Bulgarian Chemical Communications, Volume 49, Special Issue D (pp. 19– 24) 2017

Identification of mixture components by multiple linear regression and subtraction of reference spectra: searching infrared and raman libraries

S. H. Tsoneva, S. R. Nachkova, G. N. Andreev, P. N. Penchev*

Department of Analytical Chemistry and Computer Chemistry, “Paisii Hilendarski” University of Plovdiv, 24 Tzar Asen Str., 4000 Plovdiv, Bulgaria

Received April 13, 2017; Revised April 20, 2017

Dedicated to Acad. Ivan Juchnovski on the occasion of his 80th birthday

A procedure for mixture analysis by searching in infrared and Raman spectral libraries is proposed and tested with more than 60 binary mixtures of organic compounds. The procedure uses a combination of spectra subtraction and multiple linear regression. The infrared spectra appeared to be more informative and suited for that purpose than Raman spectra. The procedure is implemented into a Windows-based program developed earlier for a library search of vibrational spectra.

Key words: Raman spectra; infrared spectra; library search; mixture analysis

INTRODUCTION

Separation of mixture components and their identification and quantification is the main task of various chromatographic techniques [1] by using reference standards or by hyphenated techniques (usually GC-MS and LC-MS) that provide spectra of the separated components. Nevertheless, the number of reference standards in a lab is limited to several hundreds and smaller than a typical number of several thousands of library spectra. Furthermore, the hyphenated techniques instrumentation is much expensive and missing in most organic labs. On the other hand, the infrared (IR) and Raman spectra are well suited for the identification of organic compounds via library search [2-4]. As the mixture spectrum can be approximated by a linear combination of components spectra, it can be resolved if the reference spectra of the components are available [5]. An extensive review of the application of IR and Raman spectroscopy to mixture analysis is given in [3, 6]. No matter how the mixture analysis is applied, its math is based on multiple linear regression that evaluates the coefficients of the linear combination [6]. Even the successive subtraction (so called *spectrum stripping*) of the reference spectra–multiplied by appropriate concentration factors–from that of the mixture [5-6] is a kind of regression procedure. The last is controlled by the user who monitors and sets (or minimizes) the residuals between both spectra.

All proposed procedures are plagued by one or both of two main shortcomings: (1) the component spectra do not appear among the first hits (if the hit list but not the whole library is processed), and (2) some of the calculated regression coefficients are not statistically significant–both happen usually for the component with a low concentration. That is why, in the present paper the combination of spectra subtraction and regression is studied and the obtained results are evaluated with two objectives in mind: first, proposing a robust routine procedure and second, comparison of the performance of IR and Raman spectra. Moreover, the application of Raman spectroscopy for mixture analysis is still underdeveloped and scanty described in papers dealing with library search in general [7-8].

EXPERIMENTAL

The IR spectra were registered on a Perkin-Elmer 1750 FT-IR Spectrometer from 4000 cm^{-1} to 450 cm^{-1} at resolution 4 cm^{-1} with 16 scans and on a VERTEX 70 Spectrometer (Bruker, Germany) from 4000 cm^{-1} to 400 cm^{-1} at resolution 2 cm^{-1} with 25 scans. The Raman spectra were measured on RAM II (Bruker, Germany) with a focused laser beam of Nd:YAG laser (1064 nm). All spectra were subjected to curvilinear baseline correction by the instrument software (CDS-2 or OPUS v. 6.5). When loaded in our software IRIS, the original spectral data were converted by a smoothing procedure based on weights from a normal distribution. In the present work the library and mixture spectra of organic compounds were solely

* To whom all correspondence should be sent:
E-mail: *plamen@uni-plovdiv.net*

used. Six IR libraries were composed of 911 entries altogether and Raman one of 330 entries with most of the spectra measured in our lab [6].

The Windows-based program, IRSS, for searching in libraries of IR and Raman spectra was programmed in Delphi 1 by one of the authors (P.N.P.) and described elsewhere [6,9]. Seven different measures for comparison of IR spectra were implemented: three for peak matching and four for comparing full spectral curves. IRSS uses the spectral range from 3700 to 500 cm^{-1}, with a sampling interval of 4 cm^{-1}. All IR and Raman spectra, when loaded into IRSS, are normalized in the 0.0-1.0 interval in ordinate; here has to be mentioned that the ordinate of Raman spectra is not in absorbance units but in normalized intensity of the scattered light and future references to the ordinate will be given as dimensionless values. Furthermore, IRSS provides software tools for the import of IR spectra in JCAMP-DX format, for peak picking, and for an analysis of IR spectra of mixtures with a graphical user interface.

Methods

Peak search algorithms described in the literature [2, 6] can be generally divided into two types: (1) *forward* one used for identification of pure compounds, and (2) *reverse* one applied for identification of the components of organic mixtures. The corresponding spectra similarity measures (hit quality indices, HQIs) were implemented in IRSS and described in details in our previous paper [10].

The main requirement for the application of the mixture analysis is that all mixture components have spectra in the used libraries and there are no strong intermolecular interactions between mixture components [5]. Thus, the mixture spectrum (row vector $\mathbf{M}_{1,K}$) can be represented as a linear combination of all hit list spectra (matrix $S_{N,K}$) by Eqn. 1.

$$\mathbf{M}_{1,K} = \mathbf{C}_{1,N}\,\mathbf{S}_{N,K}\,, \tag{1}$$

where the subscripts show the matrix dimensions, i.e. K is the number of used wavenumbers, N is the number of used hits, and the **M**’s and **S**’s elements represent the corresponding absorbance values. In reality, only the component spectra participate in this linear combination but Eqn. 1 includes all hits as the user does not know in advance which hits are the components.

The mixture analysis procedure is the following. At the beginning, the mixture spectrum is searched in the libraries by reverse peak search algorithm. A multiple linear regression is performed [11] with the mixture spectrum and hit list spectra by Eqn. 2.

$$\mathbf{C}_{1,N} = \mathbf{M}_{1,K}\,\mathbf{S}_{K,N}^{T}\,(\mathbf{S}_{N,K}\mathbf{S}_{K,N}^{T})^{-1} \tag{2}$$

where the superscripts T and -1 designate a transposed and inverse matrix, respectively.

The calculated row vector $\mathbf{C}_{1,N}$ (obviously dimensionless) does not represent concentration of hit list compounds in the mixture because all spectra in IRSS are normalized in ordinate in the 0.0-1.0 range and there is no quantitative sample information recorded in the spectral libraries. That is why, $\mathbf{C}_{1,N}$ represent coefficients of linear combination of hit list spectra and are called *pseudo-concentration* by us. The statistically significant values of $\mathbf{C}_{1,N}$ can be used as a decision criterion, revealing which hits are plausible components. In our software IRSS the 95% and 99% confidence limits of the pseudo-concentrations are estimated according to the cited above Massart *et al*. book [11].

The first of these hits (with number F) with a statistically significant pseudo-concentration is assumed as one of the mixture components and its spectrum is subtracted from the mixture spectrum as designated by Eqn. 3.

$$\mathbf{R}_{1,K} = \mathbf{M}_{1,K} - f\,\mathbf{S}_{F,K}\,, \tag{3}$$

where $\mathbf{S}_{F,K}$ is the F row of matrix $\mathbf{S}_{N,K}$, f is a real number such that ‘$f\ \mathbf{S}_{F,K}$’ removes the spectral bands of this component from the mixture spectrum giving the remainder spectrum $\mathbf{R}_{1,K}$. The coefficient f can be set to 1.0 as it is done in [5] because the spectral bands of the component with a higher concentration prevail over the other components’ spectral bands and usually the main component is at the foremost beginning of the hit list. The coefficient f can also be obtained by the user when he/she is monitoring the remainder spectrum acquired by subtraction procedure and decides that the component spectral bands no longer appear in the remainder spectrum. The results with this scheme were published earlier by us [10] and the obtained f’s had values close to 1.0. The software IRSS has a user-friendly interface that facilitates the spectra subtraction. Another option is to set f to a value of $c_{1,F}$ obtained from the multiple linear regression performed initially. These three options are thoroughly studied in the present paper.

Further, the negative values of the remainder spectrum are truncated, the spectrum is normalized and then its peaks are searched in the libraries. A multiple regression is again performed with the remainder spectrum and the newly obtained hits. The first hit with a statistically significant values of $c_{1,S}$ is chosen as the second component. The cycle can be repeated for the third component and so on but our experience shows that more than three components are hard to be identified.

Described in this way, the mixture analysis looks pretty straightforward, but even for a mixture of components with quite different spectra it could fail and give erroneous results. Recommendations are not given in the literature to what extent the subtraction is performed, except that one or more selected spectral bands of the mixture spectrum have to be nullified. Another complication can arise if the mixture components have similar spectra with nearly all bands overlapping (because of their similar structures) thus leading to an over-subtraction–and as a result of it–the second component might not be among the first hits that are obtained from the second library search.

Our experience shows that to propose a procedure for mixture analysis means not only to give its steps (an algorithm) and the optimal values of its various parameters but also to elucidate all creative ways of solving the problem. Without doubt, the educated user would use a kind of versatile library search beyond any prescriptions. He/she would select various spectral intervals, probably starting with finger print region. He/she would monitor that the plausible components' spectral bands are subset of those of the mixture. He/she would vary both search tolerances (in wavenumber and absorbance) as well as would set various threshold values used by peak-picking (probably he/she would select different ones for the mixture, remainder and library spectra). The spectral match algorithms and the number of hits to be processed is also a variable that most influences the results. All these options cannot be systematically studied in the proposed procedure as it is usually done in analytical chemistry when one performs a consecutive optimization upon two or three parameters. That is why, the following parameters are set rigid to the recommended values derived by our previous experience with the library search.

(1) The threshold for peak-picking of library spectra was set to 0.03. Threshold for peak-picking of mixture spectra was set to 0.01 with the idea in mind that there could be a component with a way lower concentration. Some IR mixture spectra showed water vapor rotational bands and some Raman spectra were very noisy so a higher threshold (0.02-0.04) was used.

(2) The remainder spectrum is with a lower signal to noise ratio that is why a higher threshold (0.02-0.09) was used. Raman spectra are in principle with a poor signal to noise ratio thus the threshold value is quite higher than that used with IR spectra.

(3) Our experience showed that the optimal wavenumber tolerance for mixture peak search, $\Delta\nu$, is higher than that used for single compound identification, that is why a values of $\Delta\nu$ = 12 cm^{-1} was selected.

(4) The search tolerance in ordinate, ΔA, was set to 1.0 (maximum value, i.e. the band intensity was not accounted by peak match). Such was done in our previous studies so that all component spectra were in the hit list what is not a requirement for the present study. Despite that difference, the optimal values of ΔA was not searched upon.

(5) Only the first 40 hits were used for regression calculations but the user is advised to review the hit list entries and use lesser or bigger number.

(6) Despite that the search uses peaks in the whole spectral interval (3700 - 500 cm^{-1}), the starting regression interval was set to 1300 - 600 cm^{-1}. If no component identification was achieved, the other intervals recommended by us (and tried in this order) were 1800 - 600 cm^{-1}, 1800 - 500 cm^{-1} and 3700 - 500 cm^{-1}.

(7) The reverse peak search is specially designed for mixture analysis [10] but if the remainder spectrum is with a low signal to noise ratio (this is the case for most Raman spectra), the user can apply one of the four full-spectrum search algorithms [9,13]: for binary mixtures the remainder, that is properly calculated, is a spectrum of one component.

RESULTS AND DISCUSSION

The test of the proposed mixture analysis procedure was performed with 35 IR and 60 Raman spectra of mixtures of organic compounds: all spectra were measured by two of the authors – P.N.P and S.H.T. The spectral files were ordered randomly with a separate numbering for the IR and Raman spectra. Three series of search results were produced. First, the used subtraction procedure was governed only by the first three heuristics derived earlier [10], i.e. without setting the coefficient *f* in Eqn. 3 close to 1.0. The random IR and Raman spectra were searched and analyzed by one of the authors (S.R.N.) who knew neither the components nor the composition of the mixtures, i.e. the mixture analysis was performed as close as possible to the so called *blind experiment*. Second, the mixture analysis using all four heuristics [10] (i.e. with additionally *f* = 1.00) was applied, and third, *f* equal to the corresponding regression coefficient from Eqn. 2 was used for spectra subtraction. The last two series of mixture analysis were performed also on the randomly numbered spectra but setting *a priori f* to 1.00 or to the regression coefficient did not necessitate a kind of blind experiment. Only 20 of the mixtures, Table 1, were used to evaluate the results: the remaining spectra were some sort of 'padding' (ballast) in the present study in order to complicate the blind experiment. Those were IR and Raman spectra of the mixtures of hexane and cyclohexane, benzene and pyridine, and the Raman spectra of the mixtures of 1-octanol and 1-decanol, 1-nonanol and 1-decanol, 2-methyl-1-phenylpropan-1-one and 4'-methylpropiophenone, benzylacetone and butyrophenone, 2-ethylhexane-1-ol and 2-ethylhexane-1,3-diol, cyclopentanone and benzylacetone.

The first and third series produced comparable results that were substantially better than those produced in the second series. It appeared that the *f* value set by a spectroscopist, Eqn. 3, (i.e. first series of mixture analysis) had been less than the corresponding regression coefficient and the bands of the first-found component retained in the remainder spectrum.

Table 1. The identification of mixture components. The mixture concentration (in volume ratio) is given in the first column and the regression coefficients are designated with f_1 and f_2; (error) another compound was found as a component of the mixture

(a) 2'-methylacetophenone (A) and 3'-methylacetophenone (B)

A:B	IR				Raman			
v / v	1st found	2nd found	f_1	f_2	1st found	2nd found	f_1	f_2
1:1	B	A	0.50	1.07	A	B	0.50	0.94
1:4	B	A	0.81	0.77	B	A	0.91	0.81
1:9	B	A	0.88	0.48	B	A	0.85	0.35
4:1	A	B	0.71	0.66	A	B	0.78	0.86
9:1	A	B	0.83	0.77	A	B	0.84	0.47

(b) 1,4-dioxane (A) and tetrahydrofuran (B)

A:B	IR				Raman			
v / v	1st found	2nd found	f_1	f_2	1st found	2nd found	f_1	f_2
1:1	A	B	1.05	0.73	A	B	0.65	0.97
1:4	A	B	0.92	1.02	B	A	0.77	0.91
1:9	B	A	1.00	0.84	B	A	0.40	0.58
4:1	A	B[1]	1.08	0.74	A	B	0.85	0.37
9:1	A	–[2]	0.99	–	A	B	0.90	0.46

(c) 3-heptanone (A) and 4-heptanone (B)

A:B	IR				Raman			
v / v	1st found	2nd found	f_1	f_2	1st found	2nd found	f_1	f_2
1:1	B	A	0.35	0.40	B	A	0.38	0.93
1:4	B	A	0.59	0.31	B	A	0.77	1.44
1:9	B	A	0.71	0.38	B	A[1]	0.98	4.04
4:1	A	B	0.49	0.36	A	B[1]	0.95	1.67
9:1	A	error	0.63	1.61	A	B[1]	0.92	1.30

(d) 1-nonanol (A) and 5-nonanol (B)

A:B	IR				Raman			
v / v	1st found	2nd found	f_1	f_2	1st found	2nd found	f_1	f_2
1:1	A	B	0.16	0.36	–[1]	–	–	–
1:4	B	error[1]	0.47	0.74	B	error[1]	0.43	–
1:9	B	error[1]	0.57	0.72	B[1]	error[1]	0.44	1.12
4:1	A	B	0.23	0.09	B[1]	–	0.15	–
9:1	A	error[1]	0.22	0.60	error[1]	error[1]	0.62	0.48

[1] The original spectral interval for regression calculations was widened (see text).

[2] The component was not found, i.e. its regression coefficient is not statistically significant.

Despite that, the correct identification of the second component took place in most cases: the usage of reverse search of the remainder spectrum assisted the appearance of the second component in the second hit list. On the other side, the analysis of the failures in the second series showed that when both components have common spectral bands with high intensity, these bands sum together and the normalized mixture spectrum is a linear combination of the component spectra with coefficients quite less than 1.0.

In Table 1 are given the third series results, i.e. *f* is set to the regression coefficient. As can be seen, even the very structurally similar components as 3-heptanone and 4-heptanone can be identified from their mixture spectra. As expected, problems appear with some of the 1:9 or 9:1 v/v mixtures. On the other side, the worst results are for the mixtures of 1-nonanol and 5-nonanol and they could be explained with the presence of 1-decanol spectrum in the IR and Raman libraries: there is a very subtle difference between 1-nonanol and 1-decanol IR as well as Raman spectra. Also several other primary alcohols appeared in the hit list and their spectra are very similar to that of 1-nonanol.

Here has to be mentioned that O-H stretching band is completely missing in Raman spectra of saturated alcohols and these spectra are very close to those of the compounds with nearly the same aliphatic part. The C=O stretching band in Raman spectra is of very low intensity, and these both spectral peculiarities are the reason for several of the errors when the spectral interval used by regression is widened. The other mixtures (not presented in Table 1) showed similar results. As a whole the IR spectra gave better results than the corresponding Raman spectra.

Several particularities can be illustrated with the analysis of the mixture (1:1 v/v) of structurally similar compounds as 1-nonanol and 5-nonanol are. Peak search of the IR mixture spectrum resulted in 1-nonanol as a sixth hit and 5-nonanol as a tenth hit; first five hits are 1-hexanol, 1-decanol twice (a repetition in these libraries), dodecane and octacosane. The remainder spectrum is calculated with the coefficient in Eqn. 3 taken from the performed regression, *f* = 0.16. Fig. 1 shows the mixture and component spectra in the 1500 - 600 cm^{-1} interval, as well as the remainder spectrum. As can be seen, (1) the concentration ratio of 1:1 v/v does not mean equal pseudo-concentrations, (2) in the 1300-600 cm^{-1} interval the components spectra are most different, (3) the components have overlapped bands at 1465 and 1380 cm^{-1} as consequence of their common substructures, CH_3 and CH_2 groups, (4) the main difference between the component spectra is in C-OH stretching bands (primary and secondary alcohol, respectively) and $\rho(CH_2)$, 724 and 732 cm^{-1}, and (5) the spectral bands of 1-nonanol, ν(C-OH) = 1058 cm^{-1} and $\rho(CH_2)$ = 724 cm^{-1}, were vastly removed by subtraction, i.e. they are not present in the remainder spectrum.

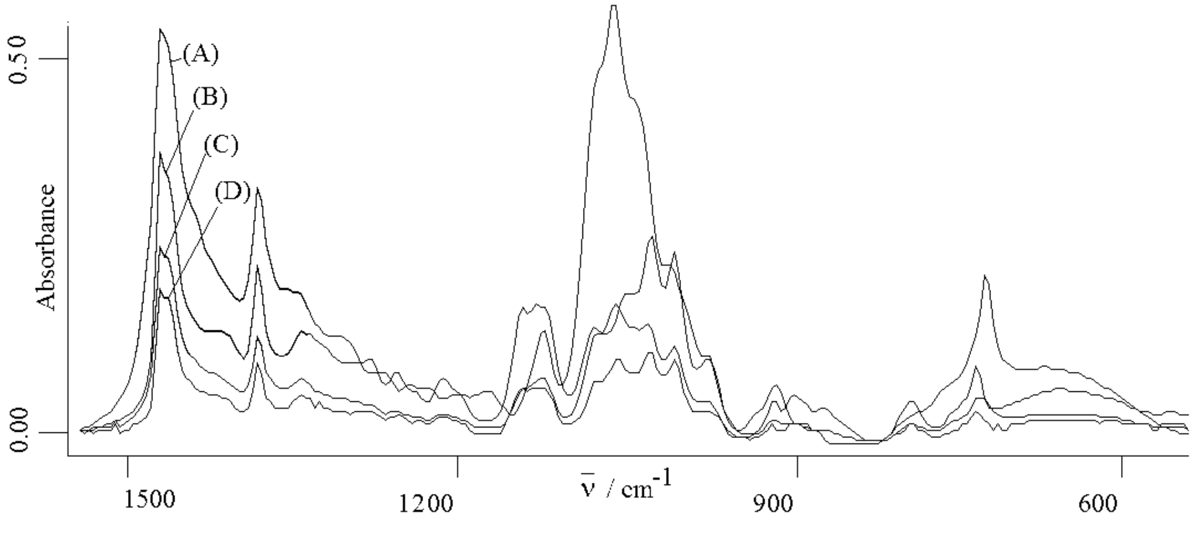

Fig. 1. Spectra of (A) the mixture (1:1 v/v) of 1-nonanol and 5-nonanol, (B) the remainder spectrum (see text), (C) 1-nonanol, (D) 5-nonanol.

CONCLUSION

The procedure for mixture analysis by searching in IR and Raman spectral libraries is implemented and tested. The components are identified by their statistically significant coefficients obtained by multiple linear regression. The user is advised to use the corresponding regression coefficient for subtraction of the spectrum of first-found component from that of the mixture. Another option is for the user to monitor that the certain spectral bands are disappearing in the remainder spectrum in the process of subtraction.

REFERENCES

1. V. Oliveri, G. Vecchio, *EJMECH.*, **120**, 252 (2016).
2. R. Freitag (ed.), Modern Advances in Chromatography. Springer-Verlag, Berlin, 2002.
3. J. T. Clerc, E. Pretsch, M. Zuercher, *Microchim. Acta*, **II**, 217 (1986).
4. H. J. Luinge, *Vib. Spectrosc.*, **1**, 3 (1990).
5. W. A. Warr, *Anal. Chem.*, **65**, A1087 (1993).
6. H. Somberg, Qualitative mixture analysis by use of an infrared library search system. Brucker Reports, 1988.
7. P. N. Penchev, Dr. Sc. Dissertation, University of Plovdiv, 2016.
8. M. B. Denton, R. P. Sperline, J. H. Giles, D. A. Gilmore, C. J. S. Pommier, R. T. Downs, *Australian J. Chem.*, **56**, 117 (2003).
9. P. Vandenabeele, *Spectrochim. Acta, Part A*, **80**, 27 (2011).
10. K. Varmuza, P. Penchev, H. Scsibrany, *J. Chem. Inf. Comput. Sci.*, **38**, 420 (1998).
11. P. N. Penchev, V. L. Miteva, A. N. Sohou, N. T. Kochev, G. N. Andreev, *Bulg. Chem. Commun.*, **40**, 556 (2008).
12. D. L. Massart, B. G. M. Vandeginste, S. N. Deming, Y. Michote, L. Kaufman, Chemometrics: A Textbook, Elsevier, Amsterdam, 1988.
13. P. N. Penchev, A. N. Sohou, G. N. Andreev. *Spectrosc. Lett.*, **29**, 1513 (1996).
14. P. N. Penchev, N. T. Kochev and G. N. Andreev, *Compt. Rend. Acad. Bulg. Sci.*, **51**, 67 (1998).

ИДЕНТИФИКАЦИЯ НА КОМПОНЕНТИТЕ НА СМЕСИ С МНОГОПРОМЕНЛИВА ЛИНЕЙНА РЕГРЕСИЯ И ИЗВАЖДАНЕ НА СПЕКТРИ: ТЪРСЕНЕ В ИНФРАЧЕРВЕНИ И РАМАН СПЕКТРАЛНИ БИБЛИОТЕКИ

С. Х. Цонева, С. Р. Начкова, Г. Н. Андреев, П. Н. Пенчев *

Катедра „Аналитична химия и компютърна химия“, Пловдивски Университет „Паисий Хилендарски”, ул. Цар Асен № 24, 4000 Пловдив

Постъпила на 13 април 2017 г.; Коригирана на 20 април 2017 г.

(Резюме)

Предложена е процедура за анализ на смеси чрез търсене в библиотеки от инфрачервени и Раман спектри. Процедурата е тествана с повече от 60 смеси от две органични съединения. За анализа се прилага комбинация от многопроменлива регресия и изваждане на спектри. Използването на инфрачервени спектри дава по-добри резултати от тези, получени с Раман спектри. Процедурата е реализирана в програма, работеща в среда на Windows.