Journal of Mathematical Chemistry, Vol. 43, No. 4, May 2008 (© 2007)
DOI: 10.1007/s10910-007-9241-y

Correlations in spectral statistics

Dorota Bielińska-Wąż*

Instytut Fizyki, Uniwersytet Mikołaja Kopernika, Grudziądzka 5, 87-100 Toruń, Poland
E-mail: dsnake@phys.uni.torun.pl

Piotr Wąż

Centrum Astronomii, Uniwersytet Mikołaja Kopernika, Gagarina 11, 87-100 Toruń, Poland

Received 24 October 2005; Revised 22 January 2007

Statistical spectroscopy is applied to the theory of molecular similarity. Statistical moments of the intensity distributions are considered as a new kind of descriptors, in particular atomic or molecular ones. A model spectrum is taken as a sum of two Gaussian distributions characterized by different parameters. The linear correlations between distribution moments and between the parameters characterizing the spectra are studied. The distributions taken under considerations have been selected using genetical algorithms.

KEY WORDS: data mining, statistical theory of spectra, molecular similarity, genetic algorithms

AMS subject classification: 76M25

1. Introduction

The theory of molecular similarity has been initialized by R. Carbo et al. [1]. In the later papers, different kinds of molecular descriptors and methods in the field of molecular similarity have been formulated [2, 3, 4, 5]. In two recent papers [6, 7] we proposed a new set of similarity indices. These indices, statistical moments of the intensity distributions, relate shapes of spectra. Using these descriptors, the appropriate similarity distances have been defined. It is assumed that the degree of similarity of systems is correlated with the degree of similarity of their spectra. Using the new kind of descriptors, the applicability of statistical spectroscopy and genetic algorithms to the similarity studies has been demonstrated and dissimilarity maps have been presented [8].

The idea of using moments of intensity distributions as descriptors comes from the statistical theory of spectra that has already been applied in many areas

*Corresponding author.

0259-9791/08/0005-1287/0 © 2007 Springer Science+Business Media, LLC

of physics [9, 10, 11]. Let us just mention, a very efficient method (not limited by the sizes of matrices) of evaluation of extreme eigenvalues of matrices representing many-electron model Hamiltonians [12], methods of determining envolopes of the molecular electronic bands [13], or studies of 'quantum chaos' [14].

According to the *principle of moments*, the degree of similarity of a pair of distrubutions is determined by the number of the lowest moments that are equal for both distributions. The similarity increases when this number is larger. This idea sounds very attractive for the theory of similarity and its application in this field seems to be quite natural.

In this paper, the linear correlations between moments of intensity distributions and between the parameters characterizing the spectra are studied. The appropriate correlation matrices are calculated. A model spectrum is taken as a sum of two Gaussian distributions characterized by different parameters. However, it is worth to notice, that using this method, any kind of atomic or molecular spectra can be studied.

2. Correlation studies and the discussion

The studies on correlations have been performed using the same model spectra as in [8], i.e., an infinite number of spectra of the type

$$I^{\gamma}(E) = N^{\gamma}\left[a_1 \exp\left[-c_1(E-\epsilon_1)^2\right] + a_2 \exp\left[-c_2(E-\epsilon_2)^2\right]\right], \tag{1}$$

where $\gamma = \{c_1, a_1, \epsilon_1, c_2, a_2, \epsilon_2\}$ is considered. The particular parameters characterize the width (c_i), the amplitude (a_i), and the locations of the maxima (ϵ_i) of the ith Gaussian component $a_i \exp\left[-c_i(E-\epsilon_i)^2\right]$ of $I^{\gamma}(E)$, where $i = 1, 2$. N^{γ} is the normalization constant. The set of parameters that defines the space in which the spectra $I^{\gamma}(E)$ are defined, is restricted to $\gamma = \{5.0, 1.0, 1.2, 5.0 + \delta c, 1.0 + \delta a, 2.7 - \delta\epsilon\}$, where

$$\delta c \in \langle 0; 20\rangle, \quad \delta a \in \langle 0; 10\rangle, \quad \delta\epsilon \in \langle 0; 1\rangle. \tag{2}$$

The analytical expressions for the moments are given in [7, 8]. The problems are solved by finding global maxima of moments as functions of δc, δa, $\delta\epsilon$. For this purpose the genetic algorithm Pikaia [15] is used. The genetic algorithm performed within the restricted space defined in (2), with the termination condition 500 generations, and the accuracy of 10^{-8} gives as a result sets of parameters $\{\delta c, \delta a, \delta\epsilon\}$. These parameters used for the creation of $I^{\gamma}(E)$ functions and their moments are presented in figures 2–5. The number of $I^{\gamma}(E)$ has been restricted to 100.

The aim of this paper is to look for the correlations between distribution moments and between the parameters characterizing the spectra. Only the linear type of correlations is considered. The effect of the correlations may be a base

for determining the number of moments that should be taken into account in order to get a proper classification of the spectra.

In order to exclude the moments that are correlated, Pearson's correlations coefficients between distributions x and y

$$C_{xy} = \frac{\sum_{k=1}^{100}(x_k - \bar{x})(y_k - \bar{y})}{\sqrt{\sum_{k=1}^{100}(x_k - \bar{x})^2}\sqrt{\sum_{k=1}^{100}(y_k - \bar{y})^2}}, \tag{3}$$

where $C_{xy} \in \langle -1, 1\rangle$ have been calculated. We are interested only in strong linear correlations, i.e., in the cases when either $C_{xy} \in \langle 0.9, 1\rangle$ or $C_{xy} \in \langle -1, -0.9\rangle$. In the latter case we have the so called negative correlation, i.e., one of the quantities increases while the second one decreases. In our case, moments of the intensity distributions and then the parameters δc, δa, $\delta\epsilon$ are treated as statistical distributions x and y. In such a way the similarity of descriptors (distribution moments) and also the similarity of parameters for particular cases, are studied.

Similar distributions are characterized by similar moments. However, it may happen that spectra that are different may have one or more similar moments. For example, distributions with different width, different assymetry, and different excess may have the same mean values. It may also happen that the width of the distributions, with different mean values, different assymetry, and different excess, is the same. Such situations are presented in figure 1. This figure shows $I^{\gamma}(E)$ functions that have the same first moment ($M_1 = 2$) (upper figure) and the same second centered moment ($M_2' = \frac{1}{2}$) (lower figure). Only four distributions are presented (the ones with different δc).

Figure 2 presents moments of 100 distributions with the same first moment (four of them are plotted in the upper part of figure 1). The distributions are labeled from 1 to 100 and they are ordered according to the increasing value of M_2'. The correlation matrix between the moments in case of constant M_1 is

$$C^{M_1} = \frac{1}{100}\begin{pmatrix} 100 & 20 & 29 & -26 \\ & 100 & 83 & -95 \\ & & 100 & -89 \\ & & & 100 \end{pmatrix}. \tag{4}$$

The elements of the matrix are the Pearson's correlation coefficients between all the considered moments, for example $C_{1\,2} \equiv C_{M_1 M_2'}$. This matrix is symmetric and therefore only the upper triangle is presented. The diagonal elements correspond to the correlation between the same moments ($C_{i\,i} = 1$, $i = 1, 2, 3, 4$).

In the case considered (constant M_1), negative strong correlation appears between the second and the fourth moment ($C_{2\,4} = -0.95$) and between the third and the fourth one ($C_{3\,4} = -0.89$). This observation, that constant M_1 values results in some other correlations of moments is also seen in figure 2:

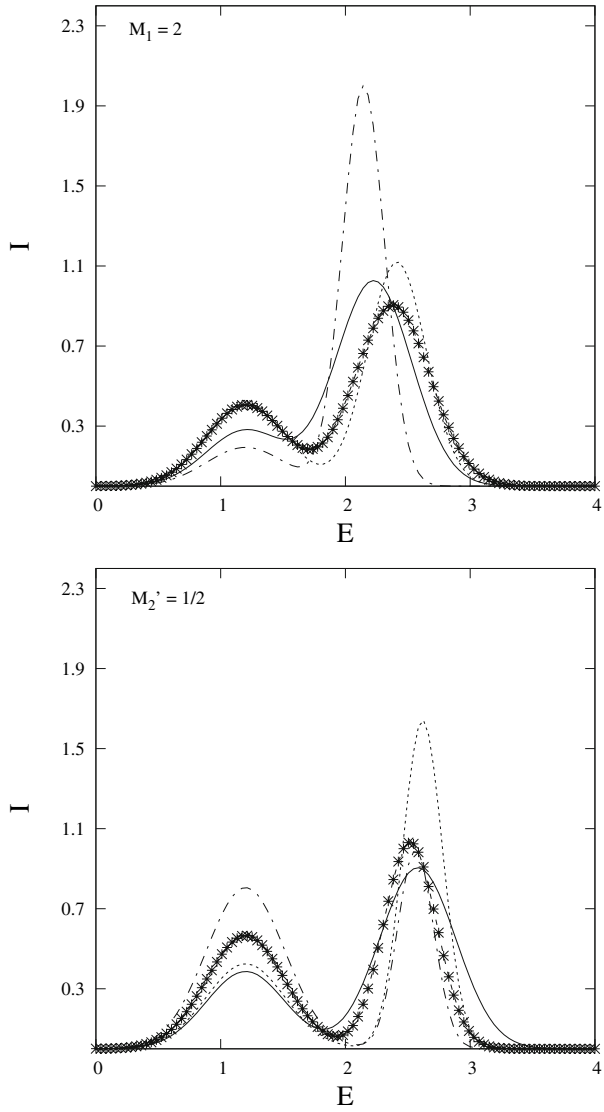

Figure 1. Intensity distributions corresponding to a constant M_1 (Upper figure) and M_2' (lower figure).

$$M_2' \sim -M_4'',$$
$$M_3'' \sim -M_4''.$$

As a result, there is only one linearly independent descriptor, which can be chosen as M_2'.

Figures 3–5 correspond to similar studies as the ones presented in figure 2, but for different moments assumed to be constant. Figure 3 presents moments

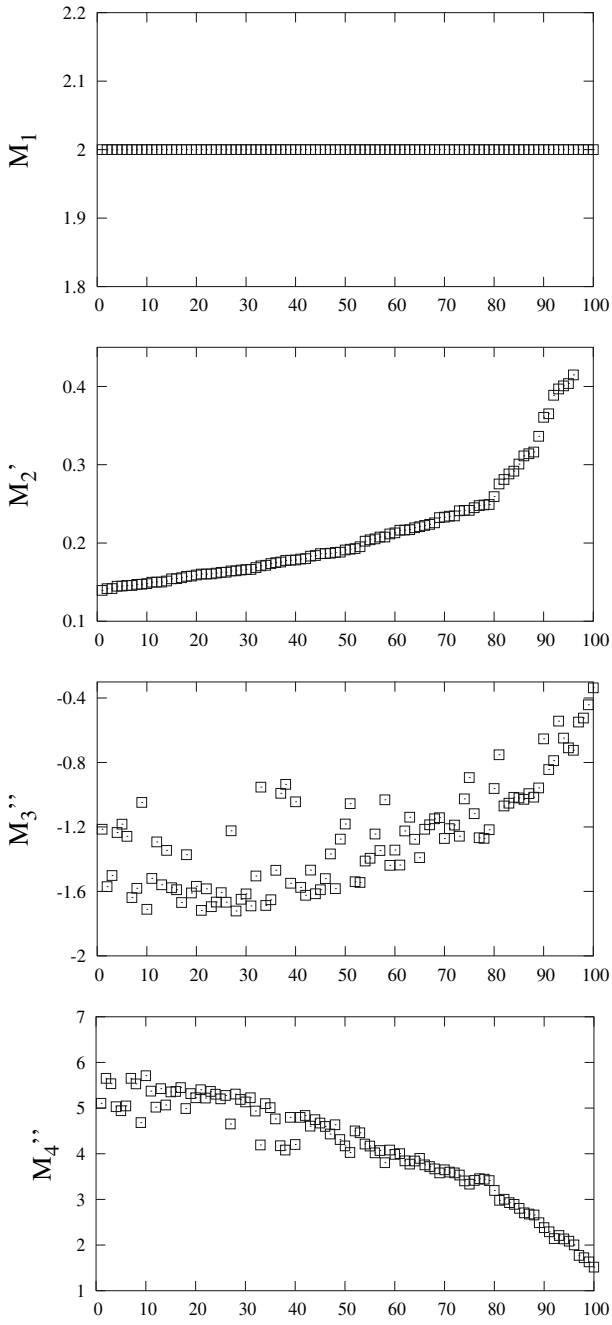

Figure 2. Moments of the intensity distributions corresponding to a constant M_1.

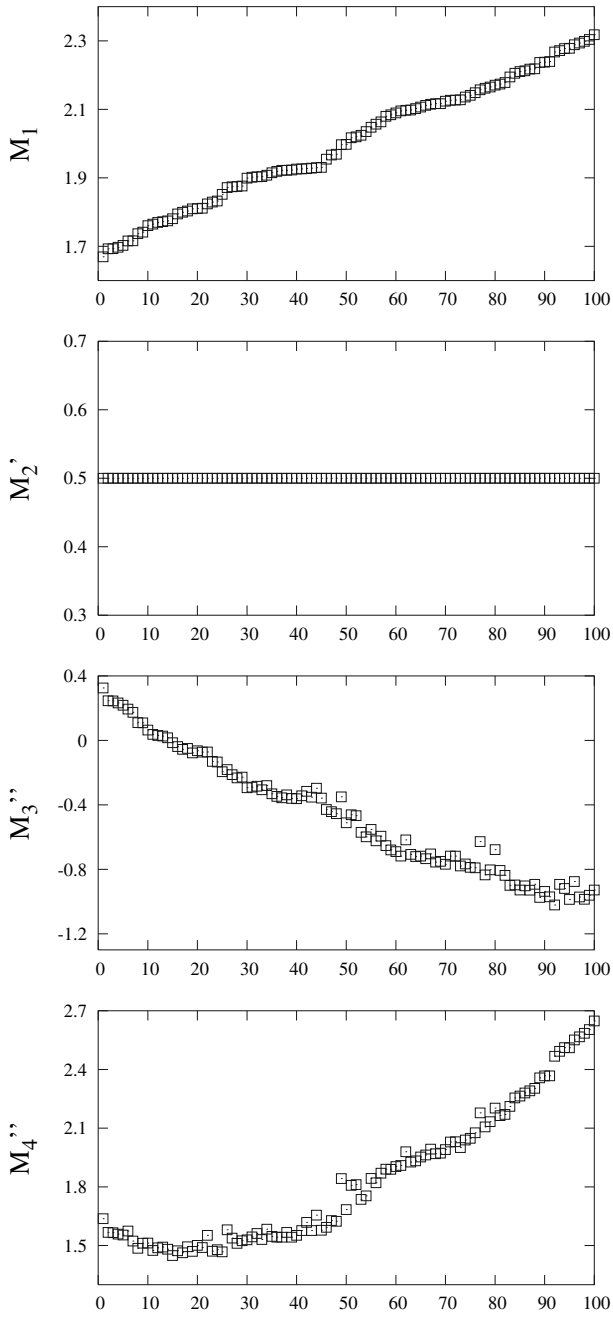

Figure 3. Moments of the intensity distributions corresponding to a constant M_2'.

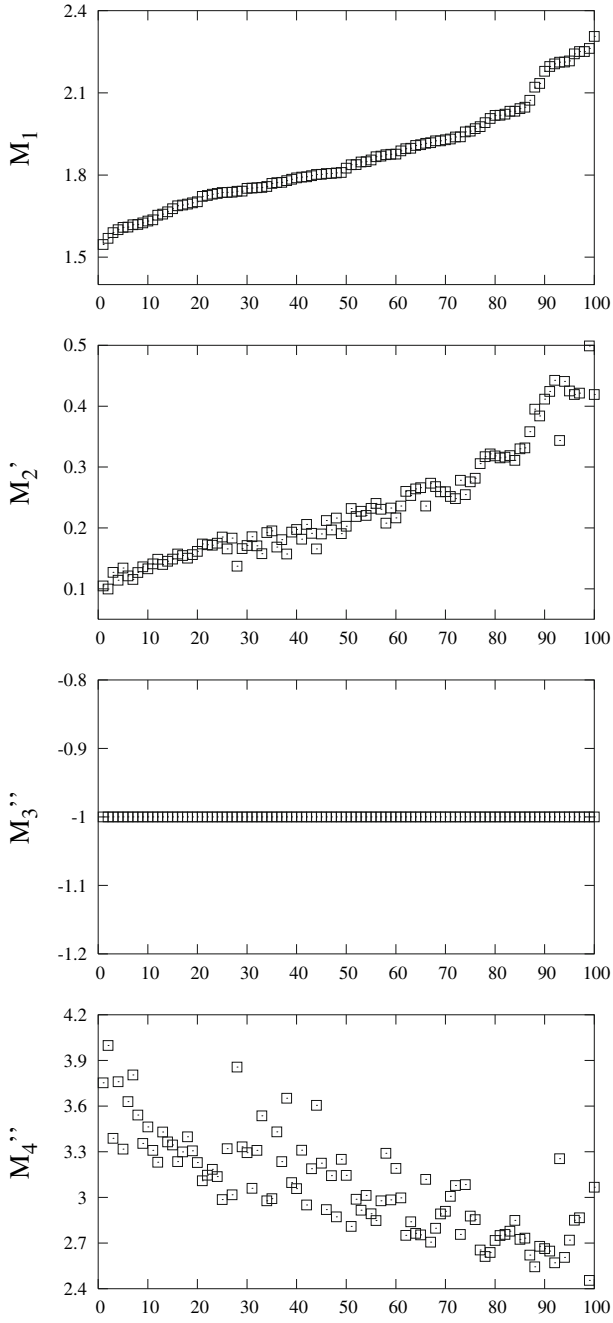

Figure 4. Moments of the intensity distributions corresponding to a constant M_3''.

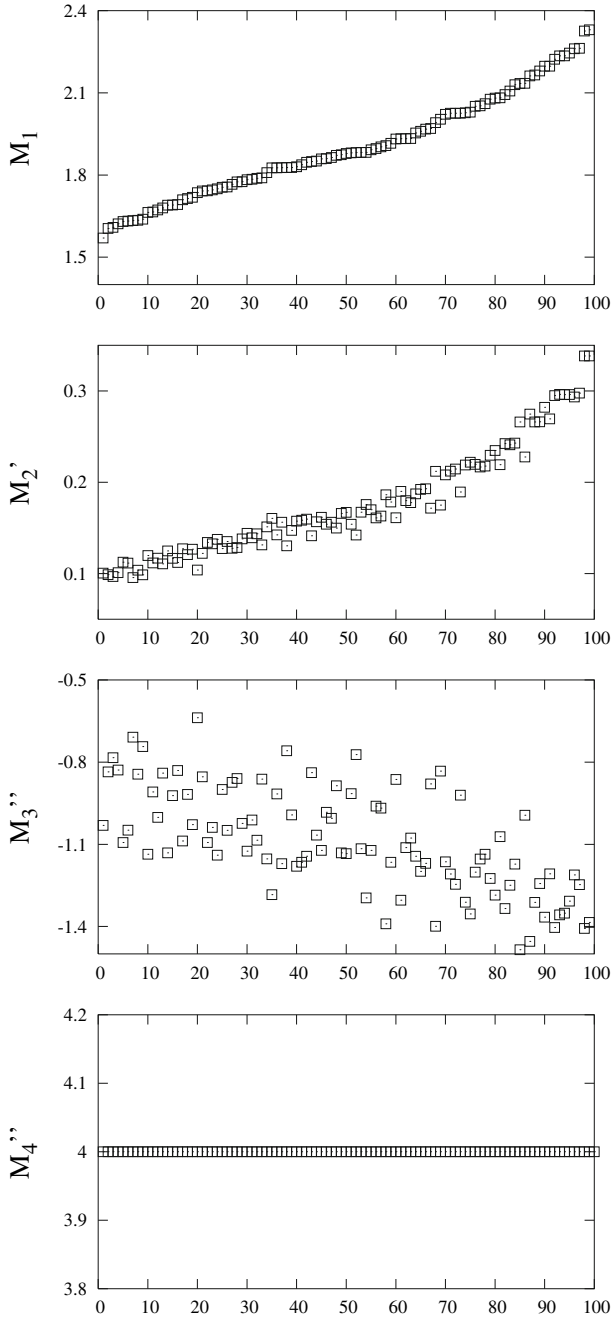

Figure 5. Moments of the intensity distributions corresponding to a constant M_4''.

of $I^{\gamma}(E)$ (four of them are presented in lower part of figure 1) in case when the secend centered moment is constant ($M_2' = \frac{1}{2}$). In figure 4, the third scaled moment is constant ($M_3'' = -1$). In figure 5 the fourth scaled moment is constant ($M_4'' = 4$).

In figure 3 (constant M_2') the distributions are ordered according to the increasing value of M_1. In this case, we observe correlations between the first, the fourth and the third moment:

$$M_1 \sim M_4'' \sim -M_3''.$$

It results in one linearly independent descriptor, for example M_1. The calculations confirm this observation. The correlation matrix in this case is

$$C^{M_2'} = \frac{1}{100}\begin{pmatrix} 100 & -4 & -99 & 97 \\ & 100 & 8 & 2 \\ & & 100 & -94 \\ & & & 100 \end{pmatrix}. \tag{5}$$

Strong correlation is for $C_{1\,3} = -0.99$, $C_{1\,4} = 0.97$, and $C_{3\,4} = -0.94$.

In figure 4 (constant M_3'') the distributions are ordered according to the increasing value of M_1. With the constant value of M_3'', we observe correlations between the first and the second moments:

$$M_1 \sim M_2'.$$

The number of linearly independent descriptors is two: (M_1', M_4''). The correlation matrix confirms these observations:

$$C^{M_3''} = \frac{1}{100}\begin{pmatrix} 100 & 99 & -42 & -76 \\ & 100 & -47 & -79 \\ & & 100 & 35 \\ & & & 100 \end{pmatrix}. \tag{6}$$

Strong correlation is for $C_{1\,2} = 0.99$.

The same correlations are observed in figure 5 (constant M_4''):

$$M_1 \sim M_2'.$$

The number of linearly independent moments is two: (M_1', M_3''). The correlation matrix confirms these observations:

$$C^{M_4''} = \frac{1}{100}\begin{pmatrix} 100 & 97 & -46 & 6 \\ & 100 & -64 & 14 \\ & & 100 & -31 \\ & & & 100 \end{pmatrix}. \tag{7}$$

Strong correlation is for $C_{1\,2} = 0.97$.

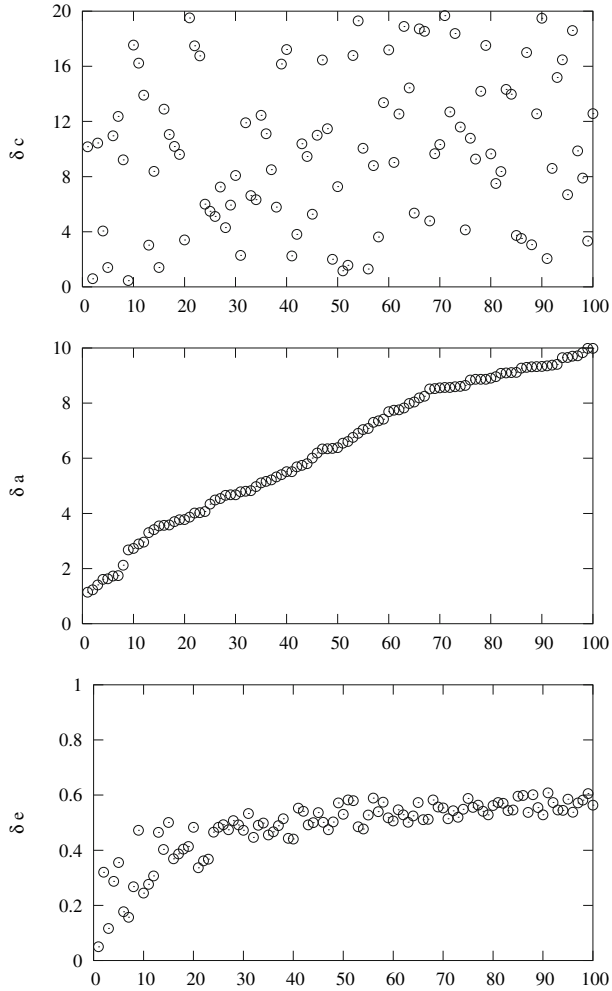

Figure 6. Parameters corresponding to a constant M_1.

Let us look for the correlations between the shapes of the spectra (characterized by parameters δc, δa, $\delta\epsilon$) for particular cases of constant moments. The Pearson's correlation matrices between parameters

$$C = \begin{pmatrix} C_{\delta c\delta c} & C_{\delta c\delta a} & C_{\delta c\delta\epsilon} \\ & C_{\delta a\delta a} & C_{\delta a\delta\epsilon} \\ & & C_{\delta\epsilon\delta\epsilon} \end{pmatrix} \tag{8}$$

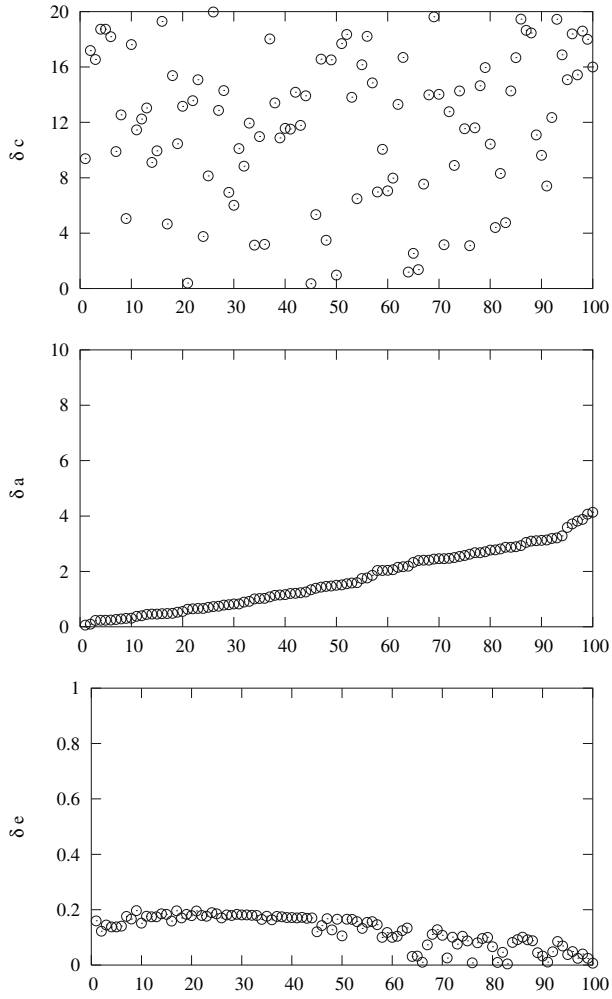

Figure 7. Parameters corresponding to a constant M_2'.

in the cases of M_1, M_2', M_3'', M_4'' constant are, respectively, equal to:

$$C^{M_1} = \frac{1}{100}\begin{pmatrix} 100 & 39 & 9 \\ & 100 & 90 \\ & & 100 \end{pmatrix}, \tag{9}$$

$$C^{M_2'} = \frac{1}{100}\begin{pmatrix} 100 & 40 & -7 \\ & 100 & -91 \\ & & 100 \end{pmatrix}, \tag{10}$$

$$C^{M_3''} = \frac{1}{100}\begin{pmatrix} 100 & 70 & 47 \\ & 100 & 36 \\ & & 100 \end{pmatrix}, \tag{11}$$

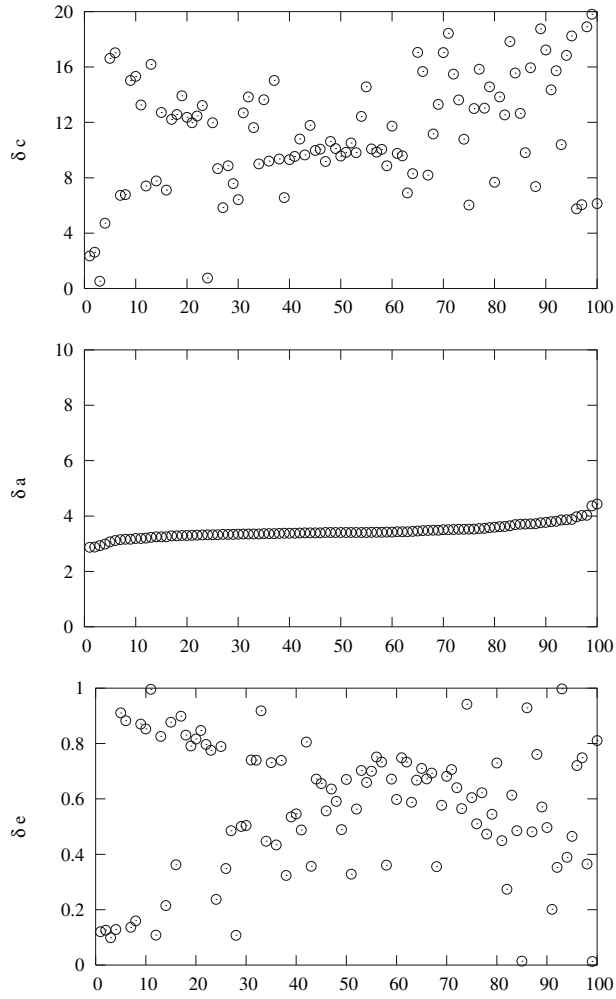

Figure 8. Parameters corresponding to a constant M_3''.

$$C^{M_4''} = \frac{1}{100}\begin{pmatrix} 100 & -64 & 63 \\ & 100 & -2 \\ & & 100 \end{pmatrix}. \tag{12}$$

As one can see, in the majority of cases there are no linear correlations between the parameters. The only strong linear correlation has been found between δa and $\delta\epsilon$ in the case of constant M_1 ($C_{\delta a\delta\epsilon} = 0.90$) and the only negative strong linear correlation has been found between δa and $\delta\epsilon$ in the case of constant M_2' ($C_{\delta a\delta\epsilon} = -0.91$). These observations are illustrated in figures 6–9 (figure 6 – constant M_1, figure 7 – constant M_2', figure 8 – constant M_3'', figure 9 – constant

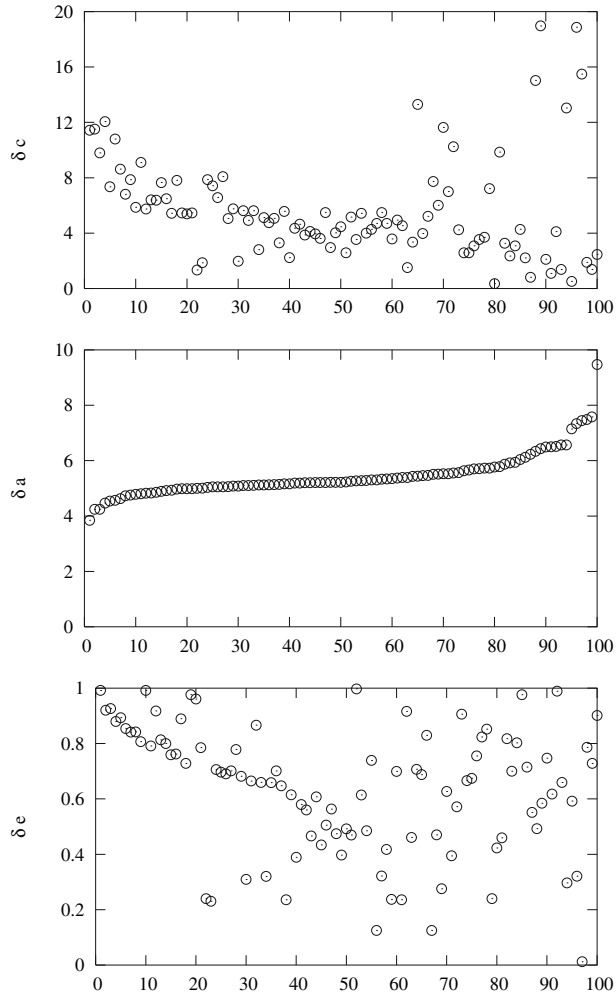

Figure 9. Parameters corresponding to a constant M_4''.

M_4''). In all cases the distributions are ordered according to the increasing value of δa.

Generally, there are no strong linear correlations between the sets of parameters corresponding to the distributions for which a selected property is set to a constant value.

Summarizing, one can observe that the standard first four lowest moments, taken usually in statistical spectroscopy are not a universal basis of descriptors. We have extracted distributions of different shapes for which a selected property (a selected moment) is constant. We have noticed that in such cases linear correlations between other properties may occur. The appropriate correlation

coefficients have been calculated. As a consequence the number of linearly independent descriptors decreases. In some cases higher-order moments have to be calculated for a more precise statistical description of spectra.

References

[1] R. Carbo, L. Leyda and M. Arnau, Int. J. Quantum Chem. 17 (1980) 1185.
[2] R. Carbo, B.Calabuig, in *Molecular Similarity*, ed. M.A. Johnson et al. (Wiley, New York, 1990).
[3] R. Carbo-Dorca and P.G. Mezey (eds.), *Advances in Molecular Similarity*, Vol. 2 (JAI Press, Stamford, CN, 1998), p. 297.
[4] J. Devillers and A.T. Balaban (eds.), *Topological Indices and Related Descriptors in QSAR and QSPR* (Gordon and Breach Science Publishers, The Netherlands, 1999), p. 811.
[5] S.C. Basak, B.D. Gute, D. Mills and D.M. Hawkins, J. Mol. Struct. (Theochem) 622 (2003) 127.
[6] D. Bielińska-Wąż, P. Wąż, S.C. Basak and R. Natarajan, in: *Symmetry, Spectroscopy and SCHUR*, ed. R.C. King et al. (Nicolaus Copernicus University Press, Toruń, 2006), pp. 27–32.
[7] D. Bielińska-Wąż, P. Wąż and S.C. Basak, Eur. Phys. J. B 50 (2007) 333, DOI: 10.1007/s10910-006-9155-0
[8] D. Bielińska-Wąż, P. Wąż and S.C. Basak, J. Math. Chem. (forthcoming issue) (2007), DOI: 10.1007/s10910-006-9155-0
[9] C.E. Porter, *Statistical Theories of Spectra: Fluctuations* (Academic, New York, 1965).
[10] T.A. Brody, J. Flores, J.B. French, P.A. Mello, A. Pandey and S.S.M. Wong, Rev. Mod. Phys. 53 (1981) 385.
[11] J. Bauche, C. Bauche-Arnoult and M. Klapisch, Adv. At. Mol. Phys. 23 (1988) 132.
[12] J. Karwowski, D. Bielińska-Wąż and J. Jurkowski, Int. J. Quantum Chem. 60 (1996) 185.
[13] D. Bielińska-Wąż, in *Symmetry and Structural Properties of Condensed Matter*, ed. T. Lulek et al. (World Scientific, Singapore, 1999), pp. 212–221.
[14] D. Bielińska-Wąż, N. Flocke and J. Karwowski, Phys. Rev. B 59 (1999) 2676.
[15] P. Charbonneau, Astr. J. Sup. Ser. 101 (1995) 309.