J Math Chem (2010) 47:1085–1105
DOI 10.1007/s10910-009-9633-2

ORIGINAL PAPER

Direct inversion in the spectral subspace: A novel method for quantitative and qualitative analysis of chemical mixtures

Gábor Pongor · János Eőri · János Rohonczy · Zsuzsanna Kolos

Received: 29 September 2009 / Accepted: 22 November 2009 / Published online: 12 December 2009

Abstract A novel method, called Direct Inversion in the Spectral Subspace (DISS), has been developed for the quantitative (and partly qualitative) analysis of chemical mixtures. The method belongs to the broad group of "supervised classification" methods: its use necessitates the components' "pure" spectra, either experimental or computed. On the basis of three simple conditions, an elegant, linearized system of equations has been deduced, taking into account a sole restriction via the Lagrange' multiplier method. This restriction is seemingly redundant but it has been shown that with its use the unknown normalization constant of the components' descriptive weighted average (CDWA) spectrum can be taken into consideration. The system of linearized equations can be solved repeatedly until convergence. Any kind of spectra can be used; the method does not require the non-negativity of spectral data points. Two versions of the new method have been developed: the normalized and the non-normalized versions regarding the components' spectra. In ideal cases, the non-normalized version of the DISS method provides a mixture's accurate composition due to the iteration for getting the correct norm of the CDWA spectrum. Realistically, the normalized version of the DISS method identifies a mixture's composition within a few molar

To the memory of Dr. Pál Császár (1944–2003).

G. Pongor (✉) · J. Eőri · J. Rohonczy · Zs. Kolos
Laboratory of Organosilicon Chemistry, Institute of Chemistry, Eötvös Loránd University,
P.O. Box 32,1518 Budapest 112, Hungary
e-mail: pongor@chem.elte.hu

J. Eőri
e-mail: eorij@chem.elte.hu

J. Rohonczy
e-mail: rohonczy@chem.elte.hu

Zs. Kolos
e-mail: kolos@chem.elte.hu

percentage points accuracy, according to the test results in IR and ^{1}H-NMR spectroscopy. The normalized method functions without any calibration measurements and needs only a control of accuracy; it is hoped that it will be a useful tool for chemical and biochemical analysis as well as for spectral databases. The DISS method is also useful for qualitative analyses in a limited sense: in the case of computed spectra of the components the set of the de facto components determined could be somewhat wider than those existing in the real system.

Keywords Decomposition of molecular spectra · Lagrange multiplier · Quantitative analysis · Qualitative analysis · IR · NMR · EPR · UV/Vis · Raman · CD · VCD · Hexa-chloro-buta-1,3-diene · Dioxane · D-Camphor · L-Menthol · Supervised classification · Spectral databases

1 Introduction

The quantitative analysis of spectra is one of the key tasks in chemical analysis (see e.g., [1–6] and references therein). In the field of molecular absorption (e.g., ultraviolet/visible, infrared) spectroscopy, the underlying principle is based on the well-known Bouguer–Lambert–Beer (B–L–B) law that establishes a linear relationship between spectral absorbance and concentration as well as path length [7–9]. That is, the absorbance is proportional to the number of particles. A deviation from linearity occurs at higher absorbances due to 'self-shielding' effects and a variety of non-linear transformations can be used to correct this problem (see e.g., [10–14] and references therein). This can be done, for example, by very elegant (but complex, CPU-time consuming) methods, as mentioned in Ref. [14], where significantly overlapped and non-linear spectral data were analysed by a Support Vector Machine learning algorithm (see e.g., Ref. [15,16] and references therein). The application of the B–L–B law raises additional obstacles: (i) its use requires the explicit knowledge of the molar absorptivity at special frequencies, or, alternatively, the application of calibration curves is needed; and (ii) the scattering of light on the samples' particles in case of suspensions also causes deviations [17]. The B-L-B law is frequently applied for the chemical analysis of multicomponent mixtures (see e.g., Ref. [5]). In this case, the 'multiwavelength spectroscopic method' [5] is used and a system of linear equations is established based on the absorbances as well as the concentrations.

Beside the absorption spectra, other topics of spectroscopy also offer important advantages for quantitative analysis; here we briefly mention nuclear magnetic resonance (NMR) and Raman spectrometry as the most frequently used methods. In the field of NMR spectrometry [4], there is a direct proportionality of the integrated resonance intensity and the concentration. Thus, absolute concentrations can be determined by using an internal intensity standard of known concentration. Raman spectrometry is similarly useful for quantitative investigations [6], with the intensity of Raman scattered light being proportional to the number of scattering molecules.

However, the problem of spectral decomposition has been encountered in chemometrics as the identification of pure component spectra from heavily overdetermined complex spectra (together with the composition). The first method for curve resolution, called 'self-modeling curve resolution' technique, was developed by Lawton and Sylvestre [18] in order to solely treat binary mixtures. Later, Borgen et al. [19,20] have generalized Lawton and Sylvestre's original procedure: this is the so-called generalized 'multivariate *N*-component (or, self-modeling) curve resolution' [19–21], which was undeservedly ignored in the literature due to its mathematical and technical difficulties [22,23]. The Principal Component Analysis (PCA, see e.g., Ref. [24]) has been applied in conjunction with self-modeling curve resolution in the field of optical spectroscopy [25–28], NMR spectroscopy [29,30], EPR spectroscopy [31], and even in liquid chromatography [32]. In case of partial or complete overlap of the signals originating from different species, the PCA method is useful in identifying the correct number of the components and is a useful tool in reconstructing the corresponding pure spectra, up to analytically investigated 3-component systems [22,23].

From the logical point of view, we face two types of problems in the aforementioned topics: (i) the spectra of the pure components (the so-called base spectra) are known and the concentrations need to be determined; (ii) the base spectra are not known either. In the field of pattern recognition (see e.g., Ref. [33] and references therein), these two types of problems are mentioned as cases of supervised (i), and unsupervised (ii) classification. In this article, we are dealing with the problem (i), that is, the base spectra are known and the concentrations have to be determined.

In this work we present a new, independent method that is mainly beneficial for quantitative analysis. As shown, it is useful even for a qualitative chemical analysis, at least in a limited sense. The concept behind the new method originated from the field of quantum chemistry, namely from Pulay's widely used Direct Inversion in the Iterative Subspace (DIIS) procedures [34–36].

Beyond the aforementioned data (path length, concentration of the sample) the actual representation of any spectrum could also depend on the values of various other parameters, for example, temperature, phase, pressure, solvent, and pH. In addition to these parameters, one must keep in mind the possible association of the components (with the solvent as well) and the effects of impurities and electronic noise. In the present article, the effects of the parameters mentioned will be eliminated in a simplified manner. We suppose that all the spectra of the pure compounds as well as those of their mixture are observed at the same values of parameters (or at least at values that are as similar to the parameters as it is possible).

2 Experimental details

The IR spectra have been recorded by a BRUKER IFS 55 spectrometer using a resolution of 4 cm^{-1} and 128 scans at room temperature. The spectra were observed in a liquid film between KBr windows in the 4,000–400 cm^{-1} region and contained $n = 3,734$ digitalized spectrum points. Both the spectroscopic grade hexa-chloro-buta-1,3-diene (spectrosc. purity) and dioxane (99.5%) were purchased from Merck. Both materials were used without further purification. Homogenization was carried out by ultrasonic technique.

The ^{1}H-NMR spectra have been recorded by a BRUKER DRX 500 spectrometer at an 11.744 T field using 5 mm Z-gradient inverse probe in dry CDCl$_3$ (Aldrich, 99.9% D-content and 1% TMS internal reference compound) solvent. The gas-chromatographic grade (>99%) D-camphor and the FCC grade L-menthol (>99%) were purchased from Merck and Sigma–Aldrich, respectively. Both crystalline compounds were used without further purifications. A standard 1D single pulse sequence was used with 30° flip angle. 64 k data points were accumulated during 3.17 s with 5.0 s recycling delay. The acquisition was repeated 16 times after 2 dummy scans. The spectrum width was 20 ppm in all cases. The spectral interval was selected between 2.60 and 0.60 ppm containing $n = 3, 172$ data points.

3 Results and discussion

3.1 Theory

Let us consider a set of chemical compounds which form a homogeneous mixture. For reasons given below, it is worth distinguishing two kinds of chemical components in the mixture: the $C_1, C_2, \ldots, C_m$ true (or, 'de facto') components and the $C_{m+1}, C_{m+2}, \ldots, C_l$ virtual (i.e., not actually present) components. Formally, all of these components can be considered as subsets of the $C_1, C_2, \ldots, C_l$ set of compounds which will be termed as 'potential components':

$$\underbrace{\overbrace{C_1, C_2, \ldots, C_m}^{de\ facto\ \text{components}}, \overbrace{C_{m+1}, C_{m+2}, \ldots, C_l}^{\text{virtual components}}}_{\text{potential components}}, \tag{1}$$

where $l \geq \mathrm{m} \geq 1$.

Let us consider the $|\Phi_I\rangle$ 'descriptive' spectra of the potential components (describing the individual spectra of the chemical components mentioned). Here, the term 'spectrum' could mean any of the usual molecular spectra, including infrared (IR), Raman, ultraviolet/visible (UV/Vis), NMR, electron paramagnetic resonance (EPR), circular dichroism (CD), vibrational circular dichroism (VCD), and so on. Naturally, we will apply one and the same type of spectrum for all the potential components and for the resulting experimental spectrum of the mixture using the same methodology (absorption, emission, Raman, reflectance, and so on) and even in the same frequency region as well. All of the mixture's components are required to have their characteristic spectra in the selected spectral region (that is, none of them is the zero function, *vide infra*). The spectra can correspond to any phase (gas, liquid, or crystalline powder); however, in the present paper we will focus on the absorption spectra of completely miscible liquid compounds.

The 'quality' of the $|\Phi_I\rangle$ descriptive spectra of the potential components could be different: $\left|\Phi_I^{\mathrm{exp}}\right\rangle$ experimental spectra, or $\left|\Phi_I^{\mathrm{comp}}\right\rangle$ computed ones (in other words: 'exact' or approximate spectra). In the case of computed spectra, the level of the quantum mechanical approximation at which the spectra were calculated is arbitrary. The natural demand is to employ at least a moderate level of theory which is inexpensive and simultaneously guarantees a 'quite accurate' computation of the spectra.

Let us consider the $|\Phi_I\rangle = \Phi_I(\nu)$ descriptive spectra of the potential components as well as the experimental spectrum of the mixture, $\left|\Phi^{\text{mix}}\right\rangle = \Phi^{\text{mix}}(\nu)$, as functions of the $\mathtt{L}_2$[a,b] Hilbert-space [37,38], where [a,b] represents the finite and closed interval of the ν frequency region considered. Nowadays, experimental spectra are recorded and saved mostly in digitalized form; thus, the spectra mentioned before can also be considered as vectors of the $\mathtt{L}_2^{(n)}$[a,b] real *n*-dimensional subspace of the Hilbert-space (with *n* typically large, about 10,000):

$$\left|\Phi^{\text{mix}}\right\rangle, |\Phi_I\rangle\ (I = 1, 2, \ldots, l) \in L_2^{(n)}[\text{a,b}] \subset L_2[\text{a,b}]. \tag{2}$$

For the construction of a mathematical procedure, we have to take into consideration the following three Conditions.

Condition I:

(a) The mixture's experimental spectrum $\left|\Phi^{\text{mix}}\right\rangle$ and all the $|\Phi_I\rangle$ $(I = 1, 2, \ldots, l)$ descriptive spectra of the potential components are known. The $\left|\Phi^{\text{mix}}\right\rangle$ spectrum is normalized in the $\mathtt{L}_2^{(n)}$[a,b] (or $\mathtt{L}_2$[a,b]) space [37,38]:

$$\mathbf{S}\Phi^{\text{mix}}(\nu) \cdot \Phi^{\text{mix}}(\nu) = 1 \tag{3a}$$

(b) In some cases (*vide infra*) we will accept other normalization conditions as well corresponding to the components' descriptive spectra in the $\mathtt{L}_2^{(n)}$[a,b] or $\mathtt{L}_2$[a,b]) space [37,38]:

$$\mathbf{S}\Phi_I(\nu) \cdot \Phi_I(\nu) = 1 \quad (I = 1, 2, \ldots, l). \tag{3b}$$

(Here the symbol $\mathbf{S}$ means definite integration over continuous variables or summation over discrete ones. For a better understanding, $\Phi_I(\nu_j)$ means the *j*-th coordinate of the *I*-th chemical component's spectrum in the $\mathtt{L}_2^{(n)}$[a,b] space).

Condition II: The interaction between the de facto components is supposed to be weak enough to use a simple additive formula for a homogenous mixture.

Thus, we can write, for instance, for the $\left|\Phi^{\text{mix}}\right\rangle$ mixture's experimental spectrum:

$$\left|\Phi^{\text{mix}}\right\rangle = \mathcal{N}^{\text{mix}} \sum_I x_I^{\text{exp}} \left|\Phi_I^{\text{exp}}\right\rangle, \tag{4}$$

where the $x_I^{\text{exp}} (I = 1, 2, \ldots, l)$ are the linear coefficients (being proportional to the actual concentrations, *vide infra*; naturally, these coefficients are zero for the $I = m + 1, m + 2, \ldots, l$ virtual components), and $\mathcal{N}^{\text{mix}}$ stands for the normalization constant of the mixture's experimental spectrum. It must be emphasized that the $\left|\Phi^{\text{mix}}\right\rangle$ mixture's experimental spectrum is empirically known and is normalized (see Condition Ia), for example, by the usual 'point-by-point' method, in case of digitalized spectra. Therefore, an explicit knowledge of the x_I^{exp} experimental concentrations is not required for this purpose:

$$\mathcal{N}^{\mathrm{mix}} = \left(\mathbf{S}\Phi^{\mathrm{mix}}(\nu)\Phi^{\mathrm{mix}}(\nu)\right)^{-1/2}. \tag{5}$$

Note that the additivity is realized at a much higher level than expected on the basis of the aforementioned breakdown of the B–L–B law (in case of absorption spectra): there are rather few frequency regions of the spectra showing high absorbance; for other wide regions, the linearity is still valid. In other words, in the simple application of the B–L–B law, the information observed only at several frequencies is used, whereas the present method applies information obtained from a wide (or 'complete') spectral region.

> Condition III: The $|\Phi_I\rangle$ descriptive spectra of the potential components are linearly independent (at both the experimental and the computed level; see Appendix A, which discusses this condition with some mathematical rigor).

During the current work, it will be assumed that Conditions I–III are fulfilled. Depending on the fact that Condition Ib is also valid (or not), we define the two basic *versions* of our new procedure: the 'normalized' (Condition Ib is also valid) and the 'non-normalized' versions (Condition Ib is omitted).

Using the notations of Appendix B, let us define the gross error vector for the case of an additive mixture of the de facto components as follows:

$$\left|\phi^{\mathrm{err}\prime}\right\rangle = \left|\varphi^{\mathrm{mix}}\right\rangle - \mathcal{N}\sum_{I}^{m} x_I\,|\Phi_I\rangle = \left|\varphi^{\mathrm{mix}}\right\rangle - \mathcal{N}\sum_{I}^{l} x_I\,|\Phi_I\rangle \tag{6}$$

where the x_I linear coeffients characterizing the composition are to be calculated (certainly, the formal inclusion of the virtual components into the summation does not change the physical situation), and the $\mathcal{N}$ normalization constant of the $\sum_I x_I\,|\Phi_I\rangle$ *components' descriptive weighted average (CDWA) spectrum* is treated as a parameter. In Eq. 6 $|\Phi_I\rangle$ will be substituted as:

$$|\Phi_I\rangle = \begin{cases} |\varphi_I\rangle & \text{(normalized version)} \quad (7a) \\ \left|\varphi_I'\right\rangle & \text{(non-normalized version)} \quad (7b) \end{cases}$$

for the components' descriptive spectra, respectively (see also Appendix B). Now, let us minimize the square of the norm of the $\left|\phi^{\mathrm{err}\prime}\right\rangle$ gross error vector [using the substitution of Eq. 7a, that is, we are dealing with the generation of the secular equation of the *normalized* version], together with the trivial

$$\sum_{I=1}^{l} x_I = 1 \tag{8}$$

constraint. Appendix B shows that the linear coefficients cannot be negative and are lower or equivalent to unity, *automatically* obeying the constraint given in Eq. 8. Consequently, these linear coefficients are the molar ratios. [On this basis one can suppose that the constraint given in Eq. 8 is unnecessary for the procedure. However, the situation is the opposite and albeit Eq. 8 contains redundant information, its explicit implementation is absolutely necessary in the present method due to the complex non-linear character of Eq. 9, *vide infra*.] Using the method of the undetermined multipliers of Lagrange, we have to minimize the following error functional with respect to the set of x_I:

$$\mathsf{F} = \left\langle \phi^{\mathrm{err}'} | \phi^{\mathrm{err}'} \right\rangle - 2\lambda \left(\sum_{I=1}^{l} x_I - 1 \right) \rightarrow \min, \tag{9}$$

where λ is a Lagrangian multiplier. Differentiating Eq. 9 by x_J and λ we get the following system of $l + 1$ linear equations:

$$\begin{bmatrix} S_{11} & S_{12} & \ldots & S_{1l} & 1 \\ S_{21} & S_{22} & \ldots & S_{2l} & 1 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ S_{l1} & S_{l2} & \ldots & S_{ll} & 1 \\ 1 & 1 & \ldots & 1 & 0 \end{bmatrix} \begin{bmatrix} x_1\mathcal{N} \\ x_2\mathcal{N} \\ \ldots \\ x_l\mathcal{N} \\ -\lambda/\mathcal{N} \end{bmatrix} = \begin{bmatrix} S_{mix1} \\ S_{mix2} \\ \ldots \\ S_{mixl} \\ \mathcal{N} \end{bmatrix}, \tag{10}$$

where (using the notations given in Appendix B again)

$$S_{JK} \equiv \langle \varphi_J | \varphi_K \rangle = \mathbf{S} \varphi_J(\nu) \cdot \varphi_K(\nu) \tag{11}$$

and

$$S_{\mathrm{mix}J} \equiv \left\langle \varphi^{\mathrm{mix}} | \varphi_J \right\rangle = \mathbf{S} \varphi^{\mathrm{mix}}(\nu) \cdot \varphi_J(\nu) \tag{12}$$

$(J, K = 1, 2, \ldots, l)$ are the corresponding scalar products in the $\mathrm{L}_2^{(n)}$[a,b] or L_2[a,b] space. It has to be emphasized that Eqs. (10–12) correspond to the *normalized* version of the new procedure; that is, where Condition Ib is also valid. In this paper we are dealing mostly with that version of the present procedure.

Appendix C provides proof that the inhomogenous system of linear equations given in Eq. 10 has a *unique* solution when parameter $\mathcal{N}$ has a certain value. Necessarily, that is the physically meaningful solution at which the convex [$0 \leq x_I \leq 1 (I = 1, 2, \ldots, l)$ and Eq. 8] conditions are automatically satisfied, so the unknown molar ratios can be evaluated. Nevertheless, the solution for Eq. 10 can only be found in an iterative procedure because the $\mathcal{N}$ normalization factor of the CDWA spectrum is also unknown; its value depends on the values of the unknown molar ratios. The iteration is repeated until convergence:

$$\mathcal{N}^{(0)} \rightarrow \left\{ \underline{x}^{(1)}, \lambda^{(1)} \right\} \rightarrow \mathcal{N}^{(1)} \rightarrow \left\{ \underline{x}^{(2)}, \lambda^{(2)} \right\} \rightarrow \mathcal{N}^{(2)} \rightarrow \ldots \tag{13}$$

where $\mathcal{N}^{(i)}$, $\underline{x}^{(i)}$ and $\lambda^{(i)}$ are the normalization coefficient, the vector of the molar ratios, and the Lagrangian multiplier corresponding to the *i*-th iterative step, respectively. Note that the $\mathcal{N}$ normalization factor of the CDWA spectrum is calculated by applying the following expression (according to the normalized and non-orthogonal basis set applied in the spectral subspace, see Appendix A) in the *i*-th step of the iteration:

$$\mathcal{N}^{(i)} = \left(\sum_I \sum_J x_I^{(i)} x_J^{(i)} S_{IJ} \right)^{-1/2}, \tag{14}$$

[c.f., Eq. 11].

There are several similarities and differences between the present method, Pulay's DIIS [34,35] procedure, and the Geometry Optimization by Direct Inversion in the Iterative Subspace (GDIIS) [36] procedure of Császár and Pulay. The main difference is as follows. In case of the DIIS (and GDIIS) procedure(s) there is an 'external' energy functional in terms of the MO coefficients (or nuclear coordinates, respectively) which has to be minimized. In both cases, the use of the formalism concludes to a purely quadratic 'internal' error functional which bears a system of linear equations that has to be solved repeatedly according to the minimization of the external functional. In case of our method there is no external functional but the error functional given in Eq. 9 is quadratic in a formal sense only, so its use results to a linearized system of equations that can necessarily be solved by an iterative method [c.f. Eq. 13]. The similarity between the procedures can be easily observed in the mathematical structure of the resulting system of equations (compare Eq. 10 of the present work to Eq. 6 of Ref. [34] and Eq. 1 of Ref. [35]). Accordingly, we propose that the present method be termed as 'Direct Inversion in the Spectral Subspace' (DISS), since our method works in the *l*-dimensional subspace of the $\mathrm{L}_2^{(n)}$[a,b] space, spanned by the $|\Phi_I\rangle$ $(I = 1, 2, \ldots l)$ descriptive spectra of the potential components. Note that the DISS method does not need the knowledge of the *molar* absorptivities or ellipticities as follows from Eq. 10. Moreover, the normalized version of the DISS procedure does not require any calibration, or, even the uniform value of the path lengths (for example, in case of absorption spectra). However, in that version the molar ratios determined by the solution of Eq. 10 can somewhat deviate from their true values due to the the unphysical normalizations of the components' spectra given in Condition Ib. The larger the differences of the norms of the non-normalized components' descriptive spectra, the larger the deviations of the molar ratios from their true values [c.f. Eqs. 7a and B.4].

The significance of the current DISS method can be illustrated in two typical application fields:

1. Typical Application I (TA I): determination of the unknown composition in case of a mixture of experimentally known compounds, i.e., $l = m$ [using the notations of Eq. 1], and the descriptive spectra of the components are of experimental ('exact') quality: $|\Phi_I\rangle = \left|\varphi_I^{\mathrm{exp}}\right\rangle$. This is a standard case of a *quantitative* analysis.
2. Typical Application II (TA II): assigning a set of the *de facto* components within a set of potential ones (e.g., in case of a performed "new" reaction producing earlier unknown compounds). That is, $l \geq m$ [c.f., Eq. 1] and the descriptive spectra of the components are of computed ('approximate') quality: $|\Phi_I\rangle = \left|\varphi_I^{\mathrm{comp}}\right\rangle$ (Naturally, even the spectra of unknown compounds are computable). This is the

case of a *qualitative* analysis, at least in a limited sense (*vide infra*, for details). However, this kind of qualitative analysis is not similar to the commonly used spectroscopic work, rather it is analogous to the problems of pattern recognition [33,38]. Note that there can be 'mixed cases' where descriptive experimental *and* computed spectra occur as well.

In the case of TA I, the normalized version of the DISS method can yield reasonable estimates of the 'accurate' values of the x_I^{exp} unknown molar ratios [c.f., Eq. 4], at least in principle, if the noise and other disturbing effects are negligible, and, if the norms of the non-normalized components' descriptive spectra are similar enough. In this ideal case, using the non-normalized version of the DISS method, the minimized value of the F error functional [see Eq. 9] will be practically zero. In contrast, in the case of TA II, the DISS method produces a minimized value for the error functional F, which does not equal zero, neither in the case of the normalized, nor in the case of the non-normalized versions. It must be emphasized that TA II of the DISS method has a very natural prerequisite: the complete set of the de facto components has to be within the set of the potential components. Since we must use approximate methods for computing the descriptive spectra, the gross error vector contains unknown *individual* spectral errors (corresponding to the components) beyond the unknown molar ratios. Consequently, the DISS method yields somewhat distorted calculated values for the components' molar ratios due to the fact that the procedure tries to fit the CDWA spectrum to the mixture's experimental one as closely as possible). It can be expected that the differences between the calculated and the 'true' values of the component ratios will be smaller as the level of the quantum mechanical approximation is raised. Note that the unreacted starting materials and/or the solvent can also be included in the set of the potential components investigated, using their experimental spectra. In this case, the set of the potential components can bear 'mixed' (partly experimental, partly computed) spectra. The distortion of the molar ratios can even result in a set of the de facto products that would be somewhat different from the true one. In summary, in this situation, the DISS method does not produce an accurate assignment of the set of the de facto products, yet its use has an advantageous feature, making the set of the potential products much more closely resembling the original one. Hopefully, this feature of the present method will be very useful in conjunction with spectral databases.

The DISS method yields small or zero x_I values for the virtual products. Note that the appearance of small *negative* values of the molar ratios shows that (a) Condition III is not completely valid, or (b) noise effects are not negligible. It goes hand in hand with an ill-conditioned or near-ill-conditioned coefficient matrix in Eq. 10. In such cases Levenberg's procedure [39–41] can be applied to make the method more robust numerically, in which the squared norm of the coefficients (multiplied by a small positive number) is added to the original functional. However, there is a difference between the use of the Levenberg's procedure used by Pulay [41] and the present situation. In the DISS method, the non-orthogonal version of the procedure has to be applied [c.f., Eq. 11]. Consequently, its effect is the addition of a small positive number, multiplied with the value of the actual matrix element, to *all* of the S_{JK} $(J, K = 1, 2, \ldots, l)$ elements of the DISS matrix.

As can be seen, our DISS method requires the pure spectra of the chemical components at the decomposition of the spectra of mixtures. In this context, the DISS method belongs to the broad class of "supervised classification" methods (see e.g., Ref. [42]).

A Fortran program has been written for the application of the present method (program DISS, Ver. 3.0, under development), which is available from the authors upon request. In its present form, it is capable of taking into consideration 10 components ($l = 10$) and the number of the discrete points in the spectra is $n = 10{,}000$. The (small) negative molar ratios mentioned before are excluded on a *lege artis* way which is equivalent with the use of S_{JJ} elements of infinite value. The program has two options according to the normalized and non-normalized versions. In both cases the solution of the secular equation starts at $\mathcal{N}^{(0)} = 1.0$ and it goes to 'self-consistency'. In the iterative solution, the norm of the difference vector, calculated between the coefficient vectors of two consecutive iteration steps, is the test for convergence. In the current version, its threshold value is 10^{-6}.

Details of the non-normalized version of the DISS method will be given in a forthcoming paper. In this case the simplicity of the normalized DISS version will be lost (namely, the lack of the need of calibration), and we need additional information. However, this version can be useful in the case of quantitative analysis of a large series of different mixtures consisting of the same set of chemical components.

3.2 Applications

Recently [38], we proposed the planning and execution of chemical reactions in four 'algorithmic steps': (i) in the first step, we hypothesize about the kind of products that could theoretically be formed (potential products/components); (ii) compute the chosen type of approximate spectra for all of the potential products; (iii) fulfil the reaction and record the corresponding experimental spectrum, and, (iv) the best agreement (proximity) between the computed spectra and the experimental one proves the structure of the de facto product(s). For the computation of the calculated spectra, we have suggested empirically corrected theoretical spectra (see e.g., Refs. [38,43–45]). The possible fields of application of the suggested procedure are the modern *in situ* spectroscopic investigations (see e.g., Refs. [46,47]). Through the further development in computer technology as well as computational methods, the use of the procedure [38] could be specifically favourable in the field of combinatorial syntheses [48].

In Ref. [38] we gave an example of the aforementioned procedure by the silylation reaction of 1,3-dihydrobenzimidazol-2-one by BSTFA [*N,O*-bis(trimethylsilyl)trifluoro-acetamid], for which we have chosen the simple and popular IR spectroscopy. Due to the two common tautomeric forms of the precursor [38], we started to search for the result of the silylation in a set consisting of the potential products as follows: 1,3-bis(trimethylsilyl)benzimidazol-2-one (**1**), 1-trimethylsilyl-benzimidazol-2-one (**2**), 1-trimethylsilyl-2-trimethylsiloxy-benzimidazole (**3**), 2-trimethylsiloxy-benzimidazole (**4**), and 1-trimethylsilyl-2-hidroxy-benzimidazole (**5**, see Scheme 1 of Ref. [38], respectively). We have determined the scaled quantum mechanical (SQM) vibrational quadratic force fields [43–45] of the potential products **1**–**5** at the B3LYP/6-31G* [49–51] level and computed the band origins and the intensities of the IR fundamentals within the harmonic approximation. Next, the SQM IR spectra [43–45] of the potential products were simulated by Lorentzian curves using a mean of the experimental halfwidths of 12 cm^{-1} [38] for all of the bands. The experimental liquid-phase IR spectrum of the product(s) of the performed reaction has been recorded and we have estimated [38] the de facto product of the reaction in the simplest base of the highest similarity between the experimental and SQM-computed spectra. For the numerical characterization of that similarity, we have proposed and used a new measure in the field of vibrational spectroscopy, the scalar product [37,38]. However, this simple procedure was not sufficiently sensitive to prove or exclude the possibility of simultaneous formation of other product(s) with any confidence. This problem can be solved in a much more elegant way according to the new, presented DISS method.

3.2.1 Decomposition of simulated mixtures (TA I)

First, we wanted to test the DISS method for *simulated* mixtures. Since simulated mixtures are without any noise, impurity or computational error, and are free of associations, they are suitable for validating the overall accuracy of the method. Simultaneously, we wanted to get some information about the properties of the iteration mentioned before. Therefore, we applied the ‘orthodox procedure’ according to Eq. 13.

The aforementioned SQM-computed IR spectra of compounds **1**–**5** were selected to represent the descriptive spectra in the 4000–400 cm^{-1} frequency region. [Note that the determinant of the ‘complete’ coefficient matrix of Eq. 10 was −0.23608, which means that the descriptive spectra are linearly independent.] Using definite simulated ‘mixtures’ of compounds **1**–**5**, we decomposed the resulting spectra into their pure components using the DISS method. As previously stated, these were simulated cases for TA I, that is, for the quantitative analysis.

Accordingly, we have created simulated mixture’s spectra by mixing the non-normalized SQM spectra of compounds (*vide supra*, and Ref. [38]) in different molar ratios, respectively. We have selected the corresponding SQM spectra of the $l = 5$ of different potential components, and used the DISS method for determining their molar ratios. Both versions of the DISS method were used. The results of the normalized version are presented in Table 1. As can be seen, the normalized version of the DISS method performs quite well, the greatest deviation is of 9.1 molar percentages. In all investigated cases the DISS method’s non-normalized version deciphered the *exact* starting composition (within 10^{-6} accuracy) with $|\lambda| < 10^{-6}$, in 10, 11, 7, and 8 iterative steps, respectively. It is trivial that the optimized value of the error functional is equal to the minus two times value of the Lagrange multiplier, c.f. Eq. 9.

Other authors have published articles with similar objectives. Magar also considered the spectra as the vectors of the L_2[a,b] Hilbert-space [52]. However, he did not take into account the importance of any constraint or the normalization of the CDWA spectrum, thus his method is a simple application of the *linear combination* of the pure spectra. In fact, it is possible to derive the secular equations without constrains as calculated by Magar [52]; his equation is similar to Eq. 10, omitting the (l+1)-th row and column of the coefficient matrix and the (l+1)-th components of the vectors on both sides of the equation. Aiming to check Magar’s method [52], we recalculated

Table 1 Decomposition of simulated mixtures using the normalized version of the DISS method

	Simulation 1	Simulation 2	Simulation 3	Simulation 4
'Experimental' molar ratios	0.1 (**1**) : 0.9 (**3**)	0.99 (**1**) : 0.01 (**3**)	0.11 (**1**) : 0.27 (**3**) : 0.62 (**5**)	0.11 (**1**) : 0.22 (**2**) : 0.33 (**3**) : 0.22 (**4**) : 0.12 (**5**)
DISS (normalized version)	0.098 (**1**) : 0.902 (**3**)	0.990 (**1**) : 0. 010 (**3**)	0.145 (**1**): 0.361 (**3**) : 0.494 (**5**)	0.125 (**1**) : 0.197 (**2**) : 0.382 (**3**) : 0.213 (**4**) : 0.083 (**5**)
Iterative steps[a]	8	8	7	7

Solutions of the DISS equations were made by using the program DISS Ver. 3.0; SQM IR spectra of the components [38] (used in the 4000–400 cm^{-1} frequency region) denoted by bold-face numbers within parentheses, see Sect. 3.2. Set of the potential components: **1** to **5** [$l = 5$, c.f. Eq. 1]. Experimental molar ratios correspond to the mixing ratios of (non-normalized) SQM IR spectra, see Sect. 3.2.1

[a] Number of iterative steps needed to reach the experimental composition of the mixture with less than 10^{-6} accuracy in the norm of the difference vector of two consecutive parameter vectors during the iterative solution; iteration started by $\mathcal{N}^{(0)} = 1.0$, see Sect. 3.1 for details

the simulated problems given in Table 1 using his procedure. The results were completely accurate, however, Magar's method [52] was not able to give any reasonable composition in case of real (not simulated) mixtures (see below). This is due to the fact that Magar's method neglects the normalizations completely (that is, not only in case of the components' descriptive spectra).

It is important to emphasize the fact that in Condition III we stated the linear independency of the descriptive spectra. This is certainly gentler than the constraint of Hennessey and Johnson [53], which requires the *orthogonality* of the pure spectra used in deducing chiral contribution of the common secondary structures from CD curves of proteins. Other authors (see e.g., Refs. [18,22,23] and references therein) required the fulfillment of the non-negativity of both spectral and concentration values in the spectral decomposition of mixtures. Our novel DISS method is more general compared to the aforementioned procedures due to the fact that it does not require the non-negativity, neither for the spectral values nor for the concentrations. Therefore, the DISS method is capable to handle even, for instance, EPR, CD, and VCD spectra as well; these contain both negative and positive data points. Nevertheless, there is no need to use any integration like in Ref. [31]. Moreover, the DISS method, rejecting the constraint of non-negativity in concentrations, does not require the use of a simplex optimization algorithm [54] like in the works of other authors (see e.g., Refs. [22,23,55–58] and references therein). The simplex procedure can be applied to linear (or *special* non-linear) target functions. However, in the case of special non-linear target functions (see e.g., Refs. [22,23,56–58]), the simplex method is suitable for functional optimizations only if the number of the components is rather small. Naturally, our DISS method is able to treat a much larger number of components. Moreover, in the simplex method, it is also problematic if a negative coefficient term would have resulted in an optimal solution. Using the DISS method, smaller negative concentration values can occur but these can be avoided (*vide supra*). The DISS method is more robust numerically than the simplex procedure.

Summing up, our method is suitable for identifying the *completely accurate* composition of chemical mixtures in ideal cases in its non-normalized version, and yields reasonable estimate of the composition in case of the normalized version.

3.2.2 Decomposition of experimental IR spectrum by SQM computed spectra (TA II)

The next test case was the reinvestigation of the silylation reaction of 1,3-dihydrobenzimidazol-2-one by BSTFA (*vide supra* and Ref. [38]). Here, we used the experimental IR spectrum of the reaction's unknown ("mixture") product(s), and the computed SQM IR (descriptive) spectra were applied in the search for the set of the de facto products (components). Obviously, this is a case of TA II, a qualitative analysis. The situation could be more complex than the former one described in Chap. 3.2.1 by the presence of possible associations. All of the mentioned spectra were recorded/computed in the 4000–400 cm^{-1} frequency region (the resolution was 2 cm^{-1}). For the computed bands, a Lorentzian shape was used; the halfwidths of all bands were considered to be 12 cm^{-1} [38]. The solution of the normalized version of the DISS equations gave the following result (using 12 steps of iteration, starting with $\mathcal{N}^{(0)} = 1.0$):

$$\begin{aligned} &x_1 = 0.665; \quad x_2 = 0.083; \quad x_3 = 0.180; \\ &x_4 = 0.072; \quad x_5 = 0.000; \quad -2\lambda = 0.606. \end{aligned} \tag{15}$$

As shown, the DISS method yields about 67 molar percentage for **1** and 18 percentage points for **3** . Other products have much smaller (<0.085) ratios. (Note the zero ratio in case of **5**, in contrast to the result derived from simple similarity in Ref. [38]). We know from our previous work [38] that the *exact* result of the silylation reaction is compound **1**, as determined by independent NMR measurements. Despite of this, the DISS method showed the formation of two 'new' compounds practically. Naturally, this is not the failure of the DISS method, rather is due to the quite low (still best harmonic) level of approximation of the computed IR spectra. However, even at this rather limited level of approximation, the DISS method yields an estimate of the *de facto* products *within* a significantly smaller list (**1** *and* **3**) of the potential products than that of the original one (**1** *to* **5**). In the case of a much larger set of potential products, this character can be of particular advantage (c.f., spectral databases).

3.2.3 Decomposition of experimental IR spectrum by pure experimental IR spectra (TA I)

Table 2 shows the results of a real experimental IR test series we performed. In these experiments, hexa-chloro-buta-1,3-diene (**HCB**) and dioxane (**D**) were mixed in differently weighted molar ratios. The wavenumber interval investigated was the 4000–400 cm^{-1} region, with a resolution of 4 cm^{-1} . The resulting experimental spectra were decomposed by the normalized version of the DISS method using the experimental IR spectra of the pure components. This is a case of TA I (quantitative analysis). This investigation series differs from the previous application in the context of the noise of the 'natural' (not simulated) spectra.

Table 2 Mixtures of hexa-chloro-buta-1,3-diene (**HCB**) and dioxane (**D**): IR spectra decomposed by the normalized version of the DISS method

	Mixture 1	Mixture 2[a]	Mixture 3[a]	Mixture 4[a]	Mixture 5	Mixture 6	Mixture 7
Exp. molar ratios	1.00 (**HCB**) : 0.00 (**D**)	0.954 (**HCB**) : 0.046 (**D**)	0.682 (**HCB**) : 0.318 (**D**)	0.501 (**HCB**) : 0.499 (**D**)	0.205 (**HCB**) : 0.795 (**D**)	0.054 (**HCB**) : 0.946 (**D**)	0.00 (**HCB**) : 1.00 (**D**)
Iterative steps[b]	2	4	7	5	11	13	2
Results:							
$x_{\mathbf{HCB}}$	1.00 (**HCB**)	1.000 (**HCB**)	0.625 (**HCB**)	0.449 (**HCB**)	0.191 (**HCB**)	0.090 (**HCB**)	0.00 (**HCB**)
$x_{\mathbf{D}}$	0.00 (**D**)	0.000 (**D**)	0.375 (**D**)	0.551 (**D**)	0.809 (**D**)	0.910 (**D**)	1.00 (**D**)

Experimental IR spectra of the components were recorded by a Bruker IFS 55 spectrometer using a resolution of 4 cm^{-1} and 128 scans. The spectra were observed in a liquid film between KBr windows, at room temperature, in the 4000–400 cm^{-1} frequency region. Set of the potential components: **HCB** and **D** [$l = 2$, c.f. Eq. 1]. Experimental molar ratios correspond to the weighted mixing ratios in Eq. 4, see Sect. 3.2.3

[a] In the case of Mixture 2, 3 and 4 the solutions were opalescent

[b] See footnote a of Table 1

As can be seen, the differences vary from −3.6 to +5.7 molar percentages. This is a quite good result considering the fact that some of the mixtures (see Mixture 2, 3 and 4 in Table 2) were opalescent in contempt of the ultrasonic homogenization. We suppose that in the case of completely miscible mixtures the difference between the weigthed and the resulting molar ratios would be even smaller ($\leq$3.6 molar percentage points, see Mixture 1, 5–7 of Table 2). We emphasize the advantage of the normalized version of the DISS method as the IR spectra were recorded using liquid film technique between KBr windows: the method does not require the use of uniform path lengths.

3.2.4 Decomposition of experimental NMR spectrum by pure experimental NMR spectra (TA I)

It is our hope that the presented DISS method will also be frequently applied for chemical analysis in the field of NMR spectrometry. Consequently, we investigated the DISS method for a better characterization on this special topic: we mixed D-camphor (**C**) and L-menthol (**M**) in given concentrations and recorded the mixtures' ^{1}H-NMR spectra. Since the resulting spectra have strongly overlapping peaks, it is reasonable to apply our method. With the use of the NMR spectra of the pure components, we decomposed the mixtures' spectra using the normalized version of the DISS method, again, a TA I (quantitative analysis) case using the terminology from the Theory chapter. The spectra were recorded in $CDCl_3$ solvent, and TMS was used as a standard. The spectral interval was between 2.60 and 0.60 ppm, corresponding to 3,172 data points. The interval of the spectra was selected in such a way that neither the solvent nor the TMS standard would give any signal in the corresponding interval. Note that the NMR signals of the pure components were slightly shifted compared to the mixture's signals. Another problem was caused by the NMR spectra's pin-style narrow peaks.

Table 3 Mixtures of D-camphor (**C**) and L-menthol (**M**): ^{1}H-NMR spectra decomposed by the normalized version of the DISS method

	Mixture 1	Mixture 2	Mixture 3	Mixture 4	Mixture 5	Mixture 6	Mixture 7
Exp. molar ratios	1.00 (**C**) : 0.00 (**M**)	0.833 (**C**) : 0.167 (**M**)	0.667 (**C**) : 0.333 (**M**)	0.500 (**C**) : 0.500 (**M**)	0.333 (**C**) : 0.667 (**M**)	0.167 (**C**) : 0.833 (**M**)	0.00 (**C**) : 1.00 (**M**)
Original							
Iterative steps[a]	2	11	8	5	7	11	2
Results:							
$x_{\mathbf{C}}$	1.00 (**C**)	0.854 (**C**)	0.702 (**C**)	0.525 (**C**)	0.350 (**C**)	0.139 (**C**)	0.00 (**C**)
$x_{\mathbf{M}}$	0.00 (**M**)	0.146 (**M**)	0.298 (**M**)	0.475 (**M**)	0.650 (**M**)	0.861 (**M**)	1.00 (**M**)
Broadened							
Iterative steps[a]	2	9	7	3	7	9	2
Results:							
$x_{\mathbf{C}}$	1.00 (**C**)	0.841 (**C**)	0.671 (**C**)	0.500 (**C**)	0.327 (**C**)	0.138 (**C**)	0.00 (**C**)
$x_{\mathbf{M}}$	0.00 (**M**)	0.159 (**M**)	0.329 (**M**)	0.500 (**M**)	0.673 (**M**)	0.862 (**M**)	1.00 (**M**)

Experimental ^{1}H-NMR spectra of the components were recorded by a Bruker DRX 500 spectrometer at 11.744 T field using 5 mm Z-gradient inverse probe in dry $CDCl_3$ solvent and 1% TMS. Set of the potential components: **C** and **M** [$l = 2$, c.f. Eq. 1]. Experimental molar ratios correspond to the weighted mixing ratios in Eq. 4. For the cases Original and Broadened see Sect. 3.2.4

[a] See footnote a of Table 1

It seemed that Eq. 10 can be ill-conditioned by the occurrence of narrow peaks, thus the NMR spectra were processed in two different ways. In the "Original" test series, the typical LB = 0.3 Hz parameter was used for the standard line broadening, since in the second case ("Broadened" series), the experimental NMR peaks were artificially broadened with LB = 2 Hz parameter by the exponential multiplication method. The experimental values and the results are shown in Table 3.

As presented in Table 3, the NMR test case was rather successful, with the largest difference between the (experimental) weighting and the DISS values being 3.5 ("Original" case) and −2.9 ("Broadened" case) molar percentages, respectively. The results shown in Table 3 prove that the use of artificially broadened NMR peaks is able to enhance the numerical accuracy of the DISS method.

4 Conclusions

In this work we have generalized our earlier procedure [38] based on the spectral proximity between the resulting mixture spectrum and the chemical components' spectra. The new DISS method has a mathematical analogy to the DIIS methods of Pulay [34,35] and of Pulay and Császár [36], methods widely used in quantum chemistry. Our DISS method operates in the spectral subspace of finite dimension of the Hilbert-space, spanned by (non-)normalized and non-orthogonal spectra of the chemical components which are supposed to be present in the mixture. It has been shown that the linear coefficients are the molar ratios whose sum is equal to unity, which is the only constraint explicitly used in the framework of the Lagrangian undetermined multipliers in DISS method. Despite the fact that the method leads to a linearized system of equations, the DISS method is iterative subsequent to the unknown value of the normalization constant of the CDWA spectrum. The iteration was convergent until ‘self-consistency’ in all test cases investigated. The DISS method is capable of finding the exact composition of mixtures in ideal cases (without associations, noise and impurities) in its non-normalized version. The applicability of the normalized version of the DISS method was shown in simple IR and NMR applications. Our method can be used in conjunction with any kind of molecular spectra (IR, NMR, Raman, EPR, CD, VCD, and so on) and can also treat spectra with negative data points. The application of the non-normalized version of the new method will be published in a forthcoming paper.

We have shown that the normalized version of the DISS method is insensitive to the mass of the samples and does not require the knowledge of the *molar* absorptivities or ellipticities and so on. This means that the use of the method is very easy: it does not need, for example, the uniform value of the path length at the recording of the mixture’s and components’ spectra. However, in order to obtain information about the confidence of the method (that is, about the validity of the conditions accepted like spectral additivity and linear independence, the role of the noise, as well as about the unknown proximity of norms of the non-normalized components’descriptive spectra), before the use of the DISS method for quantitative analysis, it is advisable to make a series of control measurement on the general accuracy.

We emphasize that the quality of the descriptive spectra used in the DISS method is not confined to the well-defined approximate levels of molecular quantum mechanics [59–61]. The well-known ‘spectrum-generators’ are also useful, similarly to the ACD/HNMR predictor well-spread in NMR spectrometry [62–64].

One question remains, namely the maximum size of the problems where the DISS method is applicable. We believe that the size is only limited by the noise of the experimental spectra (and the level of the quantum mechanical approximation, if any of them is used). In the case of very large matrices, it is not advisable to use a matrix inversion since inversion methods scale with the cube of the size of the matrix. In such cases, iterative solution for a system of linear equations are more favourable in providing a quadratic scaling.

We hope that our novel method will be extensively used in quantitative chemical and biochemical analysis. While the DISS method is supposed to be used most frequently in the IR and NMR spectra, it can also be fruitfully used in Raman spectroscopy, in particular, in aqueous media. This allows for the investigation of products and materials in food and drug chemistry as well.

Acknowledgments The authors thank Dr. Gy. Tarczay (Eötvös Loránd University, ELTE) for the recording of the infrared spectra. This work has been supported in part by the Hungarian Scientific Research Fund, OTKA, through Grants K68427 and T037658. The authors are grateful to Prof. P.G. Szalay (ELTE), Prof. P. Pulay (Univ. of Arkansas, Fayetteville), Prof. Lajos László (ELTE), and Prof. R. Rajkó (Univ. of Szeged) for helpful discussions.

Appendix A

In order to use the present method, we have to take into consideration two principles. However, since we are not able to prove these principles at this time, it is suggested that they be treated as *postulates*.

Postulate I states that the relationship between the compounds and their experimental spectra is mutually unambiguous. Here, the term ‘spectrum’ means any kind of molecular spectra (UV/Vis, IR, Raman, NMR, EPR, VCD, and so on) of the compounds, and it has to use a ‘wide enough’ interval. The verification of Postulate I is as follows. It is a well-known fact that the band origin of any spectral band is the difference of the corresponding stationary energy levels E_κ and $E_{\kappa'}$, respectively. There can be random cases where two or more different systems have the same gaps between some of their energy levels. However, the intensity of a band (e.g., absorption) depends generally on the matrix element of the $\hat{M}$ operator of the specific interaction with respect to the $\Psi\kappa$ initial and $\Psi\kappa'$ final state vectors:

$$\left\langle \Psi_{\kappa'} | \hat{\mathrm{M}} \Psi_\kappa \right\rangle. \tag{A.1}$$

Both the initial and the final state vectors are characteristic for the compound investigated. Although we can only measure the differences of the intensities, if we consider a ‘wide enough’ interval of any spectra, even at low resolution, it is very likely that it will be characteristic for one, and only one, compound.

Postulate II is somewhat stricter than the previous one. It states that the experimental spectra of the different compounds are not only different but linearly independent. Let us think for the scalar product of two spectra (corresponding to different compounds) [38]. If the scalar product is zero, the spectra would obviously be linearly independent [because none of the experimental spectra could be the zero vector, see text following Eq. 1]. However, if the scalar product is not zero (let us consider the fingerprint region of the IR spectra, for instance), even this circumstance allows for the spectra to be linearly independent. In other words, we consider the $\{|\Phi_I\rangle\}_1^l$ set of the descriptive spectra of the potential components as a normalized (or non-normalized) and always non-orthogonal basis set in the actual subspace of the Hilbert-space. Moreover, it is supposed that Postulates I and II correspond not only for the experimental but also the computed ‘descriptive’ spectra of the compounds (that is, the quality of the computed spectra is “good enough”).

Appendix B

It is easy to show that the *normalized* $\left|\Phi_I^{\exp}\right\rangle$ experimental spectrum of the *I*-th potential component ($I = 1, 2, \ldots, l$) corresponds to unit number of molecules. Let us start with N_I number of ‘spectroscopically active’ molecules in the light beam (the N_I ‘s are zero for the $I = m+1, m+2, \ldots, l$ virtual components); supposing the additivity, the resulting spectrum for the *I*-th component is:

$$N_I \left|\varphi_I^{\exp\prime}\right\rangle, \tag{B.1}$$

where $\left|\varphi_I^{\exp\prime}\right\rangle$ is the non-normalized spectrum of the *I*-th component originating from unit number of molecules (hitherto the Greek capitals mean spectra originating from an indefinite number of molecules, since Greek lower-case letters mean those originating from unit number of molecules. Also, the apostrophe stands for non-normalized spectra. Naturally, 'unit number of molecules' may mean one *mole* of the *I*-th components' molecules as well, and in that case N_I would mean the number of moles of 'spectroscopically active' molecules.) The square of the norm of this vector is:

$$\left\langle N_I\varphi_I^{\exp\prime}|N_I\varphi_I^{\exp\prime}\right\rangle = N_I^2\left\langle\varphi_I^{\exp\prime}|\varphi_I^{\exp\prime}\right\rangle \equiv N_I^2\cdot\left\|\varphi_I^{\exp\prime}\right\|^2, \tag{B.2}$$

thus the normalization constant is

$$\mathcal{N}_I = \left(N_I\left\|\varphi_I^{\exp\prime}\right\|\right)^{-1}. \tag{B.3}$$

Therefore, the normalized $\left|\Phi_I^{\exp}\right\rangle = N_I\left|\Phi_I^{\exp\prime}\right\rangle = \left|\varphi_I^{\exp}\right\rangle$ spectrum corresponds to unit number of molecules of the *I*-th chemical component. Usually, we do know neither the number of spectroscopically active molecules, nor the norm of the components' spectra originating from unit number of molecules (excepting the case of the components' *computed* spectra). Thus, the normalization constant can be determined according to the Eq. 3b in case of the normalized version of the present method. The *I*-th component's non-normalized $\left|\varphi_I^{\exp\prime}\right\rangle$ spectrum originated from unit number of molecules can be expressed as

$$\left|\varphi_I^{\exp\prime}\right\rangle = \left\|\varphi_I^{\exp\prime}\right\|\cdot\left|\varphi_I^{\exp}\right\rangle, \tag{B.4}$$

obviously. This is a suitable form for the purpose of the non-normalized version of our method.

Now, let us express the resulting spectrum of an additive mixture applying the non-normalized $\left|\varphi_I^{\exp\prime}\right\rangle$ components' experimental spectra, and suppose that there are N_I number of active molecules in the light beam ($I = 1, 2, \ldots, l$, respectively):

$$\begin{aligned}\left|\Phi^{\text{mix}\prime}\right\rangle &= \sum_I N_I\left|\varphi_I^{\exp\prime}\right\rangle = \left(\sum_J N_J\right)\left(\frac{\sum_I N_I\left|\varphi_I^{\exp\prime}\right\rangle}{\sum_J N_J}\right)\\ &= \left(\sum_J N_J\right)\sum_I x_I^{\exp}\left|\varphi_I^{\exp\prime}\right\rangle = \left(\sum_J N_J\right)\left|\varphi^{\exp\prime}\right\rangle,\end{aligned} \tag{B.5}$$

where $\left|\varphi^{\exp\prime}\right\rangle$ is the (non-normalized) 'gross descriptive' spectrum originating from the mixture's "unit number of *average* molecules" and the $x_I^{\exp}$ linear coefficients are, obviously, the molar ratios. (We denoted the normalization coefficient of $\left|\Phi^{\text{mix}\prime}\right\rangle$ as $\mathcal{N}^{\text{mix}}$, c.f., Eq. 5, thus, $\left|\varphi^{\text{mix}}\right\rangle = \mathcal{N}^{\text{mix}}\left|\Phi^{\text{mix}\prime}\right\rangle$). Consequently, if Condition Ib is also valid, due to the normalizations, the mixture's experimental spectrum and the components' experimental spectra are equally independent from the number of spectroscopically active molecules, thus, for example, from the path length and the weighting amounts (mass of the samples). This is very advantageous feature of the normalized version of the present method.

Appendix C

In Eq. 8 a trivial constraint on the molar ratios is given. Seemingly, this alone is insufficient because it also has to be valid for all of the coefficients $0 \leq x_I \leq 1[I = 1, 2, \ldots, l$; together with Eq. 8 the so called 'convex constraints']. In the original DIIS procedures [34–36], these 'additional' constraints are completely invalid because the linear coefficients can also be negative (!). However, in the current DISS procedure, the situation is the opposite: negative mixing coeffients are physically meaningless.

We can easily show that there is no need to explicitly implement all the convex constraints; it is simply enough to require the fulfillment of Eq. 8. The verification of this statement is as follows.

Eq. C.1 shows the structure of the coefficient matrix of Eq. 10:

$$\underline{\underline{A}} = \begin{pmatrix} \underline{\underline{S}} & \underline{b} \\ \underline{b}^{\dagger} & c \end{pmatrix}, \tag{C.1}$$

where $\underline{\underline{S}}$ is the (positive definite) Gram matrix of the $|\Phi_I\rangle$ $(I = 1, 2, \ldots, l)$ normalized descriptive spectra of the potential components, $\underline{b}$ is a vector, † means the adjoint, and c is a number. With the help of the Schur complement [65] of $\underline{\underline{S}}$, it could be proven that the determinant of $\underline{\underline{A}}$ is:

$$det(\underline{\underline{A}}) = det(\underline{\underline{S}}) \cdot (c - \underline{b}^{\dagger} \underline{\underline{S}}^{-1} \underline{b}). \tag{C.2}$$

In the case of Eq. 10, the value of c is zero and from this fact it follows that $\underline{\underline{A}}$ is negative definite. This means that the $\underline{\underline{A}}$ matrix is not singular, is invertible and, therefore, Eq. 10 has a unique solution. In our case, this must be the 'physically meaningful' solution with $0 \leq x_I \leq 1$ molar ratios, at least in the case of a quantitative analysis (see TA I above).

References

1. R.E. Dodd, *Chemical Spectroscopy* (Elsevier, Amsterdam, 1962)
2. H.H. Jaffé, M. Orchin, *Theory and Applications of Ultraviolet Spectroscopy* (Wiley, New York, 1962)
3. C.N.R. Rao, *Chemical Applications of Infrared Spectroscopy* (Academic Press, New York, 1963)
4. D.L. Rabenstein, D.A. Keire, in *Modern NMR Techniques and Their Application in Chemistry*, ed. by A.I. Popov, K. Hallenga. Practical Spectroscopy Series, vol. 11 (Marcel Dekker, New York, 1991), pp. 323–369
5. R. Kellner, J.-M. Mermet, M. Otto, H.M. Widmer (eds.), *Analytical Chemistry—The Approved Text to the FECS Curriculum Analytical Chemistry* (Wiley, Weinheim, 1998)
6. M.J. Pelletier, Appl. Spectrosc. **57**, 20A–42A (2003)

7. P. Bouguer, *Essai d'Optique sur la Gradation de la Lumiere* (Claude Jombert, Paris, 1729)
8. J.H. Lambert *Photometria sive de mensura et gradibus luminis, colorum et umbræ*, The widow of E. Klett, Augsburg (1760)
9. A. Beer, *Einleitung in die höhere Optik* (F. Vieweg und Sohn, Braunschweig, 1853)
10. P. Kubelka, F. Munk, Zeit. für Techn. Phys. **12**, 593–601 (1931)
11. J.L. Saunderson, J. Opt. Soc. Am. **32**, 727–729 (1942)
12. F.C. Williams, F.R. Clapper, J. Opt. Soc. Am. **43**, 595–597 (1953)
13. D.M. Haaland, R.G. Easterling, D.A. Vopicka, Appl. Spectrosc. **39**, 73–84 (1985)
14. P. Bai, J. Liu, Chin. Opt. Lett. **4**, 243–246 (2006)
15. Ch.J.C. Burges, Data Min. Knowl. Discov. **2**, 121–167 (1998)
16. http://research.microsoft.com/~cburges/papers/SVMTutorial.pdf
17. H.E. Rose, Nature **169**, 287–288 (1952)
18. W.H. Lawton, E.A. Sylvestre, Technometrics **13**, 617–633 (1971)
19. O.S. Borgen, B.R. Kowalski, Anal. Chim. Acta. **174**, 1–26 (1985)
20. O.S. Borgen, N. Davidsen, Z. Mingyang, Ø. Øyen, Microchim. Acta. **II**, 63–73 (1986)
21. A. de Juan, R. Tauler, Crit. Rev. Anal. Chem. **36**, 163–176 (2006)
22. R. Rajkó, K. István, J. Chemometrics **19**, 448–463 (2005)
23. R. Rajkó, J. Chemometrics **20**, 164–169 (2006)
24. M.A. Sharaf, D.L. Illman, B.R. Kowalski, *Chemometrics* (Wiley, New York, 1986)
25. Y.-P. Sun, D.F. Sears, J. Saltiel, Anal. Chem. **59**, 2515–2519 (1987)
26. J. Saltiel, D.F. Sears, Y.-O. Choi, Y.-P. Sun, D.W. Eaker, J. Phys. Chem. **98**, 35–46 (1994)
27. B. Hessling, G. Souvignier, K. Gerwert, Colloq. INSERM **221**, 155–158 (1992)
28. B. Hessling, G. Souvignier, K. Gerwert, Biophys. J. **65**, 1929–1941 (1993)
29. S.D. Brown, S.T. Sum, F. Despagne, B. K. Levine, Chemometrics Review, Fundamental Review Isuue. Anal. Chem. **68**, 21R–61R (1996)
30. O.M. Kvalheim, D.W. Aksnes, T. Brekke, M.O. Eide, E. Sletten , Anal. Chem. **57**, 2858–2864 (1985)
31. O. Steinbock, B. Neumann, B. Cage, J. Saltiel, S.C. Muller, N.S. Dalal, Anal. Chem. **69**, 3708–3713 (1997)
32. B. Vandeginste, R. Essers, T. Bosman, J. Reinen, G. Kateman, Fresenius' J. Anal. Chem. **57**, 971–985 (1985)
33. E. Micheli-Tzanakou, *Supervised and Unsupervised Pattern Recognition: Feature Extraction and Computational Intelligence* (CRC, Boca Raton, 2000)
34. P. Pulay, Chem. Phys. Lett. **73**, 393–398 (1980)
35. P. Pulay, J. Comp. Chem. **3**, 556–560 (1982)
36. P. Császár, P. Pulay, J. Mol. Struct. **114**, 31–34 (1984)
37. A. Navarro, J.J. López González, A. Garcia Fernández, I. Laczik, G. Pongor, Chem. Phys. **313**, 279–291 (2005)
38. Zs. Kolos, D. Knausz, J. Rohonczy, E. Vass, Gy. Tarczay, G. Pongor, Chem. Phys. **318**, 191–198 (2005)
39. K. Levenberg, Q. Appl. Math. **2**, 164–168 (1944)
40. D. Marquardt, SIAM J. Appl. Math. **11**, 431–441 (1963)
41. P. Pulay, in *Molecular Quantum Mechanics: Analytic Gradients and Beyond,* ed. by A.G. Császár, G. Fogarasi, H.F. Schaefer III, P.G. Szalay. (ELTE Institute of Chemistry, Budapest, 2007), pp. 71–73
42. B. Jähne, *Digital Image Processing* (Springer, Berlin, 2005)
43. P. Pulay, G. Fogarasi, G. Pongor, J.E. Boggs, A. Vargha, J. Am. Chem. Soc. **105**, 7037–7047 (1983)
44. G. Rauhut, P. Pulay, J. Phys. Chem. **99**, 3093–3100 (1995)
45. F. Kalincsák, G. Pongor, Spectrochim. Acta. A **58**, 999–1011 (2002)
46. ReactIR 1000 Instrument, Mettler Toledo AutoChem, Inc., USA, http://www.mt.com/autochem
47. Z. Pusztai, G. Vlád, A. Bodor, I.T. Horváth, H.J. Laas, R. Halpaap, F.U. Richter, Angew. Chem. Int. Ed. **45**, 107–110 (2006)
48. S.R. Wilson, A.W. Czarnik (eds.), *Combinatorial Chemistry* (Wiley Interscience, New York, 1997)
49. A.D. Becke, J. Chem. Phys. **98**, 5648 (1993)
50. C.T. Lee, W.T. Yang, R.G. Parr, Phys. Rev. B **37**, 785–789 (1988)
51. W.J. Hehre, L. Radom, P.v.R. Schleyer, J.A. Pople, *Ab Initio Molecular Orbital Theory* (Wiley, New York, 1986)
52. M.E. Magar, Biochemistry **7**, 617–620 (1968)
53. J.P. Hennessey Jr., W.C. Johnson Jr., Biochemistry **20**, 1085–1094 (1981)

54. T.H. Cormen, Ch.E. Leiserson, R.L. Rivest, C. Stein, *Introduction to Algorithms*, 2nd edn. (MIT Press/McGraw-Hill, Cambridge, Mass, 2001), p. 790
55. P.D. Wentzell, J.-H. Wang, L.F. Loucks, K.M. Miller, Can. J. Chem. **76**, 1–12 (1998)
56. A. Perczel, M. Hollósy, G. Tusnády, G.D. Fasman, Protein Eng. **4**, 669–679 (1991)
57. A. Perczel, K. Park, G.D. Fasman, Anal. Biochem. **203**, 83–93 (1992)
58. G. Peintler, I. Nagypál, I.R. Epstein, K. Kustin, J. Phys. Chem. A **106**, 3899–3904 (2002)
59. R. McWeeny, *Methods of Molecular Quantum Mechanics* (Academic Press, London, 1992)
60. R. J. Bartlett, J.F. Stanton in *Reviews in Computational Chemistry*, ed. by K.B. Lipkowitz, D.B. Boyd. vol. 5 (VCH Publishers, New York, 1994), pp. 65–169
61. F. Jensen, *Introduction to Computational Chemistry* (Wiley, Chichester, 1999)
62. ACD/HNMR Predictor; version 7.03 (2003), Advanced Chemistry Development, Inc. (ACD/Labs), Toronto, ON, Canada, www.acdlabs.com
63. ACD/CNMR Predictor; version 10.5 (2007), Advanced Chemistry Development, Inc. (ACD/Labs), Toronto, ON, Canada, www.acdlabs.com
64. B.L. Pagenkopf, J. Am. Chem. Soc. **127**, 3232 (2005)
65. F. Zhang, *The Schur Complement and its Application* (Springer, New York, 2005)