ISSN 1061-9348, Journal of Analytical Chemistry, 2015, Vol. 70, No. 9, pp. 1055–1061. © Pleiades Publishing, Ltd., 2015.
Original Russian Text © Yu.B. Monakhova, A.M. Tsikin, S.P. Mushtakova, 2015, published in Zhurnal Analiticheskoi Khimii, 2015, Vol. 70, No. 9, pp. 925–932.

ARTICLES

Independent Components Analysis as an Alternative to Principal Component Analysis and Discriminant Analysis Algorithms in the Processing of Spectrometric Data

Yu. B. Monakhova, A. M. Tsikin, and S. P. Mushtakova

Institute of Chemistry, Chernyshevskii State University, Saratov, Astrakhanskaya ul. 83, Saratov, 410012 Russia
e-mail: yul-monakhova@mail.ru

Received July 4, 2014; in final form, January 15, 2015

Abstract—The possibility of the application of independent component analysis (ICA) to searching patterns in the spectrometric datasets and to discriminating objects is demonstrated. The data of XRF analysis of base enamels, IR spectra of automotive lacquers, and ^{1}H NMR spectra of wines from different regions of Germany are selected for the study. In all three cases, ICA reliably separates groups of objects, increasing the percentage of correct predictions for new samples not included into the model. Moreover, ICA gives results comparable with specialized discriminant analysis methods (linear discriminant analysis, projections to latent structures discriminant analysis, and factorial discriminant analysis) in the classification of the NMR spectra of wines.

Keywords: spectroscopy, chemometrics, independent component analysis, principal component analysis, classification methods

DOI: 10.1134/S1061934815090117

The application of independent component analysis becomes more common in the practice of spectroscopic analysis of various processes [1–8]. ICA methods solve mathematically the problem of separation of individual sources and their relative contributions from the total spectroscopic signal without making any assumptions about the number of mixture components, their molecular structure, or type of spectrum. The underlying hypothesis, first applied to spectral analysis in [9], is the assumption of independence of the spectra of the mixture components. Improved ICA algorithms have been recently developed, enabling to find the least dependent (as opposed to fully independent) components, which gives a definite advantage when modeling systems with strongly overlapping signals [1, 2, 10–12].

Based on the statistical fundamentals of the method, it is obvious that ICA is rather efficient to recover the sources of signals contained in total overlapping spectra available for recording. Therefore, it is not surprising that this method is already widely used for multicomponent analysis of objects of complex composition based on various spectroscopic signals, including electronic (absorption and emission), IR, and ^{1}H NMR spectra [1–8]. It is found that the uncertainty in determining the concentrations of compounds in the mixtures normally does not exceed 10%, and the correlation coefficients between the recovered and experimental spectrum are not less than 0.90 [1, 2, 6–8]. It is also important in the modern business environment that the duration of ICA modeling is not more than 5 min, making it suitable for screening analysis.

Recently, works are carried out on testing ICA to solve other problems in analytical chemistry, for example, the study of acid-base and tautomeric equilibria and complexation reactions, including those involving hydrogen bonds [13–15].

On the other hand, the possibility of using ICA for solving discrimination problems is poorly studied. The idea is to apply the resulting matrix of the spectra of individual components and their relative contribution to the total signal as an alternative to the loadings matrix in principal component analysis (PCA) and the PCA score matrix, respectively. The usability of ICA for solving classification problems is shown by an example of discrimination of orange and grapefruit juices and their mixtures, based on ^{1}H NMR spectroscopy [16, 17]. However, in this case, the use of ICA modeling is logical, as two selected individual component reflected mainly the spectra of individual orange and grapefruit juices, and the mixtures of juices represented their linear combinations.

In another recent study, ICA in conjunction with conventional chemometric methods, such as PCA, linear discriminant analysis (LDA), factorial discriminant analysis (FDA), projections to latent structures discriminant analysis (PLS–DA), soft independent modeling of class analogy (SIMCA), has been used for discriminating samples of rice regarding their varieties

and geographical origin on the basis of ^{1}H NMR spectroscopy [18]. ICA modeling enabled a complete separation Basmati rice from other varieties of long-grain rice, which is impossible with conventional PCA. Moreover, ICA exhibits excellent sensitivity and selectivity of the classification model, being outperformed only by PLS–DA [18].

The above examples demonstrate the prospects of using ICA for discrimination of NMR spectroscopic data. It is obvious, however, that further work is required on the testing of ICA algorithms for solving classification problems of spectroscopic experiment of other types.

In this work, we selected for analysis X-ray fluorescence (XRF) data for base enamels, IR spectra of automotive lacquers, and ^{1}H NMR spectra of wines from four closely spaced wine regions of Germany. One of the above datasets (^{1}H NMR spectra) had been previously processed with PCA, which gives the possibility to compare the effectiveness of both methods [19].

EXPERIMENTAL

Equipment and samples. We studied samples of black base enamels: $n = 21$, catalog number 490, manufacturers: Vika ($n = 6$), Dynacoat ($n = 6$), Quickline ($n = 3$), and Duxone ($n = 6$). Transparent two-component lacquers were also studied: $n = 18$, manufacturers: Vika ($n = 6$), Helios ($n = 3$), OTRIX ($n = 6$), and RAND ($n = 3$). These materials are presented in the Saratov retail chain as products for repair painting of vehicles. The set of objects can be considered as representative for the Saratov region.

Sample preparation of lacquer samples for recording their IR spectra included the preparation of a mixture of lacquer with a hardener in the recommended proportions (3 : 1) and the application of the mixture with a spatula to form films on KBr crystals followed by drying under an IR lamp (60°C, 2.5 h). The spectra of the lacquer samples were recorded using an Infralum FT-801 FTIR spectrometer in the wavenumber range of 4000–500 cm^{-1} with a scanning step of 2 cm^{-1}.

Sample preparation of enamels included the application of a layer of enamel with a spatula on a polyethylene terephthalate (lavsan) film and drying under an IR lamp. The spectra of base enamels were recorded using a Shimadzu Rayny EDX-720 X-ray fluorescence spectrometer. Scanning was performed within 0.00–40.96 keV with an increment of 0.02 keV.

We selected 111 authentic samples of wines from four wine regions of Germany: Nahe (15), Moselle (31), Rheinhessen (35), and Pfalz (30). We recorded ^{1}H NMR spectra of wines by means of a Bruker Avance 400 Ultrashield spectrometer. A detailed description of the sample preparation and recording of the NMR spectra of signals is given in [19]. The test samples of wines included white and red grape varieties of Pinot Blanc (23), Pinot Noir (22), Riesling (33), Kerner (14), Müller-Thurgau (10), and Pinot Gris (9), gathered during the period from September 15 to November 11, 2009.

Preliminary processing of spectral data. Chemometric analysis of the spectral data was performed using the Matlab 2013b software package (The Math Works, United States) and a SAISIR toolbox [20]. Ellipsoids with a 95% probability were plotted using additional calculations with the SAISIR software package.

Bucketing was used to reduce the size of ^{1}H NMR spectroscopic data and to level the shift of the position of maxima in the spectra of wines [19, 21]. Furthermore, two methods for the pretreatment (autoscaling and Pareto scaling) [22] were tested for each of the three datasets to eliminate variations in the intensities of various signals.

ICA modeling. While PCA is based on finding the orthogonal axes describing the maximum variance in the data in multidimensional space of variables, the goal of ICA is to restore the "pure" sources from the observed sum signals [23]. We used the least dependent component analysis based on the mutual information based least dependent component analysis (MILCA). The MILCA algorithm has a MATLAB interface and is available on the Internet for free [24].

Full cross-validation [25] was used to determine the number of principal components (PCs) required for the development of optimal PCA models. On the other hand, before ICA decomposition, the number of significant independent components (sources) is determined by the ICA-by-Blocks procedure [26]. The method consists in dividing the original data matrix to B blocks (in this case, $B = 2$), consisting of approximately the same numbers of spectra of samples. Next, the ICA modeling of each block is performed with a different number of independent components, and then the models with the same number of independent components are compared with each other by calculating the correlation coefficients between each pair of selected signals. The optimal number of components is determined by a sharp decline in the correlation coefficients [26]. In this work, samples were randomly split 30 times.

The following procedure of chemometric data processing was used. The matrix of experimental spectra X, in which the number of rows corresponds to the number of samples in the dataset, and the columns represent variables (for example, the wave numbers for IR spectroscopy or chemical shifts for NMR spectroscopy), was set at the input for the ICA algorithm. ICA decomposition was carried out for the number of independent components, found by ICA-by-Blocks, and the number of nearest neighbors ranged from 5 to 15.

The result of the application of ICA are the signals of individual sources and their corresponding relative

concentrations (the contribution of a component in the total spectrum). Further, a matrix is derived, where each object corresponds to the vector of the relative contributions of all independent components found. In other words, the mixing matrix can be regarded as an analogue of the PCA score matrix. Therefore, the procedure for discriminating the samples on the basis ICA modeling coincides with the sequence of actions in the analysis of the PCA score matrix; namely, such a combination of independent components is visually found, where the presentation of their relative contributions in the two-dimensional coordinate system leads to the best discrimination of the groups of objects. Thus, in the use of ICA, any additional operations in the classification of the objects, with respect to PCA, are not required. Further information can be obtained by studying the recovered spectra of individual sources, which are similar to the PCA load matrix.

Discriminant algorithms. To model the most representative set of data (NMR spectra of wines), we also used LDA, FDA, and PLS–DA. These methods refer to the classification methods with training, the basis of which is to construct a mathematical model for each of the desired groups using the calibration dataset. In distinction from PCA, these methods are specifically designed for predicting the characteristics of new samples not included in the model. The models developed by ICA, LDA, FDA, and PLC–DA are validated using full cross-validation and a test set (25 of 111 samples).

RESULTS AND DISCUSSION

The need for chemometric analysis of the first two of the three test datasets is caused by difficulties in solving problems of identification in forensic expertise of paints and coatings. The solution of such problems, in particular, suggests an answer to the question of attribution of separately represented fragments of coatings of a specific object (for example, a motorcar) to a particular manufacturer at the group or type assignment. On the other hand, different nature of XRF and IR spectra is of some interest for testing ICA as a new approach for processing experimental data.

XRF spectra of base enamels. The XRF spectra of base enamels were the first analyzed dataset. The selection of spectroscopic method is explained by that the base enamels are metallized materials, which determines the specific effects of visual perception of the coatings [27, 28].

In preliminary studies, the XRF spectra of enamels of various trademarks are identical to each other, which makes it necessary to apply chemometric methods for modeling spectra for their classification. PCA was used to discriminate groups of samples in accordance with the manufacturer (Fig. 1a). Despite the almost complete separation of clusters, ellipsoids of groups Dynacoat and Duxone partially overlap each other at a 95% probability, making the assignment of a new sample to one of these two groups of producers difficult.

As an alternative approach, ICA was applied to the same dataset. Using ICA-by-Blocks, three significant independent sources were identified in the system. The resulting plot of scores in coordinates IC1–IC2 with the use of a three-component ICA model is presented in Fig. 1b. When comparing the data in Figs. 1a and 1b, it becomes apparent that ICA is more effective, because it enables not only the distinguishing of groups Dynacoat and Duxone but also a more reliable discrimination of all four groups under consideration because of the large distances between clusters.

IR spectra of automotive lacquers. Similar results were obtained in the analysis of spectroscopic signals of a different nature, namely, the IR spectra of automotive lacquers of four different brands. The plot of PCA scores is shown in Fig. 2a. As in the case of base enamels, the clusters of two brands (OTRIX and HELIOS) overlap, although only slightly. The complete separation of four clusters, however, became possible using ICA modeling (Fig. 2b). As in the previous case, the ICA model was developed based on the calculation of three independent components.

Noteworthy that in the cases of both XRF and IR data, ICA ellipsoids corresponding to different manufacturers, are significantly smaller than those obtained by the results of PCA modeling with the same degree of probability (95%) (Figs. 1 and 2). This means that ICA models have greater stability and reliability: if the values of ICA scores of a new object fall into an ellipsoid of a smaller size, this indicates a higher probability of identification of an object group affiliation in comparison with the PCA model.

^{1}H NMR spectra of wines. Determination of stable isotopes (^{2}H, ^{13}C, ^{18}O) is currently the official and standard method for determining the authenticity of many foods and beverages, including wine products [29–31]. Recently, however, alternative innovative methods for food control began to appear; one of them is NMR spectroscopy [32–34]. NMR coupled with multivariate data analysis is used to control grape varieties, geographical origin, and vintage of wines [34–37]. The modeling of NMR results by means of conventional methods of classification (LDA, PLS-DA) offered a high percentage of correct predictions for these characteristics [34–37].

To examine ICA, a representative set of NMR spectra of 111 spectra of wines produced in four regions of Germany (Nahe, Moselle, Rhainhessen, and Pfalz) was selected for this study. It should be noted that conventional PCA is principally suited for the analysis of this dataset; however, it does not offer a complete separation of clusters of Nahe and Rheinhessen (Fig. 3a). As in the case of IR and XPF spectroscopic data, ICA showed the best discriminative ability by completely separating completely all four groups at 95% probability (Fig. 3b).

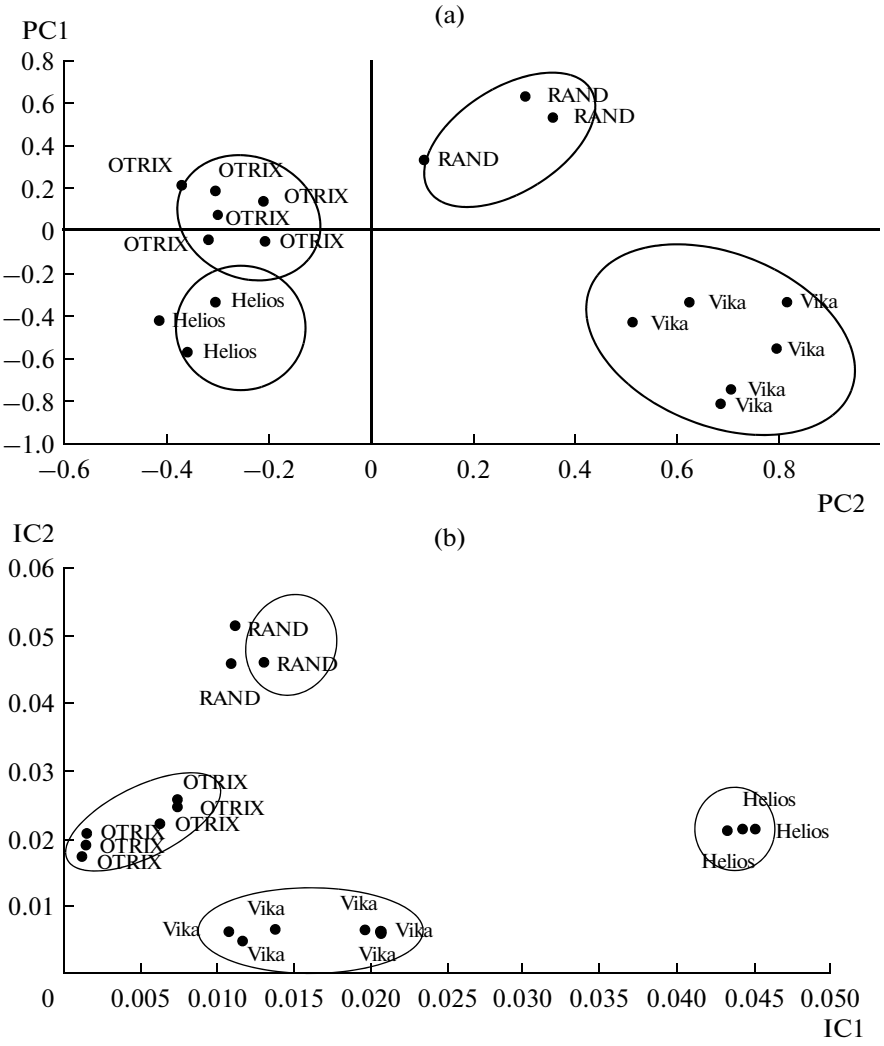

Fig. 1. Plot of (a) PCA and (b) ICA scores for the model of base enamels ($n = 21$, XRF spectroscopy); ellipsoids represent a 95% probability.

After careful examination of Fig. 3, it is seen that the best PCA model is obtained in region PC2–PC3, and for ICA, the most significant for separation are IC1 and IC2. Explanation of this fact can be found in the physical sense of PC and IC. The sequence of selection of PC in the PCA modeling is strictly determined. Each subsequent PC is orthogonal to the previous one and describes the maximum variance in the data. In ICA the order of components is random, but they all are significant and have a physical meaning. Therefore, the first PC may not correspond to the first IC. The best combination of PCs (ICs) should be determined visually using a plot of corresponding scores.

Because of its representativeness, the NMR dataset is useful to illustrate the algorithm of ICA-by-Blocks. The average values of the correlation coefficients with the standard deviation between the pairs of separated signals for ICA models with different numbers of components for ^{1}H NMR spectra of wine are shown in Fig. 4. Obviously, for models with up to six ICs, the correlation coefficients between the corresponding ICs of each block are high, indicating the separation of similar spectral profiles of each block (Fig. 4). Further addition of independent components to the model leads to a rapid deterioration of correlation coefficient themselves (and their reproducibility) between the separated signals of two blocks, which means that the seventh independent component is not significant and represents the background noise. Therefore, ICA decomposition was carried out considering six independent components. This approach ensures the separation of significant sources and can serve as a measure of quality of the decomposition of spectra in the discrimination of samples.

The undoubted advantage of ICA compared to PCA is the fact that it can be used to classify new objects, not included in the training set, using the following equation:

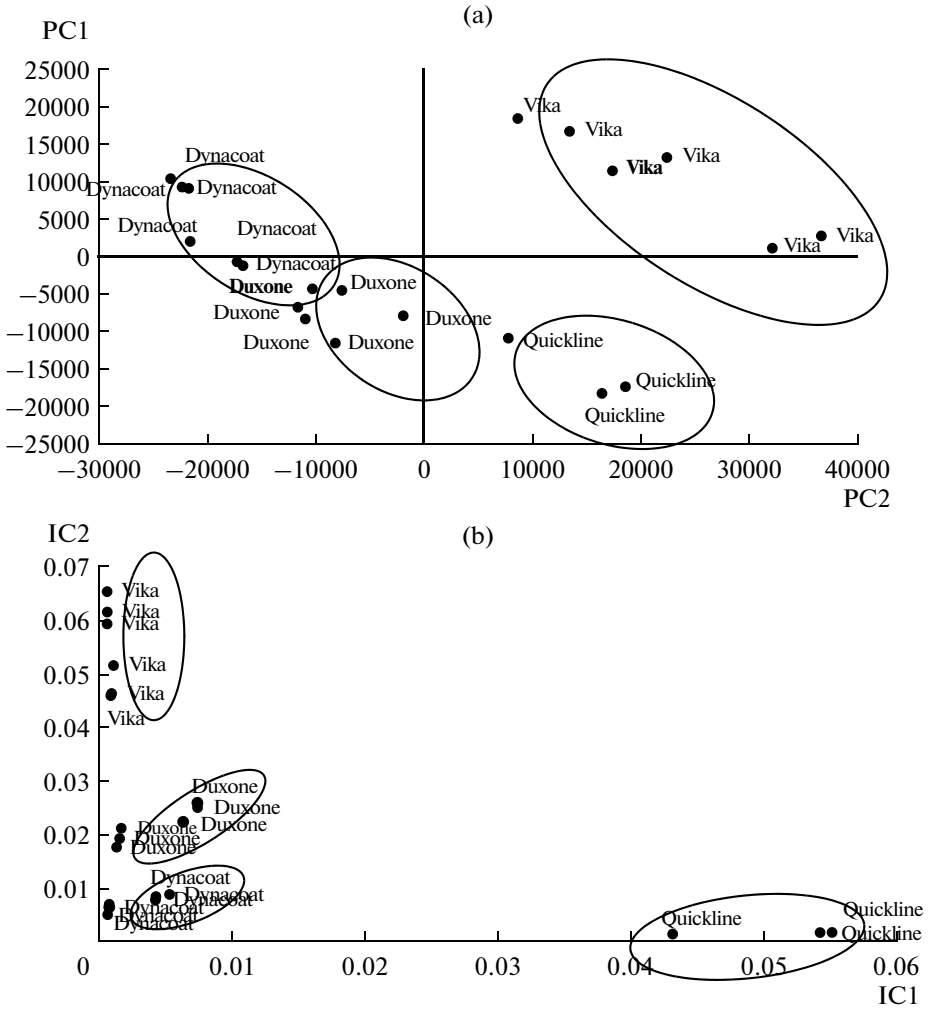

Fig. 2. Graph of (a) PCA and (b) ICA scores for the model of lacquers ($n = 18$, IR spectroscopy); ellipsoids represent a 95% probability.

$$\text{Scores} = \mathbf{X} \times \mathbf{S} \times \text{inv}(\mathbf{S}' \times \mathbf{S}),$$

where **X** is the matrix of signals of new samples and **S** is the matrix of calculated independent components. The sample is considered to be correctly classified if its scores are inside the ellipsoid of the desired groups with a 95% probability [18, 19]. Due to the small size of the IR and XRF data, the validation of the ICA model is held only by the example of the classification of geographical origin of wines by their ^{1}H NMR spectra.

The results of the full cross-validation indicate that ICA can correctly identify 92% of wine samples with respect to their geographical origin ($n = 111$). Further examination of the resulting ICA model showed that 90% of the samples from the test dataset ($n = 25$) are correctly classified correctly. Thus, the developed ICA model is accurate and stable. When analyzing a new model, the values of its scores falling in a particular cluster is an objective sign to assign this object to a particular group. It should be noted that this model enables us to distinguish the geographical origin of red and white wines from the 2009 harvest of early and late varieties. However, for other vintages, separate models must be developed.

It is interesting to compare the discriminatory ability of ICA with classification methods commonly used by analysts to model the spectroscopic data. For this dataset, the following values of correct predictions are obtained (cross-validation, %/test set, %): LDA, 93/89; PLS–DA, 93/84; and FDA, 88/76. We can conclude that ICA is superior to FDA and comparable to PLS–DA and LDA for both calibration and test datasets.

CONCLUSIONS

Thus, our calculations on real data have shown that ICA is effective enough to discriminate objects based

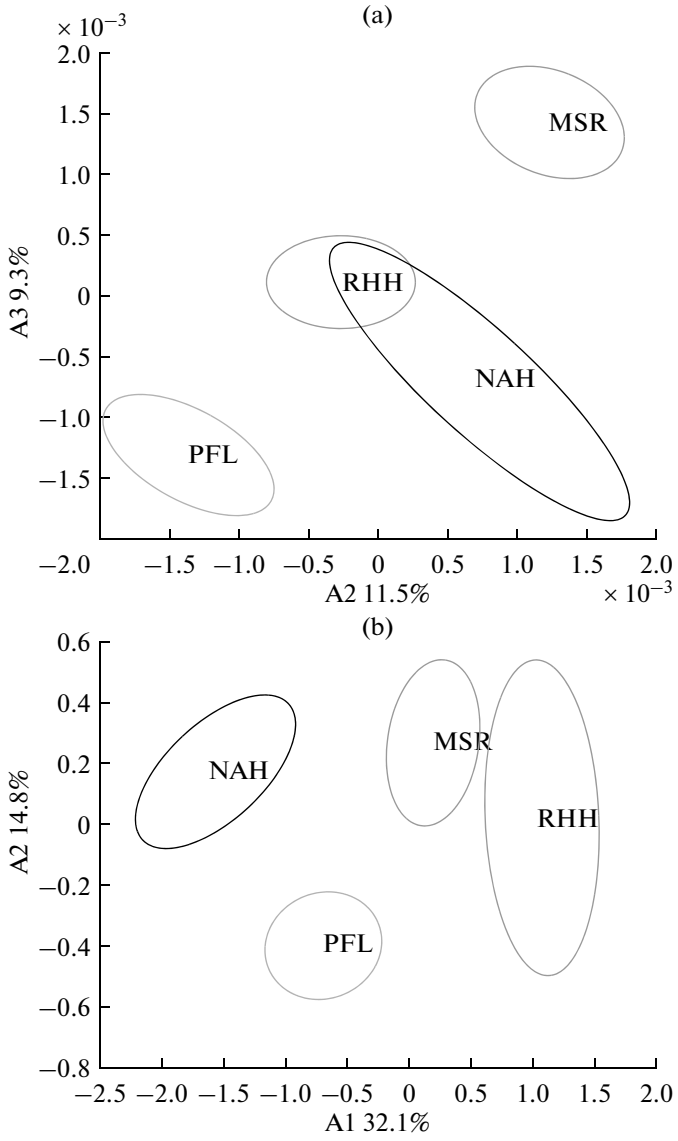

Fig. 3. Graph of (a) PCA and (b) ICA scores for the model of geographical origin of wines ($n = 111$, ^{1}H NMR spectroscopy); regions: NAH, Nahe; MSR, Moselle; RHH, Rheinhessen; and PFL, Pfalz; ellipsoids represent a 95% probability; denominations of axes: (a) PC2 and PC3 and (b) IC1 and IC2.

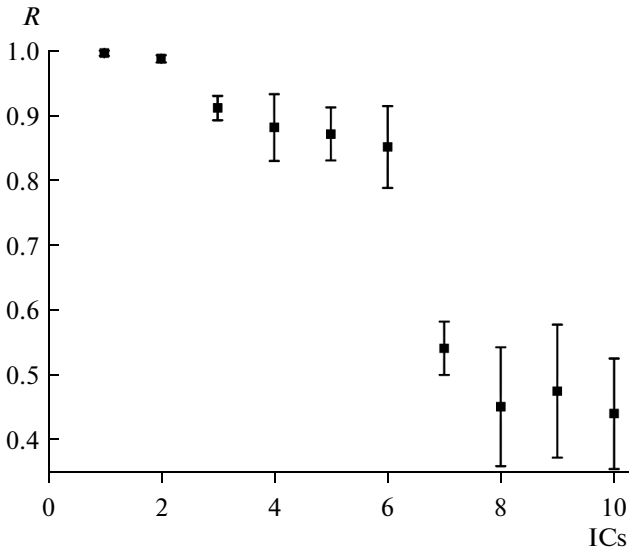

Fig. 4. ICA-by-Blocks to determine the number of significant independent components for NMR spectra of wines of the 2009 harvest ($n = 83$); the average correlation coefficients along with the standard deviation for 30 random partitions into two blocks are shown.

on spectrometric data of different types (NMR, IR, and XRF) and represents a serious alternative to PCA. In particular, ICA can solve classification problems on the assignment of lacquers and base enamels to a particular brand in the expertise of coatings and on the determination of the geographical origin of wine. ICA enables the separation of a group of objects that overlap in the space of PCA scores. The ICA "load" matrix is much easier to interpret, as it reflects the true spectra of components that contain only positive intensity, while PCA loads are abstract and difficult to find chemical interpretation [18].

It is also important that ICA shows results in the classification of new samples, comparable with conventional discriminant methods (LDA, PLS–DA, and FDA). Obviously, ICA is useful for the solutions for classification tasks in NMR, IR, and XRF spectroscopy, differing by the nature and characteristics of spectral bands, and the area of its application can be extended to other analytic signals.

A further stage of the work is to compare the effectiveness of different ICA algorithms (for example, RADICAL, JADE, or FastICA) regarding the discrimination of spectroscopic data. Moreover, ICA models should be tested with larger datasets to provide a complete validation using the test set. In-depth theoretical analysis of the ICA and PCA methods is required to find the reasons for their different efficiency for solving classification problems.

ACKNOWLEDGMENTS

The work is supported by the Ministry of Education and Science of the Russian Federation, state assignment no. 4.1212.2014/K.

REFERENCES

1. Monakhova, Y.B., Astakhov, S.A., Kraskov, A.V., and Mushtakova, S.P., *Chemometr. Intel. Lab. Syst.*, 2010, vol. 103, no. 2, p. 108.
2. Monakhova, Y.B., Mushtakova, S.P., Kolesnikova, S.S., and Astakhov, S.A., *Anal. Bioanal. Chem.*, 2010, vol. 397, no. 3, p. 1297.
3. Schelkanova, I. and Toronov, V., *Biomed. Opt. Express*, 2012, vol. 3, p. 64.

4. Mecozzi, M., Pietroletti, M., Scarpiniti, M., Acquistucci, R., and Conti, M.E., *Environ. Monit. Assess.*, 2012, vol. 184, p. 6025.
5. Hao, J., Zou, X., Wilson, M., Davies, N.P., Sun, Y., and Peet, A.C., *NMR Biomed.*, 2012, vol. 25, p. 594.
6. Monakhova, Y.B., Tsikin, A.M., Kuballa, T., Lachenmeier, D.W., and Mushtakova, S.P., *Magn. Reson. Chem.*, 2014, vol. 52, no. 5, p. 231.
7. Monakhova, Y.B., Kolesnikova, S.S., and Mushtakova, S.P., *Anal. Methods*, 2013, vol. 5, p. 2761.
8. Monakhova, Yu.B., Astakhov, S.A., Mushtakova, S.P., and Gribov, L.A., *J. Anal. Chem.*, 2011, vol. 66, no. 4, p. 351.
9. Lawton, W.H. and Sylvestre, E.A., *Technometrics*, 1971, vol. 13, p. 617.
10. Kraskov, A., Stogbauer, H., and Grassberger, P., *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.,* 2004, vol. 69, p. 066138.
11. Astakhov, S.A., Stogbauer, H., Kraskov, A., and Grassberger, P., *Anal. Chem.*, 2006, vol. 78, p. 1620.
12. Stogbauer, H., Kraskov, A., Astakhov, S.A., and Grassberger, P., *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.*, 2004, vol. 70, p. 066123.
13. Monakhova, Yu.B. and Mushtakova, S.P., *J. Anal. Chem.*, 2010, vol. 65, no. 10, p. 995.
14. Monakhova, Yu.B., Kolesnikova, S.S., Mushtakova, S.P., and Gribov, L.A., *J. Anal. Chem.*, 2011, vol. 66, no. 1, p. 53.
15. Monakhova, Yu.B., Kuznetsova, I.V., and Mushtakova, S.P., *J. Anal. Chem.*, 2011, vol. 66, no. 6, p. 565.
16. Cuny, M., Vigneau, E., Le Gall, G., Colquhoun, I., Lees, M., and Rutledge, D., *Anal. Bioanal. Chem.*, 2008, vol. 390, p. 419.
17. Cuny, M., Le Gall, G., Colquhoun, I.J., Lees, M., Rutledge, D., *Anal. Chim. Acta*, 2007, vol. 597, p. 203.
18. Monakhova, Y.B., Rutledge, D.N., Rossmann, A., Waiblinger, H.-U., Mahler, M., Ilse, M., Kuballa, T., and Lachenmeier, D.W., *J. Chemometr.*, 2014, vol. 28, no. 2, p. 83.
19. Monakhova, Y.B., Godelmann, R., Hermann, A., Kuballa, T., Cannet, C., Schafer, H., Spraul, M., and Rutledge, D.N., *Anal. Chim. Acta*, 2014, vol. 833, p. 29.
20. Cordella, C.B.Y. and Bertrand, D., *TrAC, Trends Anal. Chem.*, 2014, vol. 54, p. 75.
21. Monakhova, Yu.B., Kuballa, T., and Lakhenmaier, D.V., *J. Anal. Chem.*, 2013, vol. 68, no. 9, p. 755.
22. Berg, R.A., Hoefsloot, H.C., Westerhuis, J.A., Smilde, A.K., and van der Werf, M.J., *BMC Genomics*, 2006, vol. 7, p. 142.
23. Rutledge, D.N. and Bouveresse, D.J., *TrAC, Trends Anal. Chem.*, 2013, vol. 50, p. 22.
24. Kraskov, A.V., http://www.uclac.uk/ion/departments/sobell/Research/RLemon/MILCA/MILCA. Cited June 9, 2014.
25. Wold, S., *Technometrics*, 1978, vol. 20, no. 4, p. 397.
26. Bouveresse, J.R., Moya-Gonzalez, A., Ammari, F., and Rutledge, D.N., *Chemometr. Intel. Lab. Syst.*, 2012, vol. 112, p. 24.
27. Brock, T., Groteklaes, M., and Mischke, R., *Lehrbuch der Lacktechnologie* (Textbook of Lacquer Technology), Auflage: Vincentz Network, 2009.
28. Kosheleva, L.I., *Teor. prakt. sudebn. ekspert.*, 2012, no. 3, p. 149.
30. Aghemo, C., Albertino, A., Gobetto, R., and Spanna, F., *J. Sci. Food Agric.,* 2011, vol. 91, p. 2088.
31. Schmidt, H.-L., *Fresenius' J. Anal. Chem.*, 1986, vol. 324, p. 760.
32. Magdas, D.A., Cuna, S., Cristea, G., Ionete, R.E., and Costinel, D., *Isot. Environ. Health Stud.*, 2012, vol. 48, p. 345.
33. Le Gall, G. and Colquhoun, I.J., *NMR Spectroscopy in Food Authentication. Food Authenticity and Traceability*, Cambridge: Woodhead, 2003.
34. Spraul, M., Schutz, B., Humpfer, E., Mortter, M., Schafer, H., Koswig, S., and Rinke, P., *Magn. Reson. Chem.*, 2009, vol. 47, p. 130.
35. Godelmann, R., Fang, F., Humpfer, E., Schutz, B., Bansbach, M., Schafer, H., and Spraul, M., *J. Agric. Food Chem.*, 2013, vol. 61, p. 5610.
36. Son, H.S., Hwang, G.S., Ahn, H.Y., Park, W.M., Lee, C.H., and Hong, Y.S., *Food Res. Int.*, 2009, vol. 42, p. 1483.
37. Pereira, G.E., Gaudillere, J.P., van Leewen, C., Hilbert, G., Lavialle, O., Maucourt, M., Deborde, C., Moing, A., and Rolin, D., *J. Agric. Food Chem.*, 2005, vol. 53, p. 6382.

Translated by O. Zhukova