MULTIVARIATE SPECTROSCOPY ANALYSIS FOR CALSSIFICATION OF MOLDAVIAN MATURED WINE DISTILLATES

Khodasevich Michail[1], Scorbanov Elena[2], Cambur Elena[2], Gaina Boris[3], Degtyar Nataly[2]

[1]B.I. Stepanov Institute of Physics, National Academy of Sciences of Belarus
[2]Practical Scientific Institute of Horticulture and Food Technology, Republic of Moldova
[3]Academy of Science of Moldova, Honorary member Academy of Romanian Scientists

Scorbanov Elena: skorbanova@rambler.ru

Abstract. It was demonstrated one of the possible ways for solving the authenticity problem of matured wine distillates. Using as a reference spectral information about wine distillates of transmission spectra, angular dependence of scattering spectra will significantly improve the quality of classification and increase the correlation between the values of chemical parameters and principal components. It is proposed to use the methods of PCA (principal component analysis), classification trees and PLS (projection on latent structures) to determine their efficiency in the process of classification of wine distillates.

Key words: wine distillates, transmission spectra, principal component analysis, classification tree, projection on latent structures.

Introduction

Currently an increasing number of consumers care about their health and want to buy natural and authentic food. Authenticity (originality) is an inherent constituent part of a food quality. It defines by a set of physical, chemical and biological parameters, whose absolute quantitative values and change intervals are validated by the natural properties of raw materials and an acceptable technological influence at the ready food manufacturing. Authentication is rather critical in manufacturing and quality control of cognacs and brandies produced from matured wine distillates. The main factors preventing falsification and manufacturing the low-quality product are the control of distillates' age and geographical origin and an identification of the manufacturer.

The only conventional optical characteristics of cognacs and brandies are optical densities at wavelengths 420 nm and 520 nm [1]. However, there are different substances with similar optical properties in wine distillates. It impedes to infer about the quality and features of considered objects on the base of spectral measurements at a little number of assigned wavelengths. In this paper we apply the multivariate spectroscopy analysis to solving the problems of classification of Moldavian matured wine distillates.

Materials and methods

We have created an array of data on the physical and chemical composition of wine distillates' samples of different ages produced in Moldova by various manufacturers. The volatile components were determined by gas-liquid chromatography on the chromatograph GC HP 4890D with FID-detector, the decomposition products of lignin (aromatic aldehydes and acids) were determined on the liquid chromatograph Shimadzu LC-20A.

The typical example of distillates' transmission spectra is presented in Fig.1. It is registered by double-beam spectrophotometer Shimadzu PC 3101. Spectral resolution is

0.5 nm in the range from 190 to 480 nm and 1 nm in the range from 480 to 2600 nm. 1 mm optical path cuvette is used for spectral ranges from 190 to 480 nm and from 1100 to 2600 nm. 10 mm optical path cuvette is used for spectral range from 480 to 1100 nm. Spectra have been smoothed by 9-point cubic polynomial Savitzky-Golay filter after registration [2].

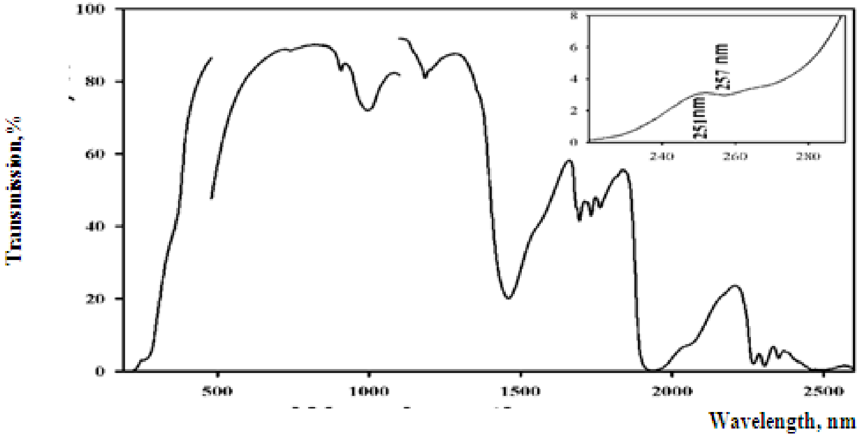

Fig.1. *Typical spectrum of matured wine distillate. Spectral region of "cognac maximum" is shown in the inset*

In the seventies of the twentieth century the appearance of high-performance computers led to the possibility of effective multivariate data processing. Traditional analytical methods demand the great time expenses, high-priced equipment and consumed materials. It was found that they can be replaced by cheap formal and indirect methods operating the multivariate data. The real breakthrough was done in infrared spectroscopy, particular in the near infrared region. Formerly this region was of little use because of the intrinsic high noise. It is caused by intense water absorption and scattering in reflection spectra. The earliest applications of multivariate data processing methods were devoted to modeling the spectroscopic data by principal component analysis (PCA) and projection on latent structures (PLS).

PCA [3] is designed to transform the original variables describing the considered set of samples in to new, uncorrelated variables called the principal components that are linear combinations of the original variables. The direction of the first principal component lies along the maximum variance in the original variables. Each subsequent principal component describes smaller variance of original data than preceding ones. In terms of matrix notation the principal components are the eigenvectors of the covariance matrix of the original variables. Depending on the field of application, it is also named as the discrete Karhunen–Loève transform, the Hotelling transform, singular value decomposition and so on. In realization through singular value decomposition the I-by-J matrix $\boldsymbol{X}$ of initial data is decomposed to product of matrices $\boldsymbol{U}$, $\boldsymbol{S}$ and transposed $\boldsymbol{P}$:

$$\boldsymbol{X} = \boldsymbol{USP}^t \quad (1)$$

Here *I* is the number of samples in the set, *J* is the number of original variables, ***U*** is the matrix from orthonormal eigenvectors $\boldsymbol{u}_r$ of the matrix ***X*** multiplied by the transposed matrix ***X***:

$$\mathbf{XX^t u_r} = \lambda_r \boldsymbol{u}_r \tag{2}$$

λ_r are the corresponding eigenvalues. ***P*** is the matrix from orthonormal eigenvectors $\boldsymbol{p}_r$ of the transposed matrix ***X*** multiplied by the matrix ***X***:

$$\mathbf{X^t X p_r} = \lambda_r \boldsymbol{P}_r \tag{3}$$

S is the diagonal matrix with square roots from λ_r in descending order. The classical presentation of PCA is $\mathbf{X = TP^t}$, where matrix ***T*** of scores in PCA is the product of matrices ***U*** and ***S*** in singular value decomposition. This matrix contains the information about the samples. Matrix ***P*** of loadings contains the information about the original variables. The main purpose of PCA is to represent the location of the samples in a reduced coordinate system where instead of *J*-axes (corresponding to *J* original variables) only A principal components ($A<J$, I) can usually be used to describe the set with maximum possible information:

$$\boldsymbol{X} = \sum_{a=1}^{A} \boldsymbol{t}_a \boldsymbol{p}_a^t + \boldsymbol{E} = \boldsymbol{TP}^t + \boldsymbol{E} \tag{4}$$

Here $\boldsymbol{t}_a$ are the principal components. Matrices ***T*** of scores and ***P*** of loadings have dimensions *I*-by-*A* and *J*-by-*A*. *E* is *I*-by-*J* matrix of remainders that contains irrelevant information.

PCA has been applied to the spectra of 42 samples of mature Moldavian wine distillates from 6 different manufacturers. Each spectrum consists of 2698 spectral data counts. PCA decomposes the multidimensional spectral counts space to low-dimensional space of principal components. Total explained variance of distillates' transmission spectra is shown to be as much as 94.5% for 4-dimensional space of principal components.

The first aim of application of PCA to the studied spectra was the identification of distillates' age. PCA cannot find the apparent dependency of scores on age of samples considered. But the great value of the total explained variance allows suggesting the presence of another factor that is modeled by PCA. Fig. 2 presents the score plot where 6 manufacturers are marked differently. You can see that our hypothesis is confirmed. PCA models the belonging to the manufacturer in the first place.

The classification trees making [4] can be applied in 3-dimensional space of principal components for identification of manufacturers. It is one of the kinds of supervised machine learning. The best results are presented in Fig. 3 and are obtained for the algorithm considering all possible combinations of 3-level predictor. Using 3 principal components this classification tree can identify 5 manufacturers from 6 ones considered.

As you could see earlier PCA cannot identify the distillates' age. We use PLS for this purpose. PLS [5] is the bilinear statistical method in contrast to the linear PCA. It projects predictors (spectra in our case) and a response (sample age) into a new low-dimensional space of latent structures simultaneously.

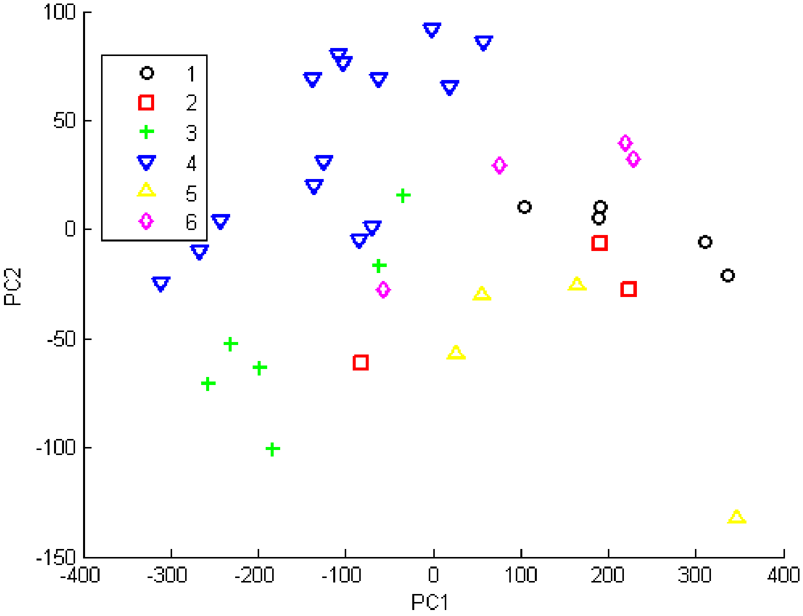

Fig. 2. *Score plots in PC1-PC2 space, distillates manufacturers are marked by different signs.*

21 latent structures give the regression factor of 0.98 on 42 samples of distillates. Results obtained by PLS are presented in Fig. 4 and show the unambiguous definition of distillates' age with relative errors being within 8% limits.

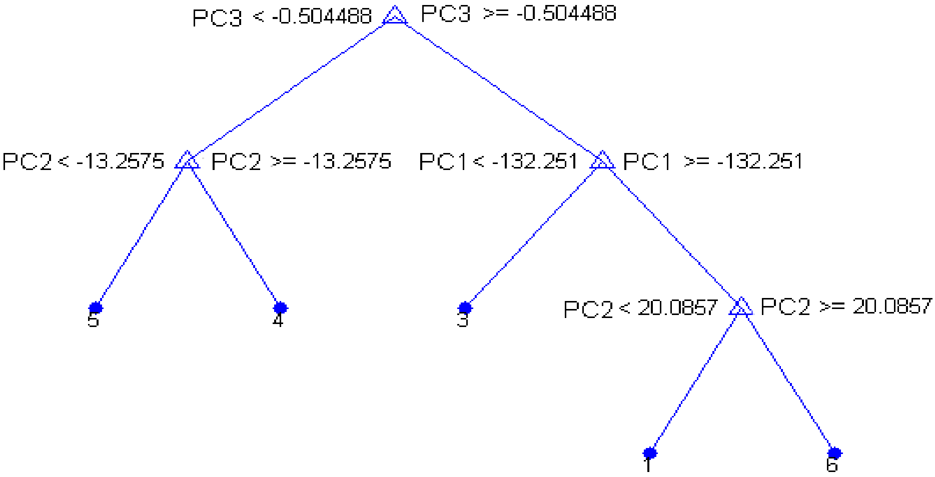

Fig.3. *Identification of distillates' manufacturers by classification tree in 3-dimensional space of principal components*

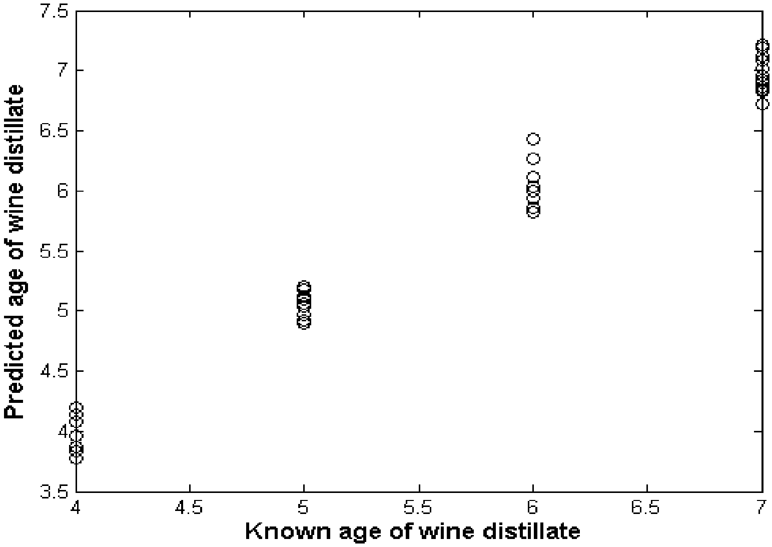

Fig. 4. *Known age of wine distillates versus predicted age as the result of application of PLS to transmission spectra.*

Conclusions

In the result of these studies it was shown the possibility of using multivariate spectroscopy analysis for identification and classification of matured wine distillates. So the application of principal component analysis, classification trees and projection on latent structures to broadband transmission spectra allows defining the manufacturer and age of wine distillates. One of the possible ways is demonstrated for solving the authenticity problem of quality cognac and brandy manufacture. These analyses, along with the physical and chemical parameters and sensory evaluation of the product, can improve the accuracy of the results of the expert opinion in arbitration disputes.

References

1. **Скурихин И. М.**, Химия коньяка и бренди. Москва. 2005.

2. **Дегтярь Н.Ф., Незальзова Е.И., Роговая М.В., Синицын Г.В., Скорбанова Е.А., Ходасевич М.А.**, // Весці Нацыянальнай АкадэміiНавук Беларусі. Серыя Фізіка-Матэматычных Навук. 2014, № 3. С. 113-117.

3. **Abdi H., Williams L. J.**, // Wiley Interdisciplinary Reviews: Computational Statistics. 2010, v. 2. P. 433–459.

4. **Brown S. D., Myles A. J.**, // Comprehensive Chemometrics: Chemical and Biochemical Data Analysis. 2009, v. 3. P. 541–569

5. **Abdi H.**, // Wiley Interdisciplinary Reviews: Computational Statistics. 2010, v. 2. P. 97–106.