2

Critical Aspects of Supervised Pattern Recognition Methods for Interpreting Compositional Data

A. Gustavo González
Department of Analytical Chemistry, University of Seville, Seville
Spain

1. Introduction

A lot of multivariate data sets of interest to scientists are called compositional or "closed" data sets, and consists essentially of relative proportions. A recent search on the web by entering "chemical compositional data", led to more than 2,730,000 results within different fields and disciplines, but specially, agricultural and food sciences (August 2011 using Google searcher). The driving causes for the composition of foods lie on four factors (González, 2007): Genetic factor (genetic control and manipulation of original specimens), Environmental factor (soil, climate and symbiotic and parasite organisms), Agricultural factor (cultures, crop, irrigation, fertilizers and harvest practices) and Processing factor (post-harvest manipulation, preservation, additives, conversion to another food preparation and finished product). But the influences of these factors are hidden behind the analytical measurements and only can be inferred and uncover by using suitable chemometric procedures.

Chemometrics is a term originally coined by Svante Wold and could be defined as "The art of extracting chemically relevant information from data produced in chemical experiments" (Wold, 1995). Besides, chemometrics can be also defined as the application of mathematical, statistical, graphical or symbolic methods to maximize the chemical information which can be extracted from data (Rock, 1985). Within the jargon of chemometrics some other terms are very common; among them, multivariate analysis and pattern recognition are often used. Chemometricians use the term Multivariate Analysis in reference to the different approaches (mathematical, statistical, graphical...) when considering samples featured by multiple descriptors simultaneously. Pattern recognition is a branch of the Artificial Intelligence that seeks to identify similarities and regularities present in the data in order to attain natural classification and grouping. When applied to chemical compositional data, pattern recognition methods can be seen as multivariate analysis applied to chemical measurements to find classification rules for discrimination issues. Depending on our knowledge about the category or class membership of the data set, two approaches can be applied: Supervised or unsupervised learning (pattern recognition).

Supervised learning methods develop rules for the classification of unknown samples on the basis of a group of samples with known categories (known set). Unsupervised learning methods instead do not assume any known set and the goal is to find clusters of objects which may be assigned to classes. There is hardly any quantitative analytical method that does not make use of chemometrics. Even if one confines the scope to supervised learning pattern recognition, these chemometric techniques are increasingly being applied to of compositional data for classification and authentication purposes. The discrimination of the geographical origin, the assignation to Denominations of Origin, the classification of varieties and cultivars are typical issues in agriculture and food science.

There are analytical techniques such as the based on sensors arrays (electronic nose and electronic tongue) that cannot be imaginable without the help of chemometrics. Thus, one could celebrate the triumph of chemometrics...so what is wrong? (Pretsch & Wilkin, 2006). The answer could be supported by a very well-known quotation attributed to the british politician D'Israeli (Defernez & Kemsley, 1997) but modified by us as follows: "There are three kinds of lies: Lies, damned lies and chemometrics". Here we have changed the original word "statistics" into "chemometrics" in order to point out the suspicion towards some chemometric techniques even within the scientific community. There is no doubt that unwarranted reliance on poorly understood chemometric methods is responsible for such as suspicion.

By the way, chemometric techniques are applied using statistical/chemometric software packages that work as black boxes for final users. Sometimes the blindly use of "intelligent problem solvers" or similar wizards with a lot of hidden options as default may lead to misleading results. Accordingly, it should be advisable to use software packages with full control on parameters and options, and obviously, this software should be used by a chemometric *connaisseur*.

There are special statistical methods intended for closed data such as compositional ones (Aitchison, 2003; Egozcue et al., 2003; Varmuza & Filzmoser, 2009), but a detailed description of these methods may be outside the scope of this chapter.

2. About the data set

Supervised learning techniques are used either for developing classification rules which accurately predict the classification of unknown patterns or samples (Kryger, 1981) or for finding calibration relationships between one set of measurements which are easy or cheap to acquire, and other measurements which are expensive or labour intensive, in order to predict these later (Naes et al., 2004). The simplest calibration problem consists of predicting a single response (y-variable) from a known predictor (x-variable) and can be solved by using ordinary linear regression (OLR). When fitting a single response from several predictive variables, multiple linear regression (MLR) may be used; but for the sake of avoiding multicollinearity drawbacks, some other procedures such as principal component regression (PCR) or partial least squares regression (PLS) are a good choice. If faced to several response variables, multivariate partial least squares (PLS2) techniques have to be used. These procedures are common in multivariate calibration (analyte concentrations from NIR data) or linear free energy relationships (pKa values from molecular descriptors). However, in some instances like quantitative structure-activity relationships (QSAR), non linear strategies are needed, such as quadratic partial least squares regression (QPLS), or regression procedures based on artificial neural networks or support vector machines. All these calibration procedures have to be suitably validated by using validation procedures based on the knowledge of class memberships of the objects, in a similar way as discussed below in this chapter that is devoted to supervised learning for classification.

Let us assume that a known set of samples is available, where the category or class membership of every sample is *a priori* known. Then a suitable planning of the data-acquisition process is needed. At this point, chemical experience, *savoir faire* and intuition are invaluable in order to decide which measurements should be made on the samples and which variables of these measurements are most likely to contain class information.

In the case of compositional data, a lot of analytical techniques can be chosen. Analytical procedures based on these techniques are then selected to be applied to the samples. Selected analytical methods have to be fully validated and with an estimation of their uncertainty (González & Herrador, 2007) and carried out in Quality Assurance conditions (equipment within specifications, qualified staff, and documentation written as Standard Operational Procedures...). Measurements should be carried out at least duplicate and according to a given experimental design to ensure randomization and avoid systematic trends.

Difficulties arise when the concentration of an element is below the detection limit (DL). It is often standard practice to report these data simply as '<DL' values. Such 'censoring' of data, however, can complicate all subsequent statistical analyses. The best method to use generally depends on the amount of data below the detection limit, the size of the data set, and the probability distribution of the measurements. When the number of '< DL' observations is small, replacing them with a constant is generally satisfactory (Clarke, 1998). The values that are commonly used to replace the '< DL' values are 0, DL, or DL/2. Distributional methods such as the marginal maximum likelihood estimation (Chung, 1993) or more robust techniques (Helsel, 1990) are often required when a large number of '< DL' observations are present.

After all measurements are done we can built the corresponding data table or data matrix. A sample, object or pattern is described by a set of "p" variables, features or descriptors. So, all descriptors of one pattern form a 'pattern vector' and accordingly, a given pattern "i" can be seen as a vector $\vec{x}_i$ whose components are $x_{i1}, x_{i2}, \dots x_{ij}, \dots x_{ip}$ in the vectorial space defined by the features. In matricial form, pattern vectors are row vectors. If we have n patterns, we can build a data matrix $X_{n \times p}$ by assembling the different row pattern vectors. A change in perspective is also possible: A given feature "j" can be seen as a column vector $\vec{x}_j$ with components $x_{1j}, x_{2j}, \dots x_{ij}, \dots x_{nj}$ in the vectorial space defined by the patterns. We can also construct the data matrix by assembling the different feature column vectors. Accordingly, the data matrix can be considered as describing the patterns in terms of features or *vice versa*. This lead to two main classes of chemometric techniques called Q- and R-modes respectively. R-mode techniques are concerned with the relationships amongst the features of the experiment and examine the interplay between the columns of the data matrix. A starting point for R-mode procedures is the covariance matrix of mean centered variables $C = X^T X$ whose elements are given by $c_{jk} = \frac{1}{n-1}\sum_{l=1}^{n}\left(x_{lj} - \bar{x}_j\right)\left(x_{lk} - \bar{x}_k\right)$ where $\overline{\mathrm{x}}_{\mathrm{j}}$ and $\overline{\mathrm{x}}_{\mathrm{k}}$ are the mean of the observations on the j*th* and k*th* feature. If working with autoscaled data, the sample correlation matrix and the covariance matrix are identical. The element r_{jk} of correlation matrix R represents the cosine between each pair of column vectors and is given by $r_{jk} = \frac{c_{jk}}{\sqrt{c_{jj}c_{kk}}}$. The diagonal elements of R are always unity. The alternative viewpoint considers the relationships between patterns, the Q-mode technique. This way normally starts with a matrix of distances between the objects in the n-dimensional pattern space to study the clustering of samples. Typical metric measurements are Euclidean, Minkowski, Manhattan, Hamming, Tanimoto and Mahalanobis distances (Varmuza, 1980).

3. Inspect data matrix

Once the data matrix has been built, it should be fully examined in order to ensure the suitable application of Supervised Learning methodology. A typical undesirable issue is the existence of missing data. Holes in the data matrix must be avoided; however some measurements may not have been recorded or are impossible to obtain experimentally due to insufficient sample amounts or due to high costs. Besides, data can be missing due to various malfunctions of the instruments, or responses can be outside the instrument range.

As stated above, most chemometric techniques of data analysis do not allow for data gaps and thereof different methods have been applied for handling missing values in data matrix. Aside from the extreme situations of casewise deletion or mean substitution, the use of iterative algorithms (IA) is a promising tool. Each iteration consists of two steps. The first step performs estimation of model parameters just as if there were no missing data. The second step finds the conditional expectation of the missing elements given the observed data and current estimated parameters. The detailed procedure depends on the particular application. The typical IA used in Principal Component Analysis can be summarized as (Walczak & Massart, 2001):

1. Fill in missing elements with their initial estimates (expected values, calculated as the mean of the corresponding row's and column's means)
2. Perform singular value decomposition of the complete data set
3. Reconstruct X with the predefined number of factors
4. Replace the missing elements with the predicted values and go to step 2 until convergence.

Replacement of missing values or censored data with any value is always risky since this can substantially change the correlation in the data. It is possible to deal with both missing values and outliers simultaneously (Stanimirova et al., 2007). An excellent revision dealing with zeros and missing values in compositional data sets using non-parametric imputation has been performed by Martin-Fernández et al. (2003).

On the other hand, a number of chemometric procedures are based on the assumption of normality of features. Accordingly, features should be assessed for normality. The well-known Kolmogorov-Smirnov, Shapiro-Wilks and Lilliefors tests are often used (González, 2007), although the data about the skewness and kurtosis of the distribution are also of interest in order to consider parametric or non parametric descriptive statistics. Some simple presentations such as the Box-and-whisker plots help in the visual identification of outliers and other unusual characteristics of the data set. The box-and-whisker plot assorted with a numerical scale is a graphical representation of the five-number summary of the data set (Miller & Miller, 2005) where it is described by its extremes, its lower and upper quartiles and the median and gives at a glance the spread and the symmetry of the data set. Box-and-whisker plots may reveal suspicious patterns that should be tested for outliers. Abnormal data can road chemometric techniques leading to misleading conclusions, especially when outliers are present in the training set. Univariate outlier tests such as Dean and Dixon (1951) or Grubbs (1969) assays are not suitable. Instead, multivariate criteria for outlier detection are more advisable. The techniques based on the Mahalanobis distance (Gemperline & Boyer, 1995), and the hat matrix leverage (Hoaglin & Welsch, 1978) have been often used for decades. Hat matrix $H = X\left(X^T X\right)^{-1} X^T$ has diagonal values h_{ii} called leverage values. Patterns having leverage values higher than 2p/n are commonly considered outliers. However these methods are unreliable for multivariate outlier detection. Numerous methods for outlier detection have been based on the singular value decomposition or Principal Component Analysis (PCA) (Jollife, 2002). Soft Independent Modelling of Class Analogy (SIMCA) has been also applied to outlier detection (Mertens et al. , 1994). Once outliers have been deleted, researchers usually remove them from the data set, but outliers could be corrected before applying the definite mathematical procedures by using robust algorithms (Daszykowski et al., 2007). Robust methods give better results, specially some improved algorithms such as resampling by the half-means (RHM) and smallest half-volume (SHV) (Egan & Morgan , 1998).

Within the field of food authentication, the wrong conclusions are, however, mostly due to data sets that do not keep all aspects of the food characterisation (partial or skewed data set) or do merge data measured with different techniques (Aparicio & Aparicio-Ruiz, 2002). A classical example is the assessment of the geographical origin of Italian virgin olive oils by Artificial Neural Networks (ANN). The differences between oils were mainly due to the use of different chromatographic columns (packed columns against capillary columns) when quantifying the free fatty acid (FFA) profile of the oils from Southern and Northern Italy respectively (Zupan et al. , 1994). The neural network thus mostly learned to recognise the chromatographic columns.

4. Data pre-treatment

Data transformation (scaling) can be applied either for statistical dictates to optimise the analysis or based on chemical reasons. Raw compositional data are expressed in concentration units that can differ by orders of magnitude (e.g., percentage, ppm or ppb), the features with the largest absolute values are likely to dominate and influence the rule development and the classification process. Thus, for statistical needs, transformation of raw data can be applied to uniformize feature values. Autoscaling and column range scaling are the most common transformation (Sharaf et al., 1986). In the autoscaling or Z-transformation, raw data are transformed according to

$$x_{ij}^{'} = \frac{x_{ij} - \overline{x}_j}{s_j} \text{ with } s_j = \sqrt{\frac{\sum_{i=1}^{n}\left(x_{ij} - \overline{x}_j\right)^2}{n-1}} \tag{1}$$

and can be seen as a parametric scaling by column standardisation leading to $\overline{x}_j^{'} = 0$ and $s_j^{'} = 1$.

Column range scaling or minimax scaling involves the following transformation

$$x_{ij}^{'} = \frac{x_{ij} - \min_j\left(x_{ij}\right)}{\max_j\left(x_{ij}\right) - \min_j\left(x_{ij}\right)} \tag{2}$$

Now, the transformed data verify $0 \le x_{ij}^{'} \le 1$. Range scaling can be considered as an interpolation of data within the interval (0,1). It is a non-parametric scaling but sensitive to outliers. When the data contain outliers and the preprocessing is necessary, one should consider robust way of data preprocessing in the context of robust Soft Independent Modelling of Class Analogy (SIMCA) method (Daszykowski et al., 2007).

The second kind of transformations are done for chemical reasons and comprise the called "constant-row sum" and "normalization variable" (Johnson & Ehrlich, 2002). Dealing with compositional data, concentrations can vary widely due to dilution away from a source. In the case of contaminated sediment investigations, for example, concentrations may decrease exponentially away from the effluent pipe. However, if the relative proportions of individual analytes remain relatively constant, then we would infer a single source scenario coupled with dilution far away from the source. Thus, a transformation is needed to normalize concentration/dilution effects. Commonly this is done using a transformation to a fractional ratio or percent, where each concentration value is divided by the total concentration of the sample: Row profile or constant row-sum transformation because the sum of analyte concentrations in each sample (across rows) sums unity or 100%:

$$x_{ij}^{'} = \frac{x_{ij}}{\sum_{j=1}^{p} x_{ij}} \tag{3}$$

leading to $\sum_{j=1}^{p} x_{ij}^{'} = 1 \text{ (or } 100\%\text{)}$. An alternative is to normalize the data with respect to a single species or compound set as reference congener, the normalization variable. This transformation involves setting the value of the normalization feature to unity, and the values of all other features to some proportion of 1.0, such that their ratios with respect to the normalization feature remain the same in the original metric.

When n > p, the rank of data matrix is p (if all features are independent). Thus, autoscaling and minimax transformations do not change data dimensionality because these treatments do not induce any bound between features. Row profiles instead build a relationship between features scores (the constant-row sum) that decreases the data dimensionality by 1 and the rank of data matrix is then p-1. Accordingly, the patterns fall on a hypersurface in the feature space and it is advisable to remove one feature to avoid problems involving matrix inversion when the rank of the matrix is less than p.

As a final remark it should be realized that when using some Supervised Learning techniques like SIMCA, the scaling of the data set is carried out only over the samples belonging to the same class (separate scaling). This is due because the own fundamentals of the methodology and has a beneficial effect on the classification (Derde et al., 1982).

5. Feature selection and extraction

Irrelevant features are very expensive ones, because they contribute to the chemical information with noise only; but even more expensive may be simply wrong features. Accordingly, they should be eliminated in order to circumvent disturb in classification. In almost all chemical applications of pattern recognition the number of original raw features is too large and a reduction of the dimensionality is necessary. Features which are essential for classification purposes are often called intrinsic features. Thus, a common practise to avoid redundant information consists of computing the correlation matrix of features R. Pair of most correlated features can be either combined or one of them is deleted. Researchers should be aware that the number of independent descriptors or features, p, must be much smaller than that of patterns, n. Otherwise, we can build a classification rule that even separates randomly selected classes of the training set (Varmuza, 1980). This assumes that the set of features is linearly independent (actually, it is the basis of the vectorial space) and the number of features is the dimensionality of the vectorial space. Accordingly, the true dimensionality should be evaluated for instance from an eigenanalysis (PCA) of the data matrix and extract the proper number of factors which correspond to the true dimensionality (d) of the space. Most efficient criteria for extracting the proper number of underlying factors are based on the Malinowski indicator function (Malinowski, 2002) and the Wold's Cross-Validation procedure (Wold, 1978). For most classification methods, a ratio n/d > 3 is advised and > 10, desirable. However, PCA-based methods like SIMCA or Partial Least Squares Discriminant Analysis (PLS-DA) can be applied without problem when p >> n. However, even in these instances there are suitable methods for selecting a subset of features and to build a final model based on it.

Therefore, when it is advisable the feature selection, weighing methods determine the importance of the scaled features for a certain classification problem. Consider a pattern vector $\vec{x}_i \equiv (x_{i1}, x_{i2}, \ldots x_{ip})$. Assuming that the data matrix X can be partitioned into a number Q of classes, let $\vec{x}_i^{(C)}$ a pattern vector belonging to class C. The averaged patterns $\vec{m}$ and $\vec{m}^{(C)}$ represent the general mean vector and the C-class mean, according to:

$$\vec{m} = \frac{\sum_{i=1}^{n} \vec{x}_i}{n} \quad \text{and} \quad \vec{m}^{(C)} = \frac{\sum_{i=1}^{n(C)} \vec{x}_i^{(C)}}{n} \tag{4}$$

When they are applied to a selected j feature we have

$$m_j = \frac{\sum_{i=1}^{n} x_{ij}}{n} \quad \text{and} \quad m_j^{(C)} = \frac{\sum_{i=1}^{n(C)} x_{ij}^{(C)}}{n(C)} \tag{5}$$

Where n(C) is the number of patterns of class C.

Accordingly, we can explore the inter-class scatter matrix as well as the intra-class scatter matrix. The total scatter matrix can be obtained as

$$T = \sum_{i=1}^{n} (\vec{x}_i - \vec{m})(\vec{x}_i - \vec{m})^T \tag{6}$$

and its elements as

$$T_{jk} = \sum_{i=1}^{n} (x_{ij} - m_j)(x_{ik} - m_k) \tag{7}$$

The within classes scatter matrix, together with its element is given by

$$\begin{aligned} W &= \sum_{C=1}^{Q} \sum_{i=1}^{n(C)} (\vec{x}_i^{(C)} - \vec{m}^{(C)})(\vec{x}_i^{(C)} - \vec{m}^{(C)})^T \\ W_{jk} &= \sum_{C=1}^{Q} \sum_{i=1}^{n(C)} (x_{ij}^{(C)} - m_j^{(C)})(x_{ik}^{(C)} - m_k^{(C)}) \end{aligned} \tag{8}$$

And the between classes matrix and element,

$$\begin{aligned} B &= \sum_{C=1}^{Q} n(C)(\vec{m}^{(C)} - \vec{m})(\vec{m}^{(C)} - \vec{m})^T \\ B_{jk} &= \sum_{C=1}^{Q} n(C)(m_j^{(C)} - m_j)(m_k^{(C)} - m_k) \end{aligned} \tag{9}$$

For a case involving two classes 1 and 2 and one feature j we have the following:

$$\begin{aligned} T_{jj} &= \sum_{i=1}^{p} (x_{ij} - m_j)^2 \\ W_{jj} &= \sum_{i=1}^{n(1)} (x_{ij}^{(1)} - m_j^{(1)})^2 + \sum_{i=1}^{n(2)} (x_{ij}^{(2)} - m_j^{(2)})^2 \\ B_{jj} &= n(1)(m_j^{(1)} - m_j)^2 + n(2)(m_j^{(2)} - m_j)^2 \end{aligned} \tag{10}$$

Weighting features in Supervised Learning techniques can be then extracted from its relative importance in discriminating classes pairwise. The largest weight corresponds to the most important feature. The most common weighting factors are:

- Variance weights (VW) (Kowalski & Bender, 1972): $VW_j = \frac{B_{jj}}{W_{jj}}$

- Fisher weights (FW) (Duda et al., 2000): $FW_j = \dfrac{(m_j^{(1)} - m_j^{(2)})^2}{W_{jj}}$
- Coomans weights (g) (Coomans et al., 1978):

$$g_j = \frac{\left| m_j^{(1)} - m_j^{(2)} \right|}{s_j^{(1)} + s_j^{(2)}} \text{ with } s_j^{(C)} = \sqrt{\frac{\sum_{i=1}^{n(C)} (x_{ij}^{(C)} - m_j^{(C)})^2}{n(C)}}$$

A multi-group criterion is the called Wilks' λ or McCabe U statistics (McCabe, 1975). This is a general statistic used as a measure for testing the difference among group centroids. All classes are assumed to be homogeneous variance-covariance matrices and the statistic is defined as

$$\lambda = \frac{\det W}{\det T} = \frac{SSW}{SST} = \frac{SSW}{SSB + SSW} \tag{11}$$

Where SSW, SSB and SST refer to the sum of squares corresponding to the scatter matrices W, B and T, respectively, as defined above. Remembering that the ratio $\eta = \sqrt{\dfrac{BSS}{WSS}}$ is the coefficient of canonical correlation, $\eta = \sqrt{1 - \lambda}$, and hence when $\eta \to 1$ for intrinsic features, $\lambda \to 0$ and more significant are the centroid difference. Before calculation of the statistic, data should be autoscaled. This later criterion as well as the largest values of Rao's distance or Mahalanobis distance is generally used in Stepwise Discriminant Analysis (Coomans et al., 1979). Certain Supervised Learning techniques enable feature selection according to its own philosophy. Thus, for instance, SIMCA test the intrinsic features according the values of two indices called discriminating power and modelling power (Kvalheim & Karstang, 1992). Using ANNs for variable selection is attractive since one can globally adapt the variables selector together with the classifier by using the called "pruning" facilities. Pruning is a heuristic method to feature selection by building networks that do not use those variables as inputs. Thus, various combinations of input features can be added and removed, building new networks for each (Maier et al., 1998).

Genetic algorithms are also very useful for feature selection in fast methods such as PLS (Leardi & Lupiañez, 1998).

6. Development of the decision rule

In order to focus the commonly used Supervised Learning techniques of pattern recognition we have selected the following methods: K-Nearest Neighbours (KNN) (Silverman & Jones, 1989), Linear Discriminant Analysis (LDA) (Coomans et al., 1979), Canonical Variate Analysis (CVA) (Cole & Phelps, 1979), Soft Independent Modelling of Class Analogy (SIMCA) (Wold, 1976), Unequal dispersed classes (UNEQ) (Derde & Massart, 1986), PLS-DA (Stahle & Wold, 1987), Procrustes Discriminant Analysis (PDA) (González-Arjona et al., 2001), and methods based on ANN such as Multi-Layer Perceptrons (MLP) (Zupan & Gasteiger, 1993; Bishop, 2000), Supervised Kohonen Networks (Melssen et al., 2006), Kohonen Class-Modelling (KCM) (Marini et al., 2005), and Probabilistic Neural Networks (PNN) (Streit & Luginbuhl, 1994). Recently, new special classification techniques arose. A procedure called Classification And Influence Matrix Analysis (CAIMAN) has been introduced by Todeschini *et al* (2007). The method is based on the leverage matrix and models each class by means of the class dispersion matrix and calculates the leverage of each sample with respect to each class model space. Since about two decades another new classification (and regression) revolutionary technique based on statistical learning theory and kernel latent variables has been proposed: Support Vector Machines (SVM) (Vapnik, 1998; Abe, 2005; Burges, 1998). The purpose of SVM is separate the classes in a vectorial space independently on the probabilistic distribution of pattern vectors in the data set (Berrueta et al., 2007). This separation is performed with the particular hyperplane which maximizes a quantity called margin. The margin is the distance from a hyperplane separating the classes to the nearest point in the data set (Pardo & Sberveglieri, 2005). The training pattern vectors closest to the separation boundary are called *support vectors.* When dealing with a non linear boundary, the kernel method is applied. The key idea of kernel method is a transformation of the original vectorial space (input space) to a high dimensional Hilbert space (feature space), in which the classes can be separated linearly. The main advantages of SVM against its most direct concurrent method, ANN, are the easy avoiding of overfitting by using a penalty parameter and the finding of a deterministic global minimum against the non deterministic local minimum attained with ANN.

Some of the mentioned methods are equivalent. Let us consider some couples: CVA and LDA and PLS-DA and PDA. CVA attempts to find linear combinations of variables from each set that exhibit maximum correlation. These may be referred to as canonical variates, and data can be displayed as scatterplot of one against the other. The problem of maximizing the correlation can be formulated as an eigenanalysis problem with the largest eigenvalue providing the maximized correlation and the eigenvectors giving the canonical variates. Loadings of original features in the canonical variates and cumulative proportions of eigenvalues are interpreted, partly by analogy with PCA. Note that if one set of features are dummy variables giving group indicators, and then CVA is mathematically identical to LDA (González-Arjona et al., 2006). PLS-DA finds latent variables in the feature space which have a maximum covariance with the y variable. PDA may be considered equivalent to PLS-DA. The only difference is that in PDA, eigenvectors are obtained from the covariance matrix $Z^T Z$ instead of $X^T X$, with $Z = Y^T X$ where Y is the membership target matrix constructed with ones and zeros: For a three classes problem, sample labels are 001, 010 and 100. Accordingly, we can consider CVA equivalent to LDA and PLS-DA equivalent to PDA.

Researchers should be aware of apply the proper methods according to the nature and goals of the chemical problem. As Daszykowski and Walczak pointed out in his excellent survey (Daszykowski & Walczak, 2006), in many applications, unsupervised methods such as PCA are used for classification purposes instead of the supervised approach. If the data set is well structured, then PCA-scores plot can reveal grouping of patterns with different origin, although the lack of these groups in the PCA space does not necessarily mean that there is no statistical difference between these samples. PCA by definition maximizes data variance, but the main variance cannot be necessarily associated with the studied effect (for instance, sample origin). Evidently, PCA can be used for exploration, compression and visualization of data trends, but it cannot be used as Supervised Learning classification method.

On the other hands, according to the nature of the chemical problem, some supervised techniques perform better than others, because its own fundamentals and scope. In order to consider the different possibilities, four paradigms can be envisaged:

1. *Parametric/non-parametric techniques*: This first distinction can be made between techniques that take account of the information on the population distribution. Non parametric techniques such as KNN, ANN, CAIMAN and SVM make no assumption on the population distribution while parametric methods (LDA, SIMCA, UNEQ, PLS-DA) are based on the information of the distribution functions. LDA and UNEQ are based on the assumption that the population distributions are multivariate normally distributed. SIMCA is a parametric method that constructs a PCA model for each class separately and it assumes that the residuals are normally distributed. PLS-DA is also a parametric technique because the prediction of class memberships is performed by means of model that can be formulated as a regression equation of Y matrix (class membership codes) against X matrix (González-Arjona et al., 1999).
2. *Discriminating (hard)/Class-Modelling (soft) techniques*: Pure classification, discriminating or hard classification techniques are said to apply for the first level of Pattern Recognition, where objects are classified into either of a number of defined classes (Albano et al., 1978). These methods operate dividing the hyperspace in as many regions as the number of classes so that, if a sample falls in the region of space corresponding to a particular category, it is classified as belonging to that category. These kinds of methods include LDA, KNN, PLS-DA, MLP, PNN and SVM. On the other hands, Class-Modelling techniques build frontiers between each class and the rest of the universe. The decision rule for a given class is a class box that envelopes the position of the class in the pattern space. So, three kinds of classification are possible: (i) an object is assigned to a category if it is situated inside the boundaries of only a class box, (ii) an object can be inside the boundaries (overlapping region) of more than one class box, or (iii) an object is considered to be an outlier for that class if it falls outside the class box. These are the features to be covered by methods designed for the so called second level of Pattern Recognition: The first level plus the possibility of outliers and multicategory objects. Thus, typical class modelling techniques are SIMCA and UNEQ as well as some modified kind of ANN as KCM. CAIMAN method is developed in different options: D-CAIMAN is a discriminating classification method and M-CAIMAN is a class modelling one.
3. *Deterministic/Probabilistic techniques*: A deterministic method classifies an object in one and only one of the training classes and the degree of reliability of this decision is not measured. Probabilistic methods provide an estimate of the reliability of the classification decision. KNN, MLP, SVM and CAIMAN are deterministic. Other techniques, including some kind of ANN are probabilistic (e.g., PNN where a Bayesian decision is implemented).
4. *Linear/Non-Linear separation boundaries*: Here our attention is focused on the mathematical form of the decision boundary. Typical non-linear classification techniques are based on ANN and SVM, specially devoted to apply for classification problems of non-linear nature. It is remarkable that CAIMAN method seems not to suffer of nonlinear class separability problems.

7. Validation of the decision rule

A very important issue is the improper model validation. This pitfall even appears in very simple cases, such as the fitting of a series of data points by using a polynomial function. If we use a parsimonic fitting where the number of points is higher than the number of polynomial coefficients, the fitting train the generalities of the data set. Overparametrized fitting where the number of points becomes equal to the number of polynimial coefficients, trains idiosyncrasies and leads to overtraining or overfitting. Thus, a complex fitting function may fit the noise, not just the signal. Overfitting is a Damocles' sword that gravitates over any attempt to model the classification rule. We are interested to an intermediate behaviour: A model which is powerful enough to represent the underlying structure of the data (generalities), but not so powerful that it faithfully models the noise (idiosyncrasies) associated to data. This balance is known as the bias-variance tradeoff . The bias-variance tradeoff is most likely to become a problem when we have relatively few data points. In the opposite case, there is no danger of overfitting, as the noise associated with any single data point plays an immaterial role in the overall fit.

If we transfer the problem of fitting a polynomial to data into the use of another functions, such as the discriminant functions of canonical variates issued from LDA, the number of discriminant functions will be p (the number of features) or Q-1 (Q is the number of classes), whichever is smaller. As a rule of thumb (Defernez & Kemsley, 1997), the onset of overfitting should be strongly suspected when the dimensionality $d > \frac{n-Q}{3}$. One of the simplest and most widely used means of preventing overfitting is to split the known data set into two sets: the training set and the validation, evaluation, prediction or test set.

Commonly, the known set is generally randomly divided into the training and validation sets, containing about P% and 100-P% samples of every class. Typical values are 75-25% or even 50-50% for training and validation sets. The classification performance is computed in average. Thus, the randomly generation of training and validation sets is repeated a number of times, 10 times for instance. Once the classification rule is developed, some workers consider as validation parameters the recalling efficiency (rate of training samples correctly classified by the rule) and, specially, the prediction ability (rate of evaluation samples correctly classified by the rule).

An alternative to the generation of training and validation sets are the cross-validation and the bootstrapping method (Efron and Gong, 1983). In the called k-fold cross validation, the know set is split into k subsets of approximately equal size. Then the training is performed k times, each time leaving out one of k the subsets, but using only the omitted subset to predict its class membership. From all predictions, the percentage of hits gives an averaged predictive ability. A very common and simple case of cross-validation is the leave-one-out method: At any given time, only a pattern is considered and tested and the remaining patterns form the training set. Training and prediction is repeated until each pattern was treated as test once. This later procedure is easily confused with jacknifing because both techniques involve omitting each pattern in turn, but cross-validation is used just for validation purposes and jacknife is applied in order to estimate the bias of a statistic.

In bootstrapping, we repeatedly analyze subsamples, instead of subsets of the known set. Each subsample is a random sample with replacement from the full sample (known set). Bootstrapping seems to perform better than cross-validation in many instances (Efron, 1983).

However, the performance rate obtained for validating the decision rule could be misleading because they do not consider the number of false positive and false negative for each class. These two concepts provide a deep knowledge of the classes' space. Accordingly, it seems to be more advisable the use of terms sensitivity (*SENS*) and specificity (*SPEC*) (González-Arjona et al., 2006) for validating the decision rule. The *SENS* of a class corresponds to the rate of evaluation objects belonging to the class that are correctly classified, and the *SPEC* of a class corresponds to the rate of evaluation objects not belonging to the class that are correctly considered as belonging to the other classes. This may be explained in terms of the first and second kind of risks associated with prediction. The first kind of errors (α) corresponds to the probability of erroneously reject a member of the class as a non-member (rate of false negative, FN). The second kind of errors (β) corresponds to the probability of erroneously classify a non-member of the class as a member (rate of false positive, FP). Accordingly, for a given class A, and setting n_A as the number of members of class A, $\overline{n}_A$ as the number of non-members of class A, $\langle n_A \rangle$ as the number of members of class A correctly classified as "belonging to class A" and $\langle \overline{n}_A \rangle$ as the number of non-members of class A classified as "not belonging to class A", we have (Yang et al., 2005):

$$\begin{array}{ll} TP = \langle n_A \rangle & FP = \overline{n}_A - \langle \overline{n}_A \rangle \\ TN = \langle \overline{n}_A \rangle & FN = n_A - \langle n_A \rangle \end{array} \qquad (12)$$

TP and TN being the number of True Positive and True Negative members of the considered class. Accordingly,

$$\begin{aligned} SENS &= \frac{\langle n_A \rangle}{n_A} = 1 - \alpha = 1 - \frac{\mathrm{FN}}{n_A} = \frac{TP}{TP + FN} \\ SPEC &= \frac{\langle \overline{n}_A \rangle}{\overline{n}_A} = 1 - \beta = 1 - \frac{\mathrm{FP}}{\overline{n}_A} = \frac{TN}{TN + FP} \end{aligned} \qquad (13)$$

It is clear that values close to unity for both parameters indicates a successfully validation performance.

With these parameters it can be built the called *confusion matrix* for class A:

$$^{C}M_A = \begin{bmatrix} TN & FP \\ FN & TP \end{bmatrix} \qquad (14)$$

As it has been outlined, the common validation procedure consists of dividing the known set into two subsets, namely training and validation set. However, the validation procedure has to be considered with more caution in case of some kinds of ANN such as MLP because they suffer a special overfitting damage. The MLP consists of formal neurons and connection (weights) between them. As it is well known, neurons in MLP are commonly arranged in three layers: an input layer, one hidden layer (sometimes plus a bias neuron) and an output layer. The number of hidden nodes in a MLP indicates the complexity of the relationship in a way very similar to the fitting of a polynomial to a data set. Too many connections have the risk of a network specialization in training noise and poor prediction ability. Accordingly, a first action should be minimizing the number of neurons of the hidden layer. Some authors (Andrea & Kalayeh, 1991) have proposed the parameter ρ which plays a major role in determining the best architecture:

$$\rho = \frac{\text{Number of data points in the training set}}{\text{Sum of the number of connections in the network}} \tag{15}$$

In order to avoid overfitting it is recommended that $1 < \rho < 2.2$.

Besides, the overfitting problem can be minimized by monitoring the performance of the network during training by using an extra verification set different from training set. This verification set is needed in order to stop the training process before the ANN learns idiosyncrasies present in the training data that leads to overfitting (González, 2007).

8. Concluding remarks

The selection of the supervised learning technique depends on the nature of the particular problem. If we have a data set composed only by a given number of classes and the rule is going to be used on test samples that we know they may belong to one of the former established classes only, then we can select a discriminating technique such as LDA, PLS-DA, SVM or some kind of discriminating ANN (MLP or PNN). Otherwise, class modelling techniques such as SIMCA, UNEQ or KCM are useful. Class modelling tools offer at least two main advantages: To identify samples which do not fall in any of the examined categories (and therefore can be either simply outlying observations or members of a new class not considered in the known set) and to take into account samples that can simultaneously belong to more than one class (multiclass patterns).

If the idiosyncrasy of the problem suggests that the boundaries could be of non-linear nature, then the use of SVM or ANN is the best choice.

In cases where the number of features in higher than the number of samples (p > n), a previous or simultaneous step dealing with feature selection is needed when non-PCA based techniques are used (KNN, LDA, ANN, UNEQ). PCA-based methods such as SIMCA and PLS-DA can be applied without need of feature selection. This characteristic is very interesting beyond of compositional analysis, when samples are characterized by a spectrum, like in spectrometric methods (FT-IR, FT-Raman, NMR...). A different behaviour of these two methods against the number of FP and FN has been noticed (Dahlberg et al., 1997). SIMCA is focused on class specificities, and hence it detects strangers with high accuracy (only when the model set does not contain outliers. Otherwise, robust SIMCA model can be used), but sometimes fails to recognize its own members if the class is not homogeneous enough or the training set is not large enough. PLS-DA, on the contrary, deals with an implicitly closed universe (since the Y variables have a constant sum) so that it ignores the possibility of strangers. However, this has the advantage to make the method more robust to class inhomogeneities, since what matters most in class differences.

In compositional data, as pointed out Berrueta et al. (2007), the main problem is class overlap, but with a suitable feature selection and adequate sample size, good classification performances can be achieved. In general, non-linear methods such as ANN or SVM are rarely needed and most classification problems can be solved using linear techniques (LDA, CVA, PLS_DA).

Sometimes, several different types of techniques can be applied to the same data set. Classification methods are numerous and then the main problem is to select the most suitable one, especially dealing with quantitative criteria like prediction ability or misclassification percentage. In order to carry out the comparison adequately, the McNemar's test is a good choice (Roggo et al., 2003). Two classification procedures A and B are trained and the same validation set is used. Null hypothesis is that both techniques lead to the same misclassification rate. McNemar's test is based on a χ^2 test with one degree of freedom if the number of samples is higher than 20. The way to obtain the McNemar's statistic is as follows:

$$\text{McNemar's value} = \frac{\left(|n_{01} - n_{10}| - 1\right)^2}{n_{01} + n_{10}} \tag{16}$$

with

n_{00}: number of samples misclassified by both methods A and B
n_{01}: number of samples misclassified by method A but not by B
n_{10}: number of samples misclassified by method B but not by A
n_{11}: number of samples misclassified by neither method A nor B
$n_{val} = n_{00} + n_{01} + n_{10} + n_{11} =$ number of patterns in the validation set

The critical value for a 5% significance level is 3.84. In order to get insight about this procedure, the paper of Roggo et al (2006) is very promising.

Finally, a last consideration about problems with the data set representativeness. As it has been claimed in a published report a LDA was applied to differentiate 12 classes of oils on the basis of the chromatographic data, where some classes contained two or three members only (and besides, the model was not validated). There is no need of being an expertise chemometrician to be aware of two or three samples are insufficient to draw any relevant conclusion about the class to which they belong. There are more sources of possible data variance than the number of samples used to estimate class variability (Daszykowski & Walczak, 2006). The requirements of a sufficient number of samples for every class could be envisaged according to a class modelling technique to extract the class dimensionality and consider, for instance, a number of members within three to ten times this dimensionality.

Aside from this representativity context, it should be point out that when the aim is to classify food products or to build a classification rule to check the authentic origin of samples, they have to be collected very carefully according to a well established sampling plan. Often not enough care is taken about it, and thus is it hardly possible to obtain accurate classification models.

9. References

Abe, S. (2005). Support vector machines for pattern classification. Springer, ISBN:1852339299, London, UK

Aitchison, J. (2003). The statistical analysis of compositional data. The Blackburn Press, ISBN:1930665784, London, UK

Albano, C.; Dunn III, W.; Edlund, U.; Johansson, E.; Norden, B.; Sjöström, M. & Wold, S. (1978). Four levels of Pattern Recognition. Analytica Chimica Acta. Vol. 103, pp. 429-443. ISSN:0003-2670

Andrea, T.A.; Kalayeh, H. (1991). Applications of neural networks in quantitative structure-activity relationships of dihydrofolate reductase inhibitors. Journal of Medicinal Chemistry. Vol. 34, pp. 2824-2836. ISSN: 0022-2623

Aparicio, R. & Aparicio-Ruíz, R. (2002). Chemometrics as an aid in authentication, In: Oils and Fats Authentication, M. Jee (Ed.), 156-180, Blackwell Publishing and CRC Press, ISBN:1841273309, Oxford, UK and FL, USA

Bishop, C.M. (2000). Neural Networks for Pattern Recognition, Oxford University Press, ISBN:0198538642, NY, USA

Berrueta, L.A.; Alonso-Salces, R.M. & Héberger, K. (2007). Supervised pattern recognition in food analysis. Journal of Chromatography A. Vol. 1158, pp. 196-214. ISSN:0021-9673

Burges, C.J.C. (1998). A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery. Vol. 2, pp. 121-167. ISSN:1384-5810

Chung, C.F. (1993). Estimation of covariance matrix from geochemical data with observations below detection limits. Mathematical Geology. Vol. 25, pp. 851-865. ISSN:1573-8868

Clarke, J.U. (1998). Evaluation of censored data methods to allow statistical comparisons among very small samples with below detection limits observations. Environmental Science & Technology. Vol. 32, pp. 177-183. ISSN:1520-5851

Cole, R.A. & Phelps, K. (1979). Use of canonical variate analysis in the differentiation of swede cultivars by gas-liquid chromatography of volatile hydrolysis products. Journal of the Science of Food and Agriculture. Vol. 30, pp. 669-676. ISSN:1097-0010

Coomans, D.; Broeckaert, I.; Fonckheer, M; Massart, D.L. & Blocks, P. (1978). The application of linear discriminant analysis in the diagnosis of thyroid diseases. Analytica Chimica Acta. Vol. 103, pp. 409-415. ISSN:0003-2670

Coomans, D.; Massart, D.L. & Kaufman, L. (1979) Optimization by statistical linear discriminant analysis in analytical chemistry. Analytica Chimica Acta. Vol. 112, pp. 97-122. ISSN:0003-2670

Dahlberg, D.B.; Lee, S.M.; Wenger, S.J. & Vargo, J.A. (1997). Classification of vegetable oils by FT-IR. Applied Spectroscopy. Vol. 51, pp. 1118-1124. ISSN:0003-7028

Daszykowski, M. & Walczak, B. (2006). Use and abuse of chemometrics in chromatography. Trends in Analytical Chemistry. Vol. 25, pp. 1081-1096. ISSN:0165-9936

Daszykowski, M.; Kaczmarek, K.; Stanimirova, I.; Vander Heyden, Y. & Walczak, B. (2007). Robust SIMCA-bounding influence of outliers. Chemometrics and Intelligent Laboratory Systems. Vol. 87, pp. 95-103. ISSN:0169-7439

Dean, R.B. & Dixon, W.J. (1951). Simplified statistics for small number of observations. Analytical Chemistry. Vol. 23, pp. 636-638. ISSN:0003-2700

Defernez, M. & Kemsley, E.K. (1997). The use and misuse of chemometrics for treating classification problems. Trends in Analytical Chemistry. Vol. 16, pp. 216-221. ISSN:0165-9936

Derde, M.P.; Coomans, D. & Massart, D.L. (1982). Effect of scaling on class modelling with the SIMCA method. Analytica Chimica Acta. Vol. 141, pp. 187-192. ISSN:0003-2670

Derde, M.P. & Massart, D.L. (1986). UNEQ: A class modelling supervised pattern recognition technique. Microchimica Acta. Vol. 2, pp. 139-152. ISSN:0026-3672

Duda, R.O.; Hart, P.E. & Stork, D.G. (2000). Pattern classification. 2nd edition. Wiley, ISBN:0471056693, NY, USA

Efron, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross validation. Journal of the American Statistical Association. Vol. 78, pp. 316-331. ISSN:0162-1459

Efron, B. & Gong, G. (1983). A leisurely look at the bootstrap, the jacknife and cross validation. The American Statiscian. Vol. 37, pp. 36-48. ISSN:0003-1305

Egan, W.J. & Morgan, S.L. (1998). Outlier detection in multivariate analytical chemical data. Analytical Chemistry. Vol. 70, pp. 2372-2379. ISSN:0003-2700

Egozcue, J.J.; Pawlowsky-Glahn, V.; Mateu-Figueros, G.; Barcelo-Vidal, C. (2003). Isometric logratio transformation for compositional data analysis. Mathematical Geology. Vol. 35, pp. 279-300. ISSN:1573-8868

Gemperline, P.J. & Boyer, N.R. (1995). Classification of near-infrared spectra using wavelength distances: Comparisons to the Mahalanobis distance and Residual Variance methods. Analytical Chemistry. Vol.67, pp. 160-166. ISSN:0003-2700

González, A.G. (2007). Use and misuse of supervised pattern recognition methods for interpreting compositional data. Journal of Chromatograpy A. Vol. 1158, pp. 215-225. ISSN:0021-9673

González, A.G. & Herrador, M.A. (2007). A practical guide to analytical method validation, including measurement uncertainty and accuracy profiles. Trends in Analytical Chemistry. Vol. 26, pp. 227-237. ISSN:0165-9936

González-Arjona, D.; López-Pérez, G. & González, A.G. (1999). Performing Procrustes discriminant analysis with HOLMES. Talanta. Vol. 49, pp. 189-197. ISSN:0039-9140

González-Arjona, D.; López-Pérez, G. & González, A.G. (2001). Holmes, a program for performing Procrustes Transformations. Chemometrics and Intelligent Laboratory Systems. Vol. 57, pp. 133-137. ISSN:0169-7439

González-Arjona, D.; López-Pérez, G. & González, A.G. (2006). Supervised pattern recognition procedures for discrimination of whiskeys from Gas chromatography/Mass spectrometry congener analysis. Journal of Agricultural and Food Chemistry. Vol. 54, pp. 1982-1989. ISSN:0021-8561

Grubbs, F. (1969). Procedures for detecting outlying observations in samples. Technometrics. Vol. 11, pp. 1-21. ISSN:0040-1706

Helsel, D.R. (1990).Less than obvious: Statistical treatment of data below the detection limit. Environmental Science & Technology. Vol. 24, pp. 1766-1774. ISSN: 1520-5851

Hoaglin, D.C. & Welsch, R.E. (1978). The hat matrix in regression and ANOVA. The American Statiscian. Vol. 32, pp. 17-22. ISSN:0003-1305

Holger, R.M.; Dandy, G.C. & Burch, M.D. (1998). Use of artificial neural networks for modelling cyanobacteria Anabaena spp. In the river Murray, South Australia. Ecological Modelling. Vol. 105, pp. 257-272. ISSN:0304-3800

Jollife, I.T. (2002). Principal Component Analysis. 2nd edition, Springer, ISBN:0387954422, NY, USA

Johnson, G.W. & Ehrlich, R. (2002). State of the Art report on multivariate chemometric methods in Environmental Forensics. Environmental Forensics. Vol. 3, pp. 59-79. ISSN:1527-5930

Kryger, L. (1981). Interpretation of analytical chemical information by pattern recognition methods-a survey. Talanta. Vol. 28, pp. 871-887. ISSN:0039-9140

Kowalski, B.R. & Bender, C.F. (1972). Pattern recognition. A powerful approach to interpreting chemical data. Journal of the American Chemical Society. Vol. 94, pp. 5632-5639. ISSN:0002-7863

Kvalheim, O.M. & Karstang, T.V. (1992). SIMCA-Classification by means of disjoint cross validated principal component models, In: Multivariate Pattern Recognition in Chemometrics, illustrated by case studies, R.G. Brereton (Ed.), 209-245, Elsevier, ISBN:0444897844, Amsterdam, Netherland

Leardi, R. & Lupiañez, A. (1998). Genetic algorithms applied to feature selection in PLS regression: how and when to use them. Chemometrics and Intelligent Laboratory Systems. Vol. 41, pp. 195-207. ISSN:0169-7439

Malinowski, E.R. (2002). Factor Analysis in Chemistry. Wiley, ISBN:0471134791, NY, USA

Marini, F.; Zupan, J. & Magrí, A.L. (2005). Class modelling using Kohonen artificial neural networks. Analytica Chimica Acta. Vol.544, pp. 306-314. ISSN:0003-2670

Martín-Fernández, J.A.; Barceló-Vidal, C. & Pawlowsky-Glahn, V. (2003). Dealing with zeros and missing values in compositional data sets using nonparametric imputation. Mathematical Geology. Vol. 35, pp. 253-278. ISSN:1573-8868

McCabe, G.P. (1975). Computations for variable selection in discriminant analysis. Technometrics. Vol. 17, pp. 103-109. ISSN:0040-1706

Melssen, W.; Wehrens, R. & Buydens, L. (2006). Supervised Kohonen networks for classification problems. Chemometrics and Intelligent Laboratory Systems. Vol. 83, pp. 99-113. ISSN:0169-7439

Mertens, B.; Thompson, M. & Fearn, T. (1994). Principal component outlier detection and SIMCA: a synthesis. Analyst. Vol. 119, pp. 2777-2784. ISSN:0003-2654

Miller, J.N. & Miller, J.C. (2005). Statistics and Chemometrics for Analytical Chemistry. 4th edition. Prentice-Hall, Pearson. ISBN:0131291920. Harlow, UK

Naes, T.; Isaksson, T.; Fearn, T.; Davies, T. (2004). A user-friendly guide to multivariate calibration and classification. NIR Publications, ISBN;0952866625, Chichester, UK

Pardo, M. & Sberveglieri, G. (2005). Classification of electronic nose data with support vector machines. Sensors and Actuators. Vol. 107, pp. 730-737. ISSN:0925-4005

Pretsch, E. & Wilkins, C.L. (2006). Use and abuse of Chemometrics. Trends in Analytical Chemistry. Vol. 25, p. 1045. ISSN:0165-9936

Rock, B.A. (1985). An introduction to Chemometrics, 130th Meeting of the ACS Rubber Division. October 1985. Available from
http://home.neo.rr.com/catbar/chemo/int_chem.html

Roggo, Y.; Duponchel, L. & Huvenne, J.P. (2003). Comparison of supervised pattern recognition methods with McNemar's statistical test: Application to qualitative analysis of sugar beet by near-infrared spectroscopy. Analytica Chimica Acta. Vol. 477, pp. 187-200. ISSN:0003-2670

Sharaf, M.A.; Illman, D.A. & Kowalski, B.R. (1986). Chemometrics. Wiley, ISBN:0471831069, NY, USA

Silverman, B.W. & Jones, M.C. (1989). E. Fix and J.L. Hodges (1951): An important contribution to non parametric discriminant analysis and density estimation. International Statistical Review. Vol. 57, pp. 233-247. ISSN:0306-7734

So, S.S. & Richards, W.G. (1992). Application of Neural Networks: Quantitative structure activity relationships of the derivatives of 2,4-diamino-5-(substituted-benzyl) pyrimidines as DHFR inhibitors. Journal of Medicinal Chemistry. Vol. 35, pp. 3201-3207. ISSN:0022-2623

Stahle, L. & Wold, S. (1987). Partial least squares analysis with cross validation for the two class problem: A monte-Carlo study. Journal of Chemometrics. Vol. 1, pp. 185-196. ISSN:1099-128X

Stanimirova, I.; Daszykowski, M. & Walczak, B. (2007). Dealing with missing values and outliers in principal component analysis. Talanta. Vol. 72, pp. 172-178. ISSN:0039-9140

Streit, R.L. & Luginbuhl, T.E. (1994). Maximun likelihood training of probabilistic neural networks. IEEE Transactions on Neural Networks. Vol. 5, pp. 764-783. ISSN:1045-9227

Todeschini, R.; Ballabio, D.; Consonni, V.; Mauri, A. & Pavan, M. (2007). CAIMAN (Classification And Influence Matrix Analysis): A new approach to the classification based on leverage-scale functions. Chemometrics and Intelligent Laboratory Systems. Vol. 87, pp. 3-17. ISSN:0169-7439

Vapnik, V.N. (1998). Statistical learning theory. Wiley, ISBN:0471030031, NY, USA

Varmuza, K. (1980). Pattern recognition in chemistry. Springer, ISBN:0387102736, Berlin, Germany

Varmuza, K. & Filzmoser, P. (2009). Introduction to multivariate statistical analysis in chemometrics, CRC Press, Taylor & Francis Group, ISBN:14005975, Boca Ratón, FL, USA

Walczak, B. & Massart, D.L. (2001). Dealing with missing data: Part I and Part II. Chemometrics and Intelligent Laboratory System. Vol. 58, pp. 15-27 and pp. 29-42. ISSN:0169-7439

Wold, S. (1976). Pattern recognition by means of disjoint principal component models. Pattern Recognition. Vol. 8, pp. 127-139. ISSN:0031-3203

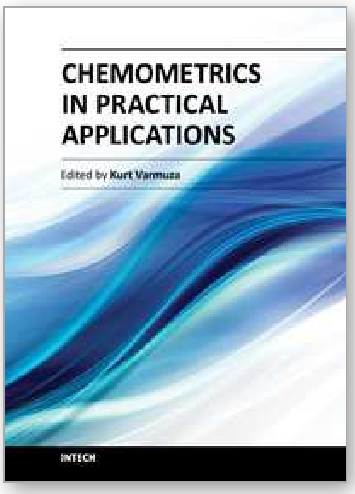
Wold, S. (1978). Cross validatory estimation of the number of components in factor and principal components models. Technometrics. Vol. 20, pp. 397-405. ISSN:0040-1706

Wold, S. (1995). Chemometrics, what do we mean with it, and what do we want from it? Chemometrics and Intelligent Laboratory Systems. Vol. 30, pp. 109-115. ISSN:0169-7439

Yang, Z.; Lu, W.;Harrison, R.G.; Eftestol, T.; Steen,P.A. (2005). A probabilistic neural network as the predictive classifier of out-of-hospital defibrillation outcomes. Resuscitation. Vol. 64, pp. 31-36. ISSN:0300-9572

Zupan, J. & Gasteiger, J. (1993). Neural Networks for chemists. VCH, ISBN:1560817917, Weinheim, Germany

Zupan, J.; Novic, M.; Li, X. & Gasteiger, J. (1994). Classification of multicomponent analytical data of olive oils using different neural networks. Analytica Chimica Acta. Vol. 292, pp. 219-234. ISSN:0003-2670



Chemometrics in Practical Applications
Edited by Dr. Kurt Varmuza

ISBN 978-953-51-0438-4
Hard cover, 326 pages
Publisher InTech
Published online 23, March, 2012
Published in print edition March, 2012

In the book "Chemometrics in practical applications", various practical applications of chemometric methods in chemistry, biochemistry and chemical technology are presented, and selected chemometric methods are described in tutorial style. The book contains 14 independent chapters and is devoted to filling the gap between textbooks on multivariate data analysis and research journals on chemometrics and chemoinformatics.

How to reference
In order to correctly reference this scholarly work, feel free to copy and paste the following:

A. Gustavo González (2012). Critical Aspects of Supervised Pattern Recognition Methods for Interpreting Compositional Data, Chemometrics in Practical Applications, Dr. Kurt Varmuza (Ed.), ISBN: 978-953-51-0438-4, InTech, Available from: http://www.intechopen.com/books/chemometrics-in-practical-applications/critical-aspects-of-supervised-pattern-recognition-for-interpreting-chemical-compositional-data

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821