Journal of Pharmaceutical & Biomedical Analysis
Vol. 9, No. 8, pp. 589–596, 1991
Printed in Great Britain

0731–7085/91 $3.00 + 0.00

Chemometrics, why, what and where to next?*

SVANTE WOLD

Department of Organic Chemistry, Umeå University, S-901 87 Umeå, Sweden

Abstract: The advantages of the application of chemometrics in pharmaceutical and biomedical analysis are discussed. Some chemometric approaches are described and the requirement for validation emphasized. Finally, possible future developments of chemometrics are assessed.

Keywords: *Chemometrics; data analysis; experimental design; multivariate.*

Introduction

Most advances in science amount to a break with 'common sense'. It was not long ago that the doctrine of analytical chemistry demanded that one single selective signal (variable) was measured with the utmost precision to provide the desired information. Chemometrics tells us that it is often better to measure many non-selective signals, and then combine them in a multivariate model [1].

We are often confronted with several variables, for instance, when comparing the chemical profiles of a group of drug treated animals with a control group. The obvious, 'common sense' approach is to scrutinize the variables one at a time; COST analysis (Consider One Separate variable at a Time). Chemometrics says the opposite; analyse everything together, multivariately [1–3].

A central question in chemistry is how to best and most efficiently make experiments. Consider as an example the optimization of a chromatographic analysis. Several factors affect the results and may be varied in the optimization, e.g. type and amount of extraction solvent, extraction time and temperature, pH and type of mobile phase, and type of stationary phase. The "common sense" way of making experiments to reach an optimal chromatographic analysis is still to Change One Separate factor at a Time (COST design). This is also the way most experimental chemistry is taught and practiced.

Fisher showed the deficiencies with this strategy in 1925. Thereafter Fisher, followed by Yule, Box, Youden, Hunter, Hunter, and numerous others have shown that laying out series of experiments where all factors are changed together, so called statistical experimental design (SED), gives much more information and reaches the optimum conditions in fewer experiments than does COST experimentation. For SED overviews, applications, and further references, see refs 4 and 5.

In retrospect, these steps of chemometrics, the recognition of the information inherent in multivariate data, and the necessity to not change one factor at a time in experimental investigations, may seem obvious, and even trivial. Accepting these ideas of chemometrics, we can now look forward at what can and should be done, to better solve practical chemical and biological problems using chemical–biological knowledge coupled with mathematics and statistics, and how we can use emerging scientific and technical developments for these purposes.

What is Chemometrics?

Chemometrics is greatly motivated by practical problem solving, utilizing experimental data efficiently and economically. A review and history of chemometrics is given by Geladi and Esbensen [6, 7]. Pertinent text books and reviews have been published [1–12].

Basically, chemometrics has two essential lines of development; that of data analysis, utilizing the inherent information in chemical data in the best way, and that of experimental design, that of planning and performing exper-

*Presented at the Symposium on "Chemometrics in Pharmaceutical and Biomedical Analysis", November 1990, Stockholm, Sweden.

iments in a way that the resulting data contain the maximum information about stated questions. These two lines of chemometrics are not separated, but intertwined.

Why multivariate modelling and analysis?

Chemometrics started around 1970 with the insight that chemical instruments were beginning to produce much more data than any existing chemical data analysis reasonably could cope with, and that the unutilized information in these data masses may be substantial.

Today the motivation for chemometrics remains very much the same, i.e. how to make intelligent use of the masses of data produced by chemical analysis and experimentation. The size of the data sets has grown, today we gather thousands of variables, second and third order data arrays, etc. However, our understanding of the situation has also improved, and we today have appropriate data analytical methodology for most routine situations in chemical, pharmaceutical and biomedical analysis.

There are two basic reasons for making a multivariate analysis instead of analysing the variables one at a time:

(1) The multivariate analysis gives an overview of all the data allowing an overall judgement and an overall evaluation of the significance of differences between groups (e.g. treated and control) and correlations (e.g. between a set of symptoms and chemical profiles).

Statistically, it is extremely difficult to judge the significance of a large number of differences between group averages or a large number of correlations in scatter plots. We know fairly well how to evaluate the significance of one difference, and one correlation. Statistics teaches us to use *t*-tests and the like so that the risk to accept a "randon result" as "real" is less than, say, 5%. With two differences or two correlations, the risk to judge at least one "random result" as real is about 10%, if the significance of each of the two is evaluated separately. With K correlations or group differences, this risk (u) is:

$$u = 1 - (0.95)^K.$$

The value of u exceeds 0.4 when $K > 10$ and exceeds 0.8 when $K > 30$. Only by making a single analysis of *all* the data — multivariate analysis — can one get this risk for spurious results under control.

(2) It is clear that information about complicated samples or processes is not associated with single variables. Such concepts as interactions, synergisms, joint influence of several factors on a biological receptor, etc., can be seen only by analysing all relevant variables *together* by multivariate analysis. And when all data are analysed together, one obtains an averaging effect, so that the systematic information is enhanced, and noise is decreased. This is the same principle as is used in the signal averaging in NMR, FT–IR, etc.

The recognition that multivariate data potentially contain much more information than few-variate has provided the incentive to design chemical analytical methods to give many signals instead of one. This, in turn, has necessitated a development of sampling and experimental design to facilitate the selection of experimental/measuring conditions with maximum information content; multivariate design [9–12].

One interesting and important property of multivariate data is their ability to capture diffuse qualities and properties such as biological activity, taste, and smell, much better than these can be quantified in a single measurement. For the investigation of biological systems this is essential, and indeed large numbers of response variables are routinely measured everywhere in drug research.

Geometric Interpretation of Chemometrics

Geometry provides a straightforward way to understand experimentation and data analysis in terms of spaces and configurations in these spaces such as points, lines, planes, and curved surfaces.

Multivariate analysis

The principles of multivariate analysis are simple. Data have been measured on a set of objects (object = sample, compound, rat, etc.). On each object, the values of K variables have been measured, e.g. concentrations of various compounds in the blood and urine of each rat, or physico-chemical properties of each compound, etc.

These data are represented as points in a multivariate space (M-space) with as many axes as there are variables (Fig. 1). One then constructs windows into this space by means of

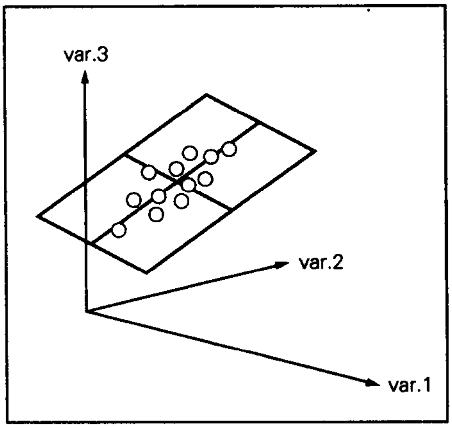

Figure 1
Multivariate space with three variables, some data, and a plane approximating the data. A window into the space, allows us to see trends, groups, outliers, etc.

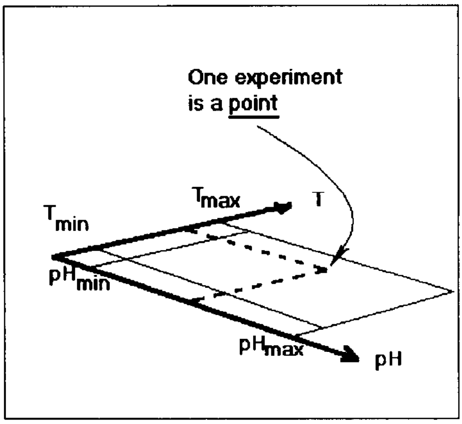

Figure 2
Two-space of two factors (pH and T). Each experiment is a point in this space.

projections on planes or hyper-planes. The information in the data is seen as "patterns" in these windows, e.g. trends, separated groups and outliers.

Multivariate analysis methods such as principal components analysis (PCA), and factor analysis (FA) usually is concerned with developing models in M-space, such as lines and planes (Fig. 1). The projection of the points down on a plane can then be displayed on a computer screen or shown as a graph, allowing the recognition of "patterns" in the data. The direction of the projection plane gives information on which variables are important and which are not, and how the important variables combine to separate groups of objects, to define the trends among objects over time, etc.

In the quantitative analysis of multivariate data, it is often practical to use two spaces, one for the factors and other predictor variables (X), and one for the response data (Y). Methods, such as PLS (partial least squares projection to latent structures) [1, 8, 11–14] and multiple regression (MR), develop models that connect the two spaces X and Y. All these models can be seen as lines and planes in the X- and Y-spaces with an interpretation very similar to that of PCA and FA.

Statistical experimental design (SED)

When we are making experiments, we manipulate a set of factors. Take, for example, a chromatographic separation where we can vary just the two factors pH and temperature (T). One experiment is now a point in the factor space (Fig. 2). The experimentation is confined to a region in this space, defined by the lower and upper limits of each factor. A measured response (y), say the distance between two important chromatographic peaks, can now be represented as a third axis, giving a three-dimensional space. So, for each setting of the two factors pH and T we can make an experiment, giving a value of the separation, y (Fig. 3). Using a simple model (see below), we can finally connect the experiments by a smooth surface, allowing predictions to be made as interpolations and mild extrapolations.

With this geometrical interpretation of experiments (factor space, X) and data (measurement space, X or Y), we can think of the task of an experimenter as that of exploring an X-space (factor space) as well as possible with as few experiments (runs) as possible with the purpose to get as good a map as possible of

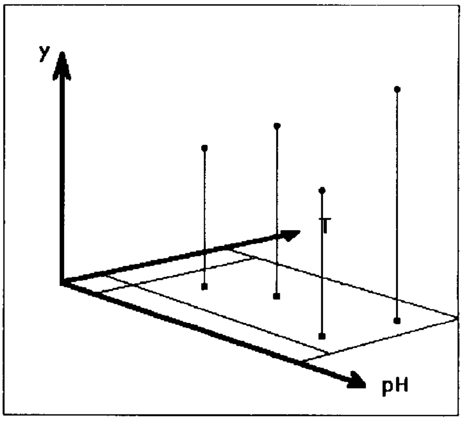

Figure 3
Two factors (x) and one response (y) form a three-dimensional space or a 2-space (X) + a 1-space (Y).

the relation between the factor space X and the response space Y. It is easy to see that the COST design leads to a poor map of this space and its relation to the Y-space, and also that COST designs lead to unnecessarily many runs. See also the discussion by Kettaneh-Wold [15] in this volume.

With SED [4, 5], a set of experiments, a plan, is laid out (see, e.g. Fig. 4), which (a) allows the precise estimation of the parameter values of the model, and (b) gives good predictions in the whole experimental region, but still (c) uses a fairly small number of experimental runs.

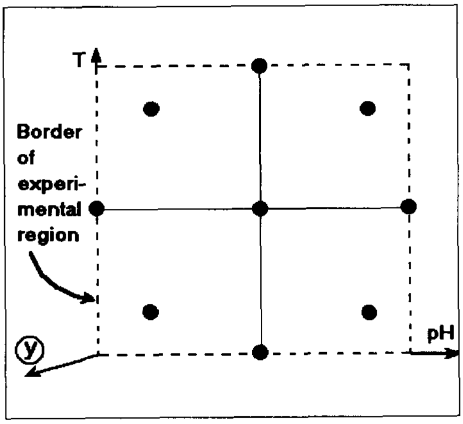

Figure 4
A statistical design (central composite) in two factors. This design supports a quadratic polynomial model:

$$y = c_0 + c_1x_1 + c_2x_2 + c_{11}x_1^2 + c_{22}x_2^2 + c_{12}x_1x_2 + \epsilon.$$

Modelling

Underlying all design and data analysis there is the concept of models. This is the necessary consequence of the theoretical developments of theoretical physics and chemistry in the beginning of this century by Planck, Bohr, Schrödinger, Heisenberg, Dirac, and others, and not the least the mathematical–philosophical results of Gödel. Basically, we are forced to realize that complete knowledge about anything is impossible to obtain, even in principle, and all we can do is to develop approximate simplified mathematical approximations of the complicated reality, i.e. models.

Nevertheless, the use of mathematical models allows us to connect the results of one experiment with those of another, to find favourable directions in the experimental space, etc., in short, to learn from experiments [4]. The models used in chemometrics usually are 'semi-empirical', such as polynomials (lines, planes, quadratic surfaces). This is because, except in very simple cases, there are no adequate fundamental models based on first principles. In cases where good fundamental models of an investigated system or process exists, they can, of course, be used for the data analysis. This, provided that the fundamental models are not 'ill posed' (ill conditioned) with numerical and statistical deficiencies [16].

Using the models, the data are separated into two parts; the systematic part 'explained' by the model, and the 'noise', the remaining unmodelled part of the data. Initially the acceptance of 'noise' is psychologically difficult, but any experiments and measurements have an inherent variability,

$$\text{Data } (Y) = \text{Model[factor values } (X),\\ \text{parameters } (\beta)] + \text{Noise.}$$

With the systematic part of the model, we can now predict the behaviour of our system for new values of the X-variables. When the X-variables are factors such as pH, T, and so on, this allows us to manipulate the system to achieve desired goals as well as possible, 'optimizing' the system. The noise part is modeled by a statistical distribution such as the normal distribution. This allows us to get an idea of the uncertainty — confidence intervals — of parameter values (β) and model predictions ($\hat{Y}$).

When the X-variables are signals, and Y measures concentrations or other system properties, we can predict Y from the values of X for new samples; we have a "multi-variate standard curve" [1].

Data Analysis and Validation

When data are related to a model, this usually corresponds to the estimation of the values of a number of parameters in the model. This is done so that the deviations between the model and the data become small using, for instance, the criterion of least squares. This model fitting balances on a fine line between on the one hand overfit — producing partly spurious correlations — and on the other hand underfit — not utilizing all the information in the data.

The risk for overfit arises because with sufficient free parameters, a model can be made to fit any data exactly if careful precautions are not taken, and if a correct model fitting is not done. We note that many models used in chemometrics, such as PLS, often have many more apparent parameters than data, but additional constraints — with PLS the modelling of the X-data — keeps the risk for overfit under control [1, 11].

To avoid the pitfalls of overfit, spurious results, and the consequential random predictions of new events and observations, some kind of model validation is essential. The best validation is always a number of new and representative observations that were not part of the model fitting. The X-values of these new observations should, when inserted in the model, predict the corresponding Y-values much better than by chance.

Often, however, new observations are not immediately available. One can then still use a type of validation that simulates the prediction of new events, so called cross-validation [1, 17]. This is based on the deletion of a few observations from the data, developing the model on the basis of the remaining data, and then predicting the deleted data from their X-values and the developed model. This is repeated a number of times, until each observation has been deleted once, and once only. The resulting measure of predictive power, PRESS (predictive residual sum of squares) provides a good measure of how well the model actually will predict new observations.

Naturally, cross-validation (and the closely related boot-strapping) can be misused too, giving a PRESS that apparently is very good even when the actual predictive power of the model is zero. This is accomplished by using PRESS as the optimization criterion, starting with a large number of variables (and parameters) with stepwise selection of the variables that make PRESS look the best. Then, finally, one forgets how many variables that actually were involved, and pretends that only a mild variable selection has been performed.

Apart from this obvious misuse, however, cross-validation works very well, and is used extensively in chemometrics for initial model validation. Martens [1] and Wold [17] give further details on the method.

Another safety belt, preventing much misuse of models, is to always display data and results as plots and graphs. Man is very good at judging patterns (and lack of patterns) in graphs, and as a rule, one should never believe results of modelling and data analysis if they are not supported by pertinent plots. With the graphical abilities of today's computers and software, there is no excuse for not showing plots!

Chemometrics, Why and What?

So, to the first question, 'why and what is chemometrics', we may answer as follows:

Why?

Because statistics and mathematics teach us that there are better ways to use the information in the data than just looking at them, plotting variables one at a time or pairwise. The two worst effects of this "traditional" approach are (i) an increased risk of spurious correlations and other results, and (ii) the real information in the data is seen less clearly.

Using appropriate modelling and analysis, the information in the data can be extracted in an optimal way, while keeping the risk for spurious results under control. This allows us to make the best interpretation, decisions, optimization, etc.

Analytical instruments such as HPLC, NMR, GC–MS, produce more and more data for a given chemical or biological sample. Chemometrics provides tools to make good use of these data, enabling the scientists to make sense of the otherwise overwhelming masses of data flooding the laboratory of today.

Second; economics, competition, ethics and regulations put strong pressure on reducing the number of experiments, in particular with animals and humans. Only with a combination of statistical experimental design and good multivariate data analysis can we solve our problems within the constraints stated by economics and society.

What?

Basically, the principles of chemometrics are simple:

(1) Use quantitative (mathematical) models to connect, rationalize, and interpret the chemical/biological data. Never forget the difference between model and truth!

(2) Include variability (ϵ) in the model, and handle it by means of distributions.

(3) When changing conditions, making experiments, optimizing etc., do not change one factor at a time, keeping the others fixed. Rather, use statistical designs, planning sets of

experiments, usually with $N = 10$–20, where all factors are varied together.

(4) Analyse all data together with an appropriate model. With the type of data measured in laboratories today, this usually involves multivariate modelling and analysis, with PCA, PLS, and other projection methods. Always use some kind of validation (at least cross-validation) to judge the predictive power of the model. Show the results as plots.

In practice, this is, of course, more difficult than it sounds, but with increasing experience, increasing understanding, better software, better graphics, and better computers, things are improving.

Thus chemometrics makes modelling, experimental design, and data analysis simpler and more rational, allowing the scientist to concentrate on the really difficult part of research and development, to translate diffuse objectives into quantitative response measurements without losing the essence of the problem, to think of all factors that may influence the investigated system, to express them in a practical and pertinent way, to make accurate measurements and good experiments, and to interpret the results of the data analysis, i.e. bring the resulting coefficients back to relate to the original objective. This will never be easy or automatic, because this is the essence of research and development.

The Future of Chemometrics

Chemometrics is relatively new and still has a number of areas that must be much better developed before more speculative futuristic visions are pursued. In particular, we still lack the following:

(1) Good and easily applied methodology for dynamic systems and processes. These systems show particular data analytical problems in that the consecutive observations are not statistically independent. Ordinary multivariate methods such as PCA and PLS must be modified if they shall be useful in this context. Also, these applications produce extremely large masses of data, often at intermittent intervals, making present methods of analysis and graphical representation insufficient.

Most methods of experimental design are also based on the assumption of independence between observations. For process optimization, better ways of experimentation than changing a few variables at a time are still not well developed.

(2) Appropriate graphic representation of large data sets and results. We do have adequate graphics for up to four or five X-variables and three or four Y-variables in regression and PLS modelling, and for up to three or four latent variables (components) in PC and PLS-modelling. However, the increasing size of the data sets in chemical and biological R&D makes today's principles for graphical representation obsolete. Possibly multi-layer graphics where a data set can be looked at with different levels of 'resolution', and zooming between these levels, is a possible direction to go.

(3) Tools for using results from multivariate analysis for the optimization of systems and processes. When we think about experimentation and optimization, we have a strong tendency to think of factors as independent. Hence, all our tools for the experimental manipulation of processes are based on these independence assumptions. However, most variables in complicated systems processes are correlated, and it is not clear how to go from a good multivariate model back to the actual change of the process conditions to improve the 'quality and performance' of the process output.

(4) Good connections between chemometric models and fundamental chemical theory. There is a clear gap of communications between chemometrics and much of the rest of chemistry. Parts of analytical chemistry provide the exception. One reason for this gap is the different types of models used in chemometrics and in physical chemistry, inorganic chemistry, organic chemistry, and biochemistry. The models of the latter branches of chemistry are often, at least apparently, based on fundamental principles, which makes them easier to interpret, and generalize. At the same time, it is clear that they are often unsuitable for analysing data and for making quantitative predictions. This is because they are often ill conditioned with strong correlations between parameters, rarely multivariate, and usually not flexible enough to model data in areas where we know less, i.e. at the frontiers of research, and in the investigation and optimization of complicated systems.

It is possible to regularize any type of 'fundamental' model to make it numerically and statistically better conditioned, and then extend the model by serial expansions in the resulting metric, and finally make the model multivariate. This would perhaps provide a bridge between chemometrics and chemistry, models that retain the shape of the 'fundamental' models, but that are more suitable for serious use with experimental data.

(5) Good general textbooks, courses, academic programs, software, etc. Chemometrics, still being in its infancy, has still long to go before it is an established chemical discipline. So far no general chemometrics textbook has been written that is based on the philosophy of modelling, and at the same time tries to start from chemical problems showing what chemometrics can do about it. Reference 1 is a notable exception, but concerns only one area of applications, albeit an important one.

The same goes for academic programs, courses, etc., all of which take time to develop, and also a much wider basis than we presently have. One area that moves faster than others is software development; we now start to have software corresponding to the state of art in multivariate analysis and experimental design. And to incorporate the developments of tomorrow into these packages should not be too difficult.

The real future: can anything be said about it except that we do not know anything about it until we see it? We can only express what we wish will happen and what we wish will not.

Undesirable developments

Chemometrics has so far been motivated very much by practice, and we hope this to continue. There is a risk, however, that chemometrics becomes more theoretical, further removed from chemical experimentation, analogous to the sad development of psychometrics and biometrics, and much of technometrics and econometrics. This will happen if chemometrics is separated from chemistry and put in separate departments, or organized as part of statistics or computer science departments. The responsibility that this does not happen rests solely with the chemometricians themselves; we must not let ourselves be impressed by pure mathematical theory without chemical substance.

Another undesirable future is connected with the present fad of expert systems and artificial intelligence. As some see it, computers will become more and more 'intelligent' and will take over tasks that we today think of as human and creative, even scientific. However, as long as chemistry and biology remain experimental sciences, the problems investigated in research and development will always have in them a substantially novel part. This follows from what was discussed above under modelling, that of our knowledge about reality as always incomplete and mere approximations, more or less crude, of the real relationships.

Therefore we can be reassured that 'AI' and expert systems will never be able to substitute real research or development, only the simplest routine tasks where human thought, creativity, and emotional input is not needed.

Desirable developments

We today begin to see chemometrics being used to construct analytical instruments that give optimally informative data. This interesting trend will hopefully continue, with chemometrics and statistical design used to plan the measuring and experimentation setup, instead of the opposite which unfortunately still remains the norm.

If, as we hope, chemometrics continues to be coupled to chemical practice, its future is strongly coupled to important practical problems emerging in the future of chemistry, of which we know little. What is clear is that experimentation becomes more and more expensive, and in pharmaceutical and biomedical research, more and more regulated. This provides a stronger motivation for using appropriate statistical design and modelling to make experimentation optimally efficient. May be we will see only designed experimentation in the near future.

At the same time, measurements continue to be cheaper and cheaper, which leads to larger and larger data sets (with respect to the number of variables) being collected in each experiment. To utilize the information in these data, appropriate multivariate modelling and analysis is essential. We can see the emerging need of hierarchical models on different levels of "data resolution" hooked to appropriate graphics, allowing us to simultaneously getting an overview and penetrating knowledge of complicated systems.

These developments will allow us new ways to question chemical and other databases. Today's search based on questions such as 'all compounds with two double bonds and one amino group' are all based on linear additive (implicit) models. When we start to use our present (and future) knowledge of multivariate modelling to database construction and retrieval, interesting patterns will emerge.

Finally, in an interesting futuristic extrapolation, Geladi [18] has pointed out the possibility of small, cheap, sensor chips hooked to a small cheap computer on the same chip programmed with multivariate analytical modelling, all powered by a photo-electric cell (on the same chip). An appropriately constructed chip will be able to analyse its environment and display or communicate the results in terms of 'quality diagnostics'. A chip can sit in a bottle of wine and tell how good, and mature, the wine is. Similar chips can tell us if a loaf of bread is still fresh, if a drug has deteriorated, and if our feet smell good.

Acknowledgements — Support from the Swedish Natural Science Research Council (NFR), the Swedish Board for Technical Development (STU), and Center for Environmental Research (CMF) is gratefully acknowledged.

References

[1] H. Martens and T. Naes, *Multivariate Calibration*. Wiley, New York (1989).

[2] D.L. Massart, B.G.M. Vandeginste, S.N. Deming, Y. Michotte and L. Kaufman, *Chemometrics: A Textbook*. Elsevier, Amsterdam (1988).

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

[4] G.E.P. Box, W.G. Hunter and J.S. Hunter, *Statistics for Experimenters*. Wiley, New York (1978).

[5] C.K. Bayne and I.B. Rubin, *Practical Experimental Design and Optimization Methods for Chemists*. Verlag Chemie, Heidelberg (1986).

[6] P. Geladi and K. Esbensen, *J. Chemometrics* **4**, 337–354 (1990).

[7] P. Geladi and K. Esbensen, *J. Chemometrics* **4**, 389–399 (1990).

[8] L. Ståhle and S. Wold, *J. Pharm. Methods* **16**, 91–110 (1986).

[9] R. Carlson, *Chemica Scripta* **27**, 545–552 (1987).

[10] S. Hellberg, M. Sjöström *et al.*, *Acta Pharm. Jugosl.* **37**, 53 (1987).

[11] S. Wold, M. Sjöström and S. Hellberg, *Bull. ISI 46th Session (Tokyo 1987)*, invited paper 30.1 (1987).

[12] S. Wold, M. Sjöström, R. Carlson, T. Lundstedt, S. Hellberg, B. Skagerberg, C. Wikström and J. Öhman, *Anal. Chim. Acta* **191**, 17–32 (1986).

[13] H. Wold, in *Systems Under Indirect Observation* (K.-G. Joreskog and H. Wold, Eds), Chapter 1, Vol. II. North-Holland, Amsterdam (1982).

[14] A. Höskuldsson, *J. Chemometrics* **2**, 211–228 (1988).

[15] N. Kettaneh-Wold, *J. Pharm. Biomed. Anal.* **9**, 605–610 (1991).

[16] G.H. Golub and C.F. Van Loan, *Matrix Computations*. Johns Hopkins University Press, Baltimore, MD (1983).

[17] S. Wold, *Technometrics* **20**, 397 (1978).

[18] P. Geladi, The Past and Future of Chemometrics. Talk at the 2nd Czech. Chemometrics Conference, Brno, September 1990.

[*Received for review 26 November 1990; revised manuscript received 10 July 1991*]