Applied Chemometrics: From Chemical Data to Relevant Information

K. Varmuza

Laboratory for Chemometrics
Vienna University of Technology, Getreidemarkt 9/160, A-1060 Vienna, Austria

kvarmuza@email.tuwien.ac.at, http://www.lcm.tuwien.ac.at

Manuscript for plenary lecture 3 March 2000

CHEM 1
1st Conference on Chemistry
Cairo University - Chemistry Department
6 - 9 March 2000, Cairo, Egypt

Version: 2 November 2001

Applied Chemometrics: From Chemical Data to Relevant Information

Kurt VARMUZA

Vienna University of Technology, Laboratory for Chemometrics
Institute of Food Chemistry, Getreidemarkt 9/160, A-1060 Vienna, Austria
Email: kvarmuza@email.tuwien.ac.at

The basic principles of multivariate data analysis in chemometrics are explained. The most used methods based on linear latent variables are discussed (principal component analysis, linear discriminant analysis) and demonstrated by examples from analytical chemistry and spectroscopy.

Introduction

Chemometrics has been defined as *„The chemical discipline that uses mathematical and statistical methods to design or select optimal procedures and experiments, and to provide maximum chemical information by analyzing chemical data".*[1,2] The most prominent part of chemometrics[3-13] is data interpretation by multivariate methods. Chemometric methods are often applied in situations when no sufficient theory is available for describing or solving problems. Typical for problems of this type is the use of many variables to describe a system; furthermore often only hidden relationships exist between the available data and the desired information and the aim of chemometrics is to find out some of these relationships (Figure 1).

Examples of such widespread problems in chemistry are: recognition of the chemical structure from spectral data (spectral classification), quantitative analyses of substances in complex mixtures (multivariate calibration), determination of the origin of samples (cluster analysis and classification), and prediction of properties or activities of chemical compounds or technological materials (quantitative structure-activity or structure-property relationships).

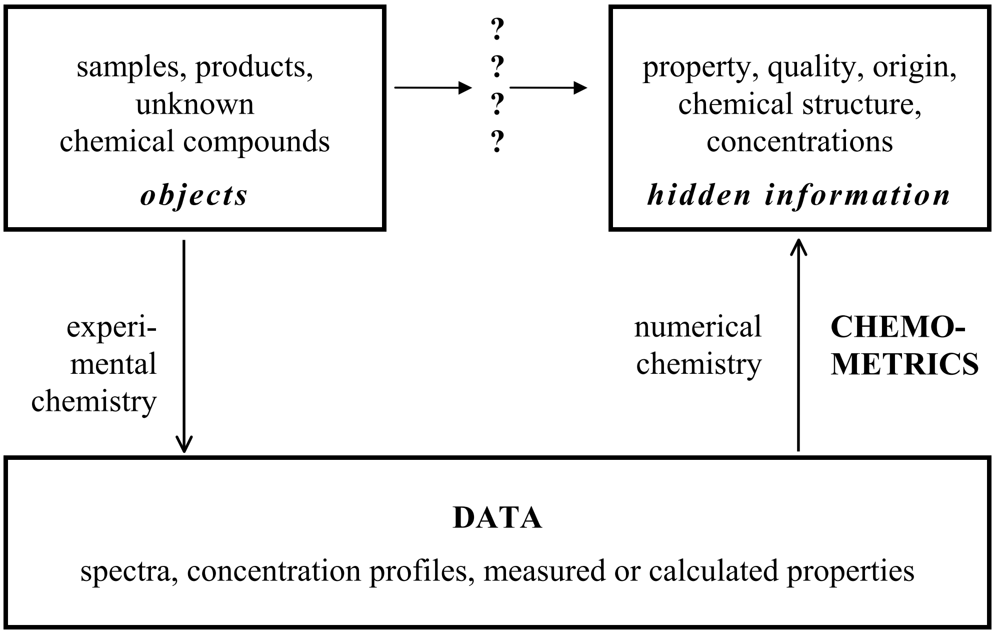

samples, products, unknown chemical compounds
objects
? ? ? ?
property, quality, origin, chemical structure, concentrations
hidden information
experi- mental chemistry
numerical chemistry
CHEMO- METRICS
DATA
spectra, concentration profiles, measured or calculated properties

Figure 1. From objects via experimental data and chemometrics to hidden information.

Chemometrics provides powerful methods to reduce the large amount of data which is produced easily by automated instruments such as chromatographs coupled to a spectrometer. Another measure for the huge amount of chemical data available today is the number of registered chemical substances by the Chemical Abstract Service which reached 22.7 million at the begin of February 2000; the increase per day is more than 4000 new compounds. Spectroscopic libraries today contain up to some 100 000 entries.

The typical chemometric strategy (Figure 2) is data-driven and consists of the following steps. (a) Collection of data. (b) Generation of a mathematical model which is usually based on multivariate statistics or neural networks. (c) Interpretation of the model parameters in terms of the underlying chemistry. (d) Application of the model to new cases, or often the search for a better model or for more appropriate variables. During this process the possibility must always be carefully considered that a significant relationship does not really exist in the given data or cannot be extracted by the applied methods. The data-driven philosophy in chemometrics avoids prejudices to some extent but on the other hand it includes the danger of finding artifact correlations. Consequently, results

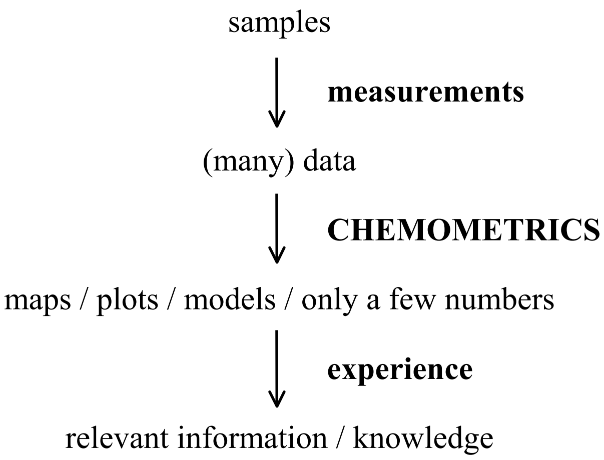

samples
measurements
(many) data
CHEMOMETRICS
maps / plots / models / only a few numbers
experience
relevant information / knowledge

Figure 2. Typical strategy in chemometrics.

from chemometric methods must not be over-interpreted and it should always be tried to explain the resulting model parameters in terms of chemistry.

Multivariate data

A set of multivariate data describes objects and their features. An object may be for instance a sample, or a spectrum, or a chemical structure. Objects are characterized by a pre-defined set of features. A feature is a numerical variable that describes an aspect of the objects; typical features are concentrations of selected substances, or intensities of spectral signals. The fruitful application of statistical methods requires a reasonable number of objects and features; typical for chemical problems are 20 to 1000 objects and 3 to 500 features. Such data are best described by an $n.p$ matrix $\boldsymbol{X}$, containing a row for each of the n objects, and a column for each of the p features.

Each feature can be considered as a coordinate of a point; each object then corresponds to a point in a p-dimensional feature space. The fundamental hypothesis for multivariate data interpretation is the existence of relationships between the locations or the distances of points (objects) and relevant properties. A *scatter plot* is a two-dimensional representation of the feature space in which for instance each point corresponds to an object (Figure 3). If two features are selected as plot coordinates a *feature plot* is obtained; such a plot, however, utilizes only a small part of the information in the data. The essential concept of multivariate data analysis is the use of so-called *latent variables* as plot coordinates. A latent variable is a mathematical function of all features and therefore may contain much more information than a pair of features. The goal of many chemometric methods is to find a mathematical function or a more general algorithm to define appropriate latent variables. The different methods for calculating latent variables (also called *components*) can be grouped into linear and non linear methods or can be described by their particular criterion that is optimized. The guiding principle is a representation of the *p*-dimensional multivariate data by a minimum number of latent variables.

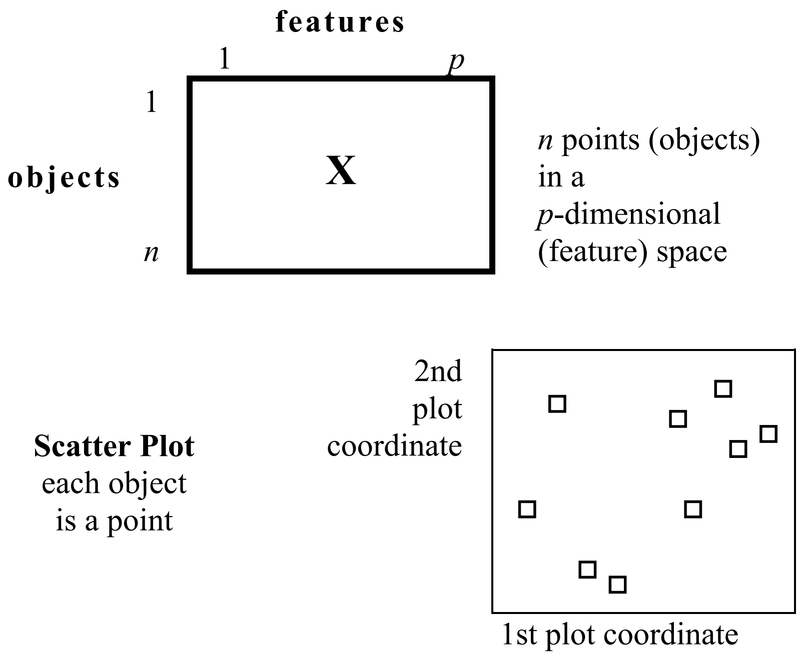


Figure 3. Multivariate data and latent variables. In a simple feature plot two selected features are used as plot coordinates. In typical multivariate methods (PCA, PLS) latent variables are used as plot coordinates. Many methods use latent variables that are linear functions of all features; this corresponds to a projection onto a straight line.

Projection of the feature space

A particular direction that defines a linear latent variable in a p-dimensional feature space is described by a vector $\boldsymbol{b}$ ($b_1, b_2, \ldots b_p$) which is usually scaled to length one. The value of the corresponding latent variable u for an object $\boldsymbol{x}$ ($x_1, x_2, \ldots x_p$) is obtained by projecting the object point onto a straight line which is defined by the direction $\boldsymbol{b}$ (Figure 4). Mathematically this is a *linear combination* of the features x_j of the object and the vector components b_j; an equivalent notation is the scalar product of the vectors $\boldsymbol{b}^T$and $\boldsymbol{x}$.

$$u = \boldsymbol{b}^T\boldsymbol{x} = b_1 x_1 + b_2 x_2 + \ldots + b_p x_p \qquad (1)$$

The value of a latent variable is called a *score*; scores are often used as plot coordinates. The vector components b_j are called *loadings*; they define the direction of the latent variable in the feature space and they describe the contributions of the individual features to the scores. Usually two orthogonal directions $\boldsymbol{b_1}$ and $\boldsymbol{b_2}$ are used as projection axes (the product $\boldsymbol{b_1.b_2}$ becomes zero) to define a projection plane.

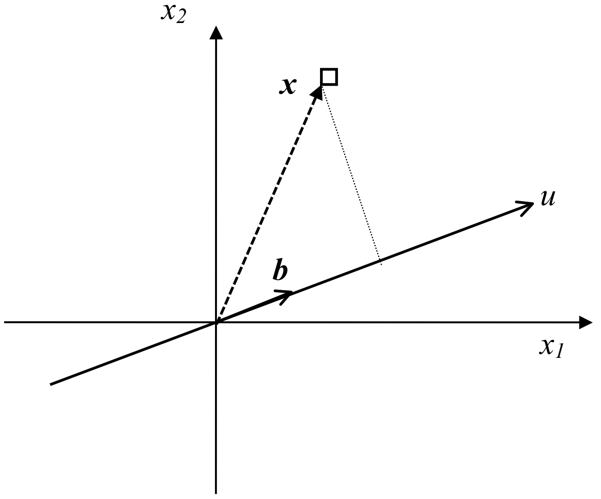


Figure 4. Projection of an object vector $\boldsymbol{x}$ onto a straight line which defines a latent variable by vector $\boldsymbol{b}$. The value (score) of the latent variable is u. The number of features, p, is two in this scheme but is in chemical applications usually between 5 and 200.

The vectors $\boldsymbol{b_1}$ and $\boldsymbol{b_2}$ can be arranged in a matrix $\boldsymbol{B}$. Scores $\boldsymbol{U}$ for objects $\boldsymbol{X}$ are calculated by a matrix multiplication.

$$\boldsymbol{U} \quad = \quad \boldsymbol{X} \, . \, \boldsymbol{B} \tag{2}$$

Two fundamental types of plots can be generated (Figure 5). In a *score plot* each point corresponds to an object; the coordinates are given by the scores. The distances between objects in the score plot are approximations of the distances in the multivariate feature space; groups (clusters) of similar objects can be detected visually. In a *loading plot* each point corresponds to a feature; the coordinates are given by the loadings of the features for the same axes as used in the corresponding score plot. The loading plot indicates the similarities and correlations between features. Furthermore this plot makes evident which features are responsible for the relative positions of the objects in the score plot. Features with small loadings are located near the origin; they have - on the average - only little influence on the data structure. A feature with a high loading for a latent variable causes that objects are placed in the corresponding region of the score plot if this feature has a large value.

The mathematical criteria for calculating appropriate directions of latent variables are characteristic for the different methods of multivariate data analysis; the methods which are most important for chemometrics are listed in Figure 6. The principal aims of multivariate data analysis methods are defined as following.

A. Exploratory data analysis. In a so-called unsupervised situation only the feature matrix $\boldsymbol{X}$ is available. The purpose of data interpretation may be a search for groups of similar objects (cluster analysis), or a search for outliers (objects that have no similar ones in the data set), or a search for relevant features. The mostly used techniques are principal component analysis (PCA), cluster analysis by dendrograms, and Kohonen maps. Applications in chemistry are for instance: evaluation of data tables obtained by automated analytical instruments, search for spectra-structure-relationships or structure-property-relationships, and cluster analysis of samples or chemical structures.

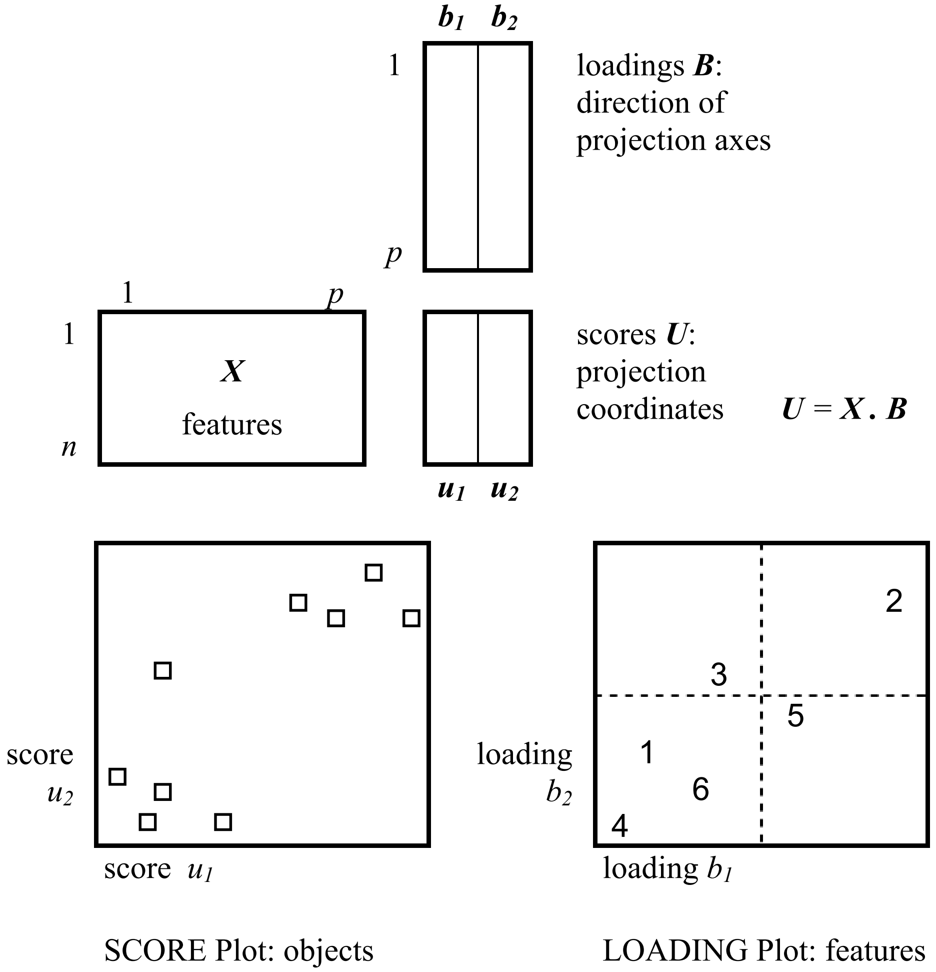
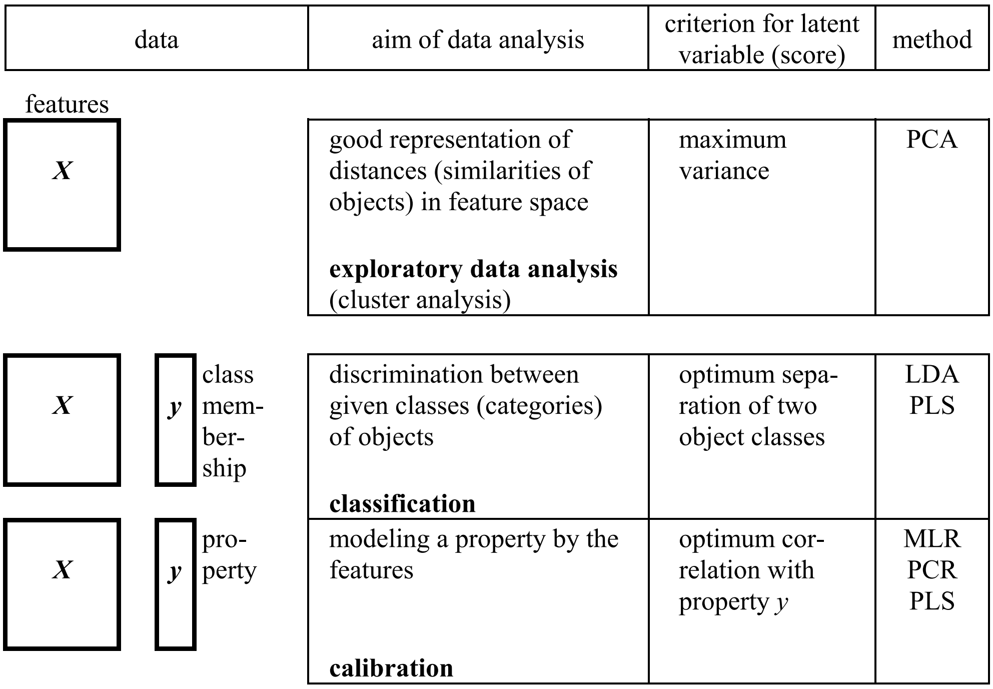


Figure 5. Score plot and loading plot. The score plot in this demo example indicates two clusters of objects and one outlier. From the loading plot follows that feature with number 2 is characteristic for objects that are located in the right upper corner of the score plot; features 1, 4, and 6 are characteristic for objects in the left lower corner; features 3 and 5 are near the origin of the loading plot and therefore have only little influence.

B. Classification. Besides a feature matrix $\boldsymbol{X}$ also a $\boldsymbol{y}$-vector is given that defines the class memberships of the objects. A training set containing objects with known class memberships is used to develop a classifier. A test set containing objects not present in the training set and also with known class memberships serves to evaluate the performance of the classifier. The most used techniques for multivariate classification are linear discriminant analysis (LDA), class modeling (for instance SIMCA), *k*-nearest-neighbor classification (KNN), and artificial neural networks (ANN). Applications in chemistry are for example the recognition of compound classes from spectral data, and the determination of the origin of samples.

C. Multivariate calibration. This method is the most frequently and routinely used multivariate technique in chemical laboratories. Aim is the development of a mathematical model that describes the relationship between a set of x-variables and one or several y-variables. The traditional technique for this purpose is multiple linear regression (MLR); it has been complemented by more robust and more powerful methods such as principal component regression (PCR), partial least squares regression (PLS), or artificial neural networks. The main applications in chemistry are infrared spectroscopy, evaluation of multi-sensor data, and modeling of structure-property relationships.



data	aim of data analysis	criterion for latent variable (score)	method
features ***X***	good representation of distances (similarities of objects) in feature space **exploratory data analysis** (cluster analysis)	maximum variance	PCA
X ***y*** class membership	discrimination between given classes (categories) of objects **classification**	optimum separation of two object classes	LDA PLS
X ***y*** property	modeling a property by the features **calibration**	optimum correlation with property *y*	MLR PCR PLS

Figure 6. Mathematical criteria for latent variables with regard to the purpose of data analysis. PCA, principal component analysis; LDA, linear discriminant analysis; PLS, partial least squares regression; MLR, multiple linear regression; PCR, principal component regression; y, dependent variable.

Principal component analysis (PCA)

The most frequently used method with multivariate data is principal component analysis (PCA). The latent variable which best describes the relative distances between the objects is given by the *direction with maximum variance*. This direction is called the *first principal component* (PC1). The second principal component (PC2) is orthogonal to PC1 and again has the maximum possible variance. Further principal components can be determined by continuing this concept. For data with n objects and p features the maximum number of principal components is given by the minimum of p and n. Determination of all principal components corresponds to a rotation of the p-dimensional coordinate system. In many cases only PC1 and PC2 are used to define a projection plane for a visual inspection of the multivariate data. Note that the correlation coefficient between the scores of any two principal components is zero; PCA is therefore often used to transform data which exhibit highly correlating features into a set of uncorrelated new variables (the PC scores).

In Figure 7 a two-dimensional example with six objects demonstrates the latent variable with maximum variance (PC1) in comparison with another latent variable that separates two given classes of objects optimally.

The relevance of a principal component is measured by the variance of the corresponding scores expressed in percent of the total variance (calculated as the sum of the variances of all features). If the sum of the variances of the two scores which are used as projection coordinates is high (for instance above 70% of the total variance) then the scatter plot represents a good two-dimensional visualization of the p-dimensional data structure.

The matrix $\boldsymbol{B}$ which contains the principal component vectors has to be calculated by iterative procedures. Most used methods in chemometrics are *Singular Value Decomposition* (SVD)[14,15] and the *NIPALS*-algorithm;[16] the traditional reference method is the calculation of eigenvectors from the covariance matrix by *Jacobi rotation.*[15,17]

An application of the PCA to chemical analytical air pollution data is shown in Figure 8. The original data consist of the concentrations of 20 polycyclic aromatic hydrocarbons measured in 70 aerosol samples (50 from city Vienna, 20 from city Linz in Austria).[18]

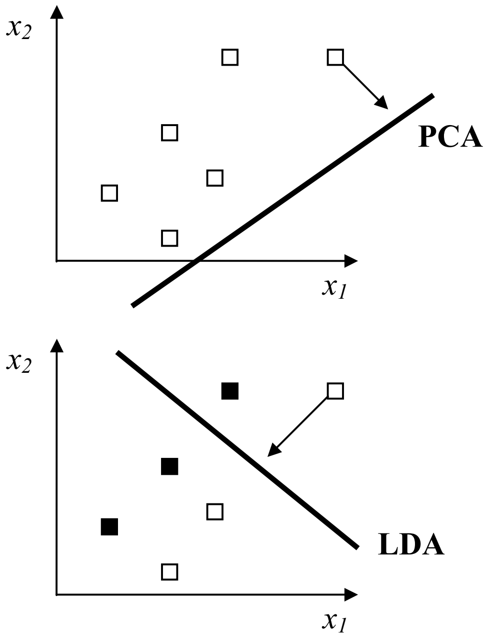


Figure 7. Different aims of defining latent variables. PCA, principal component analysis, preserves the distances between the objects by a latent variable which possesses maximum variance of the scores. LDA, linear discriminant analysis, results in a latent variable which allows maximum separation between two given classes.

The measured compounds range from anthracene to coronene; the concentrations are given in percent of the sum of all 20. The scatter plot with PC1 and PC2 shows a clear separation of the samples from Vienna and Linz. Thus the different concentration profiles of polycyclic aromatic hydrocarbons in the two cities are demonstrated: Linz has heavy chemical and iron industry while in Vienna the pollution is mainly caused by traffic and domestic heating during the cold season. The PCA plot using only the 50 Vienna samples (Figure 9) does not show clearly separated clusters, however, demonstrates the influence of domestic heating. The diameter of the circles in the plot is drawn proportional to the average temperature of the sampling days. The first principal component (horizontal axis) mainly describes the factor of domestic heating: at the right hand side all samples from summer are located while at the left hand side typical winter samples are placed.

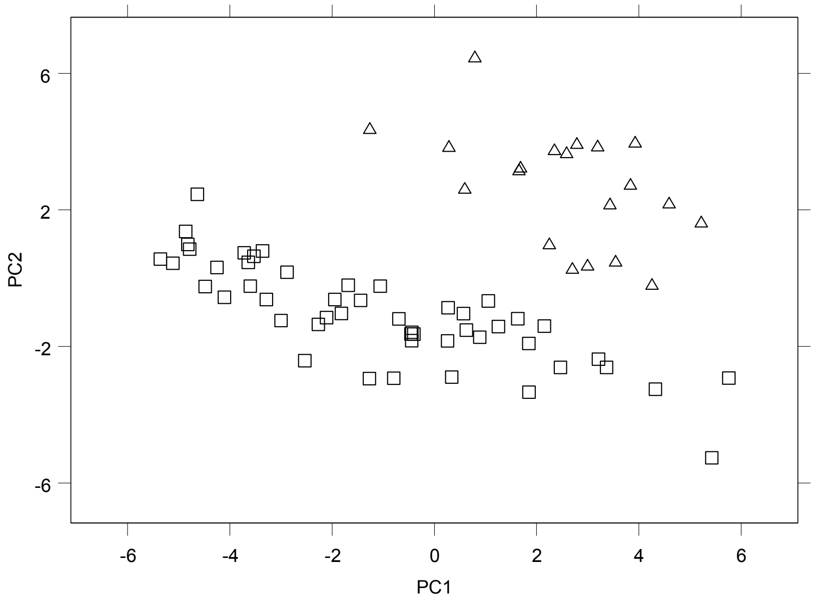


Figure 8. PCA scatter plot from aerosol data. Objects: 50 samples from city Vienna (□) and 20 samples from city Linz (Δ), Austria. Features: concentrations (% sum) of 20 polycyclic hydrocarbons, autoscaled. Variances of PC1 and PC2 are 45.9 % and 26.0 % of total variance, respectively. The samples from the two cities are clearly separated showing the different concentration profiles.

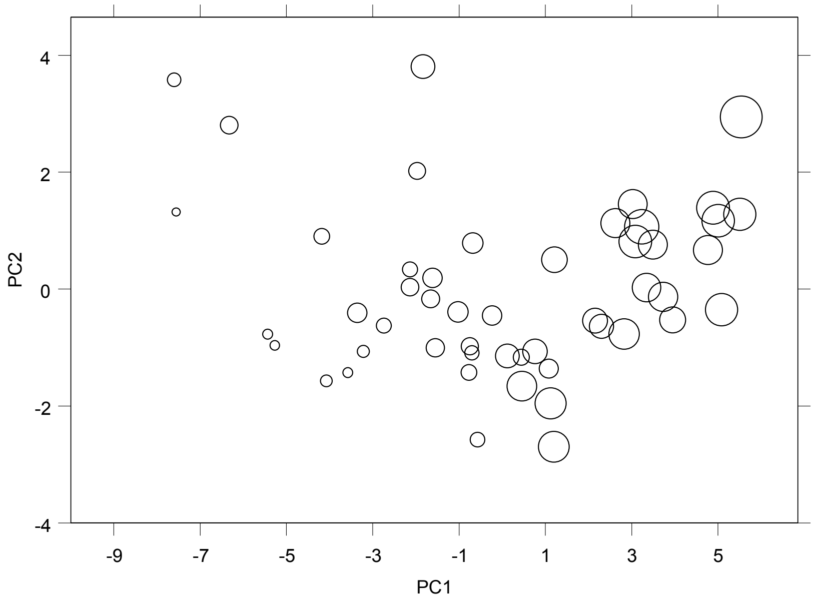


Figure 9. PCA scatter plot from aerosol data. Objects: 50 samples from city Vienna. Features: concentrations (% sum) of 20 polycyclic aromatic hydrocarbons, autoscaled. Variances of PC1 and PC2 are 60.0 % and 10.6 % of total variance, respectively. The diameter of the cycles is proportional to the average temperature at the sampling day (between -6.5 and 28.3 Celsius).

Classification by linear discriminant analysis (LDA)

Discriminant analysis has the aim to assign objects to one of several pre-determined classes; only the two-class problem will be treated here briefly. *Linear discriminant analysis* (LDA) uses a latent variable (*discriminant variable*) in the feature space that maximally separates two classes of objects (Figure 7). A widely used criterion for class separation is the *t*-value as known from the statistical t-test. The mathematical tool for calculating the direction $\boldsymbol{b}_{LDA}$ which has the maximum possible *t*-value is multiple linear regression. The *y*-variable is binary in this case and indicates the membership of an object to one of the two mutually exclusive classes (for instance with values +1 and -1). The discriminant vector $\boldsymbol{b}_{LDA}$ is calculated from the data of a training set containing objects from both classes (data matrices $\boldsymbol{X}_A$ and $\boldsymbol{X}_B$) by

$$\boldsymbol{b}_{LDA} = \boldsymbol{C}^{-1} (\boldsymbol{m}_A - \boldsymbol{m}_B) \qquad (3)$$

$$\boldsymbol{C} = [(n_A - 1)\, \boldsymbol{C}_A + (n_B - 1)\, \boldsymbol{C}_B] / (n_A + n_B - 2) \qquad (4)$$

with n_A, n_B being the number of objects in class A and B, respectively; $\boldsymbol{m}_A$, $\boldsymbol{m}_B$ being the mean vectors and $\boldsymbol{C}_A$, $\boldsymbol{C}_B$ being the covariance matrices of class A and B, respectively.[8,14] The discriminant score u is calculated by

$$u = \boldsymbol{x} \,.\, \boldsymbol{b}_{LDA} \qquad (5)$$

with $\boldsymbol{x}$ being the feature vector of the classified object. If the values +1 and -1 are used to denote class 1 and class 2, respectively, then the object is assigned to class 1 if $u > 0$ and to class 2 otherwise.

Calculation of $\boldsymbol{b}_{LDA}$ requires the inversion of the pooled covariance matrix $\boldsymbol{C}$ which is impossible if $\boldsymbol{X}$ contains highly correlated features or if $p > n$. This problems can be overcome by a preceding principal component analysis. Instead of correlating features a set of principal component scores is used as independent variables in the regression equation (*principal component regression*, PCR). Remember that PCA scores are uncorrelated by definition. An alternative method would be *partial least squares regression* (PLS)[6] in which a latent variable is determined so that a maximum covariance to the dependent variable *y* is obtained.

Multivariate classification is a standard method in many scientific fields ranging from food chemistry, spectroscopy, botanical taxonomy to archaeometry. The example given here is from mass spectrometry. Measured mass spectra are usually evaluated by a spectral library search resulting in a hitlist which contains the most similar spectra from a spectral library. If the unknown is a member of the library then often an unambiguous identification is possible. However, some substance classes exhibit very similar mass spectra and simple spectra similarity criteria cannot distinguish between them. For instance fatty acid ethyl esters have very similar mass spectra as the corresponding α-methyl-substituted methyl esters. LDA has been applied to discriminate between these two substance classes. From a mass spectral library 34 ethyl esters (class 1) and 49 α-methyl-substituted methyl esters (class 2) have been selected and used as a training set. The spectra were transformed into a set of 14 features by modulo-14 summation.[19] Next step was a PCA to obtain uncorrelated variables and then LDA was applied. Figure 10 shows a scatter plot with the discriminant variable as abscissa and the first principal component as ordinate. The two classes are completely separated; note that a PCA plot would not be able to separate the classes sufficiently. Two additional compounds from the library - not used for training - have been projected into the plot; both are correctly classified. In this successful example it was possible to transform mass spectra into appropriate features and then to apply an automatic classification procedure for a discrimination of classes of compounds that could not be distinguished by library search or by human interpretation of the spectra. Based on the same principles a set of spectral classifiers has been developed for automatic recognition of some substructures from low resolution mass spectra[20,21] and from infrared spectra.[22] Such classification results can be used together with the molecular formula of the unknown by an isomer generator software[23]. This approach for a systematic structure elucidation is capable to produce exhaustive sets of all isomeric compounds fulfilling the classification results.[19-21,24,25]

Summary and outlook

The important task in chemometrics is the extraction of relevant information from chemically oriented data. For this purpose tools from mathematics, statistics and computer science are used and especially methods for multivariate data analysis are most powerful. Recently also artificial neural networks and genetic algorithms found great interest but could justify the expected performance only in a few examples. Standard methods of

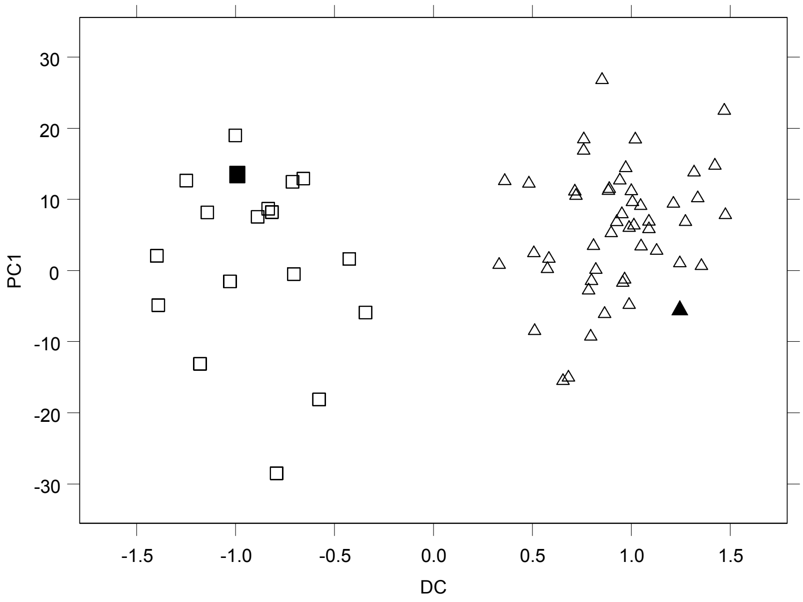


Figure 10. Separation of two types of fatty acid esters exhibiting very similar mass spectra by linear discriminant analysis. □, ethyl esters; Δ, α-methyl, methyl esters; filled symbols are classified test spectra from $C_{17}H_{35}COOC_2H_5$ and $C_{16}H_{33}CH(CH_3)COOCH_3$ (both correctly assigned). DC, discriminant variable; PC1, first principal component.

multivariate data analysis remain those procedures which are based on linear latent variables, such as PCA, PLS, LDA and MLR; nonlinear methods are certainly necessary if linear methods fail.

The need for an automatic extraction of relevant information from complex data will remain a prominent task in future application areas such as multivariate calibration, improvement of chromatographic separation, interpretation of spectra, classification of materials, investigation of relationships between chemical structures and properties, as well as for modeling and optimization of syntheses and for process monitoring.

Chemometrics is already well established in analytical chemistry, drug design and process control. In other parts of chemistry a deficit of information about the multivariate approach of data analysis is still recognizable. A number of good commercial software products for multivariate data analysis is available today and provide powerful tools not only for routine problems but also for searching "not yet detected secrets" behind chemical data.

(1) N.N. *Chemom. Intell. Lab. Syst.* **1986**, *1*.

(2) N.N. *J. Chemometrics* **1986**, *1*.

(3) Adams, M. J. *Chemometrics in analytical spectroscopy*; The Royal Society of Chemistry: Cambridge, 1995.

(4) Beebe, K. R.; Pell, R. J.; Seasholtz, M. B. *Chemometrics: A practical guide*; John Wiley & Sons: New York, 1998.

(5) Kramer, R. *Chemometric techniques for quantitative analysis*; Marcel Dekker: New York, 1998.

(6) Martens, H.; Naes, T. *Multivariate calibration*; John Wiley & Sons: Chichester, 1989.

(7) Massart, D. L.; Vandeginste, B. G. M.; Deming, S. N.; Michotte, Y.; Kaufman, L. *Chemometrics: A textbook*; Elsevier: Amsterdam, 1988.

(8) Massart, D. L.; Vandeginste, B. G. M.; Buydens, L. C. M.; de Jong, S.; Lewi, P. J.; Smeyers-Verbeke, J. *Handbook of chemometrics and qualimetrics: Part A*; Elsevier: Amsterdam, 1997.

(9) Vandeginste, B. G. M.; Massart, D. L.; Buydens, L. C. M.; De Jong, S.; Smeyers-Verbeke, J. *Handbook of chemometrics and qualimetrics: Part B*; Elsevier: Amsterdam, 1998.

(10) Varmuza, K. *Pattern recognition in chemistry*; Springer-Verlag: Berlin, 1980.

(11) Varmuza, K. Chemometrics: Multivariate view on chemical problems. In *The encyclopedia of computational chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, I. H. F., Schreiner, P. R., Eds.; Wiley & Sons: Chichester, 1998; Vol. *1*, pp 346-366.

(12) Zupan, J. *Algorithms for chemists*; John Wiley & Sons: Chichester, 1989.

(13) Zupan, J.; Gasteiger, J. *Neural networks in chemistry and drug design*; Wiley-VCH: Weinheim, 1999.

(14) Healy, M. J. R. *Matrices for statistics*; Clarendon Press: Oxford, 1995.

(15) Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. *Numerical recipes*; Cambridge University Press: Cambridge, 1986.

(16) Geladi, P. Notes on the history and nature of partial least squares (PLS) modelling. *J. Chemometrics* **1988**, *2*, 231-246.

(17) Nash, J. C. *Compact numerical methods for computer: Linear algebra and function minimisation*; Adam Hilger: Bristol, 1990.

(18) Jaklin, J.; Krenmayr, P.; Varmuza, K. Polycyclic aromatic compounds in the atmosphere of Linz (Austria). *Fresenius Z. Anal. Chem.* **1988**, *331*, 479-485.

(19) Varmuza, K.; Werther, W. Systematic structure elucidation of organic compounds based on mass spectra classification and isomer generation. In *Advances in mass spectrometry*; Karjalainen, E. J., Hesso, A. E., Jalonen, J. E., Karjalainen, U. P., Eds.; Elsevier: Amsterdam, 1998; Vol. *14*, pp 611-626.

(20) Varmuza, K.; Werther, W. Mass spectral classifiers for supporting systematic structure elucidation. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 323-333.

(21) Varmuza, K.; Penchev, P.; Stancl, F.; Werther, W. Systematic structure elucidation of organic compounds by mass spectra classification. *J. Mol. Struct.* **1997**, *408/409*, 91-96.

(22) Penchev, P. N.; Andreev, G. N.; Varmuza, K. Automatic classification of infrared spectra using a set of improved expert-based features. *Anal. chim. acta* **1999**, *388*, 145-159.

(23) *MOLGEN: Isomer Generator Software*; vers. 3.1, 1998; University of Bayreuth, Institute for Mathematics II: D-95440 Bayreuth, Germany.

(24) Varmuza, K.; Werther, W.; Stancl, F.; Kerber, A.; Laue, R. Computer-assisted structure elucidation of organic compounds, based on mass spectra classification and exhaustive isomer generation. In *Software development in chemistry*; Gasteiger, J., Ed.; Gesellschaft Deutscher Chemiker: Frankfurt am Main, 1996; Vol. *10*, pp 303-314.

(25) Gray, N. A. B. *Computer-assisted structure elucidation*; John Wiley: New York, 1986.