MATCH
Communications in Mathematical and in Computer Chemistry

no. 51, April 2004
MATCDY (51) 39- 57 (2004)

ISSN 0340 - 6253

ABOUT ORTHOGONAL DESCRIPTORS IN QSPR/QSAR THEORIES

Francisco M. Fernández,[1] *Pablo R. Duchowicz and Eduardo A. Castro*

INIFTA (Conicet, UNLP), Diag 113 y 64 S/N, Sucursal 4, Casilla de Correo 16,

1900 La Plata, Argentina.

(Received October 16, 2003)

By means of linear algebra we develop several useful mathematical expressions for multivariate linear regression that are suitable for QSPR/QSAR theory. Our equations reveal the effect of adding or subtracting orthogonal molecular descriptors and may be useful for the construction of optimal QSPR/QSAR models. We illustrate the application of our equations by means of three simple test cases studied earlier by other authors.

---

[1] Corresponding author, e-mail: fernande@quimica.unlp.edu.ar

I. INTRODUCTION

Multivariate regression analysis has proved useful in structure-property and structure-activity studies. For such applications one has to determine the best set of molecular descriptors for the given molecular property or activity. Earlier attempts to derive the best quantitative structure-property and structure-activity regression (QSPR, QSAR) models have led to the use of orthogonal descriptors.[1-5]

It is well known that the use of a set of orthogonal descriptors does not improve the global statistical parameters of the linear regression with respect to a model based on nonorthogonal descriptors, although it has been found that the standard errors of the regression coefficients are smaller in the former case.[4] Moreover, orthogonal descriptors (orthogonal predictor variables in general) offer several advantages: first, expressions for some statistical parameters (such as the correlation coefficient and standard deviation, for example) are simpler.[6] Second, orthogonal descriptors have proved more suitable for the search of the best model according to given quality criteria.[6-9] Third, the orthogonalization algorithms are useful to identify linear dependent or almost linear dependent descriptors[10,11] and make the linear regression more stable.[10] Fourth, the coefficients of the orthogonal descriptors do not change when the set is augmented while those of the nonorthogonal descriptors vary appreciably.

Several authors have developed a program that enables one to obtain the best subset (with respect to the values of selected statistical parameters) of I descriptors out of a large number of N descriptors.[6-9] This analysis requires linear regressions for all $\binom{N}{I} = N!/[(N-I)!I!]$ possible subsets. Once we have the best set of descriptors we then look for the best ordering among the $I!$ possible ones according to a dominant component analysis.[6-9] Orthogonalization also facilitates this step because it makes it easier to determine the contribution of each descriptor to the model. This systematic search for the best QSPR/QSAR model has proved useful for many chemical applications.[6-9,12-15]

Linear algebra proves suitable for the discussion of multivariate regression analysis.[11] The purpose of this paper is to develop this approach further, and derive some useful relationships that

have not been discussed earlier in this context.[11] We do this in Sec. II. In Sec. III we apply the general equations to some simple illustrative examples. Finally, in Sec. IV we summarize the main results of the paper.

II. MULTIVARIATE REGRESSION

In this section we develop the method of least squares and multivariate regression from the point of view of linear algebra.[16] We will try to go beyond a previous discussion on the subject[11] and show that some numerical results discussed previously by other authors can be proved rigorously and easily. Although some of the theoretical results derived below may be well known, we show them anyway in order to make this paper self-contained. In addition to it, the following mathematical discussion will be useful to introduce the notation that we will use throughout this paper.

Consider a vector space $V=\{f,g,\ldots\}$ on the field of real numbers, endowed with an inner product $f\cdot g$. We define the norm $\|f\|=\sqrt{f\cdot f}$ of a vector $f\in V$ and a distance between f and g, $f,g\in V$, as $D(f,g)=\|f-g\|$.

Let $B=\{f_0,f_1,\cdots,f_n\}$ be a set of $n+1$ linearly independent vectors that span a subspace $S_B\subseteq V$. It is well known that the set B is linearly independent if and only if the determinant $|M|$ of the matrix M with elements $M_{ij}=f_i\cdot f_j,\quad i,j=0,1,\ldots$ is nonzero.

It is our purpose to find the closest approximation to a given vector $f\in V$ by means of a linear combination of vectors of B:

$$\tilde{f}=\sum_{j=0}^{n}c_j f_j. \tag{1}$$

It is well known that the set of coefficients c_j that minimize the distance $D(f,\tilde{f})$, and are consequently solutions to the equations $\partial D(f,\tilde{f})^2/\partial c_j=0,\quad j=0,1,\ldots,n$ make $f-\tilde{f}$

orthogonal to the subspace S_B. Therefore, it follows from $\left(f-\tilde{f}\right)\cdot f_j = 0, \quad j=0,1,\ldots,n$ that the optimal coefficients c_j are solutions to the set of linear equations

$$\sum_{k=0}^{n} M_{jk} c_k = f \cdot f_j, \quad j=0,1,\ldots,n. \tag{2}$$

Under such conditions it follows that $D(g,f) \geq D(\tilde{f},f)$ for all $g \in S_B$. One can also prove that $(f-v)\cdot\left(\tilde{f}-v\right) = \left\|\tilde{f}-v\right\|^2$ for all $v \in S_B$ as well as $\left\|f-\tilde{f}\right\|^2 = \|f-v\|^2 - \left\|\tilde{f}-v\right\|^2$. It follows from these expressions that $f.\tilde{f} = \left\|\tilde{f}\right\|^2$, and $\left\|\tilde{f}\right\|^2 \leq \|f\|^2$.

As argued in the introduction, it is convenient to use an orthogonal basis set for the subspace S_B, from both the theoretical and practical point of view. There are many ways of orthogonalizing a finite set of linearly independent vectors; here we consider the well-known Gram-Schmidt algorithm.[10,11,16] Starting from the basis B we construct a new basis $\{u_0,u_1,\ldots u_n\}$ hierarchically according to

$$\begin{aligned} u_0 &= f_0, \\ u_j &= f_j - \sum_{k=0}^{j-1} \frac{u_k \cdot f_j}{\|u_k\|^2} u_k. \end{aligned} \tag{3}$$

The closest approximation to f in terms of the set of orthogonal vectors takes the simpler form

$$\tilde{f} = \sum_{j=0}^{n} b_j u_j, \quad b_j = \frac{f \cdot u_j}{\|u_j\|^2} \tag{4}$$

where we realize that the coefficients b_j are independent of n.

Note that according to Eq. (3) we can write

$$u_j = \sum_{k=0}^{j} d_{kj} f_k, \quad d_{jj} = 1 \tag{5}$$

where the explicit expressions for the coefficients d_{kj} are not relevant for present discussion. If we substitute Eq. (5) into Eq. (4) and take into account that the least-squares solution is unique, we conclude that $c_k = \sum_{j=k}^{n} d_{kj} b_j$ from which it follows that

$$c_n = b_n. \tag{6}$$

This result was suggested by numerical experiments conducted by Randić[2] and later discussed by this and other authors,[2,3,5,14] but it was not proved rigorously as far as we know. Eq. (6) is a consequence of the particular triangular form of the Gram-Schmidt linear combination (3) and one does not expect that it applies to other orthogonalization procedures.

If we define $u_w = u - \left(u \cdot w / \|w\|^2\right) w$ and $v_w = v - \left(v \cdot w / \|w\|^2\right) w$ for $u, v, w \in V$, and take into account the Cauchy-Schwarz inequality[16] we conclude that the ratio

$$C(u,v,w) = \frac{u_w \cdot v_w}{\|u_w\| \|v_w\|} \tag{7}$$

satisfies $-1 \le C(u,v,w) \le 1$.

In the application of these results to multivariate linear regression the vectors are N-tuples of real numbers $f = (f(1), f(2), \ldots, f(N))$ and we choose the inner product to be

$$f \cdot g = \sum_{j=1}^{N} f(j) g(j). \tag{8}$$

Typically the components of the vector f_0 are $f_0(j) = 1, \quad j = 1, 2, \ldots N$ so that $\|f_0\|^2 = N$ and

$$\frac{g \cdot f_0}{\|f_0\|^2} = \frac{1}{N} \sum_{j=1}^{N} g(j) = < g > \tag{9}$$

is the expectation value of $g \in V$. Clearly, either c_0 or b_0 plays the role of the constant or intercept in multivariate regression analysis.

It follows from the definition (7), and from $f \cdot \tilde{f} = \|\tilde{f}\|^2$ and $f \cdot f_0 = \tilde{f} \cdot f_0 = b_0 \|f_0\|^2$, that

$$C(f,\tilde{f},f_0)=\frac{\|\tilde{f}\|^2-b_0^2\|f_0\|^2}{\sqrt{\left(\|f\|^2-b_0^2\|f_0\|^2\right)\left(\|\tilde{f}\|^2-b_0^2\|f_0\|^2\right)}}$$
$$=\sqrt{\frac{\|\tilde{f}\|^2-b_0^2\|f_0\|^2}{\|f\|^2-b_0^2\|f_0\|^2}} \tag{10}$$

Therefore, the coefficient of correlation R between f and $\tilde{f}$ is simply given by

$$0\le R=C(f,\tilde{f},f_0)=\sqrt{\frac{\|\tilde{f}\|^2-b_0^2\|f_0\|^2}{\|f\|^2-b_0^2\|f_0\|^2}}\le 1. \tag{11}$$

The standard deviation S in terms of present notation reads[17]

$$S=\frac{D(f,\tilde{f})}{\sqrt{N-n-1}}. \tag{12}$$

It follows from Eq. (11) that

$$R^2=\sum_{j=1}^{n}R_j^2,\quad R_j^2=\frac{b_j^2\|u_j\|^2}{\|f\|^2-b_0^2\|f_0\|^2}=C(f,u_j,f_0)^2. \tag{13}$$

From now on we assume that the vector f represents a given property or activity for a set of N molecules, and the vectors $f_1,f_2,\ldots,f_n$ are predictor variables or descriptors in the language of QSPR/QSAR theory for those molecules. Accordingly, the vector $\tilde{f}$ is the regression model.

Some authors, like Randić,[2] do not modify one of the remaining n vectors, say f_1, during orthogonalization. In order to compare our coefficient b_0 with Randić's b_0^R we simply take into account that

$$\tilde{f}=b_0^R f_0+b_1 f_1+\sum_{j=2}^{n}b_j u_j,\quad b_0^R=b_0-b_1\frac{f_1\cdot u_0}{\|u_0\|^2}. \tag{14}$$

We calculate the error of the coefficients c_j and b_j by means of the standard formula[17] which in present notation takes the form

$$\delta c_j = \sqrt{(M^{-1})_{jj}}S, \quad \delta b_j = \frac{S}{\|u_j\|}. \tag{15}$$

The Gram-Schmidt algorithm gives us

$$f_j = \sum_{k=0}^{j} e_{kj} u_k, \quad e_{jj} = 1 \tag{16}$$

where the explicit form of the coefficients e_{kj} is not relevant for our purposes. Notice that the matrix $E = (e_{ij})$ is upper triangular. It follows from Eq. (16) and from the definition of the matrix elements of M

$$M_{ij} = f_i \cdot f_j = \sum_{k=0}^{\min(i,j)} e_{ki} e_{kj} \|u_k\|^2 \tag{17}$$

that $M = E^t U E$, where $U = \left(\|u_j\|^2 \delta_{ij}\right)$, and the superscript t stands for transpose. Therefore,

$$\begin{aligned} \left(M^{-1}\right)_{nn} &= \sum_{i=0}^{n}\sum_{j=0}^{n}\left(E^{-1}\right)_{ni}\left(U^{-1}\right)_{ij}\left(\left(E^t\right)^{-1}\right)_{jn} \\ &= \left(E^{-1}\right)_{nn}\left(U^{-1}\right)_{nn}\left(\left(E^t\right)^{-1}\right)_{nn} = \left(U^{-1}\right)_{nn} = \|u_{nn}\|^{-2} \end{aligned} \tag{18}$$

and we have

$$\delta c_n = \delta b_n \tag{19}$$

in addition to Eq. (6).

What happens when we add or remove one of the descriptors to or from our model?. If we define

$$\tilde{f}^{[k]} = \sum_{j=0}^{k} b_j u_j \tag{20}$$

and

$$S^{[k]} = \frac{D(f, \tilde{f}^{[k]})}{\sqrt{N-k-1}} \tag{21}$$

then we have

$$(N-k-2)\left(S^{[k+1]}\right)^2 = (N-k-1)\left(S^{[k]}\right)^2 - b_{k+1}^2 \left\|u_{k+1}\right\|^2 . \tag{22}$$

Arguing exactly in the same way for the case of k orthogonal descriptors given in arbitrary order we obtain

$$\left(S^{[k+1]}\right)^2 = \left(S^{[k]}\right)^2 + \frac{\left(S^{[k]}\right)^2 - b_j^2 \left\|u_j\right\|^2}{N-k-2} \tag{23}$$

or

$$\left(S^{[k]}\right)^2 = \left(S^{[k+1]}\right)^2 + \frac{b_j^2 \left\|u_j\right\|^2 - \left(S^{[k+1]}\right)^2}{N-k-1} \tag{24}$$

which clearly show the effect on S of adding or removing u_j, respectively. Eq. (24) explains why Lučić et al[6] could improve the model by removing carefully selected orthogonal descriptors. If $b_j^2 \left\|u_j\right\|^2$ is sufficiently small, then removal of u_j may decrease the value of S while slightly reducing the value of R (c.f., Eq. (13)). Taking into account that $b_j^2 \left\|u_j\right\|^2 = (f \cdot u_j)^2 / \left\|u_j\right\|^2$ plays such a relevant role in the values of R and S we consider it to be a measure of the contribution or weight of the orthogonal descriptor u_j to the model.

III. SIMPLE NUMERICAL EXAMPLES

The value of R increases and approaches unity and $D(f, \tilde{f})$ decreases and approaches zero as the number of linearly independent vectors f_j increases. When $n+1 = N$ we have an interpolation

case with $R = 1$ and $D(f, \tilde{f}) = 0$ because $S_B = V$. On the other hand, S may increase with the number of descriptors as shown by Eq. (23). A QSPR/QSAR model is successful if it gives acceptable statistical parameters with a small number of descriptors. The values of the global statistical parameters (S, R, etc.) depend on the subspace spanned by the chosen vectors f_j or u_j and are independent of the order of construction of the orthogonal vectors. As far as we know there is no direct way of selecting the subset of $m \leq n$ vectors of B that gives the optimal approximation to f. For this reason we follow other authors[6-9,12,13,15] and single out the sets of $1, 2, \ldots n$ descriptors with the smallest value of S. This strategy requires $\sum_{m=1}^{n} \binom{n}{m} = 2^n - 1$ linear regressions. We select the subset of m descriptors with the smallest value of S and then inspect all the $m!$ orthogonalization orderings of those vectors to obtain the optimum sequence according to a dominant component criterion similar to that proposed by Randić[2] and applied by other authors.[6] We base our criterion on the value of $b_j^2 \|u_j\|^2$ as argued in the preceding section. Our algorithms and main guiding ideas are similar to those proposed by Trinajstić, Lučić and coworkers.[6-9,12,13,15]

The first test case for our numerical investigation is the fitting of Hosoya's Z index by means of connectivity and higher connectivity indices ${}^{J}X$ discussed by Randić.[2] In this case we choose $f = Z$ and $f_j = {}^{J}X$, $j = 1, 2, 3, 4$ given in Table II of that paper.[2] Consequently, our orthogonal descriptors u_j should be compared with Randić's ${}^{J}\Omega$. The coefficients b_j that we obtain by means of the procedure outlined above agree (up to round-off errors) with those generated by correlations and residuals[2]. The coefficient b_0 (the constant in the language of linear regression) should be modified according to Eq. (14) in order to have complete agreement.

Table I

Statistical parameters for our best sets of 1, 2, 3 and 4 connectivity indices and for two models proposed by Randić[2]

Descriptors	R	S
$\{u_2\}$	0.99288	0.34865
$\{u_1, u_3\}$	0.99803	0.19844
$\{u_1, u_3, u_4\}$	0.99903	0.15293
$\{u_1, u_2, u_3, u_4\}$	0.99906	0.16800
$\{u_1, u_2, u_3\}$	0.99872	0.17512
$\{u_1, u_2, u_4\}$	0.99865	0.17987

In Table I we show the sets of 1, 2, 3, and 4 descriptors with the smallest values of S. Notice that the combination $\{u_1, u_3, u_4\}$ gives a smaller value of S than the complete available set $\{u_1, u_2, u_3, u_4\}$. The set of three descriptors obtained by Randić by means of his systematic procedure $\{u_1, u_2, u_3\}$, and the one that he mentioned to be the best $\{u_1, u_2, u_4\}$ yield greater values of S as shown in the last two rows of Table I. Table I also shows that in this case it is profitable to remove one of the descriptors $\left(f_2 = {}^2X\right)$ because we thus obtain a smaller value of S and almost the same value of R with less predictor variables.

Table II

Coefficients of the nonorthogonal and orthogonal descriptors for the model with smallest S

Descriptor	Coefficient	Error	Descriptor	Coefficient	Error
f_0	-44.24820797	1.336605724	u_0	17	0.0509757195
f_1	18.93451086	0.4736646298	u_1	17.96670982	0.3594210985
f_3	0.7807444357	0.2042389764	u_3	1.102350730	0.1464411551
f_4	-0.6857907910	0.3035836814	u_4	-0.6857907910	0.3035836814

In Table II we show the coefficients of the nonorthogonal and orthogonal descriptors and their respective standard errors for the best model. At this stage we do not perform a dominant component analysis and simply orthogonalize the descriptors in the given order. We appreciate that the errors are smaller for the orthogonal descriptors as argued earlier by Randić,[4] and that $c_4 = b_4$ and $\delta c_4 = \delta b_4$ according to equations (6) and (19), respectively. Our calculations show that the order of orthogonalization affects the coefficients b_j, their standard errors δb_j, and the absolute relative errors $|\delta b_j / b_j|$.

The magnitude of the coefficients b_j depends on the orthogonalization order; for example, the ordering $\{f_0, f_4, f_1, f_3\}$ yields the model with the largest coefficient b_j:

$$\begin{aligned} Z = &(17 \pm 0.05097571950) f_0 + (19.65031864 \pm 0.4350803632) u_1 \\ &+ (0.7807444357 \pm 0.2042389764) u_3 - (4.045288078 \pm 0.1798196005) u_4 \end{aligned} \tag{25}$$

with coefficients of u_1 and u_4 larger than the corresponding ones in Table II obtained from the orthogonalization sequence $\{f_0, f_1, f_3, f_4\}$.

The next example is the fitting of the standard deviation S by means of a polynomial function of the coefficient of correlation R proposed by Randić.[3] Those values of S and R were obtained from several single-variable regressions for 18 octane isomers.[3] In our notation $f = S$

and $f_j = R^j$, $j = 0,1,2,3,4$, where the values of S and R are given in Table IV of that paper.[3] By application of the method outlined above we obtain the optimum model

$$\begin{aligned} S = &(4.696 \pm 0.001592282751) f_0 - (2.610009295 \pm 0.0333750566) u_2 \\ &+ (7.238292482 \pm 0.6107060177) u_3 - (4.85257838 \pm 0.007179313592) u_4 . \end{aligned} \tag{26}$$

with standard deviation $S' = 0.01007$ and correlation coefficient $R' = 0.99996$. The model in Eq. (26) corresponds to the orthogonalization order $\{f_0, f_4, f_2, f_3\}$ that gives the largest coefficients b_j. Notice that we obtain a smaller standard deviation when we omit the linear term. If we carry out the calculation within the same subspace with nonorthogonal vectors we obtain the model as a polynomial function of R:

$$\begin{aligned} S = &6.393791575 \pm 0.01289288983 - (5.721961166 \pm 0.2646728651) R^2 \\ &+ (7.238294377 \pm 0.6107060138) R^3 - (6.499904447 \pm 0.3778579316) R^4 . \end{aligned} \tag{27}$$

The errors of the coefficients are larger but the standard deviation and correlation coefficient are exactly the same.

Our last example is the modeling of the boiling points of several octanes by means of connectivity indices given in tables 2 and 1, respectively, of Lučić et al.[6] In this case we choose f_0 as usual and $f_{j+1} = {}^j\chi$, $j = 0-6$. We have tested all possible sets of 1,2,3,4,5,6 and 7 connectivity indices, and found that the smallest standard deviation is given by the model

$$\begin{aligned} bp = &(113.7132222 \pm 0.2070758473) u_0 - (24.74180191 \pm 1.170657363) u_1 \\ &+ (70.786048 \pm 5.466486863) u_2 - (3.221924496 \pm 1.412762703) u_3 \\ &- (10.50036138 \pm 0.9155462828) u_5 - (12.14038512 \pm 1.543384974) u_6 \\ &- (7.664196453 \pm 7.358202687) u_7 \end{aligned} \tag{28}$$

with $S = 0.87855$ and $R = 0.99331$. In this case we have not tried a dominant component analysis keeping the given order of descriptors. Notice that our optimum model is given by 6 descriptors and not by 5 as found by Lučić et al.[6] The disagreement is due to misprints in two entries of Table 1 in that paper.[18]

Table III

Corrected connectivity indices for the molecules considered by Lučić et al[6,18]

${}^0\chi$	${}^1\chi$	${}^2\chi$	${}^3\chi$	${}^4\chi$	${}^5\chi$	${}^6\chi$
6.24264	3.91421	2.41421	1.45710	0.85355	0.47855	0.25000
6.40577	3.84606	2.47119	1.85162	1.10517	0.40824	0.00000
6.40577	3.80806	2.68252	1.56294	1.12993	0.28867	0.14433
6.40577	3.80806	2.65564	1.74740	0.75671	0.49279	0.14433
6.40577	3.77005	2.88962	1.38502	0.80258	0.43301	0.28867
6.56891	3.71874	**2.77106**	2.25930	0.80473	**0.16666**	0.00000
6.56891	3.71874	2.82059	1.99156	1.23148	0.00000	0.00000
6.62132	3.68198	2.87132	2.56066	0.75000	0.00000	0.00000
6.56891	3.68073	3.00997	1.88208	0.78867	0.33333	0.00000
6.56891	3.66390	3.14296	1.57069	0.97140	0.33333	0.00000
6.56891	3.62589	3.36504	1.32136	0.66666	0.66666	0.00000
6.62132	3.62132	3.26776	1.88388	0.85355	0.17677	0.00000
6.62132	3.56066	3.66421	1.28033	0.70710	0.53033	0.00000
6.73205	3.55341	3.34715	2.10313	0.76980	0.00000	0.00000
6.78445	3.50403	3.49683	2.47417	0.40824	0.00000	0.00000
6.78445	3.48138	3.67532	2.09077	0.61237	0.00000	0.00000
6.78445	3.41650	4.15863	1.02062	1.22474	0.00000	0.00000
7.00000	3.25000	4.50000	2.25000	0.00000	0.00000	0.00000

Table III shows the connectivity indices with the two corrected entries in boldface.[18] If we take into account this corrected table of connectivity indices the agreement is complete. For example, our best model with nonorthogonal descriptors is

$$
\begin{aligned}
&(2583.416346 \pm 450.923869) f_0 - (182.7562811 \pm 25.59325052) f_1 \\
&-(310.4975894 \pm 67.62496045) f_2 - (46.14952271 \pm 12.54416987) f_3 \\
&+(8.304446637 \pm 1.872065541) f_4 - (5.666366773 \pm 1.703086193) f_6
\end{aligned} \tag{29}
$$

with $R = 0.9926$ and $S = 0.8867$. On the other hand, for orthogonal descriptors in the given order we have

$$
\begin{aligned}
&(113.7132222 \pm 0.2090046467) u_0 - (24.74180192 \pm 1.181561403) u_1 \\
&+(70.78604760 \pm 5.517404228) u_2 - (89.06184140 \pm 7.014153148) u_3 \\
&+(8.851964612 \pm 1.864818580) u_4 - (5.666367297 \pm 1.703086193) u_6
\end{aligned} \tag{30}
$$

which does not agree with those reported by Lučić et al[6] because the orthogonalization order is different.

Table IV

Coefficients of the orthogonal descriptors obtained by orthogonalization in the order indicated by their indices

Descriptor Indices				Orthogonal Descriptor coefficients			
0	2	5	6	113.7132222	30.32810932	-10.26946598	-12.77271243
0	2	6	5	113.7132222	30.32810932	-8.028451822	-13.12861838
0	5	2	6	113.7132222	2.87934073	40.7257032	-12.77271243
0	5	6	2	113.7132222	2.87934073	6.368681705	53.41560197
0	6	2	5	113.7132222	6.731371914	36.48492833	-13.12861838
0	6	5	2	113.7132222	6.731371914	2.262083058	53.41560198

In Table IV we show the coefficients of the orthogonal descriptors for all the possible orderings of the subset $m = 3$ with smallest S.

The agreement with the results of Lučić et al[6] is complete except for the first entry because of a misprint in those author's Table 4.[18]

Table V

Weights of the descriptors for several orderings of the optimal model

Descriptor Indices	Descriptor weigths				
2 3 6 1 4	0.6743500578	0.1346742520	$9.980236002 \cdot 10^{-5}$	0.1517627958	0.02429655627
3 2 6 1 4	0.7773041441	0.03172016579	$9.980236002 \cdot 10^{-5}$	0.1517627958	0.02429655627
2 3 6 4 1	0.6743500578	0.1346742520	$9.980236002 \cdot 10^{-5}$	0.1131001591	0.06295919298
3 2 6 4 1	0.7773041441	0.03172016579	$9.980236002 \cdot 10^{-5}$	0.1131001591	0.06295919298

In Table V we show the weights of the orthogonal descriptors in the subset with the smallest value of $S = 0.8867$, for four different orthogonalization orderings. The entries in that table suggest that it may be reasonable to remove the orthogonal vector u_6 with the smallest weight. If we do that we obtain four new models with four descriptors each that yield $S = 0.85481$ and $R = 0.99251$. This value of S is smaller than the smallest one that we can get by removal of nonorthogonal descriptors. This notable advantage of orthogonal predictor variables that help us to remove insignificant descriptors was first pointed out by Lučić et al.[6]

IV. CONCLUSIONS

The main equations developed in Sec. II by means of linear algebra prove useful to explain many results in QSPR/QSAR theory. For example, we have rigorously derived the relation between the coefficients of the last descriptor in equivalent models with nonorthogonal and orthogonal descriptors Eq. (6) already mentioned before by other authors[2,3,5,14]. In addition to it, we have shown that the errors of those coefficients are exactly the same (Eq. (19)). More important are present expressions for the contribution of each orthogonal descriptor to the correlation coefficient Eq. (13) and to the standard deviation equations (23) and (24). They reveal the effect on the statistical parameters of addition or subtraction of orthogonal descriptors. The knowledge of the contribution or weight of each descriptor is relevant for removing insignificant descriptors from a

model. In fact, numerical results in Sec. III show that removal of orthogonal descriptors with negligible weight may result in a smaller value of S and an almost similar value of R. Although this kind of results were mentioned before[6] most discussions were based on numerical investigation on particular examples and not on rigorous equations like (13), (23) and (24). It is also important to notice that the optimum models derived in the preceding section from Table V are not exactly those proposed by Lučić et al[6] who were the first authors in realizing that one may obtain better models from orthogonal descriptors than from nonorthogonal descriptors.

We have also addressed the problem of construction of optimal models by considering subsets of descriptors chosen from a larger set. This investigation can be carried out either with nonorthogonal or orthogonal descriptors.

Finally, it is worth mentioning that computer algebra systems (CAS) like Derive[19] and Maple[20] proved extremely useful for the calculations in this paper. Although CAS are rather slow for massive computations, they are however extremely useful for algebraic and numerical investigation on relatively small problems because they allow us to set arbitrary precision, and even to carry out multivariate regression in exact rational arithmetic, thus avoiding round off errors.

Acknowledgments

One of the authors (F.M.F.) would like to thank Dr. R. H. Tipping (University of Alabama, Tuscaloosa), Dr. B. Lučić (The Rugjer Bošković Institute) and the derivians Valeriu Anisiu and Enric Puig.

References

(1) Randić, M. Search for Optimal Molecular Descriptors. *Croat. Chem. Acta* **1991**, *64*, 43-45.

(2) Randić, M. Resolution of Ambiguities in Structure-Property Studies by Use of Orthogonal Descriptors. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 311-320.

(3) Randić, M. Fitting of Nonlinear Regressions by Orthogonalized Power Series. *J. Comput. Chem.* **1993**, *14*, 363-370.

(4) Randić, M. Curve-Fitting Paradox. *Int. J. Quantum Chem.* **1994**, *21 S*, 215-225.

(5) Randić, M. Retro-Regression--Another Important Multivariate Regression Improvement. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 602-606.

(6) Lučić, B.; Nikolic, S.; Trinajstić, N.; Juretić, D. The Structure-Property Models Can Be Improved Using the Orthogonalized Descriptors. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 532-538.

(7) Lučić, B.; Trinajstić, N. New Developments in QSPR/QSAR Modeling Based on Topological Indices. *SAR QSAR Environ. Res.* **1997**, *7*, 45-62.

(8) Lučić, B.; Trinajstić, N. Multivariate Regression Outperforms Several Robust Architectures of Neural Networks in QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 121-132.

(9) Lučić, B.; Trinajstić, N.; Sild, S.; Karelson, M.; Katritzky, A. R. A New Efficient Approach for Variable Selection Based on Multiregression: Prediction of Gas

Chromatographic Retention Times and Response Factors. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 610-621.

(10) Draper, N. R.; Smith, H. Applied Regression Analysis. Second Edition ed., John Wiley & Sons: New York, 1981.

(11) Klein, D. J.; Randić, M.; Babić, D.; Lučić, B.; Nikolic, S.; Trinajstić, N. Hierarchical Orthogonalization of Descriptors. *Int. J. Quantum. Chem.* **1997**, *63*, 215-222.

(12) Amić, D.; Davidović-Amić, D. Structure-Activity Correlation of Flavone Derivatives for Inhibition of cAMP Phosphodiesterase. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1034-1038.

(13) Amić, D.; Davidović-Amić, D.; Bešlo, D.; Lučić, B.; Trinajstić, N. The Use of the Ordered Orthogonalized Multivariate Linear Regression in a Structure-Activity Study of Courmarin and Flavonoid Derivatives as Inhibitors of Aldose Reductase. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 581-586.

(14) Šoškić, M. Link Between Orthogonal and Standard Multiple Linear Regression Models. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 829-832.

(15) Šoškić, M.; Plavšić, D.; Trinajstić, N. Inhibition of the Hill Reaction by 2-Methylthio-4,6-bis(monoalkylamino)-1,3,5-triazines. *J. Molec. Struct. (Theochem)* **1997**, *394*, 57-65.

(16) Apostol, T. M. Calculus. Second Edition ed., Blaisdell Publishing Co.: Waltham, Mass., 1969; Vol. II.

(17) Hildebrand, F. B. Introduction to Numerical Analysis. , McGraw-Hill Book Company, Inc: New York, 1956.

(18) Lučić, B., Personal Communication.

(19) Derive 5: http://education.ti.com/us/product/software/derive/features/features.html.

(20) Maple 7: http://www.maplesoft.com/.