APPLICATION OF MULTIVARIATE ANALYSIS IN PHARMACEUTICAL DEVELOPMENT WORK

N-O Lindberg* and T Lundstedt**
Pharmacia AB

* Pharmaceutical Research and Development
Oncology Immunology
P.O. Box 941
S-251 09 Helsingborg
Sweden
Correspondence

** Structure-Property
Optimization Centre
P.O. Box 839
S-201 80 Malmö
Sweden

INTRODUCTION

Regression is a valuable technique when analyzing the relationship between a dependent variable (response, Y-variable) and independent variables (X-variables, objects, factors). Multivariate analysis (MVA) looks for interdependence among all variables. This entails extracting information from data with many variables and using all the variables simultaneously. MVA is concerned with ways of obtaining information out of existing multivariate data, a process which often involves the use of matrices. Consequently, MVA is a suitable approach when investigating complicated relationships.

This survey mainly deals with principal components analysis (PCA), partial least squares projection to latent structures (PLS), and classification techniques.

MVA has been used in chemistry for many years.

In a review of the application of chemometrics to the characterization of macromolecules, the chemometric methods were limited to MVA (1). PCA and PLS were described. A classification technique, SIMCA, was presented as a product control method.

COMPUTATIONAL METHODS

The PCA Method

Experimental data were evaluated by means of PCA employing a suitable computer program (2). The program was used in order to fit the scaled data (variance 1) to a principal components model. The procedure can be illustrated in the following manner:

Assume that each of the variables defines a coordinate axis (m). Then m axes will define an m-dimensional space in which each object can be described by a point (with coordinates according to the m-dimensional variable space). The whole set of objects will define a swarm of points in the m-dimensional variable space.

PCA constitutes a projection of this point swarm down to a space of a lower dimension in such a way that the first principal component (PC) describes the direction through the swarm which evinces the largest variation in the data. The second component shows the second largest variation, etc. The components are mutually orthogonal. The principles are illustrated in three dimensions in Fig. 1.

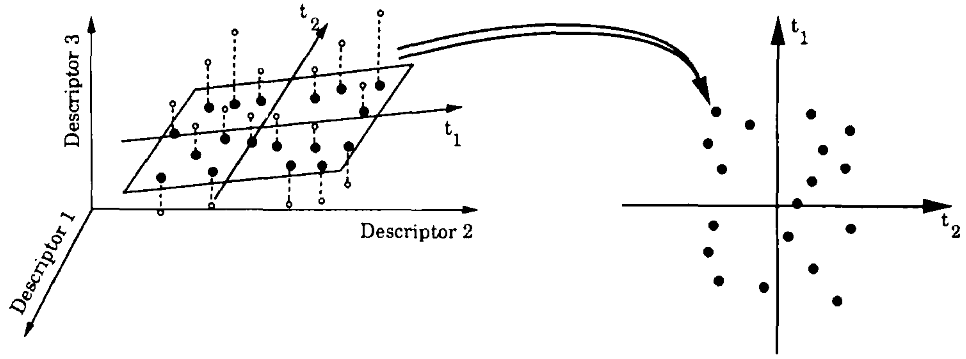

FIGURE 1

A geometrical illustration of the principles of PC modelling with three descriptors and two principal components t_1 and t_2.

The mathematical expression of a PC model is then

$$x_{ik} = a_i + \sum_{j=1}^{A} b_{ij} t_{jk} + e_{ik}$$

Here x_{ik} denotes the scaled value of variable *i* in object *k*. The analysis corresponds to a least squares fitting of a straight line (A=1) or an A-dimensional hyperplane to the data points in the m-dimensional variable space. The parameters a_i determine the centre of the data set; b_{ij} are the direction coefficients (one for each variable and component) of the line/ hyperplane. For each object *k*, the parameters t_{jk} describe the position of the object point projected down to the model. Hence, t-values can be used to relate objects to one another. The b-values (loadings) can, together with the residual variance e_{ik}, provide information as to how much each variable contributes to the model. An important result of this empirical modelling is that the systematic variation in the data can be described with fewer variables than in the original data set.

Determination of the significant number of components (product terms in the equation above) is performed by way of cross validation (3).

Classification Techniques

As was pointed out above, an important outcome of empirical modelling is that the systematic variation in the data can be described with fewer variables than in the original data set. Then the objects that are similar can be characterized by means of a PC model with few components. Pattern recognition is based on a "training set" of objects belonging to a "known" class. This classification does not need to be perfect. The method tolerates a limited number of erroneously assigned objects, outliers.

In the first phase of pattern recognition, the training set is used to develop a mathematical model describing the class. In the second phase, that of prediction, the model is put to use in predicting which class "new" objects will belong to (that is, objects not included in the model).

When a new object is available and the variables are measured, it is possible to project the data (x_j) down to the PC model which results in the scores t_j describing the place of the new object with respect to the ones used in the construction of the PC model. The residual standard deviation (RSD) measures the degree of fit to the model. If the RSD for the new object is large compared to the total RSD for the class of approved objects, then the new object is dissimilar to the objects which were included in the model before. As the loading matrix **B** is fixed, this projection corresponds to the multiple linear regression (after scaling):

$$\boldsymbol{x_j} - \boldsymbol{a} = \boldsymbol{t_j} \boldsymbol{B} + \boldsymbol{e_j}$$

thus

$$t_j = B'(x_j - a)$$

where **a** is the row vector of the averages of the variables.

SIMCA

Provided that most of the variables express a genuine similarity, a class can be well approximated by way of a PC model with few components. From this model, which is the basis of the SIMCA method, tolerance intervals can be constructed around the PC hyperplanes (Fig. 2). New objects are assigned to a certain class if they are inside its class-tolerance "cylinder", or characterized as outliers if they are outside all "cylinders".

The pattern recognition problem is often formulated in such a manner that the resulting data are asymmetric (4). This is the case when a well-specified class of objects with specified variables in a given test system is contrasted to other objects with other properties. The well-specified class is then homogenous, occupies a narrow domain in the variable space, and can be modelled by reference to a small number of principal components. By contrast, the other objects are more or less randomly disseminated in the variable space (Fig. 3).

The PLS Method

PLS is a statistical method of relating multivariate descriptor data sets to multivariate response data sets. What follows here is only a brief sketch of the method (for a detailed description, see reference 5).

In the present case, the descriptor set (the **X**-matrix) consists of the experimental variables and the response data set (the **Y**-matrix) consists of the observed results. A quantitative relation between these matrices

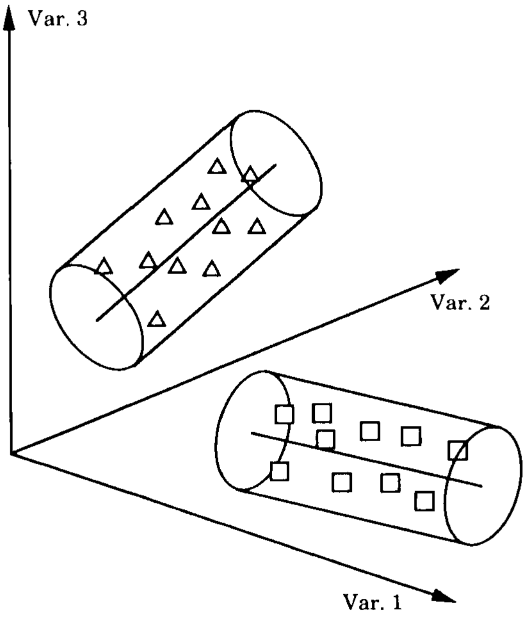

FIGURE 2

On the basis of the dissemination of the objects around the class model, a tolerance interval is constructed for any probability level, usually 95 percent.

is established by way of a PC-like ("principal-components"-like) decomposition of the matrices **X** and **Y**, and a correlation between the models thus obtained. These models are slightly tilted (biased) from the ordinary PC models in order to achieve maximum correlation between the components of the **X**-block and the components of the **Y**-block. A geometrical illustration is supplied in Fig. 4.

APPLICATIONS

Formulation and Processes

In an early paper (6), the role of PCA in the selection of a tablet formulation was presented. Ten response variables -- disintegration

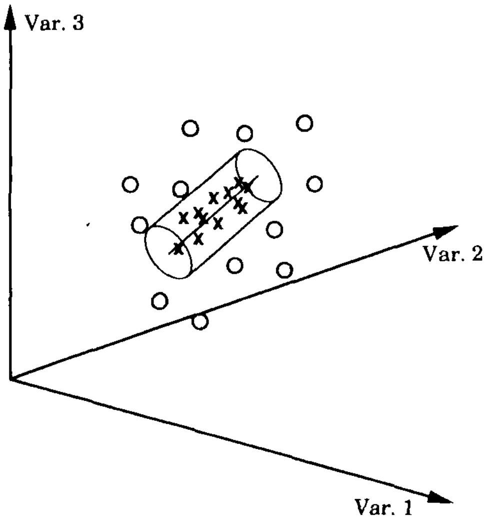

FIGURE 3

The asymmetric pattern recognition data structure. One class is well defined and occupies a small and regular part of the variable space, while the other objects are randomly spread in the variable space outside the class model. With SIMCA, the proper class is modelled and contained in a class interval. New batches are classified as belonging to this class if they are inside the tolerance interval. Batches outside this interval are classified as outliers.

time, breaking strength, dissolution, friability, thickness uniformity, porosity, mean pore diameter, weight uniformity, tablet breakage and granulation mean diameter -- were measured in each of 27 experiments from a half-factorial design for five factors at two levels, including centre point and star points. These five factors were: diluent ratio (calcium phosphate - lactose ratio), compression pressure, amount of starch disintegrant, binder level (amount of gelatin), and amount of magnesium stearate. The first principal component, dissolution, contributed 95.4 % to the overall information about the formulations, while the first two components (dissolution and disintegration) contributed 99.3 %. This

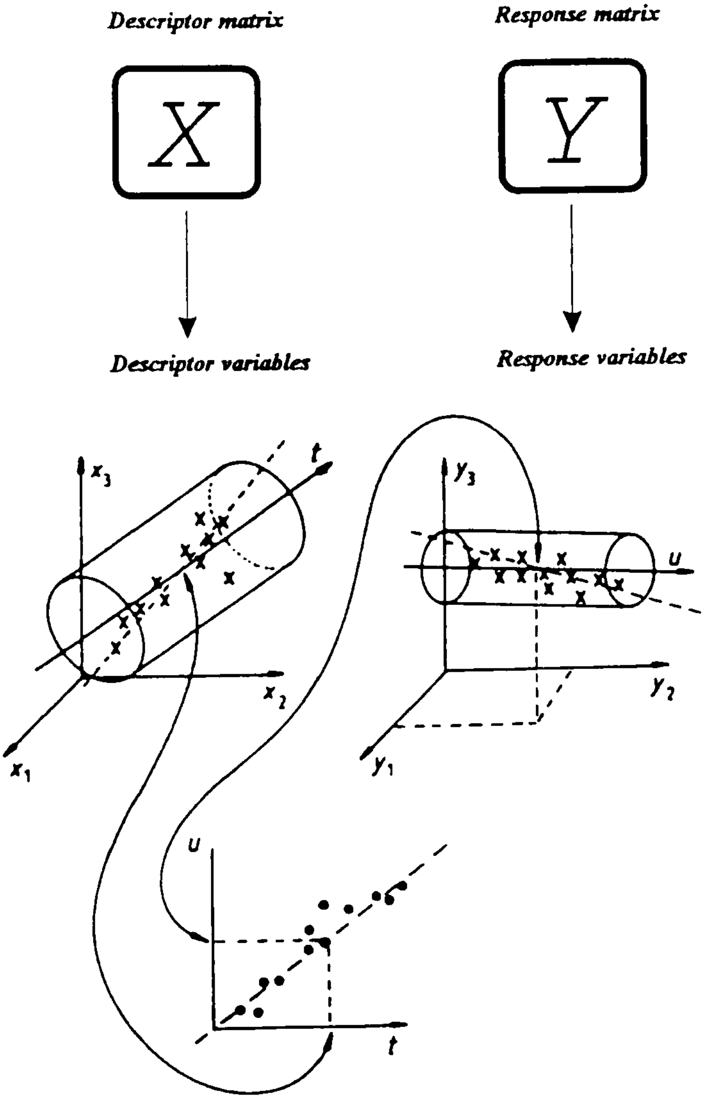

FIGURE 4

The PLS-components t and u are marked by solid vectors. They are slightly tilted from the PC-vectors (dashed) to achieve a maximum correlation between the response component, u, and the descriptor component, t.

meant that eight out of ten response variables contributed nothing further to the overall information. As the response variable dissolution was the predominant parameter of the system this parameter alone could effectively be used in the comparison of candidate formulations based on the formulation in question.

The influence of diluents and the concentration and viscosity of the wetting liquid on granule and tablet characteristics was studied by means of PCA (7). Three formulae with different proportions of tricalcium phosphate, lactose, and guar gum solution were investigated, including two brands of gum of three concentrations. The two brands differed in viscosity. Overall, 18 formulae were realized with each type of gum. Each formula was represented by 12 response variables, seven granule characteristics -- percentage of fines, bulk volume before tapping, settling rate, flow rate, granule strength, micropore volume, and total pore volume -- and five tablet characteristics -- ratio of axially transmitted pressure to applied pressure, hardness for constant pressure, friability, tablet porous volume and "massic area". The first two components accounted for nearly 90 % of the information from the data set.

Different lubricants -- varying qualities of sodium benzoate, sodium stearyl fumarate and PEG 6000 -- were investigated in three different concentrations when tabletting at three compression forces (8). Fifteen variables regarding the powder mixtures, tabletting process, and tablet properties -- lubricant concentration, compression force, flow rate, flow-rate variation, angle of repose, Hausner ratio, force transmission ratio R, ejection force, remaining force, tablet thickness, tablet weight, weight variation, crushing strength, friability, and disintegration time -- were measured. By means of PCA, the fifteen variables were represented by four principal components which jointly accounted for about 82 % of the total variation. The first component was characterized by the influence of compression force, crushing strength, R-ratio and friability. A high value on the part of the first three variables resulted, not unexpectedly, in a low value of friability. Flow rate, angle of repose, variation in flow rate, and tablet weight dominated the second principal component; all four variables were positively correlated. The third principal component was mainly influenced by the lubricant concentration and by the force of ejection.

By way of superimposing two score plots the best combination was identified. A lubricant consisting of conebulized sodium benzoate and PEG 6000 possessed the best properties in the study.

Two types of tablet formulations based on cellulose powder or lactose, Aerosil, potato starch, and magnesium stearate, were investigated (9). Eighteen powder and tablet parameters were measured. PCA indicated that two components explained almost all the variance for both base formulations. With the cellulose formulation, the first component accounted for about 67 % of the information, which was due to the influence from magnesium stearate and cellulose powder. The second component was influenced by cellulose powder and starch. About 100 % of the information was accounted for by the first two components. Regarding the lactose formulation, the first component explained approximately 64 % of the variance, which was influenced by magnesium stearate and lactose. The second component was influenced by lactose and starch.

PCA was employed for the purpose of examining the relationships between tablet properties (10). The experimental design was a five-factor, orthogonal, central composite, second-order design. The five factors were: quantity of water in the granulating solution (starch paste), wet massing, screen size for dry grinding, quantity of magnesium stearate, and compression pressure. Twenty-seven formulations were manufactured. Four response variables -- granulation surface area, mean dissolution rate, mean disintegration time, and mean hardness (crushing strength) -- were contained in the study. Three principal components accounted for 95 % of the correlation structure. The first component was mainly associated with the first three response variables, i.e. granulation surface area, mean dissolution rate and mean disintegration time, whereas the second principal component seemed to be associated with mean hardness only.

An optimized direct-compression formulation of a conventional theophylline tablet was developed with the aid of response surface methodology and successive quadratic programming (11). The two independent variables studied consisted of compression force and the percentage of disintegrant. Nine experiments were performed. Five response variables -- ejection force, hardness (crushing strength), friability, disintegration time, and mean in-vitro dissolution time (MDT) -- were determined. PCA was performed for the tablet properties crushing strength, friability, disintegration time and MDT. The first principal component was found to account for approximately 93 % of the overall information about the studied tablet formulations. Measurements of the four tablet properties were of approximately equal importance in defining the characteristics of these tablet formulations.

Sustained-release tablets of ibuprofen were optimized by means of PCA (12). Three formulation variables, namely ibuprofen level, polymer level and diluent ratio were included in the study. In the investigation where Eudragit was used as polymer, 13 variables were studied: drug concentration, polymer concentration, dissolution of the drug at pH 1.5 and 7.5, tablet crushing strength (hardness), friability, hardness/friability ratio, time to release 50 and 80 % drug, drug-release rate constant of first order kinetics at pH 1.5 and 7.5, drug-release rate constant according to Higuchi, and dissolution efficiency. Four principal components jointly made up 91 % of the information in the system. The first principal component mainly consisted in hardness, hardness/ friability ratio and friability, whereas the second component indicated dissolution, according to the authors concerned. In the score plot five groups were noted. By superimposing the score plots from components 1 and 2, as well as 1 and 3, respectively, it was possible to select the most interesting experiments in terms of good compression characteristics, satisfactory technical tablet properties, and efficient polymers for retarding the drug release.

Similar tests were performed with cellulose polymers.

Suitable formulations were described.

A multivariate analysis was performed in order to reveal the main differences between, as well as the common properties of different granulations/mixers (13). Two main parameters -- quantity of granulating liquid and kneading time -- were studied when a mixture of lactose, starch, and polyvinylpyrrolidone was granulated with aqueous ethanol. Different types and sizes of granulators were tested: planetary mixers 12 (Ours) and 60 (Collette) litres, and high-shear mixers -- Lödige 50 litres, Moritz 10 and 50 litres. A large number of response variables were registered before, during and after the compression of the granulations into tablets. The 1760 results obtained for all the 55 different granules produced with the five studied granulators were processed by means of discriminant analysis, which is a form of PCA.

Two components accounted for 86 % of the whole discriminant power. Therefore, the 32 response variables can be analyzed on these two axis of the loading plot. The first component was mainly made up of bulk volume, friability of granules and tablets, and granule fractions larger than 160 µm. This group of variables was opposed by tablet hardness, residual and ejection forces, and granule fractions smaller than 160 µm. The second component was explained by flowability and uniformity of size distribution.

The score plot indicated four groups: one for each of the small planetary mixer, and Moritz 10 and 50 litre respectively, and a group with both the large planetary mixer and Lödige. The granules produced in the large planetary mixer and the Lödige granulator seemed to have quite similar characteristics. In order to verify that, a second analysis was performed where the Lödige granules were taken as outliers. Discriminant axes were recalculated and all the different granules except the Lödige granules were represented on the first discriminant plane. This indicated that Collette 60-litre and Lödige-50 litre mixers produce the same kind of granules. Consequently, transposition of wet granulation between these two granulators will be an uncomplicated process.

Particle Characterization

The suitability of PCA for particle-shape analysis was examined (14). Shape-sorted citric-acid monohydrate and sodium-perborate tetrahydrate of different size fractions were studied by means of optical microscopy. 100 particles were examined. For each particle, the length, width, projected area and projected perimeter were determined using a microcomputer-based image analyzer, and particle thickness was determined by means of a specially mounted miniature displacement transducer. PCA was performed on the lengths, widths and thicknesses of each size- and shape-sorted fraction. Data from PCA for 64 size- and shape-sorted fractions were presented. Three principal components were obtained.

Regarding citric acid sorted by the Jeffrey-Galion vibratory shape sorter, the first principal component had a high length coefficient. This component was effectively a size descriptor accounting for about 60-70 % of the variance. The fact that 60-70 % of the within-sample variance resulted from size variation rather than shape variation could be seen as a measure of the success of the shape-sorting operation.

Even though PCA turned out to be a powerful technique for studying the variance within a sample, it is important to realize that it cannot describe particle shape as such, only the variation in shape. The method can yield different trends and patterns for the same material sorted by different methods, or for different materials sorted by the same method.

Dissolution Rate

The improvement in the dissolution of phenacetin was studied by means of PCA (15). Solid dispersions of drug and PEG 6000 or urea in different proportions were performed and the in-vitro dissolution tested in different media. The dissolution kinetics was described by the Weibull distribution. Two principal components jointly made up approximately 93 % of the total variance. The loading plot demonstrated that the first component was influenced by the quantities of drug dissolved after 240 min and infinite time, i.e. S_{240} and F_∞ respectively. As to the second component, the time for dissolving 5, 30 and 50 % of drug, i.e. $t_{5\%}$, $t_{30\%}$ and $t_{50\%}$ respectively, had the largest impact on that component.

The score plot revealed four groups with the 400 mg dose, where groups I and II represented an acceleration of the dissolution kinetics and an improvement of the dissolved drug. Solid dispersions with 80 % urea showed the most suitable dissolution behaviour. Group IV corresponded to a slow dissolution. Pure phenacetin was found in this group.

The quantitative relationship between the release rate of griseofulvin and the chemical and physical properties of a series of polymers, used for the preparation of solid dispersions, was investigated by the application of i.a. PCA and PLS (16). Twenty-seven polymers or viscosity types of polymers were investigated. The amount of griseofulvin dissolved after 2, 4, 6, 8 and 10 min was determined; these are the y-variables. Six chemical/physical characteristics, x-variables -- apparent degree of crystallinity of griseofulvin in sample powder; wetting of samples by water; logarithm of viscosity of 1 % polymer solution; pH of 1 % polymer solution; solubilizing effect of polymer on griseofulvin; and apparent dissolution rate of polymer -- were used in order to establish a quantitative relationship with the release rate of the drug.

The score plot (PCA) indicated dramatic differences between three groups of polymers: the HPCs, the PVM/MAS and the other ones. Because of the high degree of correlation between the y-variables of a two-components model accounting for 65 % of the variance, the release rate after 6 min was selected as the most appropriate one.

One single PLS component accounted for almost 51 % of the y-variance, and the main factor responsible for the release rate was the solubility of the polymer in methanol.

In a study of the variables influencing the dissolution rate of prednimustine, PCA as well as PLS was applied to the data (17). Seven small-scale batches and four production campaigns with 35 batches altogether were examined. 66 variables -- including, among other things, three dissolution-rate variables, 18 particle-size parameters, 12 gas-adsorption parameters, impurity and process variables -- were included in the data set. From the score plot of pooled data, Fig. 5, five regions can be observed to coincide with the small-scale batches and four full-scale production batches. PCA indicated that the dissolution rate of the drug was influenced by particle-size variables as well as by impurity, process and gas-adsorption variables. Consequently, there was no general relationship between dissolution rate and particle size only, which was to be expected on the basis of the Noyes-Whitney equation. PLS indicated the existence of relationships between dissolution rate and particle-size, impurity, process and gas-adsorption variables. It can be observed from Fig. 6 that the first component (WC 1) is mainly influenced by some particle-size variables (e.g. fccl, ccl, dmed, hund, femt), process variables (e.g. cych, bach, bapr) and dissolution-rate parameters (diss, dixx). The dissolution rate was reasonably well predicted, the correlation coefficient being ≥ 0.90.

Permeability

Four-component non-aqueous microemulsions containing lecithin, taurodeoxycholic acid, ethyl oleate and 1,2-propylene glycol were studied with chemometric techniques in order to emphasize the effects of the components on the release of a drug (18). Retinol was used as model drug. The permeation of retinol through a hydrophilic membrane can be related to the qualitative and quantitative composition of the

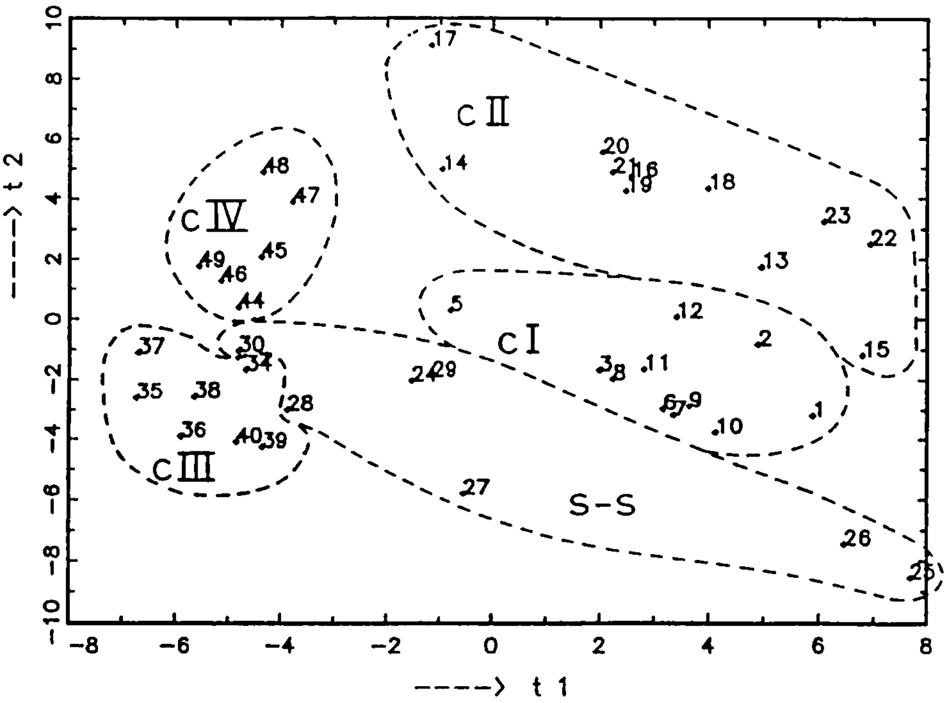

FIGURE 5

Score plot of whole data set.
X-axis: principal component no. 1
Y-axis: principal component no. 2
Figures refer to object numbers (batch numbers); S-S means small-scale batches, cI means campaign I, etc.

microemulsion, whose internal phases behave as a reservoir for the lipophilic drug. The use of PLS was established and a cross-validation was performed.

In the case of set A, full-rank second-degree Cox model, the final model accounted for about 84 % of the variance in prediction with one latent variable (PLS component). This latent variable contained two variables, taurodeoxy cholic acid and (taurodeoxycholic acid)^2.

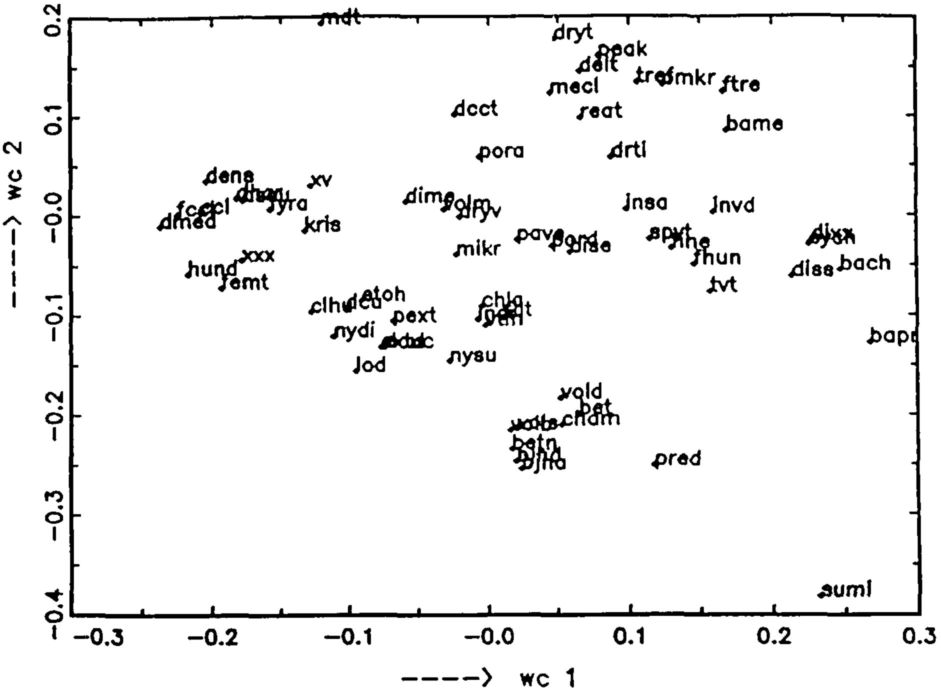

FIGURE 6

Weights plot, X-and Y-weights, of whole data set.
X-axis: component no. 1
Y-axis: component no. 2
Variables are denoted by a combination of three or four letters.

With set B, full-rank third-degree Cox model, the final model contained four variables. This model explained about 87 % of the variance in prediction.

Both calculated models showed good predictive abilities as regards retinol permeability from microemulsions through a hydrophilic membrane. Therefore, they could be used to select proper mixture compositions to achieve a desired drug-release performance.

Ointments

A suspension ointment of ketophenylbutazone was formulated with varying proportions of liquid and solid paraffin (19). Eight formulations were tested regarding four liberation parameters -- parameters k and q respectively of the Lippold model, parameter a of the exponential model of liberation, and Higuchi´s diffusion coefficient D -- as well as seven rheometric characteristics -- parameter K of the practical boundary tension, flow index n, surface of the thixotropic loop S, apparent viscosity Q, the coefficients of thixotropic destruction B and M respectively, and the penetrometric characteristic P.

The application of PCA did not yield any interpretable results. Factor analysis, which can be described as a version of PCA, indicated that practically all information contained in the matrix could be reproduced by means of two factors. The first factor included all the liberation characteristics, the penetrometric characteristic P, and the flow factor n. The second factor showed a close connection with the remaining rheometric characteristics. Cluster analysis indicated two separate groups. The first group included the liberation parameters except parameter q, but included the flow factor n and the penetrometric characteristic P. Included in the second group were liberation parameter q and the rest of the remaining rheometric characteristics.

SUMMARY

As this review indicates, multivariate data analysis has been applied within pharmaceutical research and development for about 20 years. However, it is only from the late 1980s onwards that more than occasional papers have been published.

Principal components analysis has been successfully employed in investigations of the relationships between response variables of tablet formulations; between formulations and response variables; between raw-material qualities and response variables; and between granules produced from different types of granulators. The same technique was applied to studies of the variation in shape of powders. The selection of proper microemulsion compositions, with a view to achieving a desired drug-release performance, was also performed by means of principal component analysis. An application to a suspension ointment is also mentioned.

The influence on dissolution rate exerted by solid dispersions was studied by way of principal components analysis, too.

Partial least squares projection to latent structures was applied with reference to the quantitative relationship between the dissolution rate of the drug and the properties of the polymers of the dispersion, or the properties of the drug itself.

REFERENCES

1. A. Hagman and S. Jacobsson, Drug Dev. Ind. Pharm. 16, 2527 (1990).
2. S. Wold, Pattern Recognition 8, 127 (1976).
3. S. Wold, Technometrics 20, 397 (1978).
4. W.J. Dunn and S. Wold, J. Med. Chem. 25, 595 (1980).
5. R. Carlson, L. Hanson and T. Lundstedt, Acta Chem. Scand. B40, 444 (1986).
6. N.R. Bohidar, F.A. Restiano and J.B. Schwartz, J. Pharm. Sci 64, 966 (1975).

7. L. Benkerrour, D. Duchêne, F. Puisieux and J. Maccario, Int. J. Pharm. 19, 27 (1984).

8. P. Wehrlé, P. Nobelis and A. Stamm, S.T.P. Pharma 4, 275 (1988).

9. F. Podczeck and U. Wenzel, Pharm. Ind. 52, 348 (1990).

10. T. Schofield, J.F. Bavitz, C.M. Lei, L. Oppenheimer and P.K. Shiromani, Drug. Dev. Ind. Pharm. 17, 959 (1991).

11. S. Dawoodbhai, E.R. Suryanarayan, C.W. Woodruff and C.T. Rhodes, Drug Dev. Ind. Pharm. 17, 1343 (1991).

12. A.A. Abdel Rahman, A.E. Aboutaleb, A. Stamm, S.I. Abdel Rahman and E.M. Samy, Bull. Pharm. Sci., Assiut Univ. 15, 63 (1992).

13. P. Wehrlé, P. Nobelis, A. Cuiné and A. Stamm, Drug Dev. Ind. Pharm. 19, 1983 (1993).

14. M. Whiteman and K. Ridgway, Spec. Publ. R. Soc. Chem. 102, 340 (1992).

15. S. Raynaud, H. Maillols, N. Moscoso Pinel, J.P. Laget and H. Delonca, S.T.P. Pharma 1, 523 (1985).

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

17. N-O. Lindberg and T. Lundstedt, in "Conference Book of the 12th Pharmaceutical Technology Conference", March 30th - April 1st, 1993, Elsinore, Denmark, Vol. 2 p. 37.

18. F. Pattarino, E. Marengo, M.R. Gasco and R. Carpignano, Int. J. Pharm. 91, 157 (1993).

19. H. Zacek and P. Dolezal, Folia Pharm. 12, 61 (1987).