CHEMOMETRICS IN ANALYTICAL CHEMISTRY

DARINKA BRODNJAK VONČINA

Faculty of Chemistry and Chemical Engineering, University of Maribor,
Smetanova 17, 2000 Maribor, Slovenia (darinka.brodnjak@uni-mb.si)

Abstract: Chemometrics is a scientific discipline closely connected with statistics and mathematics. It has an important role in analytical chemistry. Modern analytical methods provide opportunity to collect large amounts of data for various samples. For handling analytical results different chemometric methods are employed, such as basic statistical methods for the determination of mean and median values, standard deviations, minimal and maximal values of measured parameters and their mutual correlation coefficients, the *principal component analysis* (PCA), *cluster analysis* (CA), and *linear discriminant analysis* (LDA). The objectives of chemometrics in analytical chemistry are focused on characterization and chemometrical classification of different samples. The quality of environmental samples such as water, sediment, soil, air samples etc. can be determined according to measured physical and chemical parameters, which represent the individual samples. Chemometric methods give information regarding measured parameters about similarity between sampling locations, sources of pollution, seasonal behavior and time trends. Monitoring of general pollution of environmental samples and following measuring parameters which are above permitted level given by legislation can be used for searching of pollution source and for planning prevention measures from pollution. Food samples can also be characterized by chemometrical methods. Chemometrics can be used for fast and efficient determination of food sample categories, such as edible oils, wines, fruits and fruit juices etc. Classification can also be performed according to the origin, source or season. From all these facts it is evident that the aim of chemometrics in analytical chemistry is high and extensive.

Key words: analytical chemistry, chemometrics, principal component analysis, classification, characterization

1. Introduction

Chemometrics is a scientific discipline between measurement oriented chemistry and applied statistics. Analytical chemistry is the most significant discipline where chemometrics plays an important role (ADAMS 1990; WILKINS 1990).

Chemometrics uses mathematical and statistical methods to choose optimal procedures and experiments and to provide chemical information by analysing chemical data (WILLET 1987; NILSSON 1965). In many fields of chemistry there is a need for solving practical oriented problems. Chemists usually can not measure directly the parameters they want to determine, but they are measuring signals on the instruments which then should be calculated into individual parameters, like concentration of components in samples, or different physical parameters.

Modern analytical methods provide large amounts of data. In environmental chemistry monitoring of pollutants in waters, soil and air, provides different multidimensional data, which comprise measurements of parameters for the individual sample, named also an object. Each sample is revealed by many measurements, named also variables. Using multivariate analysis can give us hidden information about quality of environmental samples or about similarity or dissimilarity between different environmental samples (ŠNUDERL *et al.*, 2007; VONČINA *et al.*, 2007). One use of multivariate data in chemistry is discrimination, another is classification. Each object is characterised by a set of measurements. It can be presented as a vector in a multidimensional space. Multivariate methods can help in visualising different objects. Chemometrics includes exploratory data analysis (MASSART *et al.,* 1997; BRERETON 1990), experimental design (VARMUZA 1980; VAN DER WATERBEEMD 1995; DEMMING *et al*., 1987) and modelling. Chemometrics methods are used to find hidden information present in multivariate data, coming from analytical instruments. It can also be used for instrument control, or for multivariate calibration (ADAMS 1990).

There is also possibility to use chemometrics for determining different properties of individual samples in biology, food chemistry, forensic chemistry, geochemistry, archaeology etc. (LANKMAYR *et al*., 2004; ŠNUDERL *et al*., 2005; BRODNJAK-VONČINA *et al*., 2005). It has to be stressed that measurements should be of good quality, accurate and reliable, as this is necessary for obtaining high quality information from multivariate data.

Chemometrics methods are divided mainly into two groups. First group are so called "unsupervised learning methods" where problems are solved by finding similarities in the data. The second group are "supervised learning methods", where also quantitative information about known classes of different samples is available.

Unsupervised learning methods group analytical data using clustering methods, or by projecting high dimensional data onto lower dimensional space, usually onto one plane (Fig. 1).

Supervised learning is a technique for providing a mathematical function from training data. Training consists from pairs of input, objects (vectors) and desired outputs (class label).

The most used techniques for exploratory data analysis are *principal component analysis*, *cluster analysis* (MASSART *et al.,* 1997; BRERETON 1990), and *Kohonen maps (*KOHONEN 1998). Training set with known class memberships is used to calculate a classifier. A prediction set, containing objects not used in the training and also with known class memberships, serves to test the performance of the classifier. The most used techniques for classification are *linear discriminant analysis, K nearest neighbour classification* and *artificial neural networks*, (HECHT-NIELSEN 1987; KOHONEN 1998; DAYHOF 1990; ZUPAN 1989; ZUPAN *et al.,* 1993). The application of supervised learning methods in chemistry is: recognition of compound classes, determination of origin of samples, detection of low/high quality products, etc. Multivariate calibration is the multivariate technique, used in chemical laboratories (ADAMS 1990). Traditional techniques are multiple linear regression and non linear neural networks. The goal in chemometrics methods is to find a mathematical function of the multivariate data for defining new latent variables, possessing maximum problem relevant information. Methods can be grouped in linear and nonlinear methods.

2. Cluster analysis (CA)

Clustering techniques have been applied to a wide variety of research problems, usually for classification or characterisation. Cluster analysis belongs to exploratory data analysis. It allows grouping of samples on the basis of their similarities or differences. Cluster analysis uses all the variance or information contained in original data set. The starting point is a distant matrix containing the distances between all possible pairs of objects. The pairs with the smallest distance are merged and thus number of object is reduced. Result is a hierarchical tree, named “dendrogram”.

3. Principal component analysis (PCA)

Principal component analysis PCA is by far the most important method in multivariate data analysis. It has two main applications: visualisation of the multivariate data and data reduction and transformation. Mathematical pre-processing of the variables, features may have a great influence on the result of data interpretation. Usually data transformation should be performed for variables, i.e. columns in data matrices. Best known are two methods, *column centering* and *column standardisation*. *Column centering* data means that the mean value of each column is subtracted from individual elements of all objects. *Column standardisation or autoscaling* is performed in such way, that the mean of the column elements is subtracted from individual elements and divided by the column standard deviation. Consequently, each column has zero mean and unit variance. The column standardisation method is used when units of individual variables are incomparable.

Principal component analysis PCA is also the most frequently used concept for defining a latent variable. It tries to represent the multidimensional data structure optimally. Direction which best describes the relative distance between the objects is the direction with maximum variance. This direction is called first principal component PC1. The second principal component (PC2) is orthogonal to PC1 and has the second maximum possible variance. PCA can transform data from multidimensional space into two dimensions without losing considerable amount of information. PCA estimates the correlation structure of the variables and hence the possible structure of latent variables, (principal component influencing data structure). Detection of outliers is also possible since they do not belong to a certain class of objects. The projection of the data on the plain of the new latent variables (principal components) is called score of the objects. The scores of the principal component are weighed sum of original variables and weights are called loadings. The scores contain information about objects, and loadings about variables. Similar object are placed closely together in the PC1/PC2 plane.

In Fig. 1 the classification of different vegetable oils according to the content of fatty acids is presented. It is evident that the clusters are formed, corresponding to seven different oil classes. The samples of mixed or unknown oil type labeled with “0” are distributed into seven classes according to their composition of oils..

4. Linear discriminant analysis (LDA)

Linear discriminant analysis belongs to supervised pattern recognition methods (JURS *et al*., 1975, VARMUZA 1980, MASSART *et al.,* 1997, BRERETON 1990) and has the aim to assign object to several predetermined classes. LDA uses latent variables (discriminant variables), that maximally separate different classes of objects between each other. Using data from training set linear functions named discriminant functions are calculated that can be used to predict a class for samples with unknown class label.

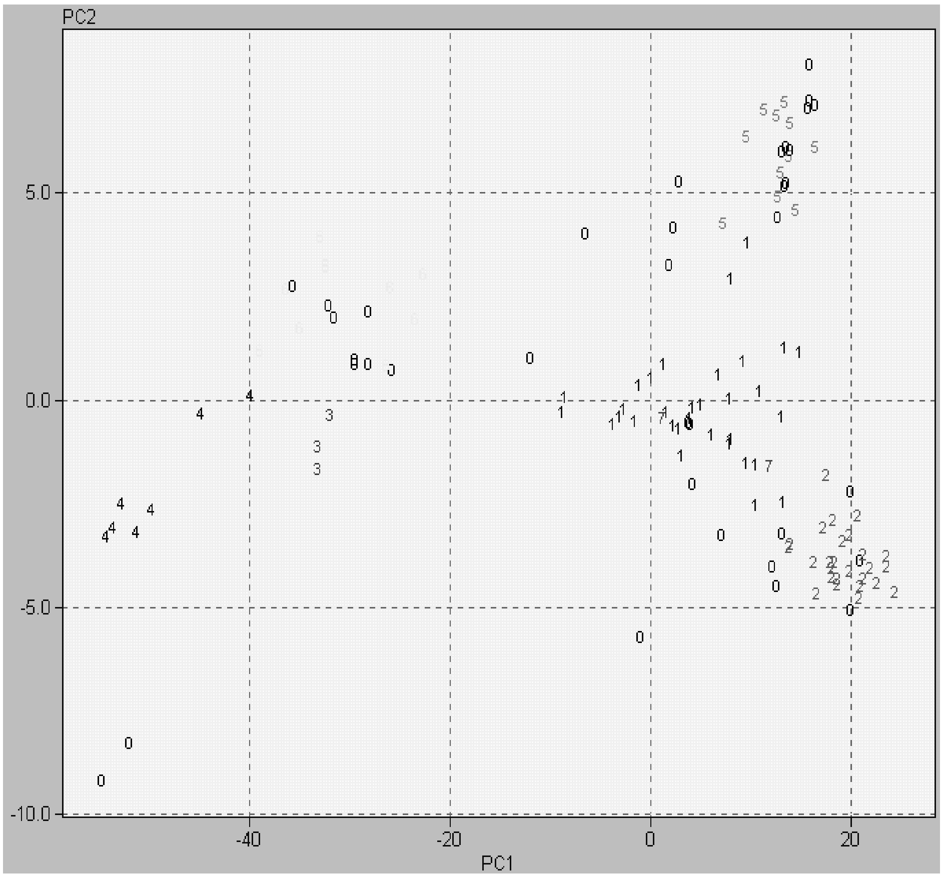

Fig. 1. 132 oil samples: pumpkin (1), sunflower (2), peanut (3), olive (4), soybean (5), rapeseed (6), corn (7), mixed or unknown type (0) in the PC1-PC2 co-ordinate system.

5. Conclusions

It is evident that importance of chemometrics methods in analytical chemistry is exceptional. It helps in solving problems and in obtaining hidden information from complex data and in evaluating different analytical results. It can be applied to develop and implement automated methods for classification of different samples in routine control laboratories. It is used for classification and characterisation of different samples from environmental chemistry, food chemistry, biology, forensic chemistry, geochemistry, archaeology and many other scientific disciplines.

References

ADAMS, M.J.: Chemometrics in Analytical Spectroscopy. The Royal Society of Chemistry, Cambridge 1990.

BRERETON, R.G.: Chemometrics. Application of mathematics and statistics to Laboratory Systems, Ellis Horword, New York, 1990.

BRERETON, R.G.: Multivariate Pattern Recognition in Chemometrics. Elsevier, Amsterdam 1990.

BRODNJAK-VONČINA, D., CENCIČ-KODBA, Z., NOVIČ, M.: Multivariate data analysis in classification of vegetable oils characterized by the content of fatty acids. Chemometr. Intell. Lab. Syst., 75, 2005, 31-43.

BRODNJAK-VONČINA, D., DOBČNIK, D., NOVIČ, M., ZUPAN, J.: Determination of concentrations at hydrolytic potentiometric titrations with models made by artificial neural networks. Chemometr. Intell. Lab. Syst., 47, 1999, 79-88.

DAYHOF, J.: Neural Network Architectures, an Introduction. Van Nostrand Reinhold, New York, 1990, p.192.

DEMMING, S. N., MORGAN, S.L.: Experimental Design: a Chemometric Approach. Elsevier, Amsterdam, 1987.

HECHT-NIELSEN, R.: Counterpropagation networks. Appl. Optics, 26, 1987, 4979-4984.

KOHONEN, T.: Self-Organization and Associative Memory. Springer-Verlag, Berlin, 1988.

JURS, P.C., ISENHOUR, T.L.: Chemical Application to Pattern Recognition. Wiley, New York, 1975.

LANKMAYR, E., MOCAK, J., ŠNUDERL, K., BALLA, B., WENZL, T., BANDOINENE, D., GFRERER, M., WAGNER, S.: Chemometrical classification of pumpkin seed oils using UV-Vis, NIR and FTIR spectra. J. Biochem. Biophys. Methods, 61, 2004, 95-106.

MASSART, D.L., VANDEGINSTE, B.G.M., BUYDENS, L.M.C., DE JONG, S. LEWI, P.J., VERBEKE, J.S.: Handbook of Chemometrics and Qualimetrics: Part A. Elsevier, Amsterdam, 1997.

NILSSON, N.J.: Linear Learning Machines. Mc.Graw-Hill, NewYork, 1965

SHARAF, M.A., ILLMAN, D. L., KOWALSKI, B.R.: Chemometrics. Wiley, New York, 1986.

ŠNUDERL, K., NETRIOVÁ, J., MOCAK, J., LEHOTAY, J., BRODNJAK-VONČINA, D.: Chemometric analysis of biochemical laboratory data of oncology patients after morphine treatment. Sci. Pap. Univ. Pardubice. Ser. A, Fac. Chem. Technol., 11, 2005, 315-330.

ŠNUDERL, K., SIMONIČ, M., MOCAK, J., BRODNJAK-VONČINA, D.: Multivariate data analysis of natural mineral waters. Acta Chim. Slov., 54, 2007, 33-39.

TEACH/ME, SDL - Software Development Lohninger. Teach/Me DataLab 2.002 © 1999 Springer, Berlin, Developed by H. Lohninger and the Teach/Me people.

VAN DER WATERBEEMD, H.: Chemometric Methods in molecular Design. Methods and Principles in Medical Chemistry. VCH, Weinheim, 1995.

VARMUZA, K.: Pattern Recognition in Chemistry. Springer, Berlin, 1980.

VONČINA, E., BRODNJAK-VONČINA, D., SOVIČ, N., NOVIČ, M.: Chemometric characterisation of the quality of ground waters from different wells in Slovenia. Acta Chim. Slov., 54, 2007, 119-125.

WILKINS C.L, Computer- Enhanced Analytical Spectroscopy. The Royal society of chemistry, Cambridge, 1990.

ZUPAN, J.: Algoritms for Chemists. Wiley , Chichestert, 1989.

ZUPAN, J.: GASTEIGER, J.: Neural Networks for Chemists. VCH, Weinheim, 1993.

WILLET, P.: Similarity and Clustering in Chemical Information. Wiley, New York, 1987.