Retro-Regression—Another Important Multivariate Regression Improvement†

Milan Randić‡

Department of Mathematics and Computer Science, Drake University, Des Moines, Iowa 50311, and National Institute of Chemistry, Hajdrihova 19, Ljubljana, Slovenia

Received July 22, 2000

We review the serious problem associated with instabilities of the coefficients of regression equations, referred to as the MRA (multivariate regression analysis) "nightmare of the first kind". This is manifested when in a stepwise regression a descriptor is included or excluded from a regression. The consequence is an unpredictable change of the coefficients of the descriptors that remain in the regression equation. We follow with consideration of an even more serious problem, referred to as the MRA "nightmare of the second kind", arising when optimal descriptors are selected from a large pool of descriptors. This process typically causes at different steps of the stepwise regression a replacement of several previously used descriptors by new ones. We describe a procedure that resolves these difficulties. The approach is illustrated on boiling points of nonanes which are considered (1) by using an ordered connectivity basis; (2) by using an ordering resulting from application of greedy algorithm; and (3) by using an ordering derived from an exhaustive search for optimal descriptors. A novel variant of multiple regression analysis, called retro-regression (RR), is outlined showing how it resolves the ambiguities associated with both "nightmares" of the first and the second kind of MRA.

INTRODUCTION

Multivariate regression analysis (MRA), one of the oldest data reduction techniques, remains still a method that is widely used, occasionally misused, and, we could also add, most of the times not used to its full potential. An important improvement of MRA, the construction of stable regression equations,[1−4] has found application, particularly among researchers in chemical graph theory.[2−13] Here we will consider another problem associated with multivariate regression equations illustrated on the boiling points of 35 isomers of nonane using five connectivity indices as descriptors.[14,15]

Orderly Stepwise Regression. In Table 1 (the top part) we have listed stepwise regression equations using from one to five connectivity indices. The connectivity indices are ordered according to the length of the underlying paths as ${}^0\chi$, ${}^1\chi$, ${}^2\chi$, ${}^3\chi$, ${}^4\chi$, ${}^5\chi$, ${}^6\chi$. The numerical values of the connectivity indices ${}^k\chi$ ($k = 0-6$) of nonane isomers can be found in a book of Kier and Hall[16] and also in a more recent paper by Lučić and Trinajstić.[17] The boiling points are taken as reported in refs 17 and 18. The statistical parameters associated with the multivariate regressions of Table 1 are listed in Table 2 (the top part). As we see from Table 2 each successive step increases the coefficient of the regression (r), reduces the standard error (s), and increases the Fisher ratio (F). All the above three features of the regression suggest an improvement of the quality of the regression with each step, although the question of how many descriptors could justifiably be used on $n = 35$ data points remains open. Because in this paper we will focus attention on how to improve MRA we will not be concerned with the problem of determining the permissible maximal number of descriptors that one can use—a topic that deserves more attention.

Table 1: Stepwise Regression Equations for Five Connectivity Indices[a]

ordered connectivities
$BP = -19.015\ {}^0\chi + 276.979$
$BP = +40.510\ {}^0\chi + 68.953\ {}^1\chi - 444.628$
$BP = -156.743\ {}^0\chi - 387.164\ {}^1\chi - 72.645\ {}^2\chi + 3144.538$
$BP = -92.497\ {}^0\chi\_ - 92.259\ {}^1\chi - 6.177\ {}^2\chi + 14.052\ {}^3\chi + 1189.433$
$BP = -284.008\ {}^0\chi - 451.881\ {}^1\chi - 50.913\ {}^2\chi + 20.746\ {}^3\chi + 13.106\ {}^4\chi + 4210.087$
greedy algorithm
$BP = -8.354\ {}^2\chi + 167.166$
$BP = -7.533\ {}^2\chi + 4.743\ {}^3\chi - 154.083$
$BP = +12.066\ {}^2\chi + 16.687\ {}^3\chi - 64.175\ {}^0\chi + 530.232$
$BP = +15.356\ {}^2\chi + 16.719\ {}^3\chi - 82.449\ {}^0\chi - 10.033\ {}^5\chi + 656.559$
$BP = -14.763\ {}^2\chi + 12.347\ {}^3\chi - 130.411\ {}^0\chi - 10.487\ {}^5\chi - 153.177\ {}^1\chi + 1757.520$
exhaustive search
$BP = -8.354\ {}^2\chi + 167.166$
$BP = -26.141\ {}^0\chi + 9.496\ {}^3\chi + 309.167$
$BP = +46.937\ {}^1\chi - 12.574\ {}^4\chi - 20.024\ {}^5\chi - 34.473$
$BP = -109.338\ {}^0\chi - 80.649\ {}^1\chi + 14.703\ {}^3\chi - 10.743\ {}^5\chi + 1245.878$
$BP = -0.665\ {}^0\chi + 3.103\ {}^3\chi - 12.101\ {}^4\chi - 21.760\ {}^5\chi - 11.762\ {}^1\chi + 526.142$

[a] (Top part) ordered according to the length of the underlying paths used in their calculation; (middle part) ordered according to greedy algorithm; and (bottom part) selected by an exhaustive search as (reported by Lučić and Trinajstić).[17]

THE MRA "NIGHTMARE OF THE FIRST KIND"

Throughout this article we will assume that regressions using up to five connectivity descriptors are legitimate for the characterization of the boiling points of the 35 nonane isomers. The problem we wish to consider is the dramatic change of the coefficients of regressions at subsequent steps

† This paper is dedicated to Professor F. A. Cotton, on the occasion of his 70th birthday.

‡ Corresponding author.

10.1021/ci000106d CCC: $20.00
Published on Web 03/31/2001

Table 2: Statistical Parameters: the Regression Coefficient *r*, the Standard Error *s*, and the Fisher Ratio *F* for Connectivity Indices[a]

n	descriptors	*r*	*s*	*F*
	Ordered Connectivities			
1	$^0\chi$	0.5617	5.146	15.2
2	$^0\chi$, $^1\chi$	0.7632	4.082	22.3
3	$^0\chi$, $^1\chi$, $^2\chi$	0.8937	2.880	41.0
4	$^0\chi$, $^1\chi$, $^2\chi$, $^3\chi$	0.9263	2.459	45.3
5	$^0\chi$, $^1\chi$, $^2\chi$, $^3\chi$, $^4\chi$	0.9643	1.757	76.9
	Greedy Algorithm			
1	$^2\chi$	0.7764	3.920	50.1
2	$^2\chi$, $^3\chi$	0.8552	3.274	43.6
3	$^2\chi$, $^3\chi$, $^0\chi$	0.9244	2.448	60.7
4	$^2\chi$, $^3\chi$, $^0\chi$, $^5\chi$	0.9618	1.786	92.6
5	$^2\chi$, $^3\chi$, $^0\chi$, $^5\chi$, $^1\chi$	0.9667	1.698	82.7
	Exhaustive Search			
1	$^2\chi$	0.7764	3.920	50.1
2	$^0\chi$, $^3\chi$	0.8998	2.757	68.0
3	$^1\chi$, $^4\chi$, $^5\chi$	0.9563	1.877	110.4
4	$^0\chi$, $^1\chi$, $^3\chi$, $^5\chi$	0.9656	1.697	103.3
5	$^0\chi$, $^3\chi$, $^4\chi$, $^5\chi$, $^6\chi$	0.9676	1.67	85.2

[a] (Top part) for ordering based on length of the paths; (middle part) ordered according to greedy algorithm; and (bottom part) selected by an exhaustive search.

in stepwise regression. This chaotic behavior of the coefficients of stepwise regression equations is typical for MRA and can be traced back to the strong interrelationship of the descriptors.[1] We will refer to this as "the MRA nightmare of the first kind", because, as we will see later, there is another MRA "nightmare" associated with the selection of optimal descriptors, to which we will refer as "the MRA nightmare of the second kind". When we use instead of the descriptors d_1, d_2, d_3, ... their orthogonalized counterparts d_1*, d_2*, d_3*, ..., the stepwise regression equation become[1]

$$\text{PROPERTY} = c_{11}\, d_1^* + \text{const.}_1$$

$$\text{PROPERTY} = c_{11}\, d_1^* + c_{22}\, d_2^* + \text{const.}_1$$

$$\text{PROPERTY} = c_{11}\, d_1^* + c_{22}\, d_2^* + c_{33}\, d_3^* + \text{const.}_1$$

....

It is interesting to observe that the coefficients that appear in orthogonalized equations are the same as the "diagonal" coefficients of the "unstable" stepwise regressions equations corresponding to orthogonalized descriptors shown before! Hence, one can extract the "stable" stepwise regression equations from the set of ordinary stepwise regression equations based on nonorthogonalized descriptors without constructing orthogonalized descriptors d_k*. Thus the orthogonalized stepwise regression corresponding to the equations of Table 1 are

$$\text{BP} = -19.0148\ ^0\chi^* + 276.9794$$

$$\text{BP} = -19.0148\ ^0\chi^* + 68.9534\ ^1\chi^* + 276.9794$$

$$\text{BP} = -19.0148\ ^0\chi^* + 68.9534\ ^1\chi^* - 72.6448\ ^2\chi^* + 276.9794$$

$$\text{BP} = -19.0148\ ^0\chi^* + 68.9534\ ^1\chi^* - 72.6448\ ^2\chi^* + 14.0521\ ^3\chi^* + 276.9794$$

$$\text{BP} = -19.0148\ ^0\chi^* + 68.9534\ ^1\chi^* - 72.6448\ ^2\chi^* + 14.0521\ ^3\chi^* + 13.1059\ ^4\chi^* + 276.9794$$

It is also interesting to find out that orthogonal descriptors constructed as residuals of successive correlations among descriptors can also be viewed to be orthogonal in the sense of Gram-Schmidt procedure for orthogonalization of vectors.[19]

Greedy Algorithm. In Table 1 (the middle part) we show the regression equations and in Table 2 (the middle part) the statistical parameters of stepwise regressions using from one to five connectivity descriptors selected according to the greedy algorithm. Here we choose as the first descriptor one which gives the highest correlation coefficient (or the smallest standard error or the largest Fisher ratio). The selection of descriptors which determines the greedy algorithm is illustrated in Table 3, in which at each step are given *r*, *s*, and *F*. In each row the best result has been emphasized, and the corresponding descriptor kept in all the successive stepwise regressions. As a result we obtained the following ordering of descriptors: $^2\chi$, $^3\chi$, $^0\chi$, $^5\chi$, $^1\chi$.

An advantage of the orderly algorithm is that it allows a comparison of regression equations for different properties of the same compounds (e.g., ref 20). This is not possible in the case of greedy algorithm, because different properties will induce different order and different composition of the set of the connectivity indices. An advantage of greedy algorithm is that it gives better regressions for the same number of descriptors. However, the greedy algorithm does not necessarily lead to the best combination of descriptors. To find the best combination of *n* descriptors one would have to examine all combinatorial possibilities of descriptors. This task has not been practical, particularly when one consider combinations of several descriptors form a rather large pool of descriptors. However, recently Lučić, Trinajstić,[16] and coworkers outlined an efficient procedure which examines regression coefficients of all combinations of descriptors without actually performing the regressions.[21−27] This approach opens the possibility for an exhaustive selection of the best descriptors but as we will see such procedure introduces novel difficulties for an interpretation of the results.

THE MRA "NIGHTMARE OF THE SECOND KIND"

We will illustrate difficulties associated with an exhaustive search for the best combination of descriptors in MRA to be addressed here by considering one of the recent papers of Lučić and Trinajstić.[17] They searched for best descriptors for the correlation of the boiling points of the 35 nonane isomers using up to seven connectivity indices. Among 10^9 combinations in reasonable time they found the best single, the best two, the best three, etc., descriptor.[27] In Table 1 (the bottom part) we show the regression equations, and in Table 2 (the bottom part) we show the statistical parameters of associated regressions using from one to five connectivity descriptors.

As expected, the statistical parameters shown in Table 2 (bottom) at each step (except, of course, when a single

Table 3: Stepwise Selection of the Best Descriptors[a] by the Greedy Algorithm

	$^0\chi$	$^1\chi$	$^2\chi$	$^3\chi$	$^4\chi$	$^5\chi$	$^6\chi$
r	0.5617	0.6848	**0.7764**	0.5123	0.0931	0.0087	0.3175
s	5.15	4.53	**3.92**	5.34	6.19	622	5.90
F	15.2	29.1	**50.1**	11.7	0.3	0.0	3.7
r	0.8036	0.8199		**0.8552**	0.8166	0.8447	0.7768
s	3.76	3.62		**3.27**	3.65	3.38	3.98
F	29.2	1.8		**43.6**	32.0	39.8	24.3
r	**0.9244**	0.9086			0.8722	0.8643	0.9086
s	**2.45**	2.68			3.14	3.23	2.68
F	**60.6**	48.9			32.9	30.5	48.9
r		0.9263			0.9390	**0.9618**	0.9297
s		2.46			2.24	**1.79**	2.40
F		45.3			55.9	**92.6**	47.8
r		**0.9667**			0.9646		0.9618
s		**1.70**			1.75		1.81
F		**82.7**			77.6		71.6

[a] Emphasized in boldface print.

descriptor is used) surpass the statistical parameters associated with the greedy approach (Table 2, middle). The regression equation at the bottom of Table 1 at first sight appears no different than those in the middle or at the top of Table 1, but there is an *important* difference: Several of the best descriptors found in step k of the stepwise regression often are no longer included as the best descriptors selected in step k+1. As we see from the bottom of Table 1, in fact, at every step at least one and sometimes several descriptors from the previous step no longer appear in the successor step. We refer to this apparent ambiguity as the "nightmare of the second kind" in view that here even descriptors are not the same, so the orthogonalization procedure as outlined earlier cannot be performed. References 20−26 illustrate additional cases displaying the "nightmare" of the second kind.

How is one to interpret the results of such analyses? What to do? We will outline in the next section how to proceed and how to reduce the "nightmare of the second kind" to the "nightmare of the first kind", for which, as we have seen, there is a cure: construction of orthogonalized descriptors.

STRUCTURE−PROPERTY SUBSPACE

When considering interpretation of MRA results the question to consider is as follows: how does one decide what are the best descriptors for a particular property. Should one include all descriptors used in different steps, or should one discard the initial descriptors, despite that sometimes they can account for a large part of the variability of the property? Most MRAs report calculated properties and residuals without discussing the *regression equations* on which the calculations are based. Usually *individual* descriptors are highlighted and attempts have been made to interpret their use. But such interpretations are ambiguous, to say the least, because of the ambiguities arising from the interrelation of descriptors. Moreover, in our view focusing attention to individual descriptors continues to be misleading. Instead of considering the role of individual descriptors that occur in a regression equation one should consider *all* descriptors that occur in the regression equation. What is important for characterization of the results is the *subspace* spanned by the *set of the descriptors* occurring in the equation, rather then individual descriptors. As we see from the bottom part of Table 1 at different steps of the stepwise regression one "jumps" from one structure−property subspace (of smaller dimension) to another structure−property subspace (of higher dimension), but the higher subspace does not contain the smaller subspace as its own part.

Table 4

	$^0\chi$	$^3\chi$	$^4\chi$	$^5\chi$	$^6\chi$
r	0.8722	0.9626	0.9399	0.9232	**0.9626**
s	3.19	1.79	2.23	2.51	**1.77**
F	23.8	94.6	56.8	43.3	**94.6**
r	0.6140	0.9066	**0.9250**	0.9024	
s	5.07	2.71	**2.44**	2.77	
F	6.3	47.7	**61.2**	45.3	
r	0.5957	0.7632		**0.8998**	
s	5.07	4.08		**2.76**	
F	8.8	22.3		**68.0**	
r	0.5123	**0.5617**			
s	5.34	**5.15**			
F	11.7	**15.2**			
r	**0.5123**				
s	**5.34**				
F	**11.7**				

RETRO-REGRESSION

A way out of this quandary is retro-regression, to be outlined here. Let us reexamine the last regression equation of Table 1:

$$\mathrm{BP} = -50.665\ ^0\chi + 3.4103\ ^3\chi - 12.101\ ^4\chi - 21.760\ ^5\chi - 11.762\ ^6\chi + 526.141$$

Let suppose that this equation represents an acceptable solution. This equation defines structure−property subspace given by the set of descriptors $\{^0\chi, ^3\chi, ^4\chi, ^5\chi, ^6\chi\}$. For this structure−property subset we need to find a suitable basis. This means to find the order in which to choose descriptors for a sequence of stepwise regressions that give the above subspace as the final answer. To get there we view the regression steps at the bottom of Table 1 as the history of arriving at the final solution. However, by having found the structure-subspace of interest we are longer interested in its history! Instead, one searches for a route that leads to the *same* structure−property subspace but in such a way that we stay all the time in the subspace $\{^0\chi, ^3\chi, ^4\chi, ^5\chi, ^6\chi\}$. Retro-regression is the answer. In the retro-regression one starts with the five *final* descriptors that define the solution set and seeks the least important descriptor to be eliminated in a stepwise fashion. The least important descriptor is taken to be one that makes the smaller increase of the standard errors when removed from the set.

In Table 4 we show all the steps of retro-regression. The descriptors are indicated at the head of the columns of Table 4 and the statistical parameters at the beginning of each row. In each section the emphasized entry signifies the descriptor selected to be deleted at that stage. Thus the last entry in the first row (under the heading $^6\chi$) gives the statistical parameters for the combination of descriptors $\{^0\chi, ^3\chi, ^4\chi, ^5\chi\}$. As we see, this particular combination among the five possibilities obtained by excluding one of the five descriptors gives the smallest standard error. This indicates that descriptor $^6\chi$ is the least important of the five descriptors. The process continues, and eventually we obtain the connectivity descriptors ordered according to their importance in the regression as $^0\chi, ^3\chi, ^5\chi, ^4\chi, ^6\chi$. This procedure provides an

Table 5: Stepwise Regression Equations for the Five Connectivity Indices Selected by the Exhaustive Search of Descriptors and Ordered by the Retro-Regression[a] and the Corresponding Data for Orthogonal Regression Equations for the Same Five Connectivity Indices

					const
Stepwise Regression Equations					
$-19.01480\ {}^{0}\chi$					276.97938
± 4.87598					± 35.89972
$-26.14091\ {}^{0}\chi$	$9.49556\ {}^{3}\chi$				309.16858
± 2.72625	± 1.04210				± 19.55042
$-32.37795\ {}^{0}\chi$	$7.97357\ {}^{3}\chi$	$-7.91831\ {}^{5}\chi$			361.29401
± 3.12252	± 1.04121	± 2.51793			± 23.95542
$-39.98663\ {}^{0}\chi$	$5.67065\ {}^{3}\chi$	$-16.05473\ {}^{5}\chi$	$-7.77910\ {}^{4}\chi$		433.89672
± 2.66871	± 0.86763	± 2.36990	± 1.444499		± 21.98843
$-50.66500\ {}^{0}\chi$	$3.10324\ {}^{3}\chi$	$-21.75956\ {}^{5}\chi$	$-12.10063\ {}^{4}\chi$	$-11.76230\ {}^{6}\chi$	526.14158
± 5.66563	± 1.47024	± 3.51811	± 2.46661	± 5.58509	± 48.50062
Corresponding Data for Orthogonal Regression Equations					
$-19.01480\ {}^{0}\chi$					276.97938
± 4.87598					± 35.89972
$-19.01489\ {}^{0}\chi$	$9.49555\ {}^{3}\chi^*$				276.98003
± 2.61167	± 1.04210				± 19.22860
$-19.01489\ {}^{0}\chi$	$9.49554\ {}^{3}\chi^*$	$-7.91832\ {}^{5}\chi^*$			276.97961
± 2.59290	± 0.92189	± 2.51794			± 17.01048
$-19.01479\ {}^{0}\chi$	$9.49553\ {}^{3}\chi^*$	$-7.91834\ {}^{5}\chi^*$	$-7.77910\ {}^{4}\chi^*$		276.97930
± 1.67498	± 0.66834	± 1.82544	± 1.44500		± 12.33214
$-19.01474\ {}^{0}\chi$	$9.49551\ {}^{3}\chi^*$	$-7.91840\ {}^{5}\chi^*$	$-7.77905\ {}^{4}\chi^*$	$-11.76339\ {}^{6}\chi^*$	276.97891
± 1.58658	± 0.63307	± 1.72910	± 1.36873	± 5.58502	± 11.68128

[a] Under each coefficient the standard errors are shown. The statistical parameters with each regression, the regression coefficient *r*, the standard error *s*, and the Fisher ratio *F*, are shown in bold in Table 4.

ordering for the descriptors occurring in the last row of Table 1. Because we ordered descriptors, we can now construct a stepwise regression in which descriptors from the previous step will always be included in the following steps. The resulting equations are listed in Table 5 (the top part), where we included also the standard errors for the coefficients of the equations.

We succeeded to eliminate the "nightmare of the second kind", associated with the occurrence of different descriptors in successive stepwise regressions. All that is left is only the "nightmare of the first kind", that is reflected in variations of the magnitudes of the coefficients for individual descriptors at different steps of the regression. As we mentioned earlier the stability of the regression equations can now be obtained by using orthogonalized descriptors. The final orthogonalized regression equation is

$$\mathrm{BP} = -19.0148\ {}^{0}\chi^* + 9.4956\ {}^{3}\chi^* - 7.9183\ {}^{5}\chi^* - 7.7791\ {}^{4}\chi^* - 11.7623\ {}^{6}\chi^* + 276.9794$$

Here the asterisk indicates orthogonalized descriptors. Thus ${}^{0}\chi^*$ is the same as ${}^{0}\chi$, while ${}^{3}\chi^*$ is the residual of a regression of ${}^{3}\chi$ against ${}^{0}\chi^*$, and so on.

ORTHOGONAL DESCRIPTORS

In the lower part of Table 5 we show the standard errors for the coefficients of the orthogonalized equations, while in Table 6 we show the correlation matrices for nonorthogonal and orthogonalized descriptors of Table 5. As we see the correlation matrix for orthogonal descriptors is the identity submatrix, confirming that indeed we have orthogonal descriptors. Finally, let us draw the attention of our readers to standard errors of the coefficients of the regression equation. As we see from Table 5 the standard errors associated with the coefficients of orthogonalized equations are reduced at every successive step of the regression.

Table 6: Correlation Matrix for Nonorthogonal and Orthogonalized Descriptors of the Retro-Regression

	BP	${}^{0}\chi$	${}^{3}\chi$	${}^{5}\chi$	${}^{4}\chi$	${}^{6}\chi$
Upper Matrix						
BP	1.000	−0.562	0.512	−0.009	−0.093	0.317
${}^{0}\chi$	−0.562	1.000	0.287	−0.669	−0.146	−0.687
${}^{3}\chi$	0.512	0.287	1.000	−0.523	−0.190	−0.474
${}^{5}\chi$	−0.009	−0.669	−0.523	1.000	−0.259	0.482
${}^{4}\chi$	−0.093	−0.146	−0.190	−0.259	1.000	−0.191
${}^{6}\chi$	0.317	−0.687	−0.474	0.482	−0.191	1.000
	BP	1	2/1	3/2	4/3	5/4
Lower Matrix						
BP	1.000	−0.562	0.703	−0.215	−0.266	−0.099
1	−0.562	1	0	0	0	0
2/1	0.703	0	1	0	0	0
3/2	−0.215	0	0	1	0	0
4/3	−0.266	0	0	0	1	0
5/4	−0.099	0	0	0	0	1

CONCLUDING REMARKS

The approach of Lučić and Trinajstić, which represents an exhaustive search for best set of descriptors, allows one to find the best combination of descriptors from a large pool of topological indices. However, their approach only provides a "history" of arriving at the best structure−property subspace and does not allow one to construct orthogonalized descriptors which would yield stable regression equations, an essential prerequisite for interpretation of the model. We have outlined here a procedure, called retro-regression, which allows one to construct a basis for the best structure−property subspace, by ordering descriptors appearing in the last stepwise regression, that can be readily orthogonalized.

The technique outlined here, although it may appear to be of limited interest to the researcher in QSAR (quantitative structure−activity relationship) and LFER (linear free energy relationship), has broader applications. Orthogonalization of descriptors is important, for example, not only in such diverse applications as is similarity/dissimilarity analysis[28] but also

even for those using pattern recognition[29] and those using PCA (the Principal Component Analysis of Hotelling).[30] Although the principal components themselves are orthogonal by virtue of being eigenvalues of a matrix, it is often overlooked that the descriptors used for their construction are not orthogonal, which then represents a "nightmare" of a kind, and an interpretation of individual principal components remains ambiguous.

ACKNOWLEDGMENT

I would like to thank Professor A. T. Balaban for discussions and valuable comments on the manuscript.

REFERENCES AND NOTES

(1) Randić, M. Orthogonal molecular descriptors. *New J. Chem.* **1991**, *15*, 517−525.
(2) Randić, M. Resolution of ambiguities in structure−property studies by use of orthogonal descriptors. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 311−370.
(3) Randić, M. Fitting of non linear regressions by orthogonalized power series. *J. Comput. Chem.* **1993**, *14*, 363−370.
(4) Randić, M. Curve fitting paradox. *Int. J. Quantum Chem: Quantum Biol. Symp.* **1994**, *21*, 215−225
(5) Soškić, M.; Plavšić, D.; Trinajstić, N. Link between orthogonal and standard multiple linear regression models. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 829−832.
(6) Pogliani, L. Structure−property relationships of amino acids and some dipeptides. *Amino Acids* **1994**, *6*, 141−153
(7) Pogliani, L. Molecular connectivity descriptors of the physicochemical properties of the α-amino acids. *J. Phys. Chem.* **1994**, *98*, 1494−1499.
(8) Pogliani, L. Molecular modeling by linear combination of connectivity indexes. *J. Phys. Chem.* **1995**, *99*, 925−937.
(9) Pogliani, L. A strategy for molecular modeling of a physicochemical property using a linear combination of connectivity indexes. *Croat. Chem. Acta* **1996**, *69*, 95−109.
(10) Pogliani, L. Properties of molecular connectivity terms and physicochemical properties. *J. Mol. Struct. (THEOCHEM)* **1999**, *466*, 1−19.
(11) Pogliani, L. Modeling properties with higher-level molecular connectivity descriptors. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 104−111.
(12) Pogliani, L. From molecular connectivity indices to semiempirical connectivity terms: Recent trends in graph theoretical descriptors. *Chem. Rev.* **2000**, *100*, 3827−3858.
(13) Ivanciuc, O.; Ivanciuc, T.; Carbos-Bass, D.; Balaban, A. T. Comparison of weighting schemes for molecular descriptors: Application on quantitative structure − retention relationship models for alkylphenols in gas−liquid chromatography. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 732−743.
(14) Randić, M. On the characterization of molecular branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609−6615.
(15) Kier, L. B.; Murray, W. J.; Randić, M.; Hall, L. H. Molecular connectivity V. Connectivity series concept applied to density. *J. Pharm. Soc.* **1976**, *65*, 1226−1230.
(16) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research;* Academic Press, New York, 1976.
(17) Lučić, B.; Trinajstić, N. New developments in QSPR/QSAR modeling based on topological indices. *SAR QSAR Environ. Res.* **1997**, *7*, 45−62.
(18) Randić, M.; Trinajstić, N. Isomeric variations in alkanes: Boiling points of nonanes. *New J. Chem.* **1994**, *19*, 179−188.
(19) Randić, M. A comment on construction of orthogonal molecular descriptors, *J. Chem. Inf. Comput. Sci.* Submitted for publication.
(20) Randić, M.; Seybold, P. G. Molecular shape as a critical factor in structure−property-activity studies. *SAR QSAR Environ. Res.* **1993**, *1*, 77−85.
(21) Lučić, B.; Trinajstić, N.; Sild, S.; Karelson, M.; Katritzky, A. R. A new efficient approach for variable selection on Multiregression: Prediction of gas chromatographic retention ties and response factors. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 610−61.
(22) Lučić, B.; Nikolić, S.; Trinajstić, N.; Jurić, D.; Mihalić, Z. A structure-property study of the solubility of aliphatic alcohols in water. *Croat. Chem. Acta* **1995**, *68*, 417−434.
(23) Lučić, B.; Nikolic, S.; Trinajstić, N.; Juretic, D.; Juric, A. A novel QSAR approach to physicochemical properties of the α-amino acids. *Croat. Chem. Acta* **1995**, *68*, 435−450.
(24) Amić, D.; Davidović-Amić, D.; Jurić, A.; Lučić, B.; Trinajstić, N. Structure−activity correlation of flavone derivatives for inhibition of cAMP phosphodiesterase. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1034−1038.
(25) Lučić, B.; Nikolić, S.; Trinajstić, N.; Juretic, D. The structure−property models can be improved using the orthogonalized descriptors. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 532−538.
(26) Amić, D.; Davidović-Amicc, D.; Beslo, D.; Lučić, B.; Trinajstić, N. The use of ordered orthogonalized multivariate linear regressions in a structure−activity study of coumerin and flavonoid derivatives as inhibitors of aldose reductase. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 581−586.
(27) 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.
(28) Randić, M. Orthosimilarity, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1092−1097.
(29) Kovalski, B. R.; Bender, C. F. Pattern recognition. A powerful approach to interpeting chemical data. *J. Am. Chem. Soc.* **1972**, *94*, 4, 5632−5639.
(30) Hotelling, H. Analysis of a complex of statistical variables into principal componentns. *J. Educ. Psychol.* **1933**, *24*, 417−441 and 489−520.

CI000106D