DRUG DEVELOPMENT AND INDUSTRIAL PHARMACY, 16(17), 2527-2545 (1990)

APPLICATIONS OF CHEMOMETRICS FOR CHARACTERIZATION OF MACROMOLECULES

Anders Hagman and Sven Jacobsson
KABI Pharma, Home Market Products,
Research and Development Department, Box 1828,
S-171 26 Solna, Sweden.

ABSTRACT

The possibilities of employing methods of chemometrics in order to characterize macromolecules are described. The review has been limited to chemometric methods concerning multivariate data analysis. Principal component analysis (PCA) has shown to be very useful for pattern recognition problems arising from chromatographic and spectroscopic data. An example of using a classification technique, SIMCA (Soft Independent Modelling of Class Analogy), as a product control method is presented. The suitability of Partial Least Squares (PLS) for relating data of different natures, e.g. chemical data with biological data, is discussed. Moreover, examples ranging from spectroscopic determinations to QSAR (Quantitative Structure Activity Relationships) are illustrated.

INTRODUCTION

Chemometrics has during the latest years become a well established discipline of chemistry and it is defined as the application of mathematical and statistical methods to design or select optimum procedures and experiments and to provide maximum chemical information by analyzing chemical data (1). A good overview of the subject can be provided by two books by Kowalski (2,3) and in recent years there has been a number of overview/review-articles published concerning chemometrics related to pharmaceutical analysis (4-7).

The process of analysis can be divided roughly in two sub-disciplines from the point of view of chemometrics; experimental design/optimization and multivariate data analysis (Fig 1). The first discipline includes the design and implementation of analyses so that the most informative experiments are carried out and the measurement step optimized for maximum information and efficiency.

Experimental design (e.g. factorial design) together with optimization methods (e.g. simplex) are important methods in order to obtain an efficient experiment. Massart and co-workers (ref 8,9) have shown excellent examples for extraction and determination of drugs in formulations and biological samples.

Multivariate data-analysis methods ensure that all available information from the low level data are

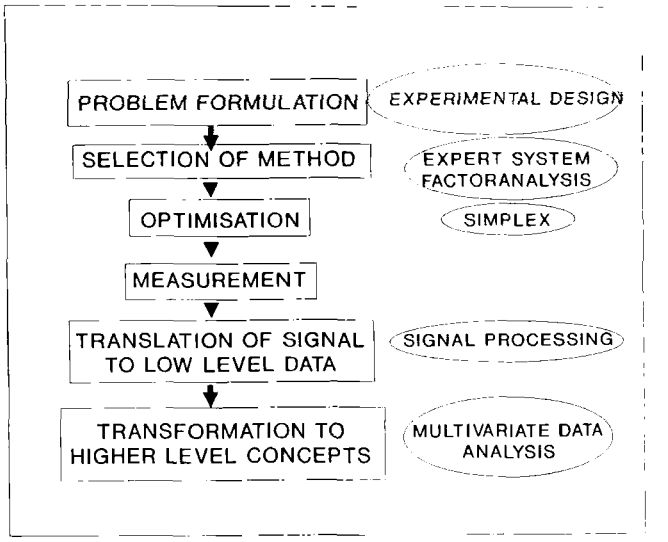


FIGURE 1

Chemometric methods which can be involved in a measurement process.

extracted and transformed to a higher level concept, which is easier to interpret. This is accomplished by analyzing all variables at the same time, compared to traditional ways of changing one variable at a time, which minimizes the risk for spurious correlations. The possibilities of employing the methods of chemometrics in order to characterize the active component or the excipients in the pharmaceutical preparation are many. However, this article will only focus on some multivariate data-analysis methods. Applications of these methods for characterization of macromolecules will be presented and the potential for chemometrics in the pharmaceutical field will be discussed.

PATTERN RECOGNITION

Many methods, such as chromatographic or spectroscopic methods, that are used to characterize the pharmaceutical substance itself or excipients used in the finished formulation often generate a lot of data. Pattern recognition (PARC) can play a role in transforming the complex information into a more systematic form which is easier to interpret. The scope of PARC is to enable visualization of the data-matrix, consisting of e.g. the number of peaks in the chromatogram and the number of samples, by reducing the multidimensional space to a two-dimensional picture. Principal component analysis (PCA) is one of the best known methods of achieving this. To better understand how the method of principal component analysis operates, one can consider a simple situation, where we have only 3 variables, representing the three axes in this data-space (fig 2). Each individual sample/object is represented as a point in this space. In order to "open a window" into this space, a new set of orthogonal coordinate axes, principal components, must be generated. These have their origin at the center of the gravity of the data-set. The first principal component has a direction of the line so that it takes into account as much variance in the data as possible or, to put it another way, so that one looses as little information as possible. The next principal component, orthogonal to the first, has a direction where the second largest variance occurs. The objects are then projected down to the plane of the two principal components. A large data-set may therefore be represented by only a few

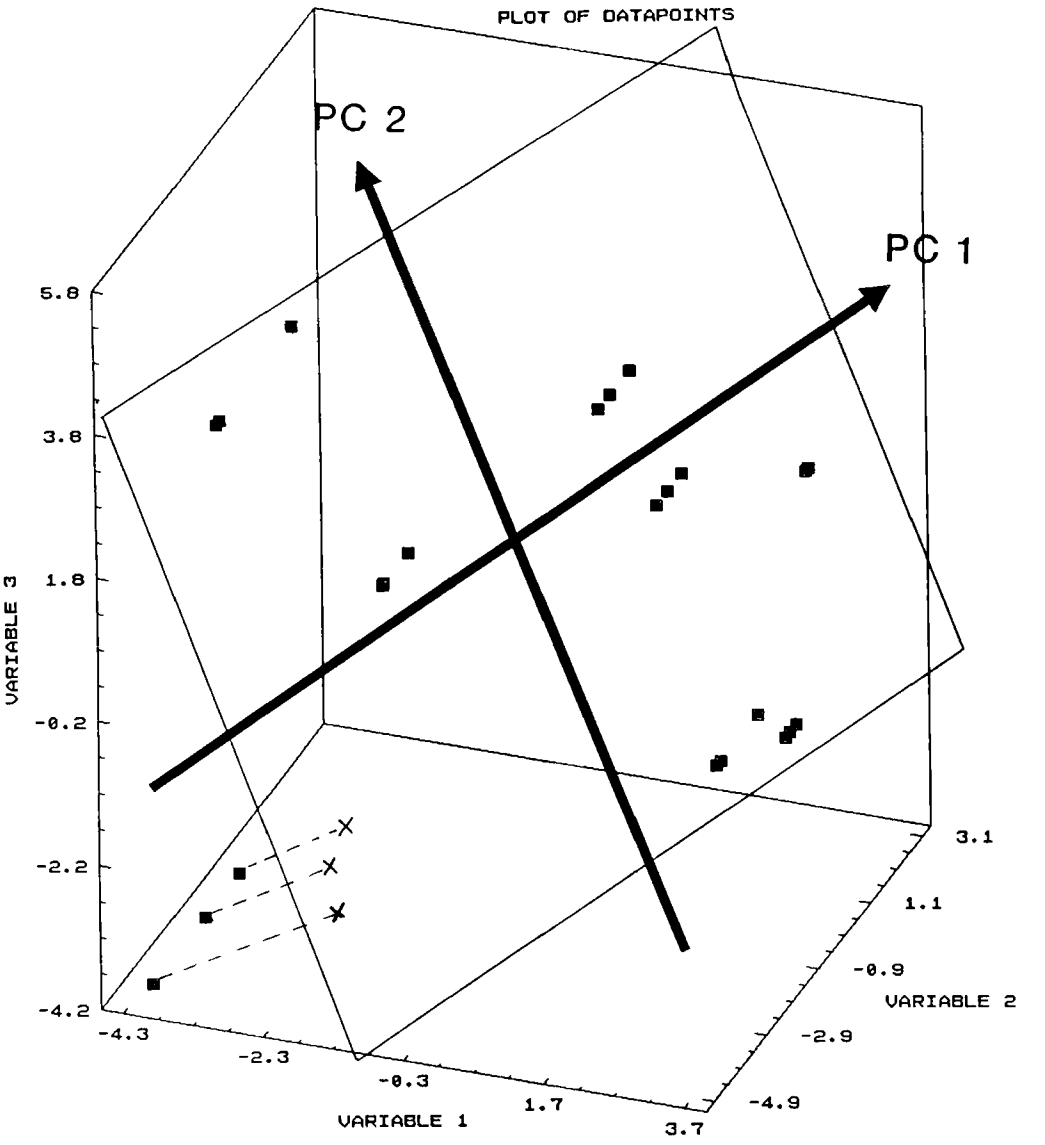


FIGURE 2

The first and second principal component define a plane where the objects can be projected down to.

latent variables, principal components, which describe a large part of the variance in the data as a linear combination of the original data. Clusters, trends or outliers can now easily be detected by a visible inspection of the plot.

The most common application of PARC is using the chromatogram as a pattern (10). Fewster and Walden (11) have used PCA for visualization and compilation of data from electrophorogram of two-dimensional gel electrophoresis of proteins. The variables were chosen by putting a four times four grind pattern over the electrophorogram, where each square represented one variable. With this method a good overview over the data-set was obtained, along with insight about finer details, compared to traditional methods. A similar approach has been used for interpreting electron microscopy images of biological macromolecules (12,13).

The experimental observations were the grind pattern of the elements that represented the digitized image of the molecule projection and each image represented one object. Van Heel and Frank (12) could, with this technique, separate molecule images of hemocyanin into groups on the basis of structure-related differences.

CLASSIFICATION

Classification of samples is one of the goals of pattern recognition. Methods for classification can be divided into unsupervised and supervised approaches The difference between these methods is that the supervised approach requires a reference set (training

set), a set of data with known effects, while unsupervised methods require no prior tests. Cluster analysis is an example of unsupervised approach and Bratchell (14) has given a detailed description of different clustering techniques. One of the most popular methods of supervised classification is SIMCA (Soft Independent Modelling of Class Analogy) (15). The basic strategy of SIMCA is to classify a sample according to a specific property by using several measurements that are directly or indirectly related to that property. An empirical relationship, or model, can then be derived from a set of data (reference set) using samples whose specific property of interest is known. The model consists of principal components, which are assigned for each group of objects or classes. New samples with unknown property can then be tested against the model with traditional statistics (F-tests) and thus classified. A tolerance region can be constructed around the model by means of the residual variance at a certain tolerance level (fig 3). The test-objects are assigned to a class if the degree of fitness between the object data vector and the corresponding class model is sufficiently good, i.e. if the object is on the inside of the tolerance region. It is essential to decide how many principal components adequately describe each class and this is performed by a method called crossvalidation (16).

The SIMCA method has been used to differentiate batches according to sensory qualities, in this case odor, of the final packing product and the changes in polymer pellets production (17). The raw-material, pellets, were analyzed by dynamic headspace /gas

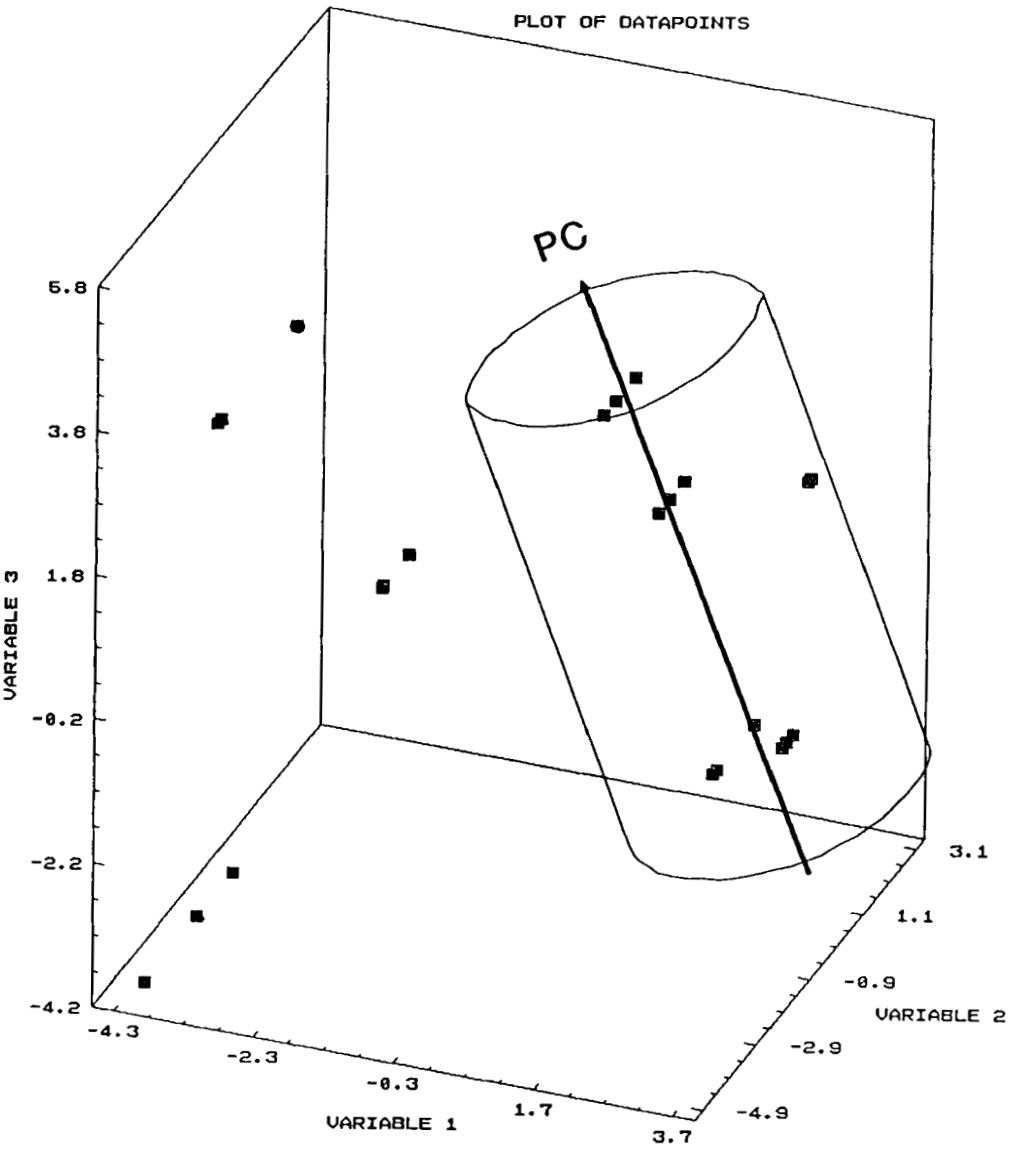


FIGURE 3

Objects are classified in SIMCA as belonging to a class if they are inside the tolerance interval, otherwise as outliers.

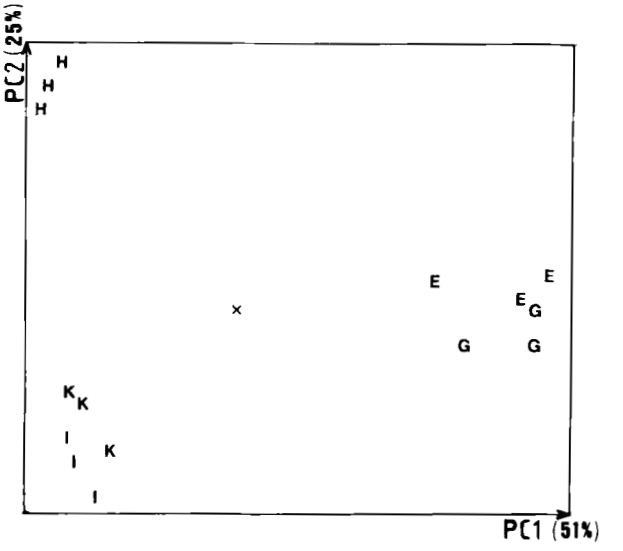


FIGURE 4

Principal component plot of data obtained from Dynamic Headspace/ gas chromatography/ mass spectrometry of Polypropylene pellets. The reference-set was contained of E and G (not acceptable samples from a product quality view) and H, I and K (acceptable batches). (from ref. 17)

chromatography. It was very difficult to determine if a batch was acceptable or not by a visible inspection of the chromatogram because the batches differed from each other, even batches that gave the same odor-results. The chromatographic profile was transferred to a data-matrix and after appropriate scaling, PCA was performed for the reference set (fig 4).

The principal component plot shows three different classes, one non-acceptable and two classes of acceptable. The difference between the acceptable batches was a result of change in processing

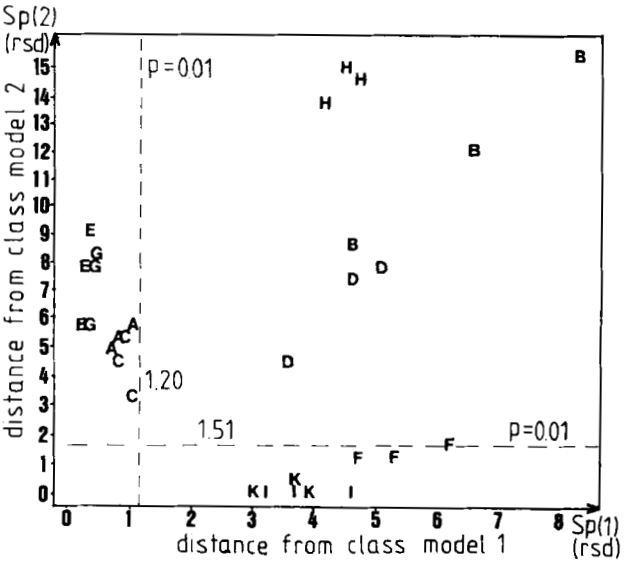


FIGURE 5

A Coomans plot of the SIMCA classification of the polypropylene test-set. Dashed lines indicated the tolerance level at 1 % probability around the classes. (from ref. 17)

conditions for the batches delivered later than the problem batches. All batches in the test-set were correctly classified, see fig 5.

MULTIVARIATE CALIBRATION

Multivariate calibration is applicable to the determination of both major and minor constituents and for a wide range of instrument types (e.g. diode array). The only criteria is that the component is characterized or measured by a number of parameters. One of the best benefits with multivariate calibration is that the need for sample preparation is reduced,

due to the fact that selective input measurements are no longer needed.

Partial Least Squares (PLS) (18) is a very useful multivariate calibration-technique to relate two or more blocks of data with each other, especially as a predicative tool. Compared to multi linear regression (MLR) PLS can handle more than one dependent variable and is not critically influenced by correlation between the independent variables. Furthermore, it can tolerate missing values in the data-matrix. In the PLS method X (independent) variables are related to a block of Y (dependent) through a process where the variance in Y-block influences the calculation of the principal components (PLS-components) of the X-block and vice versa. An iteration of this criss-cross process gives geometrically a tilting of PLS-components for X respectively Y, so that the correlation between principal component for block X and block Y is optimized (fig 6). This is the difference between PLS and PCA, where PCA is optimized with respect to the variance and covariance description and therefore PLS has not the same properties as PCA , eg. orhogonality. It is important that the number of PLS-components are correct so that an overfitting of the model is avoided. This can be done by cross-validation.

Application of PLS to characterize and quantify three different forms of carrageenan in a production batch by using Pyrolysis/Gas chromatography had been demonstrated (19). In this study a number of carrageenan production batches have different properties due to variations in the polymer

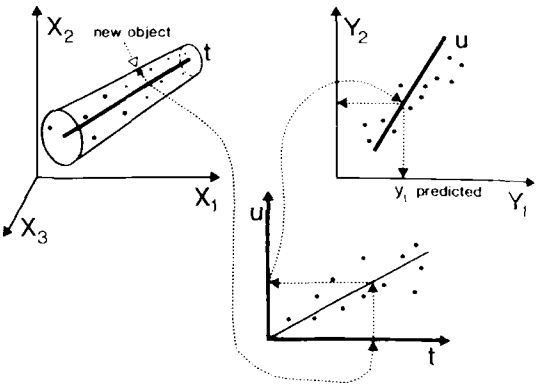
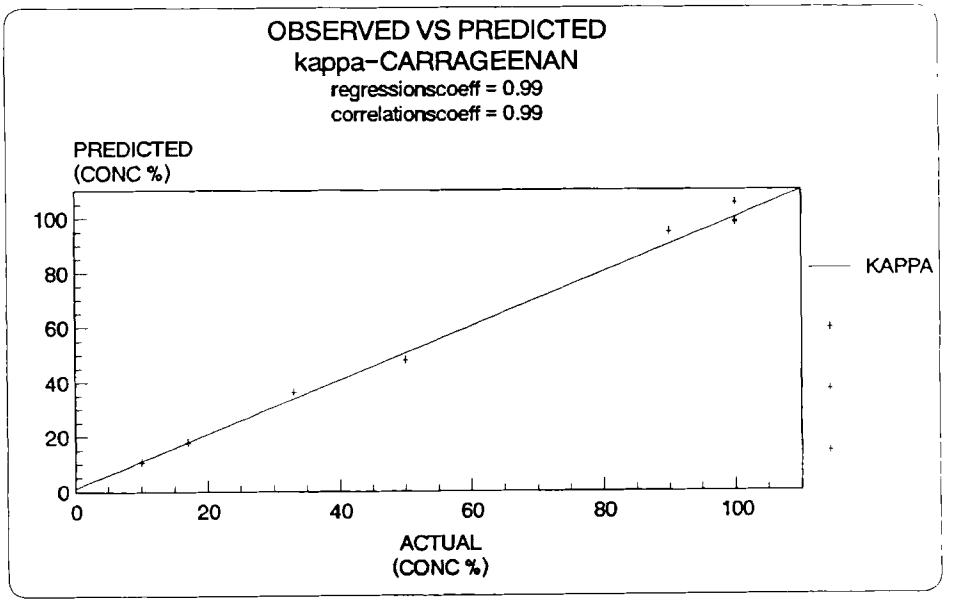


FIGURE 6

Geometrical illustration of the PLS method. The upper left coordinate system is the independent block (X) and the right the dependent block (Y). The principal component (PLS-component) t and u are correlated to each other via an inner relation (lower coordinate system). The dotted lines indicate the path taken when the variables in Y for a new object is predicted.

composition. Therefore it was desirable to determine the structural variations. The chromatographic profile from pyrolysis/gas chromatography reflects indirectly the structure of the compound. However, the results for a production batch were often very hard to interpret, due to overlapping chromatographic profiles of the three different carrageenan forms. The chromatographic profile was used as X-block and the concentration of different carrageenans was Y-block. A very good calibration model was obtained (fig 7) , despite the interferences in the chromatogram. The results were correlated with data from NMR and showned good correspondence.

Spectroscopic methods have often difficulties of finding frequency regions where the constituents of



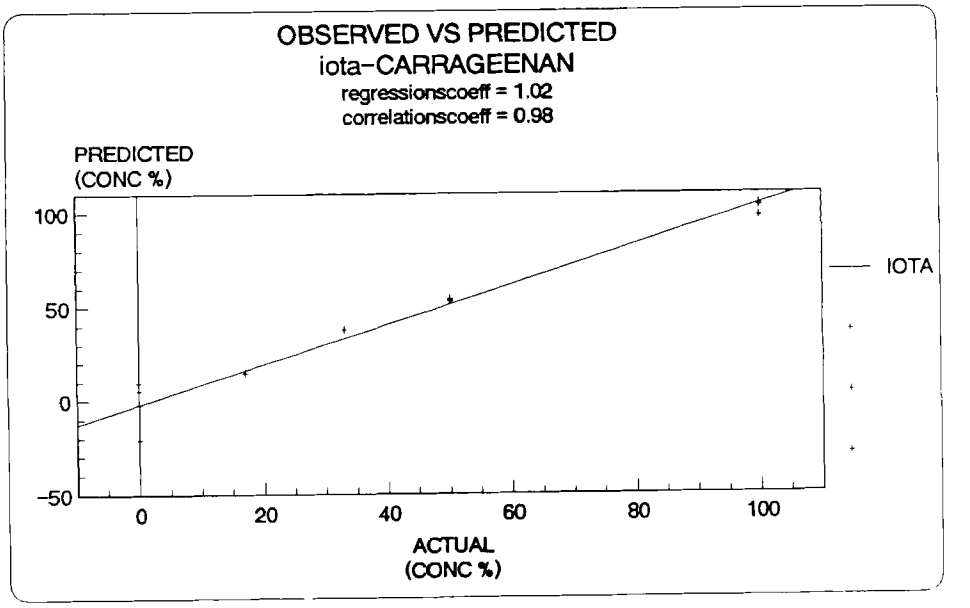


FIGURE 7

PLS-predicted versus actual calibration results from pyrolysis/gas chromatography of two carrageenan forms together with the regression line. (from ref 19)

interest selectively absorb or emit light. However, in combination with multivariate data-analysis the whole spectra can be analyzed simultaneously, which gives a more selective and accurate quantitative determination. Hartauer and Guillory (20) have used FTIR/ATR/PLS for a simultaneous determination of two components in an intravenous pharmaceutical formulation. Despite the fact that the spectra for the two formulations, one with and one without the two active substances, were almost identical, PLS could extract the useful information for an accurate calibration. The most informative parts of the data are often those containing the relation between the variables, so called covariances.

Another possibility with PLS is that you can determine the importance of each variable for the correlation with the calibration-model. An important variable contributes very much to the direction of the PLS-component. Bonelli et. al. (21) have been studying the relationship between the dissolution rates in water of Griseofulvin and chemical and physical properties of a number of polymer carriers, e.g. carrageenans and polyethylene glycol. By using PLS, the factors that contributed the most to the releasing rate could be determined; that were the degree of crystallinity of Griseofulvin and the amount water absorbed by the powder sample (fig 8).

An area where chemometric methods have been very successful, is QSAR (Quantitative Structure Activity Relationships). In a number of publications the use of PLS to characterize polypeptides for their biological

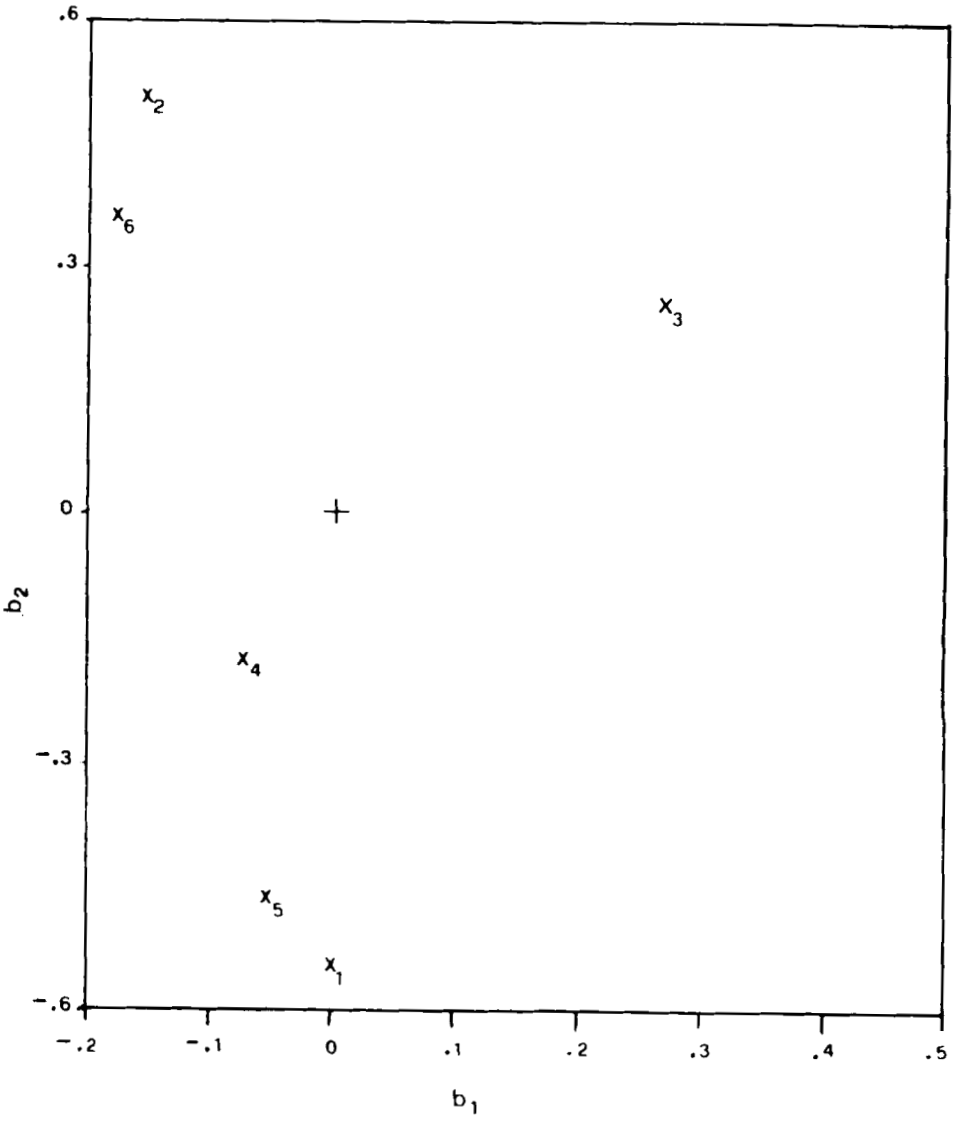


FIGURE 8

Loading plot indicating the relative information content of some physicochemical properties of a serie polymers for the dissolution rate of grieofulvin in water. The important variables for the dissolution rate are those who have the largest distance from origin in Y-direction. x_1= degree of crystallinity of griseofulvin, x_2= wetting of samples by water, x_3= viscosity of the polymer solution, x_4= pH, x_5= solubilizing effect of polymer on griseofulvin, x_6= dissolution rate of the polymer in water. (from ref 21)

activity by measuring the physicochemical properties of the amino acids are described (22,23). Wold et. al. (22) developed a multivariate PLS model to describe principal property scales of amino acids and extended these models to uncoded amino acids.

CONCLUSIONS

Chemometrics can have an increasing impact on the characterization of macromolecules and can be applied to almost every chemical situation. Chromatographic fingerprinting of polymers, electron micrographic images and interfering spectra can give rise to very complicated data which are difficult to interpret. Complex data can be more easy to interpret by just making them "visible" in a more systematic form. This can be achieved by reducing the dimensionality of the data-matrix by means of principal component analysis.

Quality control of raw materials, intermediates and final products is one of the crucial points in the development of a pharmaceutical product. Therefore, by using classification-methods, e.g. SIMCA, it is feasible to classify the type of polymer sample or it's source as acceptable/non-acceptable from a qualitative point of view.

In many situations, it is necessary to correlate a block of data, e.g chemical data, with another, e.g. biological data, or to predict a value of one block by using the data from the other block, which is easier to measure. Partial Least Squares (PLS), a multivariate calibration/regression method, has shown

to be a powerful tool for this kind of problems. However, it must be stressed that a chemometric method itself cannot solve problems which origin from a bad raw-data, but it can give a better opportunity to find it out. The authors hope that this survey of some chemometric methods can give some ideas of using them for other applications and show the benefits of a chemometric approach.

REFERENCES

(1) B.R. Kowalski, Trends Anal. Chem., 1, 71 (1981)

(2) B.R. Kowalski, "Chemometrics-Mathematics and Statistics in Chemistry", NATO ASI Series C, vol 138, Reidel, Germany, 1984

(3) M.A. Sharaf, D.L. Illman and B.R Kowalski, "Chemometrics", Wiley, New York, 1986

(4) J.C. Berridge, Analyst, 112, 385 (1987)

(5) B.V. Fisher and R. Jones, J. Pharm. Biomed. Anal., 5, 455 (1987)

(6) D.L. Massart and L. Buydens, J. Pharm. Biomed. Anal., 6, 535 (1988)

(7) J.C. Berridge, Anal. Chim. Acta, 223, 149 (1989)

(8) G. Hoojewijs and D.L. Massart, J. Pharm. Biomed. Anal. , 2, 449 (1984)

(9) M. De Smet, G. Hoojewijs, M. Puttemans and D.L. Massart, Anal. Chem., 56, 2662 (1984)

(10) P.C. Jurs, B.K. Lavine and T.R. Stouch, J. res. Natl. Bur. Stand., 90, 543 (1985)

(11) P.H. Fewster and D.B. Walden, Comput. Biol. Med., 17, 29 (1987)

(12) M. Van Heel and J. Frank, Ultramicroscopy, 6, 187 (1981)

(13) J. Frank, Ultramicroscopy, 9, 3 (1982)

(14) N. Bratchell, Chemometrics Intell. Lab. Syst., 6, 105 (1989)

(15) C. Albano, W. Dunn III, U. Edlund, E Johansson, B. Norden, M Sjöström and S. Wold, Anal. Chim. Acta, 103, 429 (1978)

(16) S. Wold, Technometrics, 20, 397, (1978)

(17) A. Hagman, K. Karlsson and S. Jacobsson, J. High Resolut. Chromatogr., Chromatogr. Commun., 11, 46 (1988)

(18) R.W. Gerlach, R.R. Kowalski and H. Wold, Anal. Chim. Acta, 112, 417 (1979)

(19) A. Hagman, in preparation

(20) K.J. Hartauer and J.K. Guillory, Pharm. Research, 6, 608, (1989)

(21) D. Bonelli, S. Clementi, C. Ebert, M. Lovrecich and F. Rubessa, Drug Dev. Ind. Pharm., 15, 1375 (1989)

(22) S. Wold, L. Eriksson, S. Hellberg, J. Jonsson, M. Sjöström, B. Skagerberg and C. Wikström, Can. J. Chem., 65, 1814 (1987)

(23) B. Skagerberg, M. Sjöström and S. Wold, Quant. Struct.-Act. Relat., 6, 158 (1987)