use of new descriptors, as the quantum topological indices [27], will be published elsewhere.

5. Conclusions and expectatives

Many of the numerical values reported here relate to the linear cross-validation technique. In this way, the present approach is well suited to measure the predictive capabilities of the proposed methodology. The observed general trends rely in the fact that, for every property, there exists an optimal PCs set. Several numerical results showed the benefits to use a property-oriented vector classification. Thus, the methodology overcomes the intention of the classical PCA and it is closer to the technique of Partial

Least Squares. Actual studies in our laboratory are carried out in order to compare both property-dependent methodologies.

Acknowledgements

One of the authors (L.V.) thanks the "Dirección General de Investigaciones y Cooperación Técnica de la Universidad Católica del Norte" for the help received to perform this work. The other author (E.B.) thanks the "Agencia Española de Cooperaciön Internacional (AECI)" for giving a grant which allowed to visit the "Universidad Católica del Norte", in Antofagasta. In this place, many studies related with the topic of QSPR were initiated. Financial resources from the European Community Commission contract #ENV4-CT97-0508, the CICYT Research Project: #SAF 96-0158 and the Fundació Maria Francisca de Roviralta are also acknowledged.

Appendix: Quantum similarity matrices and indices

The cornerstone of the field of the quantum (molecular) similarity theory [16–21] consists on measuring in some way the similarities between the density functions of the molecules. It is expected that the numerical similarities and differences between density functions can be translated to similarities and differences among the properties. For every pair of molecules, A and B, with respective wavefunctions Ψ_A and Ψ_B, it its possible to obtain the related first order density functions, ρ_A and ρ_B. Then, an estimation of the similarities between the electronic distributions of A and B can be obtained from the mathematical Euclidean distance between the functions ρ_A and ρ_B, which is defined as the norm of the difference of the two density functions:

$$d_{AB}^2 = \int \left|\rho_A - \rho_B\right|^2 \mathrm{d}\tau \geqslant 0. \tag{A.1}$$

Then,

$$d_{AB}^2 = \int \rho_A^2 \,\mathrm{d}\tau + \int \rho_B^2 \,\mathrm{d}\tau - 2\int \rho_A \rho_B \,\mathrm{d}\tau,$$

and because only the last term of the previous equation is sensible to the relative molecular spatial positions, it is common to use, as a measure of molecular similarity of the two molecules the term

$$Z_{AB} = \max\left(\int \rho_A \rho_B \,\mathrm{d}\tau\right), \tag{A.2}$$

which is obtained by means of the integral maximisation with respect to the relative molecular arrangements.

The most known quantum similarity index is due to Carbó [16]. The index related to a pair of molecules is defined as

Table 4
Molecular quantum similarity measures and indices attached to the family of the octane isomers. In the superior triangle, Carbó indices are shown and in the inferior one and in the diagonal the similarity measures are tabulated. The molecular numbering is the same as the one appearing in table 1.

1	2	3	4	5	6	7	8	9
445.512	.313	.248	.406	.343	.260	.475	.399	.354
139.191	444.817	.619	.635	.727	.708	.650	.615	.561
110.381	275.313	445.012	.646	.587	.570	.613	.649	.468
180.629	282.620	287.303	445.040	.620	.561	.631	.628	.547
152.807	323.691	261.066	275.850	445.230	.598	.674	.552	.478
115.822	315.076	253.577	249.657	266.370	445.121	.585	.736	.646
211.578	289.205	272.924	280.645	300.002	260.272	445.103	.627	.468
177.425	273.744	288.541	279.204	245.647	327.540	278.869	444.728	.368
157.540	249.846	208.232	243.773	212.699	287.562	208.591	163.883	445.474
142.895	249.958	236.620	210.205	231.400	262.035	247.175	222.517	182.258
121.139	234.660	234.231	183.144	223.199	238.299	257.399	286.690	183.046
179.928	271.417	246.699	254.791	243.127	248.583	250.185	222.842	221.721
165.024	232.937	196.821	233.750	221.433	181.569	218.631	219.446	321.338
132.308	179.458	212.929	217.804	167.695	164.940	193.429	169.452	254.961
239.505	223.991	177.188	250.998	185.954	169.572	208.257	203.172	250.304
118.571	161.991	189.378	206.180	159.577	130.136	172.481	154.770	271.448
159.419	179.850	225.034	228.891	205.070	141.558	204.042	193.222	236.919
200.692	168.737	167.618	229.302	161.023	246.277	197.263	189.195	183.885

10	11	12	13	14	15	16	17	18
.321	.272	.404	.370	.297	.538	.266	.358	.450
.561	.527	.610	.523	.403	.503	.364	.404	.379
.531	.526	.554	.442	.478	.398	.425	.505	.376
.472	.411	.572	.525	.489	.564	.463	.514	.515
.519	.501	.546	.497	.377	.418	.358	.460	.362
.588	.535	.558	.408	.370	.381	.292	.318	.553
.555	.578	.562	.491	.434	.468	.387	.458	.443
.500	.644	.501	.493	.381	.456	.348	.434	.425
.409	.411	.498	.721	.572	.562	.609	.532	.413
445.854	.676	.613	.502	.502	.413	.400	.530	.466
301.348	445.755	.410	.485	.419	.548	.436	.409	.332
273.330	182.862	445.284	.519	.477	.648	.436	.511	.622
223.906	216.049	231.337	445.526	.554	.496	.654	.492	.445
223.851	186.519	212.296	246.983	445.508	.427	.642	.582	.478
183.975	244.319	288.726	221.051	190.174	445.453	.351	.625	.560
178.506	194.445	194.254	291.511	285.951	156.478	445.624	.506	.407
236.220	182.224	227.622	219.050	259.299	278.523	225.516	445.485	.697
207.881	148.145	276.827	198.171	212.822	249.413	181.308	310.494	445.508

$$R_{AB} = \frac{Z_{AB}}{(Z_{AA} \cdot Z_{BB})^{1/2}}. \tag{A.3}$$

This parameter has the property to be normalised between the values 0 (molecule A totally dissimilar to molecule B) and 1 (molecules A and B identical).

In this way, when studying a set of n molecules, a symmetric $n \times n$ similarity matrix is obtained when collecting the similarity measures or the indices between all the molecular pairs. The ith column can be taken as a theoretical molecular descriptor vector.

Despite the basic foundation described here, there is no a unique way to extract selectively from the density functions information related to the molecular properties. Usually, the linear techniques seem to be adequate to obtain models with predictable capabilities.

The calculations relative to the 18 octane isomers where performed with the MOPAC 7.0 [28] program under the PM3 approach and global geometry optimisation. The precision parameters where set to the best available ones (SCFCRT $= 10^{-25}$, GNORM $= 0.0$). The mean CPU timing for every molecule was 300 s in a personal computer Pentium Aptiva at 100 MHz. The similarity integral computations and optimisation where performed with the MOLSIMIL-93 [29,30] program, which uses as input some of the MOPAC program data output. The total CPU time consumed was 1 hour. The obtained similarity measures and indices are shown in table 4.

More details related to the computation of the quantum similarity matrix related to the studied molecular family are described in a previous article [22].

References

[1] Y.C. Martin, *Quantitative Drug Design*, Medicinal Research Series, Vol. 8 (Marcel Dekker, New York, 1978).

[2] E.J. Lien, SAR, *Side effects and Drug Design*, Medicinal Research Series, Vol. 11 (Marcel Dekker, New York, 1987).

[3] Y.C. Martin, E. Kutter and V. Austel (eds.), *Modern Drug Research*, Medicinal Research Series, Vol. 12 (Marcel Dekker, New York, 1989).

[4] C.G. Wermuth (ed.), *Trends in QSAR Molecular Modelling 92* (Escom, Leiden, 1993).

[5] H. Kubinyi (ed.), *3D QSAR in Drug Design* (ESCOM, Leiden, 1993).

[6] H. van de WaterBeemd (ed.), *Structure-Property Correlations in Drug Research* (Academic Press, San Diego, CA, 1996).

[7] M. Charton (ed.), *Advances in Quantitative Structure-Property Relationships*, Vol. 1 (Jai Press, Greenwich, CT, 1996).

[8] D.C. Montogmery and E.A. Peck, *Introduction to Linear Regression Analysis* (Wiley, New York, 1992).

[9] I.T. Jolliffe, *Principal Component Analysis* (Springer, New York, 1986).

[10] L. Pla, *Análisis Multivariado: Método de Componentes Principales*, Monography no. 27 (Secretaría General de la OEA, Washington, 1986).

[11] R.B. Cattell, Multivar. Behab. Res. 1 (1966) 245.

[12] S. Wold, E. Johansson and M. Cocchi, PLS-partial least-squares projections to latent structures, in: *3D QSAR in Drug Design*, ed. H. Kubinyi (ESCOM, Leiden, 1993), pp. 523–550.

[13] D.M. Allen, The relationship between variable selection and data augmentation and a method for prediction, Technometrics 16 (1974) 125–127.

[14] C.M. Cuadras, C. Arenas and J. Fortiana, Some computational aspects of a distance-based model for prediction, Commun. Statist. Simula. 25 (1996) 593–609.

[15] C.M. Cuadras and C. Arenas, A distance based regression model for prediction with mixed data, Commun. Statist. Theor. Meth. 19 (1990) 2261–2279.

[16] R. Carbó, M. Arnau and L. Leyda, Int. J. Quant. Chem. 17 (1980) 1185.
[17] R. Carbó and E. Besalú, Theoretical foundations of quantum similarity, in: *Molecular Similarity and Reactivity: From Quantum Chemical to Phenomenological Approaches*, ed. R. Carbó (Kluwer, Amsterdam, 1995) p. 3.
[18] R. Carbó-Dorca, E. Besalú, Ll. Amat and X. Fradera, *Quantum Molecular Similarity Measures: Concepts, Definitions and Applications to QSAR*, Advances in Molecular Similarity, Vol. 1, eds. R. Carbó-Dorca and P.G. Mezey (JAI Press, London, 1996).
[19] R. Carbó-Dorca, Ll. Amat, E. Besalú and M. Lobato, *Quantum Molecular Similarity*, Advances in Molecular Similarity, Vol. 2, eds. R. Carbó-Dorca and P.G. Mezey (JAI Press, London, 1998).
[20] R. Carbó-Dorca, L. Amat, E. Besalú, X. Gironés and D. Robert, Quantum mechanical origin of QSAR: theory and applications, J. Mol. Struct. (Theochem) 504 (2000) 181–228.
[21] R. Carbó, B. Calabuig, L. Vera and E. Besalú, Adv. Quantum Chem. 25 (1994) 253.
[22] L. Vera, M. Guzmán and P. Ortega, Bol. Soc. Chil. Quím. 42 (1997) 341.
[23] R.L. Harvey, *Neural Network Principles* (Prentice Hall, New York, 1994).
[24] R.C. Weast (ed.), *Handbook of Chemistry and Physics*, Edition 57 (CRC Press, Boca Raton, FL, 1977) p. D-84.
[25] Organic Compounds Database, available at http://www.colby.edu/chemistry/cmp/ cmp.html.
[26] M. Abramowitz and I.A. Stegun, *Handbook of Mathematical Functions* (Dover, New York, 1965).
[27] M. Lobato, Ll. Amat, E. Besalú and R. Carbó-Dorca, Quant. Struct.-Act. Relat. 16 (1997) 465–472.
[28] Available via internet at ftp://ccl.osc.edu/pub/chem/software/MS-DOS/mopac_for_dos/mopac7/.
[29] R. Carbó and B. Calabuig, MOLSIMIL, Version 90, Comp. Phys. Commun. 55 (1989) 117.
[30] R. Carbó-Dorca, B. Calabuig and E. Besalú, MOLSIMIL, Version 93, Institute of Computational Chemistry, University of Girona, Girona, Spain, 1993.