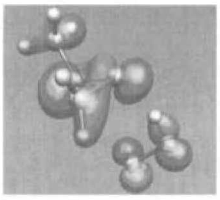

Computational Chemistry Column

Column Editors:
Prof. Dr. H. Huber, University of Basel
Prof. Dr. K. Müller, F. Hoffmann-La Roche AG, Basel
Prof. Dr. H.P. Lüthi, Univ. of Geneva, ETH-Zürich

Chimia 55 (2001) 70–80
© Neue Schweizerische Chemische Gesellschaft
ISSN 0009–4293

Chemometrics and Modeling

Frédéric Estienne, Yvan Vander Heyden, and D. Luc Massart*

Abstract: Chemometrics is a chemical discipline in which mathematical and statistical techniques are applied to design experiments or to analyze chemical data. An important part of chemometrics is modeling, in which one tries to relate two or more characteristics in such a way that the obtained model represents reality as closely as possible. In this article some less known but useful regression methods such as orthogonal least squares, inverse and robust regression are introduced and compared with the well-known classical least squares regression method. Genetic algorithms are described as a means of carrying out feature selection for multivariate regression. Regression methods such as principal component regression and partial least squares are introduced as well as the use of N-way principal components.

Keywords: Analytical chemistry · Chemical data analysis · Chemometrics · Modeling · QSAR · Regression methods

Introduction

Chemometrics has been defined [1] as a chemical discipline that uses mathematics, statistics and formal logic (a) to design or select optimal experimental procedures, (b) to provide the maximum relevant chemical information by analyzing chemical data, and (c) to obtain knowledge about chemical systems.

In this article we will focus on how chemometrics is used for modeling purposes. However, first we should note that, while modeling is probably the most important area of chemometrics, there are many other applications such as method validation, optimization, statistical process control, signal processing, *etc.*

Modeling is applied when two or more characteristics of the same objects are measured or calculated and then related to each other, for example the concentration of a chemical compound to an instrumental signal, the chemical structure of a drug to its activity or instrumental responses to sensory characteristics. The purpose of the modeling usually is to make predictions (*e.g.* predict the concentration of a certain analyte in a sample from a measured signal), but sometimes simply to verify the nature of the relationship.

The expertise of the authors is in the use of chemometrics for analytical chemical purposes and most examples will therefore come from that area.

Classical Univariate Least Squares: Straight Line Models

Before introducing some of the more sophisticated methods such as genetic algorithms, latent variable procedures or neural nets, we should look shortly at the classical univariate least squares methodology (often called ordinary least squares – OLS), which is what analytical chemists generally use to construct a (linear) calibration line. In most analytical techniques the concentration of a sample cannot be measured directly but is derived from a measured signal that is in direct relation to the concentration. Suppose x represents a concentration and y the corresponding measured instrumental signal. To be able to define a model $y = f(x)$ a relationship between x and y has to exist. The simplest and most convenient situation is when the relation is linear which leads to a model of the type $y = b_0 + b_1x$ and which represents a straight line. The coefficients b_0 and b_1 represent the intercept and the slope of the line. Relationships between y and x that follow a curved line can for instance be represented by a regression model of the type $y = b_0 + b_1x + b_{11}x^2$.

The least squares regression analysis is a methodology that allows the coefficients of a given model to be estimated. For calibration purposes one usually focuses on straight line models which we also will do in the rest of this section.

*Correspondence: Prof. Dr. D.L. Massart
Farmaceutisch Instituut
Vrije Universiteit Brussel
Laarbeeklaan 103
B–1090 Brussels
Tel.: +32 2 477 47 34
Fax: +32 2 477 47 35
E-Mail: fabi@fabi.vub.ac.be

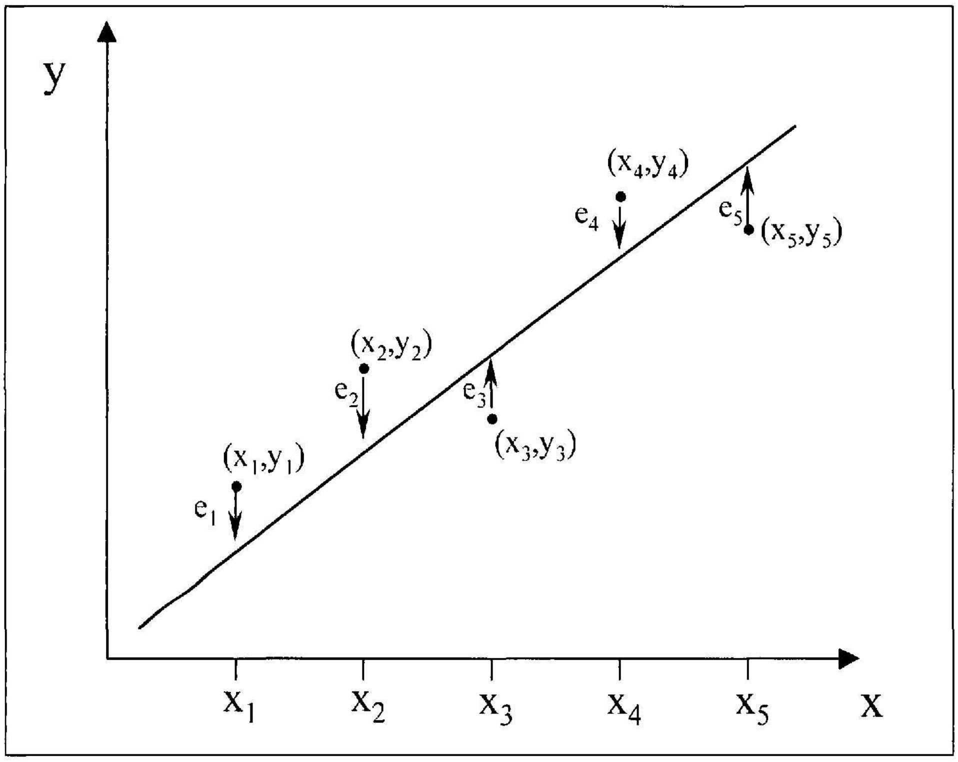

Fig. 1. Straight line fitting through a series of measured points.

Conventionally the x values represent the so-called controlled or independent variable, *i.e.* the variable that is considered not to have a measurement error (or a negligible one), which is the concentration in our case. The y values represent the dependent variable, *i.e.* the measured response, which is considered to have a measurement error. The least squares approach allows b_0 and b_1 values to be obtained such that the model fits the measured points (x_i, y_i) best (Fig. 1).

The true relationship between x and y is considered to be $y = \beta_0 + \beta_1 x$ while the relationship between each x_i and its measured y_i can be represented as $y_i = b_0 + b_1 x_i + e_i$. The signal y_i is composed of a component predicted by the model, $b_0 + b_1 x$, and a random component, e_i, the residual (Fig. 1). The least squares regression finds the estimates b_0 and b_1 for β_0 and β_1 by calculating the values b_0 and b_1 for which $\Sigma e_i^2 = \Sigma\ (y_i - b_0 - b_1 x_i)^2$, the sum of the squared residuals, is minimal. This explains the name 'least squares'. Standard books about regression, including least squares approaches, are given in [2][3]. Analytical chemists can find information in [4][5].

Some Variants of the Univariate Least Square Straight Line Models

A fundamental assumption of OLS is that there are only errors in the direction of y. In some instances, two measured quantities are related to each other and the assumption then does not hold, because there are also measurement errors in x. This is for instance the case when two methods are compared to each other. Often one of these methods is a reference method and the other a new method, which is faster or cheaper and a demonstration is required that the results of both methods are sufficiently similar. A certain number of samples are analyzed with both methods and a straight line model relating both series of measurements is obtained. If β_0 as estimated from b_0 is not more different from 0 than an *a priori* accepted bias and β_1 as estimated by b_1 is not more different from 1 than a given amount, then one can accept that for practical purposes y = x. In its simplest statistical expression, this means that it is tested that $\beta_0 = 0$ and $\beta_1 = 1$ or to put it in another way that b_0 is statistically different from 0 and/or b_1 is statistically different from 1. If this is the case then it is concluded that the two methods do not yield the same result but that there is a constant (intercept) or proportional (slope) systematic error or bias.

This means that one should calculate b_0 and b_1 and at first sight this could be done by OLS. However both regression variables (not only y_i but now also x_i) are subject to error, as already mentioned. This violates one of the key assumptions of the OLS calculations.

It has been shown [5–8] that the computation of b_0 and b_1 according to the OLS-methods leads to wrong estimates of β_0 and β_1. Significant errors in the least squares estimate of b_1 can be expected if the ratio between the measurement error on the x values and the range of the x values is large. In that case OLS should not be used. To obtain correct values for b_0 and b_1 the sum of least squares must now be obtained in the direction given in Fig. 2. Such methods are sometimes called *errors in variables models* or *orthogonal least squares*. Detailed studies of the application of models of this type can be found in [9][10].

Another possibility is to apply *inverse regression*. The term inverse is applied in opposition to the usual calibration procedure. Calibration consists of measuring samples with a known characteristic and deriving a calibration line (or more generally a model). A measurement is then carried out for an unknown sample and its concentration is derived from the measurement result and the calibration line. In view of the assumptions of OLS, the measurement is the y-value and the concentration the x-value, *i.e.*

$$\text{Measurement} = \text{f (concentration)} \quad (1)$$

This relationship can be inversed to become

$$\text{Concentration} = \text{f (measurement)} \quad (2)$$

OLS is then applied in the usual way, meaning that the sum of the squared residuals is minimized in the direction of y, which is now the concentration. This may appear strange, since, when the calibration line is computed, there are no errors in the concentrations. However, if it is taken into account that there will be an error in the predicted concentration of the unknown sample, then minimizing in this way means that one minimizes the prediction errors, which is what is important to the analytical chemist. It has been shown indeed that better results are obtained in this way [11–13]. The analytical chemist should therefore really apply Eqn. (2), instead of the usual Eqn. (1). In most cases the difference in prediction quality between both approaches is very small in practice, so that there is generally no harm in applying Eqn. (1). We will see however that when multivariate calibration is applied, inverse regression is the rule. It should be noted that, when the aim is not to predict y-values, but to obtain the best possible estimates of β_0 and β_1, inverse regression performs less well than the usual procedure.

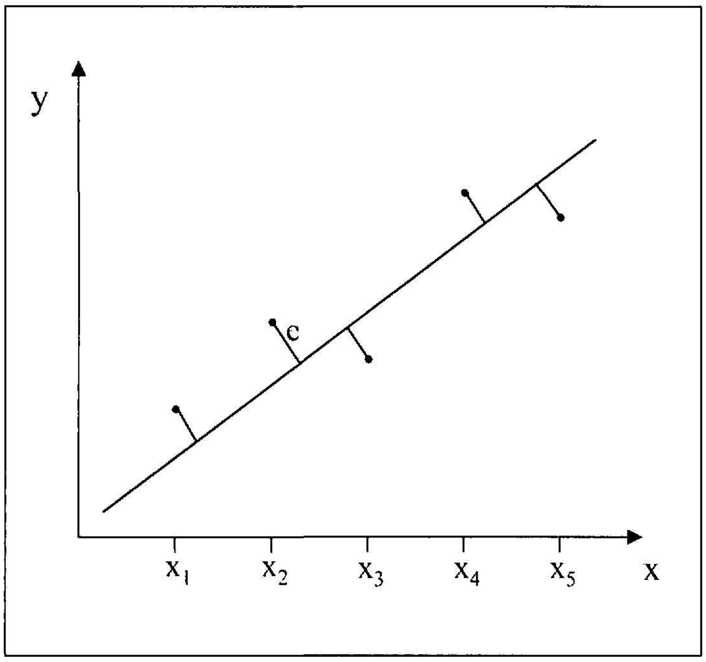

Fig. 2. The errors-in-variables model.

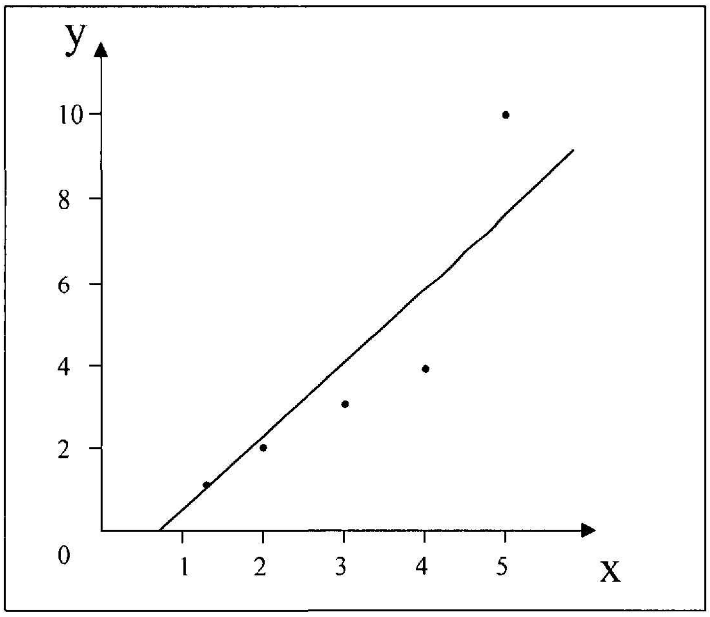

Fig. 3. The leverage effect.

Robust Regression

One of the most frequently occurring difficulties for an experimentalist is that of the presence of outliers. The outliers may be due to experimental error or to the fact that the proposed model does not represent the data well enough. For example, if the postulated model is a straight line, and measurements are made in a concentration range where this is no longer true, the measurements obtained in that region will be model outliers. In Fig. 3 it is clear that the last point is not representative for the straight line fitted by the rest of the data. The outlier attracts the regression line computed by OLS. It is said to exert *leverage* on the regression line. One might think that outliers can be discovered by examining the residuals towards the line. As can be observed this is not necessarily true: the outlier's residual is not much larger than that of some other data points.

To avoid the leverage effect, the outlier(s) should be eliminated. One way to achieve this is to use more efficient outlier diagnostics than simply looking at residuals. Cook's squared distance or the Mahalanobis distance can for instance be used.

A more elegant way is to apply so-called robust regression methods. The easiest to explain is the *single median method* [14]. The slope between each pair of points is computed. For instance the slope between points 1 and 2 is 1.10, between 1 and 3 1.00, between 5 and 6 6.20. The complete list is 1.10, 1.00, 1.03, 0.95, 2.00, 0.90, 1.00, 0.90, 2.23, 1.10, 0.90, 2.67, 0.70, 3.45, 6.20. These are now ranked and the median slope (here the 8th value 1.03) is chosen. All pairs of points of which the outlier is one point have high values and end up at the end of the ranking, so that they do not have an influence on the chosen median slope: even if the outlier was still more distant, the selected median would still be the same. A similar procedure for the intercept, which we will not explain in detail, leads to the straight line equation y = 0.00 + 1.03 x, which is close to the line obtained with OLS after eliminating the outlier. The single median method is not the best robust regression method. Better results are obtained with the least median of squares method (LMS) [15], the iteratively reweighted [16] or biweight regression [17]. Comparing results of calibration lines obtained with OLS and with a robust method is one way of finding outliers towards a regression model [18].

Multivariate (Multiple) Regression

Multivariate regression, also often called multiple regression or multiple linear regression (MLR) in the linear case, is used to obtain values for the b coefficients in an equation of the type

$$y = b_0 + b_1x_1 + b_2x_2 + \dots b_mx_m \qquad (3)$$

where x_1, x_2, ..., x_m are different variables. In analytical spectroscopic applications, these variables could be the absorbencies obtained at different wavelengths, y being a concentration or other characteristic of the samples to be predicted, in QSAR (the study of quantitative structure-activity relationships) they could be variables such as hydrophobicity (log P), the Hammett electronic parameter σ, with y being some measure of biological activity. In experimental design, equations of the type

$$\begin{aligned} y=&b_0+b_1x_1+b_2x_2 \\ &+b_{12}x_1x_2+b_{11}x_1^2+b_{22}x_2 \end{aligned} \qquad (4)$$

are used to describe a response y as a function of the experimental variables x_1 and x_2. Both Eqn. (3) and (4) are called linear, which may surprise the non-initiated, since the shape of the relationship between y and (x_1,x_2) is certainly not linear. The term linear should be understood as linear in the regression parameters. An equation such as $y = b_0 + \log (x - b_l)$ is non-linear [2].

It can be observed from the applications cited above that multiple regression models occur quite often. We will first consider the classical solution to estimate the coefficients. Later we will describe some more sophisticated methodologies introduced by chemometricians, such as those based on latent vectors.

As for the univariate case, the b-values are estimates of the true b-parameters and the estimation is done by minimizing a (sum of) squares. It can be shown that

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

where **b** is the vector containing the b-values from Eqn. (3), **X** is an nxm matrix containing the x-values for n samples (or objects as they are often called) and m variables and **y** is the vector containing the measurements for the n samples.

One difficulty is that the inversion of the $\mathbf{X}^T\mathbf{X}$ matrix leads to unstable results when the x-variables are very correlated. As we will explain later, this happens for instance with spectroscopic data. There are two ways to avoid this problem. One is to select variables (variable selection or feature selection) such that correlation is reduced, the other is to combine the variables in such a way that the resulting summarizing variables are not correlated (feature reduction). Both feature selection and feature reduction lead to a smaller number of variables than the initial number of variables, which by itself has important advantages.

The classical approach, which is found in many statistical packages, is the so-called stepwise regression, a feature selection method. The so-called forward selection procedure consists of first selecting the variable that is best correlated with y. Suppose this is found to be x_i. The model at this stage is restricted to $y=f(x_i)$. Then, one tests all other variables by adding them to the model, which then becomes a model in two variables $y=f(x_i,x_j)$. The variable x_j which is retained together with x_i is the one which when added to the model leads to the largest improvement compared to the original model $y = f(x_i)$. Then it is tested whether the observed improvement is significant. If not, the procedure stops and the model is restricted to $y = f(x_i)$. If the improvement is significant, x_j is incorporated definitively in the model. It is then investigated which variable should be added as the third one and whether this yields a significant improvement. The procedure is repeated until finally no further improvement is obtained. The procedure is based on analysis of variance and several variants such as backwards elimination (starting with all variables and eliminating successively the least important ones) or a combination of forward and backward methods has also been proposed. It should be noted that the criteria applied in the analysis of variance are such that variables are automatically selected that are less correlated. In certain contexts such as in experimental design or QSAR, the reason for applying feature selection is not only to avoid the numerical difficulties described higher, but also to explain relationships. The variables that are included in the regression equation have a chemical and physical meaning and when a certain variable is retained it is considered that the variable influences the y-value, *e.g.* the biological activity, which then leads to proposals for causal relationships. Correct feature selection then becomes very important in those situations to avoid making wrong conclusions. A discussion comparing different strategies for feature selection in QSAR is given in [19]. One of the problems is that the procedures involve regressing many variables on y and chance correlations may then occur [20].

There are other difficulties, for instance, the choice of experimental conditions, the samples or the objects. These should cover the experimental domain as well as possible and, where possible, follow an experimental design. This is demonstrated, for instance, in [21]. Outliers can also cause problems. Detection of multivariate outliers is not evident. As for the univariate regression, robust regression is possible [15][22]. An interesting example in which multivariate robust regression is applied concerns an experimental design [23] carried out to optimize the yield of an organic synthesis.

Wide Data Matrices

Chemists often produce wide data matrices, characterized by a relatively small number of objects (a few tens to a few hundred) and a very large number of variables (many hundreds, at least). For instance, analytical chemists now often apply very fast spectroscopic methods, such as near infrared spectroscopy (NIR). Because of the rapid character of the analysis, there is no time to dissolve the sample or separate certain constituents. The chemist tries to extract the information required from the spectrum as such and to do so he has to relate a **y**-value such as an octane number of gasoline samples or a protein content of wheat samples to the absorbance at 500 to, in some cases, 10 000 wavelengths. The *e.g.* 1000 variables for 100 objects constitute the **X** matrix. Such matrices contain many more columns than rows and are therefore often called wide.

Very wide matrices are also encountered in QSAR. For instance, in comparative molecular field analysis (CoMFA), developed by Cramer [24], three-dimensional grids are laid over a set of molecules whose properties in reacting with other molecules or receptors one wants to predict. At each of the resulting lattice points electrostatic, hydrophobic and steric fields are computed. A typical grid of 5 x 2 x 2 nm^3 with a spacing of 0.02 nm yields 2 500 000 such lattice points for each molecule. This huge set of data constitutes the **X** matrix and must be related to *e.g.* biological activity data.

Feature selection/reduction then takes on a completely different complexity compared to the situations described in the preceding sections. It should be noted that variables in such matrices are often very correlated. This can for instance be expected for two neighboring wavelengths in a spectrum or the fields measured at adjacent locations in the CoMFA lattice. In what follows, we will explain which methods chemometricians use to model very large, wide and highly correlated data matrices.

Genetic Algorithms for Feature Selection

Genetic algorithms are general optimization tools aiming at selecting the fittest solution to a problem. Suppose that, to keep it simple, nine variables are measured. Possible solutions are represented in Fig. 4. Selected variables are indicated by a 1, non-selected variables by a 0. Such solutions are sometimes, in analogy with genetics, called chromosomes in the jargon of the specialists.

By random selection a set of such solutions is obtained (in real applications often several hundreds). For each solution an MLR model is built using an equation such as (3) and the sum of squares of the residuals of the objects towards that model is determined. In the jargon of the field one says that the fitness of each solution is determined: the smaller the sum of squares the better the model describes the data and the fitter the corresponding solutions are. Then follows what is described as the selection of the fittest (leading to names such as genetic algorithms or evolutionary computation). For instance out of the, say 100 original solutions, the 50 fittest are retained. They are called the parent generation. From these a child generation is obtained by reproduction and mutation.

Reproduction is explained in Fig. 5. Two randomly chosen parent solutions produce two child solutions by cross over. The cross over point is also chosen randomly. The first part of solution 1 and the second part of solution 2 together yield child solution 1'. Solution 2' results from the first part of solution 2 and the second of solution 1.

The child solutions are added to the selected parent solutions to form a new generation. This is repeated for many generations and the best solution from the final generation is retained. Each generation is additionally submitted to mutation steps. Here and there randomly chosen bits of the solution string are changed (0 to 1 or 1 to 0). This is applied in Fig. 6.

The need for the mutation step can be understood from Fig. 5. Suppose that the best solution is close to one of the child solutions in that Fig., but should not include variable 9. However, because the value for variable 9 is 1 in both parents, it is also unavoidably 1 in the children. Mutation can change this and move the solutions in a better direction.

Genetic algorithms were first proposed by Holland [25]. They were introduced in chemometrics by Lucasius *et al.* [26] and Leardi [27]. They were applied for instance in QSAR and molecular modeling [28], conformational analysis [29], multivariate calibration for the determination of certain characteristics of polymers [30] or octane numbers [31]. Reviews about applications in chemistry can be found in [32][33]. There are several competing algorithms such as simulated annealing [34] or the immune algorithm [35].

Latent Variables for Feature Reduction: Principal Components

The alternative to feature selection is to combine the variables in what we called earlier summarizing variables. Chemometricians call this latent variables and the obtention of such variables is called feature reduction. It should be understood that in this case no variables are discarded. The type of latent variable most commonly used is the principal component (PC). To explain it we will first consider the simplest possible situation. Two variables (x_1 and x_2) were measured for a certain number of objects and the number of variables should be reduced to one. In principal component analysis (PCA) this is achieved by defining a new axis or vari-able on which the objects are projected. The projections are called the scores, s1, along principal component 1, PC1 (Fig.7).

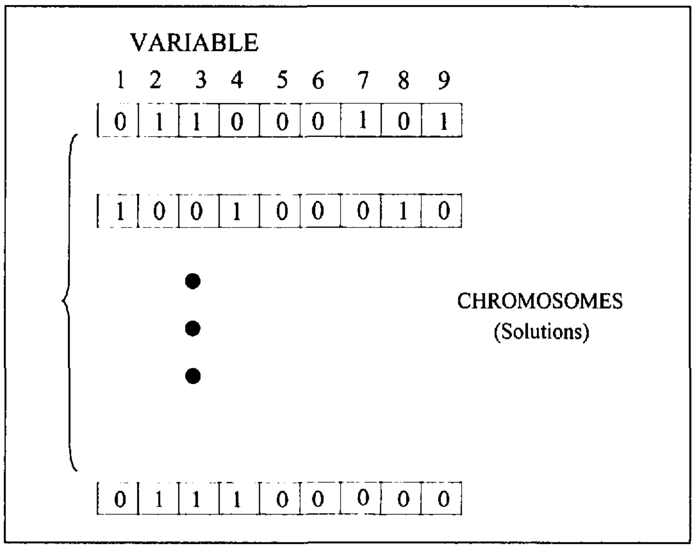

Fig. 4. A set of solutions for feature selection from nine variables for MLR.

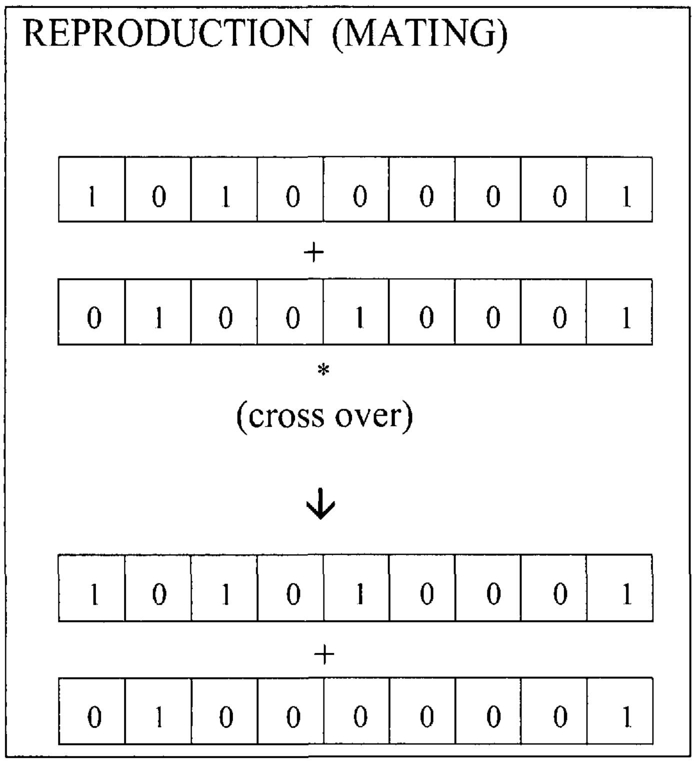

Fig. 5. Genetic algorithms: the reproduction step.

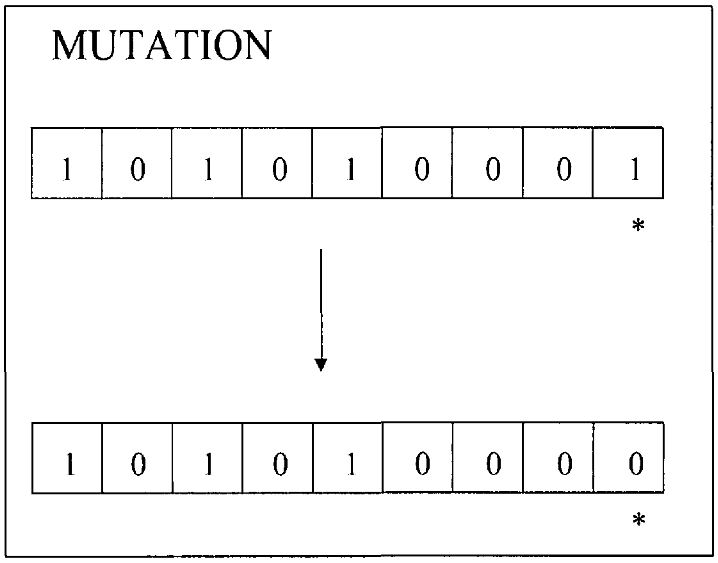

Fig. 6. Genetic algorithms: the mutation step.

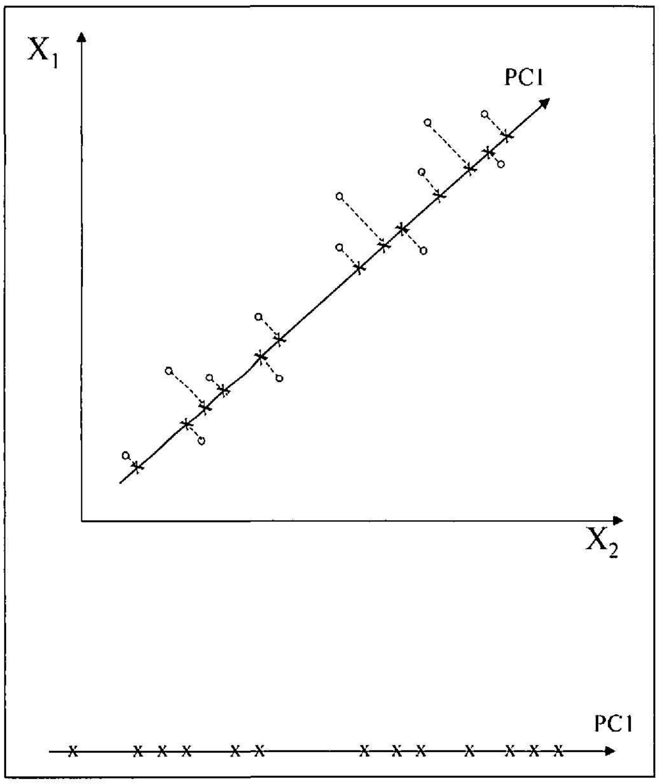

Fig. 7. Feature reduction of two variables, x_1 and x_2, by a principal component.

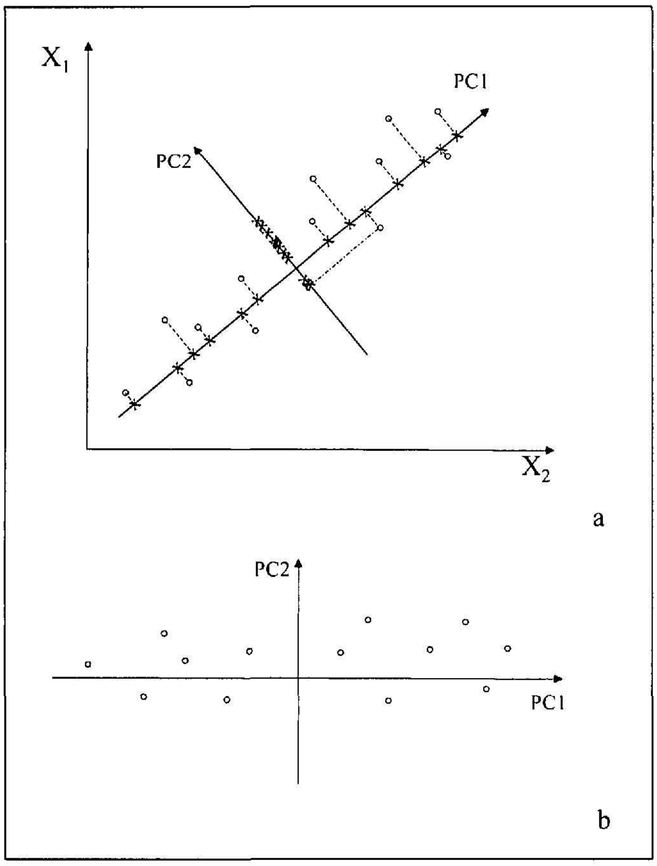

Fig. 8. a) Second PC and b) score plot of the data in Fig. 1.

The projections along PC1 preserve the information present in the x_1–x_2 plot, namely that there are two groups of data. By definition, PC1 is drawn in the direction of the largest variation through the data. A second PC, PC2, can also be obtained. By definition it is orthogonal to the first one (Fig. 8a). The scores along PC1 and along PC2 can be plotted against each other yielding what is called a score plot (Fig. 8b).

The reader observes that PCA decorrelates: while the data points in the x_1–x_2 plot are correlated they are no longer so in the s_1–s_2 plot. This also means that there was correlated and therefore redundant information present in x_1 and x_2. PCA picks up all the important information in PC1 and the rest, along PC2, is noise and can be eliminated. By keeping only PC1, feature reduction is applied: the number of variables, originally two, has been reduced to one. This is achieved by computing the score along PC1 as:

$$s = w_1x_1 + w_2x_2 \quad (5)$$

In other words the score is a weighted sum of the original variables. The weights are known as loadings and plots of the loadings are called loading plots.

This can now be generalized to m dimensions. In the m-dimensional space, PC1 is obtained as the axis of largest variation in the data, PC2 is orthogonal to PC1 and is drawn into the direction of largest remaining variation around PC1. It therefore contains less variation (and information) than PC1. PC3 is orthogonal to the plane of PC1 and PC2. It is drawn in the direction of largest variation around that plane, but contains less variation than PC3. In the same way PC4 is orthogonal to the hyperplane PC1, PC2, PC3 and contains still less variation, *etc.* For a matrix with dimensions n x m, N = min (n, m) PCs can be extracted. However, since each of them contains less and less information, at a certain time they contain only noise and the process can be stopped before reaching N. If only d << N PCs are obtained, then feature reduction is achieved.

A very important application of principal components is to visually display the information present in the data set and most multivariate data applications start therefore with score and/or loading plots. The score plots give information about the objects and the loading plots about the variables. Both can be combined into a biplot, which are all the more effective after certain types of data transformation, *e.g.* spectral mapping [36]. In Fig. 9 a score plot is shown for an investigation into the Maillard reaction, a reaction between sugars and amino acids [37]. The samples consist of reaction mixtures of different combinations of sugars and amino acids. The variables are the areas under the peaks of the reaction mixtures. The reactions are very complex: 159 different peaks were observed. Each of the samples is therefore characterized by its value for 159 variables. The PC1-PC2 score plot of Fig. 9 can be seen as a projection of the samples from 159-dimensional space to the two-dimensional space that best preserves the variance in the data. In the score plot different symbols are given to the samples according to the sugar that was present and it is ob-

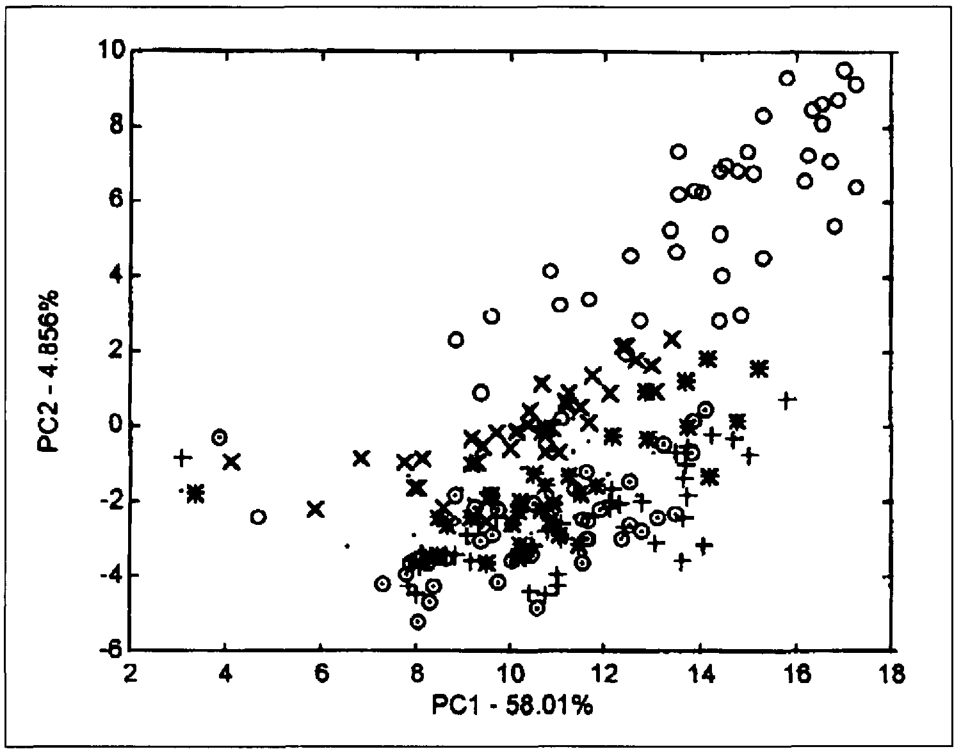

Fig. 9. PCA score plot of samples from the Maillard reaction. The samples with rhamnose have symbol O.

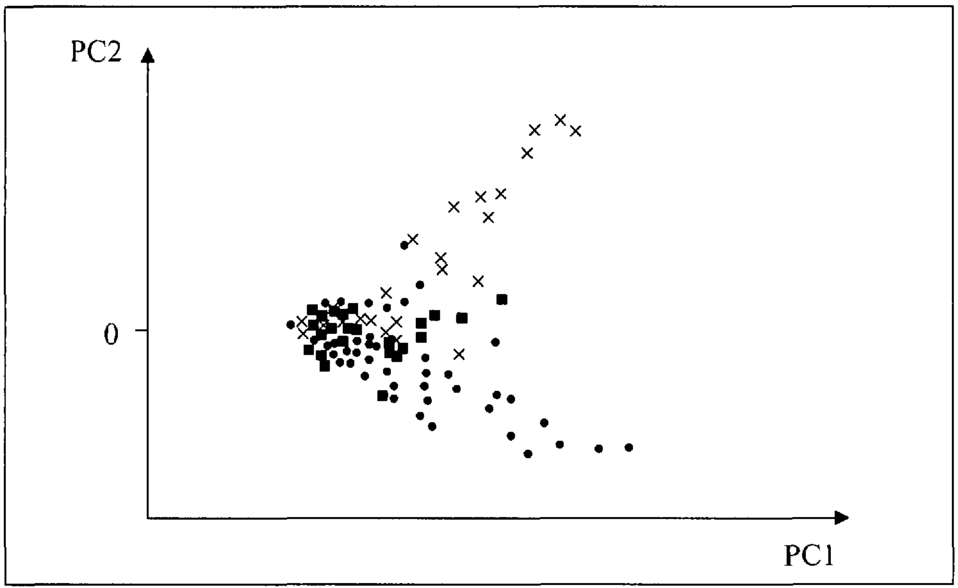

Fig. 10. PCA score plot of air samples.

served, for instance, that samples with rhamnose occupy a specific location in the score plot. This is only possible if they also occupy a different place in the original 159-dimensional space, *i.e.* their GC chromatogram is different. By studying different parts of the data and by including the information from the loading plots, it is then possible to understand the effect of the starting materials on the reaction mixture obtained.

Principal components have been used in many different fields of application. Whenever a table of samples x variables is obtained and some correlation between the variables is expected, a principal components approach is useful. Let us consider an environmental example [38]. In Fig. 10 the score plot is shown. The data consist of air samples taken at different times in the same sampling location. For each of the samples a capillary GC chromatogram was obtained. The different symbols given to the samples indicate different wind directions prevailing at the time of sampling. Clearly the wind direction has an effect on the sample compositions. To understand this better Fig. 11 gives a plot of the loadings of a few of the variables involved. It is observed that the loadings on PC1 are all positive and not very different. Referring to Eqn. (5), and remembering that the loadings are the weights (the w-values) this means that the score on PC1 is simply a weighted sum of the variables and therefore a global indicator of pollution. The samples with highest score on PC1 are those with the highest degree of pollution. Along PC2 some variables have positive loadings and others negative loadings. Those of the aliphatic variables are positive and those of the aromatic variables are negative. It follows that samples with positive scores contain more aliphatic than aromatic variables.

Combining PC1 and PC2, one can then conclude that samples with symbol x have an aliphatic character and that the total content increases with higher values on PC1. The same reasoning can be held for the samples with symbol •: they have an aromatic character. In fact, one could define new aliphaticity and aromaticity factors as in Fig. 12. This can be done in a more formal way using what is called *factor analysis*.

Other Latent Variables

There are other types of latent variables. In projection pursuit [37][39] a latent variable is chosen such that, instead of largest variation in the data set, it describes the largest inhomogeneity. In this way clusters or outliers can be observed more easily. Fig. 13 shows the result applied to the Maillard data of Fig. 9 and it can be observed that the cluster of rhamnose samples can now be observed more clearly.

If the y-values are not characteristics observed for a set of samples, but the class affiliation of the samples (*e.g.* samples 1–10 belong to class A, samples 11–25 to class B), then a latent variable can be defined that describes the largest discrimination between the classes. Such latent variables are called *canonical variates* or sometimes *linear discriminant functions* and are the basis for *supervised pattern recognition* methods such as linear discriminant analysis. In the partial least squares (PLS) section, a further type of latent factor will be introduced.

N-way Methods

Some data have a more complex structure than the classical 2-way matrix or table. Typical examples are met, for instance, in environmental chemistry [40]. A set of n variables can be measured

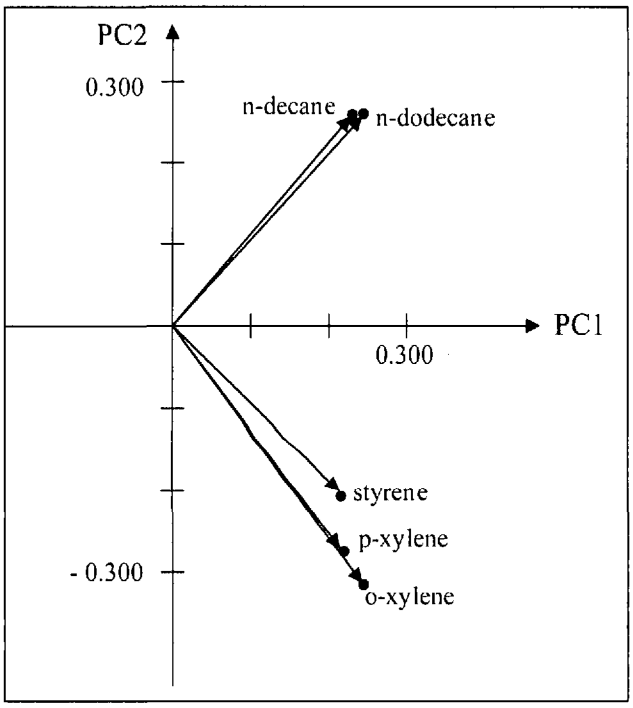

Fig. 11. PCA loading plot of a few variables measured on the air samples in Fig. 12. New fundamental factors discovered on a score plot.

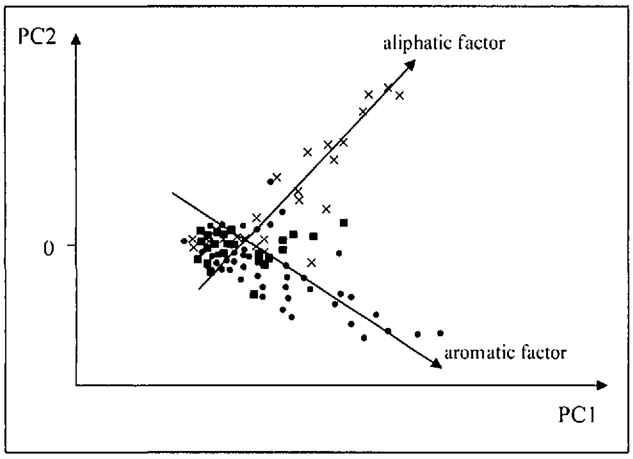

Fig. 12.

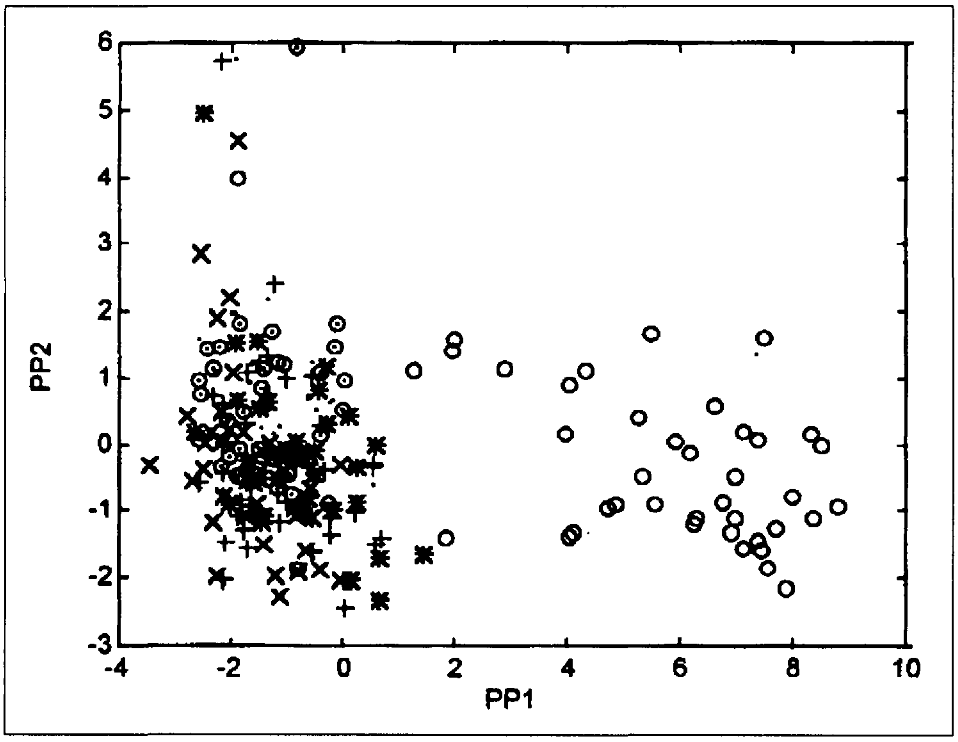

Fig. 13. Projection pursuit plot of samples from the Maillard reaction. The samples with rhamnose have symbol O.

in m different locations at p different times. This leads to a 3-way data set with dimensions n x m x p. The three ways (or modes) are the variable mode, the location mode and the time mode. This can of course be generalized to a higher number of modes, but for the sake of simplicity we will restrict here to 3-way. The classical approach to study such data is to perform what is called unfolding. Unfolding consists of rearranging a 3-way matrix into a 2-way matrix. The 3-way array can be considered as several 2-way tables (slices of the original matrix), and these tables can be put next to each other, leading to a new 2-way array (Fig. 14). This rearranged matrix can be treated with PCA. Considering the example of Fig. 14, the scores will carry information about the locations, and the loadings mixed information about the two other modes.

Unfolding can be performed in different directions so that each of the three modes is successively preserved in the unfolded matrix. In this way, three different PCA models can be built, the scores of each of these models giving information about one of the modes. This approach is called the Tuckerl model. It is the first of a series of Tucker models [41]. The most important of these is the Tucker3 model. Tucker3 is a true n-way method as it takes into account the multiway structure of the data. It consists in building, through an iterative process, a score matrix for each of the modes, and a core matrix defining the interactions between the modes. As in PCA, the components in each mode are constrained to be orthogonal. The number of components can be different in each mode. A graphical representation of the Tucker3 model for 3-way data is given in Fig. 15. It appears as a sum, weighted by the core matrix **G**, of outer products between the factors stored as columns in the **A**, **B** and **C** score matrices.

Another common n-way model is the Parafac-Candecomp model that was proposed simultaneously by Chan and Harchman [42][43]. Information about n-way methods (and software) can be found in [44–46]. Applications in process control [47][48], environmental chemistry [40][49], food chemistry [50], curve resolution [51] and several other fields have been published.

Principal Component Regression (PCR)

Until now we have applied latent variables only for display purposes. Princi-

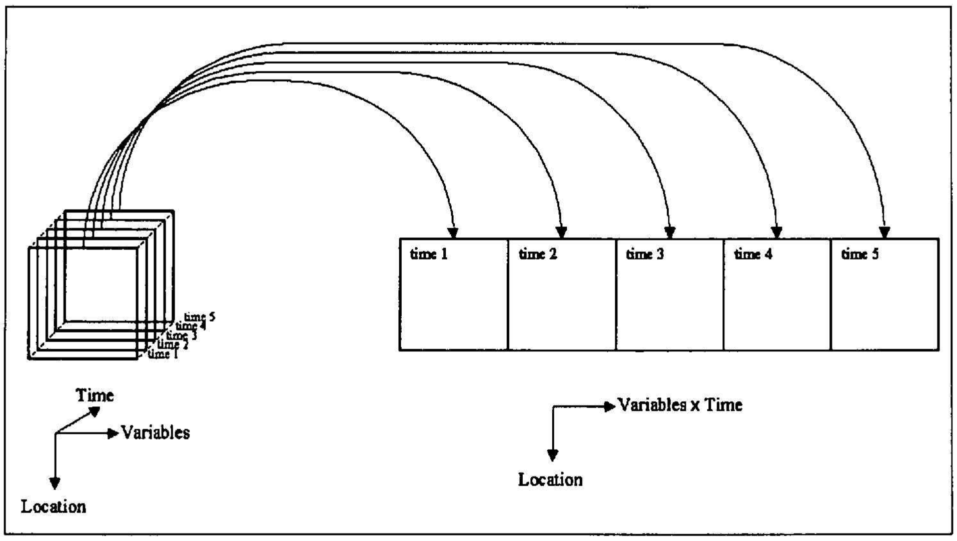

Fig. 14. Unfolding of a 3-way matrix, performed preserving the 'Location' dimension.

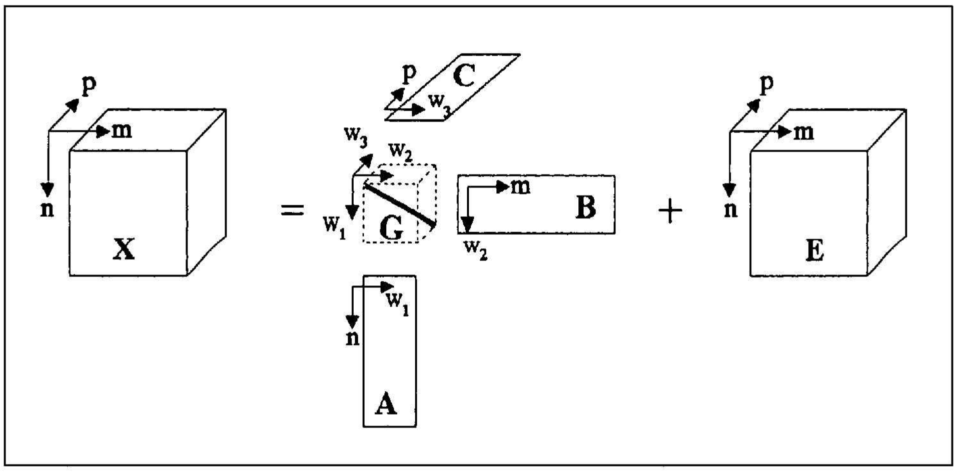

Fig. 15. Graphical representation of the Tucker3 model. n, m, and p are the dimensions of the original matrix **X**. w_1, w_2, and w_3 are the number of components extracted on mode 1, 2 and 3, respectively, corresponding to the number of columns of the loading matrices **A**, **B** and **C**, respectively.

pal components can however also be used as the basis of a regression method. It is applied among others when the x-values constitute a wide X-matrix, *e.g.* for NIR calibration (see earlier). Instead of the original x-values one applies the reduced ones, the scores. Suppose m variables (*e.g.* 1000) were measured for n samples (e.g. 100). As explained earlier this requires either feature selection or feature reduction. The latter can be achieved by replacing the m x-values by the scores on the k significant PC scores (*e.g.* 5). The X matrix now no longer consists of 100 x 1000 absorbance values but of 100 x 5 scores since each of the 100 samples is now characterized by five scores instead of 1000 variables. The regression model is:

$$y=a_1s_1 + a_2s_2+\ldots+a_5s_5 \qquad (6)$$

Since:

$$s=w_1x_1 + w_2x_2+ \ldots w_{1000}x_{1000} \qquad (7)$$

Eqn (6) becomes:

$$y=b_1x_1 + b_2x_2 + \ldots b_{1000}x_{1000} \qquad (8)$$

By using the principal components as intermediates it is therefore possible to solve the wide **X** matrix regression problem. It should be noted also that the principal components are by definition not correlated, so that the correlation problem mentioned earlier is therefore also solved.

Partial Least Squares (PLS)

The aim of PLS is the same as that of PCR, namely to model a set of y-values with the data contained in an (often) wide matrix of correlated variables. However the approach is different. In PCR one works in two steps. In the first the scores are obtained and only the **X** matrix is involved, in the second y is related to the scores. In PLS this is done in one step. The latent variables are obtained, not with the variation in **X** as criterion as is the case for principal components, but such that the new latent variable shows maximal covariance between **X** and y. This means that the latent variable is now built immediately in function of the relationship between y and **X**. In principle one therefore expects that PLS would perform better than PCR, but in practice they often perform equally well. A tutorial can be found in [52]. Several algorithms are available. A very performant one requiring the least computer time according to our experience is SIMPLS [53].

Applications of PCR and PLS

PCR and PLS have been applied in many different fields. The following references constitute a somewhat haphazard selection from a very large literature. There are many analytical applications in the pharmaceutical industry [54], the petroleum industry [55], food science [56], environmental chemistry [57]. The methods are used with near or mid infrared [58], chromatographic [59], Raman [60], UV [61], potentiometric [62] data. A good overview of applications in QSAR is found in [63].

PLS2 and Other Methods that Describe the Relationship between Two Tables

Instead of relating one y-value to many x-values, it is possible to model a set of y-values with a set of x-values. This means that one relates two matrices **Y** and **X**, or in other words two tables. For instance, one could measure for a certain set of samples a number of sensory characteristics on the one hand and obtain analytical measures on the other. This would yield two tables as depicted in Fig. 16. One could then wonder if it is possible to predict the sensory characteristics from the (easier to measure) chemical measurements or at least to understand which (combinations) of analytical measurements are related to which sensory characteristics. At the same time one wants to obtain information about the

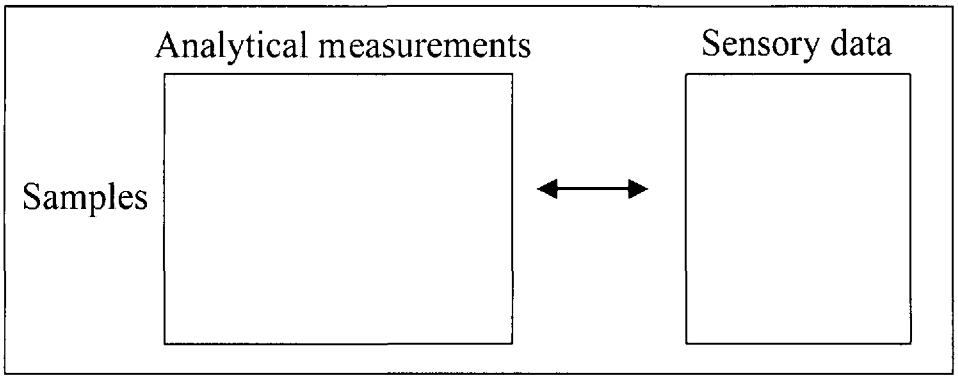

Fig. 16. Relating two 2-way tables.

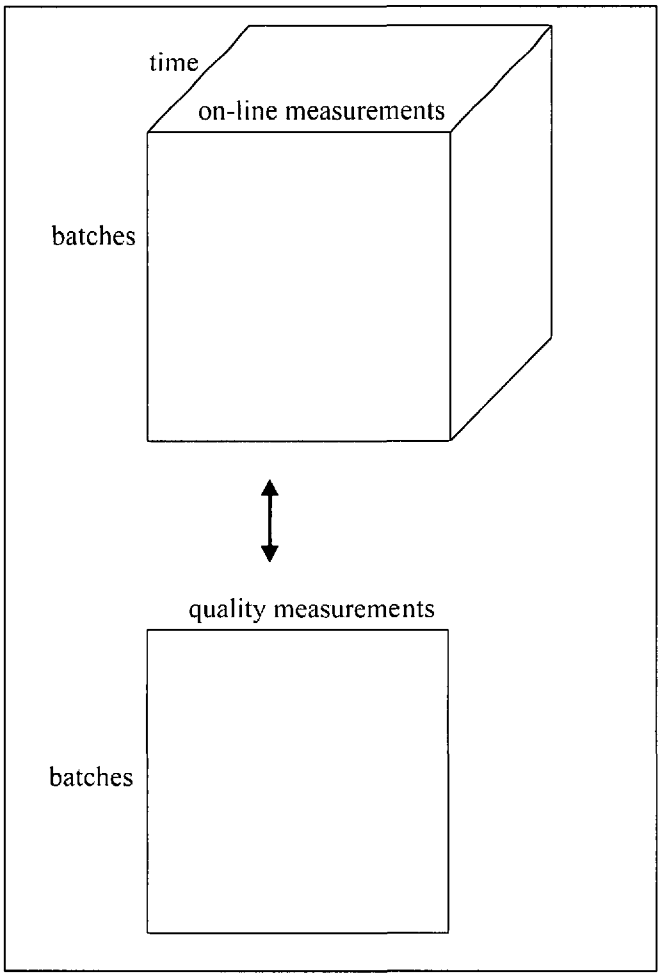

Fig. 17. Relating a two-way and a three-way table.

Fig 17. In process analysis, one is concerned with the quality of finished batches and this can be described by a number of quality parameters. At the same time for each batch a number of variables can be measured on the process in function of time [68]. This yields a two-way table on the one hand and a three-way one on the other. Relating these tables allows to predict the quality of a batch from the measurements made during the process.

structure of each of the two tables (*e.g.* which analytical variables give similar information). PLS2 can be used for this purpose. Other methods that can be applied are, for instance, canonical correlation and reduced rank regression. An example relating 20 measurements of mechanical strength of meat patties to the sensory evaluation of textural attributes can be found in [64] and a comparison of methods in [65].

Generalization

It is also possible to relate multi-way models to a vector of y-values or to 2-way tables. The same way as with 2-way data, the latent variables obtained in multi-way models are then used to build the regression models. The multi-way analog to PCR would consist in modeling the original data with Tucker3 or Parafac, and then regress the dependent y variable on the obtained scores. A more sophisticated n-way version of PLS (N-PLS) was also developed. The principle of N-PLS is to fit a model similar to Parafac, but aiming at maximizing the covariance between the dependent and independent variables instead of fitting a model in a least squares sense. The usefulness of such approaches will be apparent from

Acknowledgements

Y. Vander Heyden is a postdoctoral fellow of the Fund for Scientific Research (FWO) – Vlaanderen.

Received: December 21, 2000

[1] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. de Jong, P.J. Lewi, J. Smeyers-Verbeke, 'Handbook of Chemometrics', Elsevier, Amsterdam, **1997**.

[2] N.R. Draper, H. Smith, 'Applied Regression Analysis', Wiley, New York, **1981**.

[3] J. Mandel, 'The Statistical Analysis of Experimental Data', Dover reprint, **1984**, Wiley&Sons, New York, **1964**.

[4] D.L. MacTaggart, S.O. Farwell, *J. Assoc. Off. Anal. Chem.* **1992**, *75*, 594.

[5] J.C. Miller, J.N.Miller, 'Statistics for Analytical Chemistry', Ellis Horwood, Chichester, 3rd ed., **1993**.

[6] W.E. Deming, 'Statistical Adjustment of Data', Wiley, New York, **1943**.

[7] P.T. Boggs, C.H. Spiegelman, J.R. Donaldson, R.B. Schnabel, *J. Econometrics* **1988**, *38*, 169.

[8] P.J.Cornbleet, N.Gochman, *Clin. Chem.* **1979**, *25*, 432.

[9] C. Hartmann, J. Smeyers-Verbeke, D.L. Massart, *Analusis* **1993**, *21*, 125.

[10] J. Riu, F.X. Rius, *J. Chemometr.* **1995**, *9*, 343.

[11] R.G. Krutchkoff, *Technometrics* **1967**, *9*, 425.

[12] V. Centner, D.L. Massart, S. de Jong, *Fresenius J. Anal.Chem.* **1998**, *361*, 2.

[13] B. Grientschnig, *Fresenius J. Anal.Chem.* **2000**, *367*, 497.

[14] H. Theil, *Nederlandse Akademie van Wetenschappen Proc., Scr. A* **1950**, *53*, 386.

[15] P.J. Rousseeuw, A.M. Leroy, 'Robust Regression and Outlier Detection', Wiley, New York, **1987**.

[16] G.R. Phillips, E.R. Eyring, *Anal. Chem.* **1983**, *55*, 1134.

[17] F. Mosteller, J.W. Tukey, 'Data Analysis and Regression', Addison-Wesley, Reading, **1977**.

[18] P. Van Keerberghen, J. Smeyers-Verbeke, R. Leardi, C.L. Karr, D.L. Massart, *Chemom. Intell. Lab. Syst.* **1995**, *28*, 73.

[19] H. Kubinyi, *Quant. Struct.-Act. Relat.* **1994**, *13*, 285.

[20] J.G. Topliss, R.J. Costello, *J. Med. Chem.* **1972**, *15*, 1066.

[21] M. Sergent, D. Mathieu, R. Phan-Tan-Luu, G. Drava, *Chemom. Intell. Lab. Syst.* **1995**, *27*, 153.

[22] A.C. Atkinson, *J. Am. Stat. Soc.* **1994**, *89*, 1329.

[23] S. Morgenthaler, M.M. Schumacher, *Chemom. Intell. Lab. Syst.* **1999**, *47*, 127.

[24] R.D. Cramer III, D.E. Patterson, J.D. Bunce, *J. Am. Chem. Soc.* **1988**, *110*, 5959.

[25] J.H. Holland, 'Adaption in Natural and Artificial Systems', University of Michigan Press, Ann Arbor, MI, **1975**, revised reprint, MIT Press, Cambridge, **1992**.

[26] C.B. Lucasius, M.L.M. Beckers, G. Kateman, *Anal. Chim. Acta* **1994**, *286*, 135.

[27] R. Leardi, R. Boggia, M. Terrile, *J. Chemom.* **1992**, *6*, 267.

[28] J. Devillers ed., 'Genetic Algorithms in Molecular Modeling', Academic Press, London, **1996**.

[29] M.L.M. Beckers, E.P.P.A. Derks, W.J. Melssen, L.M.C. Buydens, *Comput. Chem.* **1996**, *20*, 449.

[30] D. Jouan-Rimbaud, D.L.Massart, R. Leardi, O.E. de Noord, *Anal. Chem.* **1995**, *67*, 4295.

[31] R. Meusinger, R. Moros, *Chemom. Intell. Lab. Syst.* **1999**, *46*, 67.

[32] P. Willet, *Trends. Biochem*, **1995**, *13*, 516.

[33] D.H. Hibbert, *Chemom. Intell. Lab. Syst.* **1993**, *19*, 277.

[34] J.H. Kalivas, *J. Chemom.* **1991**, *5*, 37.

[35] X.G. Shao, Z.H. Chen, X.Q. Lin, *Fresenius J. Anal. Chem.* **2000**, *366*, 10.

[36] P.J. Lewi, *Arzneim. Forschung* **1976**, *26*, 1295.

[37] Q. Guo, W. Wu, F. Questier, D.L.Massart, C. Boucon, S. de Jong, *Anal. Chem.* **Year?** *72*, 2846.

[38] J. Smeyers-Verbeke, J.C. Den Hartog, W.H. Dekker, D. Coomans, L. Buydens, D.L. Massart, *Atmos. Environ.*, **1984**, *18*, 2471.

[39] J.H. Friedman, *J. Am. Stat. Soc.* **1987**, *82*, 249.

[40] P. Barbieri, C.A. Andersson, D.L. Massart, S. Predonzani, G. Adami, G.E. Reisenhofer, *Anal. Chim. Acta* **1999**, 398, 227.

[41] L.R. Tucker, *Psychometrika* **1966**, *31*, 279.

[42] R. Harshman, *UCLA working papers in phonetics* **1970**, *16*, 1.

[43] J.D. Carrol, J. Chang, *Psychometrika*, **1970**, *45*, 283.

[44] C.A. Andersson, R. Bro, *Chemom. Intell. Lab. Syst.* **2000**, *52*, 1.

[45] M. Kroonenberg, 'Three-mode Principal Component Analysis. Theory and Applications', DSWO Press, Leiden, **1983**, reprint **1989**.

[46] R. Henrion, *Chemom. Intell. Lab. Syst.* **1994**, *25*, 1.

[47] P. Nomikos, J.F. MacGregor, *AIChE Journal*, **1994**, *40*, 1361.

[48] D.J. Louwerse, A.K. Smilde, *Chem. Eng. Sci.* **2000**, *55*, 1225.

[49] R. Henrion, *Chemom. Intell. Lab. Syst.* **1992**, *16*, 87.

[50] R. Bro, *Chemom. Intell. Lab. Syst.* **1998**, *46*, 133.

[51] A. de Juan, S.C. Rutan, R. Tauler, D.L. Massart, *Chemom. Intell. Lab. Syst.* **1998**, *40*, 19.

[52] P. Geladi, B.R. Kowalski, *Anal. Chim. Acta* **1986**, *185*, 1.

[53] S. de Jong, *Chemom. Intell. Lab. Syst.* **1993**, *18*, 251.

[54] K.D. Zissis, R.G. Brereton, S. Dunkerley, R.E.A. Escott, *Anal. Chim. Acta* **1999**, *384*, 71.

[55] C.J. de Bakker, P.M. Fredericks, *Appl. Spect.* **1995**, *49*, 1766.

[56] S. Vaira, V.E. Mantovani, J.C. Robles, J.C. Sanchis, H.C. Goicoechea, *Anal. Lett.* **1999**, *32*, 3131.

[57] V. Simeonov, S. Tsakovski, D.L. Massart, *Toxicological & Environmental Chemistry*, **1999**, *72*, 81.

[58] J.B. Cooper, K.L. Wise, W.T. Welch, M.B. Summer, B.K. Wilt, R.R. Bledsoe, *Appl. Spect.* **1997**, *51*, 1613.

[59] M.P. Montana, N.B. Pappano, N.B. Debattista, J. Raba, J.M. Luco, *Chromatographia* **2000**, *51*, 727.

[60] O. Svensson, M. Josefson, F.W. Langkilde, *Chemom. Intell. Lab. Syst.* **2000**, *49*, 49.

[61] F. Vogt, M. Tacke, M. Jakusch, B. Mizaikoff, *Anal. Chim. Acta* **2000**, *422*, 187.

[62] M. Baret, D.L. Massart, P. Fabry, C. Menardo, F. Conesa, *Talanta* **1999**, *50*, 541.

[63] S. Wold in 'Chemometric Methods in Molecular Design', Ed.: H. van de Waterbeemd, VCH, Weinheim, **1995**.

[64] S. Beilken, L.M. Eadie, I. Griffiths, P.N. Jones, P.V. Harris, *J. Food Sci.* **1991**, *56*, 1465.

[65] B.G.M. Vandeginste, D.L. Massart, L.M.C. Buydens, S. de Jong, P.J. Lewi, J. Smeyers-Verbeke, 'Handbook of Chemometrics and Qualimetrics' Part B, Chapter 35, Elsevier, Amsterdam, **1998**.

[66] R. Bro, H. Heimdal, *Chemom. Intell. Lab. Syst.* **1996**, *34*, 85.

[67] R. Bro, *J. Chemom.* **1996**, *10*, 47.

[68] C. Duchesne, J.F. McGregor, *Chemom. Intell. Lab. Syst.* **2000**, *51*, 125.