Eur. Phys. J. B **50**, 333–338 (2006)
DOI: 10.1140/epjb/e2006-00124-y

THE EUROPEAN PHYSICAL JOURNAL B

Statistical theory of spectra: statistical moments as descriptors in the theory of molecular similarity

D. Bielińska-Wąż[1,a], P. Wąż[2], and S.C. Basak[3]

[1] Instytut Fizyki, Uniwersytet Mikołaja Kopernika, Grudziądzka 5, 87-100 Toruń, Poland
[2] Centrum Astronomii, Uniwersytet Mikołaja Kopernika, Gagarina 11, 87-100 Toruń, Poland
[3] Natural Resources Research Institute, 5013 Miller Trunk Highway, Minnesota 55811-1442, USA

Received 30 September 2005 / Received in final form 27 November 2005
Published online 12 April 2006 –

Abstract. Statistical moments of the intensity distributions are used as molecular descriptors. They are used as a basis for defining similarity distances between two model spectra. Parameters which carry the information derived from the comparison of shapes of the spectra and are related to the number of properties taken into account, are defined.

PACS. 29.85.+c Computer data analysis – 07.05.Kf Data analysis: algorithms and implementation; data management – 33.20.-t Molecular spectra – 33.70.-w Intensities and shapes of molecular spectral lines and bands

1 Introduction

The periodic table of elements is the simplest and the most important construct establishing similarity between atoms. Defining measures of similarity between molecules is much more complicated and, though very important in both theoretical and practical dimensions, it was not seriously attempted until some 25 years ago [1]. Since then, many indices of molecular similarity have been defined and successfully used in establishing criteria of molecular similarity [2]. In the present work we propose a new set of molecular similarity indices. These indices relate shapes of molecular spectra. We assume that the degree of similarity of molecules is correlated with the degree of similarity of their spectra. On the other hand, as it is known from the statistical spectroscopy [3–5], spectra are similar if their intensity distribution moments are close. Since the evaluation of these moments is easy, their using as molecular descriptors seems to be an attractive option.

Moments of the intensity distribution, $\mathcal{I}^\gamma(E)$, belong to a set of fundamental concepts of statistical theory of spectra. In the case of discrete spectrum, the n-th statistical moment is defined as:

$$M_n^\gamma = \frac{\sum\limits_i \mathcal{I}_i^\gamma(E) E_i^n}{\sum\limits_i \mathcal{I}_i^\gamma(E)}, \tag{1}$$

where $\mathcal{I}_i^\gamma$ is the intensity of the i-th line and E_i is the corresponding energy difference. If the spectral lines are sufficiently close to each other, then the spectrum may be approximated by a continuous function. Then, the n-th moment of the intensity distribution is defined as:

$$M_n^\gamma = \frac{\int\limits_{C(E)} \mathcal{I}^\gamma(E) E^n dE}{\int\limits_{C(E)} \mathcal{I}^\gamma(E) dE}, \tag{2}$$

where $C(E)$ is the range of the energy for which the integrand does not vanish. It is convenient to consider normalized spectra $I^\gamma(E) = N^\gamma \mathcal{I}^\gamma(E)$, where $N^\gamma = (\int_{C(E)} \mathcal{I}^\gamma(E) dE)^{-1}$, for which the area below the distribution function is equal to 1. Convenient characteristics of the distributions may be derived from the properly scaled distribution moments. Moments normalized to the mean value equal to zero ($M_1^{\gamma\prime} = 0$) are referred to as the *centered moments*. The n-th centered moment reads:

$$M_n^{\gamma\prime} = \int\limits_{C'(E)} I^\gamma(E)(E - M_1^\gamma)^n dE. \tag{3}$$

The moments, for which additionally the variance is equal to 1 ($M_2^{\gamma\prime\prime} = 1$) are defined as

$$M_n^{\gamma\prime\prime} = \int\limits_{C''(E)} I^\gamma(E) \left[\frac{(E - M_1^\gamma)}{\sqrt{M_2^\gamma - (M_1^\gamma)^2}}\right]^n dE. \tag{4}$$

In this work the model spectra are approximated by continous functions taken as linear combinations of *max* unnormalized Gaussian distributions centered at ϵ_i with dispersions σ_i, defined by the parameters $c_i = 1/(2\sigma_i^2)$,

[a] e-mail: dsnake@phys.uni.torun.pl

$i = 1, 2, \ldots max$:

$$I^{\gamma}(E) = N^{\gamma} \sum_{i=1}^{max} a_i \exp\left[-c_i(E-\epsilon_i)^2\right]. \quad (5)$$

The normalization constant N^{γ} is determined so that the zeroth moment of the distribution $I^{\gamma}(E)$ is equal to 1.

The n-th moment of the distribution is equal to:

$$M_n^{\gamma} = N^{\gamma} \sum_{i=1}^{max} \int\limits_{C(E)} a_i \exp\left[-c_i(E-\epsilon_i)^2\right] E^n dE. \quad (6)$$

After some algebra we get the expressions for the moments as functions of the parameters describing the height (a_i), the width (c_i) and the locations of the maxima (ϵ_i). In particular,

$$N^{\gamma} = \left(\sum_{i=1}^{max} a_i \sqrt{\frac{\pi}{c_i}}\right)^{-1} \quad (7)$$

and, for $q = 1, 2, 3$,

$$M_q^{\gamma} = N^{\gamma} \sum_{i=1}^{max} \epsilon_i a_i Q_i^{(q)} \sqrt{\frac{\pi}{c_i}}, \quad (8)$$

where $Q_i^{(1)} = 1$, $Q_i^{(2)} = \epsilon_i + 1/(2c_i\epsilon_i)$ and $Q_i^{(3)} = \epsilon_i^2 + 3/(2c_i)$.

According to the so called *principle of moments* [5] we expect that if we identify the lower moments of two distributions, we bring these distributions to approximate identity. In this paper we apply this principle to the theory of molecular similarity. We assume that molecules have similar properties if their intensity distributions and, consequently the corresponding moments, are approximately the same.

We propose that a set of statistical moments of the intensity distributions can be treated as a new kind of molecular descriptors. A very clear meaning has the first moment, M_1, which describes the mean value of the distribution. In a similar sense a colour index has been introduced in astronomy [6] — its value allows us to compare spectra of different stars (it carries an information about molecules forming the star). The second centered moment, M_2', is the variance, which gives the width of the distribution. M_3'' is the skewness coefficient which describes the asymmetry of the spectrum. The kurtosis coefficient M_4'' is connected to the excess of the distribution.

2 Theory and the model spectra

According to the method of moments, the shapes of two distributions are more similar if the number of identical moments is larger. Similarity of distributions in two- and three-moment approximations, in the context of the construction of envelopes of electronic bands, has been analyzed in references [7–10]. Analogously, we define similarity parameters $\mathcal{S}_k^{i_1 i_2 \ldots i_k}$ (k is the number of properties taken into account in the process of comparison) as a normalized information derived from a comparison of two distributions, referred to as α and β:

$$\mathcal{S}_k^{i_1 i_2 \ldots i_k} = \sqrt{\frac{1}{k}\left(D_{i_1}^2 + D_{i_2}^2 + \ldots D_{i_k}^2\right)}, \quad (9)$$

where $i_1 < i_2 < \ldots i_k$. Here n is the total number of properties taken into account in the comparison of the two spectra and $i_k = 1, 2, \ldots n$ $(k = 1, 2, \ldots n)$, correspond to a specific property. In particular, as the property number one ($i_k = 1$) we take the first moment, as the property number two ($i_k = 2$) we take the second centered moment, number three ($i_k = 3$) — the asymmetry coefficient, number four ($i_k = 4$) — the kurtosis coefficient. In this paper we take $n = 4$ and the corresponding similarity distances are defined as follows:

$$D_q = 1 - \exp\left[-\left(P_{(q)}^{\alpha} - P_{(q)}^{\beta}\right)^2\right], \quad (10)$$

where $P_{(1)}^{\gamma} = M_1^{\gamma}$, $P_{(2)}^{\gamma} = M_2^{\gamma\prime}$, $P_{(3)}^{\gamma} = M_3^{\gamma\prime\prime}$, $P_{(4)}^{\gamma} = M_4^{\gamma\prime\prime}$ and $\gamma = \alpha, \beta$. The values of all the descriptors may vary from 0 (identical properties) to 1.

We also define an additional parameter which may be evaluated if both spectra we are going to compare are available:

$$\mathcal{D} = \frac{1}{2} \int\limits_{C'(E)} |I'^{\alpha}(E) - I'^{\beta}(E)| dE. \quad (11)$$

This parameter is given by the integral of the module of the difference between the compared distributions and is not related to the moments. In the definition of $\mathcal{D}$, I' denotes the distributions transformed so that their averages are the same. If we compare two distributions of the same shape then $\mathcal{D} = 0$. If two distributions do not overlap at all, then $\mathcal{D} = 1$. It is important to note that the distribution moments are defined as numbers attached to a given spectrum and the similarity distances D_q are easily derived from the knowledge of these numbers. The parameter $\mathcal{D}$, though it gives accurate information about similarity of two spectra, is rather cumbersome since it may be derived only if the complete spectra are given.

If two model molecules (or rather their spectra) are identical, up to the accuracy determined by the considered properties, then all $\mathcal{S}_k^{i_1 i_2 \ldots i_k}$ are equal to 0. The maximum value of $\mathcal{S}_k^{i_1 i_2 \ldots i_k}$ is 1 and corresponds to two spectra with no common features within the considered set of properties.

The result of a comparison of two different objects depends not only on the number of properties taken into account but also on their choice (i_1 or i_2 or $\ldots$ i_n). Therefore the quantities $\mathcal{S}_k^{i_1 i_2 \ldots i_k}$ defined in equation (9) should be averaged by taking all combinations of the indices i_k. Thus, we define parameters S_k as the appropriate averages of $\mathcal{S}_k^{i_1 i_2 \ldots i_k}$:

$$S_k = \binom{n}{k}^{-1} \sum_{i_1 < i_2 < \ldots < i_k}^{n} \mathcal{S}_k^{i_1 i_2 \ldots i_k}. \quad (12)$$

In our case $n = 4$ and $k = 1, 2, 3, 4$.

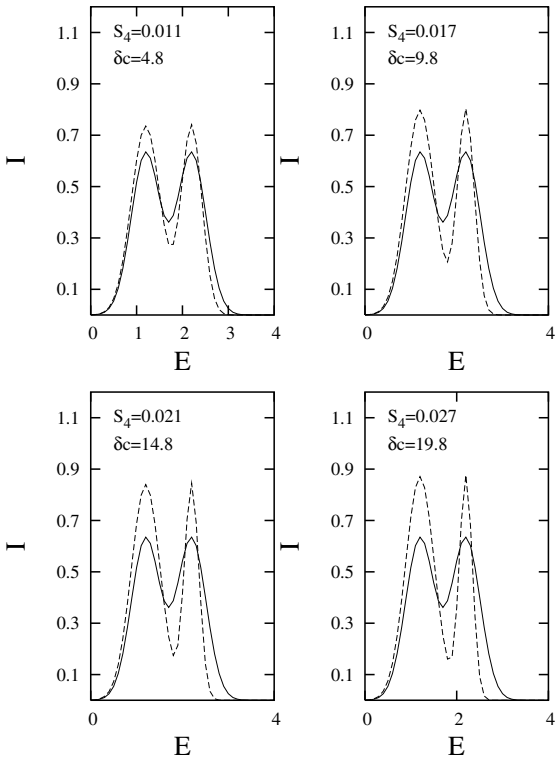

Fig. 1. Two intensity distributions (solid and dashed lines) and the corresponding similarity parameters S_4 (sequence I).

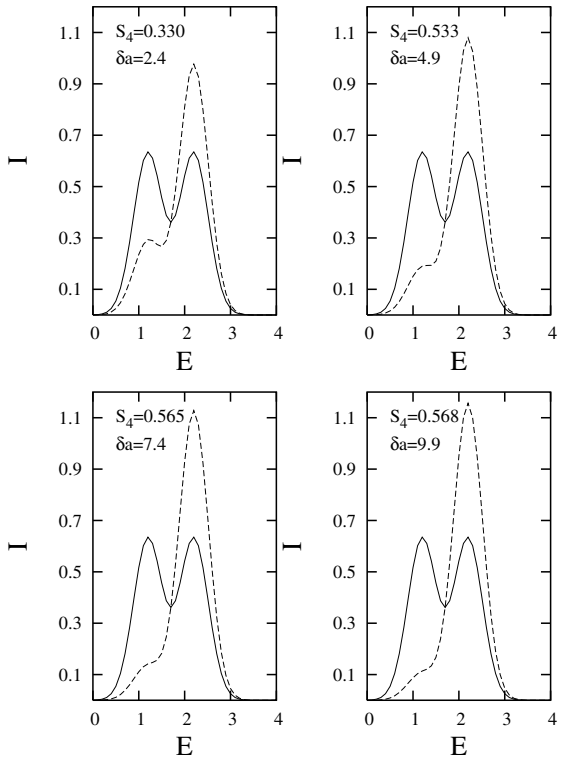

Fig. 2. Two intensity distributions (solid and dashed lines) and the corresponding similarity parameters S_4 (sequence II).

3 Results and discussion

In order to illustrate our approach, we took model spectra consisting of two bands, i.e. having two maxima ($max = 2$):

$$I^{\gamma}(E) = N^{\gamma}\left[a_1 \exp\left[-c_1(E-\epsilon_1)^2\right] + a_2 \exp\left[-c_2(E-\epsilon_2)^2\right]\right], \quad (13)$$

where $\gamma = \{c_1, a_1, \epsilon_1, c_2, a_2, \epsilon_2\}$. In order to see relations between molecular spectra, defined in equation (14) and the similarity indices defined in equations (10–12) in a simple and transparent way, we study three sequences of spectra, where in each sequence only one parameter has been modified: c_2 in sequence I, a_2 in sequence II, ϵ_2 in sequence III.

(a) Sequence I corresponds to the situation when a symmetric spectrum consisting of two identical Gaussian distributions shifted relative to each other by $\epsilon_2 - \epsilon_1 = 1$ ($a_1 = a_2 = 1.0$, $\epsilon_1 = 1.2$, $\epsilon_2 = 2.2$, $c_1 = c_2 = 5.0$) transforms to a distribution in which the width of one of the Gaussians changes due to the change of the parameter $c_2 = 5.0 + \delta c$, where $\delta c \in \langle 0; 19.8\rangle$. Then, we compare shapes of intensity distributions $I^{\alpha}(E)$ and $I^{\beta}(E)$, where $\alpha = \{5.0, 1.0, 1.2, 5.0, 1.0, 2.2\}$, $\beta = \{5.0, 1.0, 1.2, 5.0 + \delta c, 1.0, 2.2\}$.
In Figure 1 spectra corresponding to $\delta c = 0$ (solid lines) and $\delta c > 0$ (dashed lines) are compared. In each case values of δc and S_4 are given. A correlation between these two numbers and between shapes of the spectra is clearly seen. The value of S_4 increases when the two spectra become less similar.

(b) Sequence II corresponds to the same symmetric spectrum as before ($a_1 = a_2 = 1.0$, $\epsilon_1 = 1.2$, $\epsilon_2 = 2.2$, $c_1 = c_2 = 5.0$) transforming to the distributions in which the height of one of the Gaussians changes due to the changes of $a_2 = 1.0 + \delta a$, where $\delta a \in \langle 0; 9.9\rangle$. Then, we compare shapes of intensity distributions $I^{\alpha}(E)$ and $I^{\beta}(E)$, where $\alpha = \{5.0, 1.0, 1.2, 5.0, 1.0, 2.2\}$, $\beta = \{5.0, 1.0, 1.2, 5.0, 1.0 + \delta a, 2.2\}$.
In Figure 2 spectra corresponding to $\delta a = 0$ (solid lines) and $\delta a > 0$ (dashed lines) are compared. In each case values of δa and S_4 are given. The conclusions are similar to those in the case of Figure 1.

(c) Sequence III corresponds to a similar situation as before, except that the maxima in I^{α} are shifted by 1.5 rather than by 1 ($a_1 = a_2 = 1.0$, $\epsilon_1 = 1.2$, $\epsilon_2 = 2.7$, $c_1 = c_2 = 5.0$). I^{α} transforms to the distribution I^{β} for which one of the Gaussian distribution changes the location of the second maximum $\epsilon_2 = 2.7 - \delta\epsilon$, where $\delta\epsilon \in \langle 0; 0.99\rangle$. Then, we compare shapes of intensity distributions $I^{\alpha}(E)$ and $I^{\beta}(E)$, where $\alpha = \{5.0, 1.0, 1.2, 5.0, 1.0, 2.7\}$, $\beta = \{5.0, 1.0, 1.2, 5.0, 1.0, 2.7 - \delta\epsilon\}$.
In Figure 3 spectra corresponding to $\delta\epsilon = 0$ (solid lines) and $\delta\epsilon > 0$ (dashed lines) are compared. In each case values of $\delta\epsilon$ and S_4 are given. The conclusions are similar to those in the cases described by Figures 1 and 2.

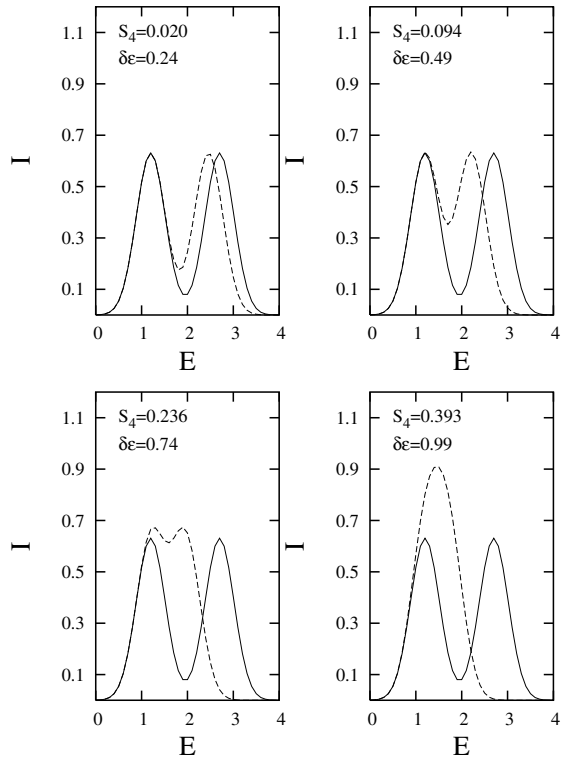

Fig. 3. Two intensity distributions (solid and dashed lines) and the corresponding similarity parameters S_4 (sequence III).

The molecular descriptors [statistical moments of $I^{\beta}(E)$] are plotted in Figure 4 versus δc (sequence I), δa (sequence II), $\delta\epsilon$ (sequence III). In case of sequence I, it is clear that the considered change of the spectrum leads to a decrease of the first moment (the intensity is shifted towards smaller energies). The dispersion of the whole distribution also decreases ($M_2^{\beta\prime}$). The asymmetry of the spectrum changes from totally symmetric ($M_3^{\beta\prime\prime}=0$) to asymmetric ($M_3^{\beta\prime\prime}\neq 0$). The kurtosis coefficient $M_4^{\beta\prime\prime}$ changes as it is presented, in a non-monotonic way. It is interesting that for $M_3^{\beta\prime\prime}$ and $M_4^{\beta\prime\prime}$ minima appear for $\delta c\neq 0$. In the case of sequence II, with an increase of δa the first moment is shifted towards higher values and the dispersion of the whole spectrum decreases. The asymmetry of the spectrum decreases and the kurtosis parameter increases. In case of sequence III, shifting the second maximum ϵ_2 to the smaller energies results in a distribution with one maximum instead of two and the intensity is shifted towards smaller energies. In consequence the first moment decreases. The whole distribution becomes more narrow and, consequently, we observe decreasing of $M_2^{\beta\prime}$. For all $\delta\epsilon$ distributions are symmetric ($M_3^{\beta\prime\prime}=0$) and the kurtosis parameter increases.

Figure 5 presents D defined in equations (10) and (11). In the case of sequence I, if $\delta c=0$, we compare two identical distributions and all the descriptors are equal to zero. The most sensitive to the changes of δc is in this case $\mathcal{D}$, contrary to the other descriptors which are nearly constant. The two distributions are rather similar in sense of the average value, of the width, of the asymmetry and of the kurtosis (the values of D_1, D_2, D_3, D_4 are small and the corresponding curves cross). In case of sequence II, we observe small values of D_2 and D_1, that indicates large similarity of the two distributions in sense of the width and of the average values. For small values of δa we observe crossings between D_3, D_4 and $\mathcal{D}$. The most sensitive to the changes of δa is D_4. In case of sequence III, the behaviour of D_1 and D_2 is very similar. Both spectra are totally symmetric ($M_3^{\alpha\prime\prime}=M_3^{\beta\prime\prime}=0$). Therefore $D_3=0$ for all $\delta\epsilon$. D_4 and $\mathcal{D}$ cross and change very substantially contrary to D_1 and D_2 which are nearly constant.

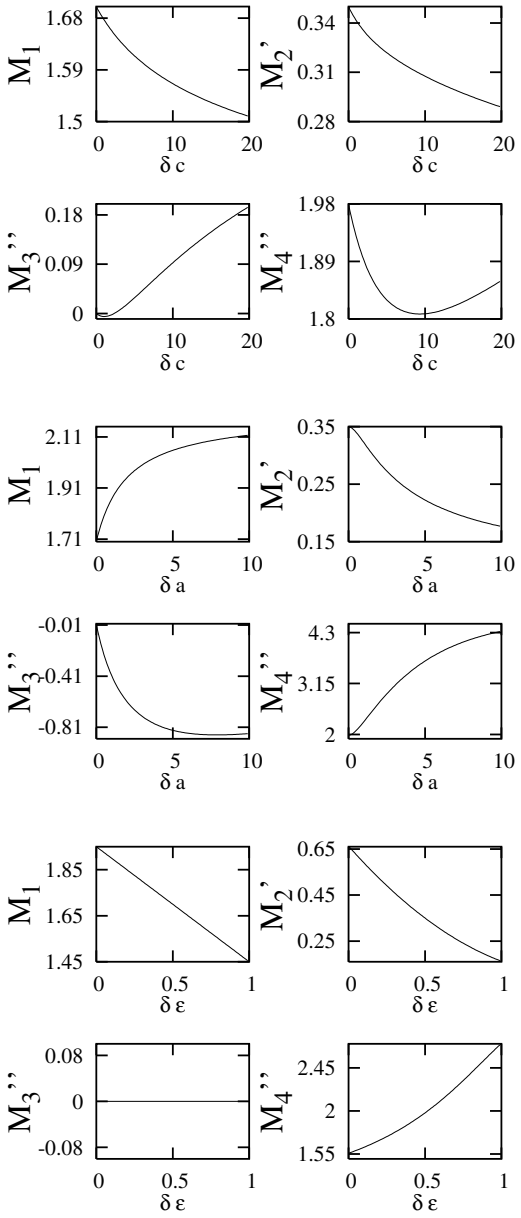

Fig. 4. Moments of the distributions as functions of δc (sequence I), δa (sequence II), $\delta\epsilon$ (sequence III).

Figure 6 presents similarity parameters S_k for $k=1,2,3,4$ (Eq. (12)). Small values of S correspond to high similarity of the model spectra. In particular, if $\delta c=0$

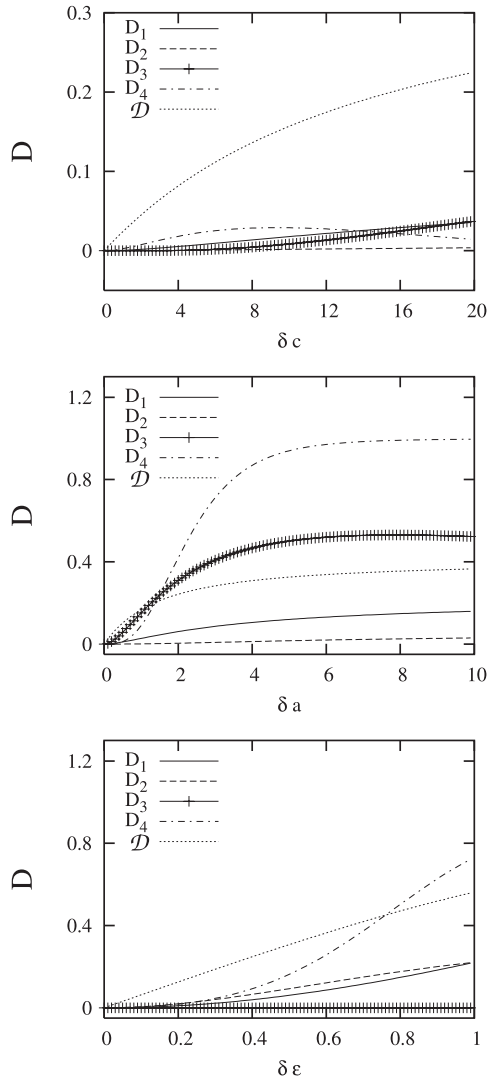

Fig. 5. Parameters D as functions of δc (sequence I), δa (sequence II), $\delta\epsilon$ (sequence III).

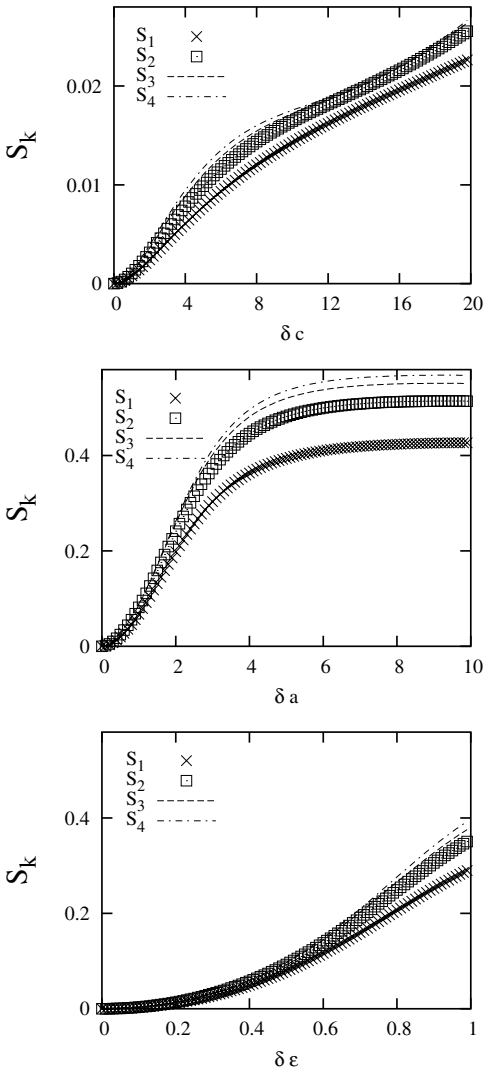

Fig. 6. Parameters S_k as functions of δc (sequence I), δa (sequence II), $\delta\epsilon$ (sequence III).

(sequence I) then $S_k = 0$ for all k. As we can see, S is the smallest for $k = 1$ and increases with increasing k. Analogously to the sequence I, $S_1 < S_2 < S_3 < S_4$ for all δa (sequence II) and for all $\delta\epsilon$ (sequence III). Intuitively, we expect that two systems which are similar to each other when only one property is considered may exhibit more differences if we look at the systems in more detail, taking into account more properties. These features can be seen in Figure 6.

4 Conclusions

Statistical moments describe in an adequate way the degree of similarity of two-band model spectra. Though the mathematical model describing shapes of the spectra is relatively simple, it reflects the behaviour of real molecular spectra. Three parameters: c_2, a_2 and ϵ_2, influence different aspects of the shapes of spectra and the resulting values of D. In particular, parameters D and corresponding S are the smallest if a_2 and ϵ_2 are constant (sequence I). In these cases spectra are only slightly modified by δc (Fig. 1). Larger differences of spectra are caused by parameter δa , while c_2 and ϵ_2 are constant (sequence II). The influence of ϵ_2 on spectra is also large (sequence III). The additional parameter $\mathcal{D}$ introduces some independent information about spectra. Contrary to the case of single-band model spectra studied in our previous paper [11], where its behaviour is very similar to D_4, here it appears to be the most sensitive index (sequence I).

Summarizing, we demonstrated that spectral density distribution moments can be used for defining similarity indices of spectra. By grouping molecules according to the spectral density distribution moments we can get a chance to discover new characteristics in the field of molecular similarity and in particular it may be a tool for studies in the area of computational toxicology [12–14].

This work has been supported by Polish Ministry of Science and Information Society Technologies, grant no 2 PO3B 033 25.

References

1. R. Carbó, L. Leyda, M. Arnau, Int. J. Quantum Chem. **17**, 1185 (1980)
2. *Molecular Similarity*, edited by M.A. Johnson et al. (John Wiley & Sons, New York, 1990)
3. T.A. Brody, J. Flores, J.B. French, P.A. Mello, A. Pandey, S.S.M. Wong, Rev. Mod. Phys. **53**, 385 (1981)
4. J.B. French, V.K. Kota, *Annual Review of Nuclear and Particle Science*, edited by J.D. Jackson et al. (Palo Alto, CA, 1982), p. 35
5. V.S. Ivanov, V.B. Sovkov, Opt. Spectrosc. **74**, 30 (1993); Opt. Spectrosc. **74**, 52 (1993)
6. B.W. Carroll, D.A. Ostlie, *An Introduction to Modern Astrophysics* (Addison-Wesley Publishing Company Inc., 1996)
7. D. Bielińska-Wąż, J. Karwowski, Phys. Rev. A **52**, 1067 (1995)
8. D. Bielińska-Wąż, J. Karwowski, Advances in Quantum Chemistry **28**, 159 (1997)
9. D. Bielińska-Wąż, J. Karwowski, J. Quant. Spec. Rad. Transfer **59**, 39 (1998)
10. D. Bielińska-Wąż, in *Symmetry and Structural Properties of Condensed Matter*, edited by T. Lulek et al. (World Scientific, Singapore, 1999), pp. 212–221
11. D. Bielińska-Wąż, P. Wąż, S.C. Basak, R. Natarajan, in *Symmetry, Spectroscopy and SCHUR*, edited by R.C. King et al. (Nicolaus Copernicus University Press, Toruń, 2006)
12. S.C. Basak, B.D. Grunwald, G.E. Host, G.J. Niemi, S.P. Bradbury, Environ Toxicol. Chem. **17**, 1056 (1998)
13. S.C. Basak, K. Balasubramanian, B.D. Gute, D. Mills, J. Chem. Inf. Comput. Sci. **43**, 1103 (2003)
14. S.C. Basak, B.D. Gute, D. Mills, D. Hawkins, J. Mol. Struct. (Theochem) **622**, 127 (2003)