Journal of Pharmaceutical & Biomedical Analysis
Vol. 6, Nos 6–8, pp. 535–545, 1988
Printed in Great Britain

0731–7085/88 $3.00 + 0.00

Chemometrics in pharmaceutical analysis*

D. L. MASSART† and L. BUYDENS

Pharmaceutical Institute, Vrije Universiteit Brussel, Laarbeeklaan 103, 1090 Brussels, Belgium

Abstract: Chemometrics is a science where chemistry and pharmaceutical science meet statistics and software. The primary focus of chemometrics involves the use of mathematical or software procedures in particular, both to develop analytical methods and to analyse the signals and results obtained. This paper focusses on chromatography and on how chemometrics has been applied to chromatographic problems in pharmaceutical and biomedical analysis. Examples of several chemometric techniques are given and recent developments in the use of optimization methods, regression methodology, multivariate analysis and expert systems are discussed.

Keywords: *Chemometrics; chromatography; pharmaceutical analysis; expert systems.*

Introduction

A slightly adapted version of the official definition of chemometrics defines it as "the chemical discipline that uses mathematical, statistical and other methods of formal logic to design or select optimal procedures and experiments, and to provide maximum chemical information by analysing chemical data". This indicates that chemometricians consider chemical analysis as a process in which the chemical determination is only one part, and that they are interested in what happens before, after and during the actual measurement process itself. To explain this by way of concrete examples, it is proposed to consider one of the more important pharmaceutical and biomedical methods, namely chromatography.

The measurement process consists of several steps, as shown in Fig. 1. First one specifies the problem, then one develops a suitable method and carries out the procedure and finally one interprets the results. Some of these steps can be further subdivided. For instance, the development of a method can really be considered to consist of two substeps. One first makes the initial selection of a method; for instance, one may first decide to carry out high-performance liquid chromatography (HPLC) using reversed-phase with methanol–water (50:50, v/v) and a buffer of pH 4. Then one may optimize these conditions to find eventually that some acetonitrile should be added to the eluent and that the pH should really be 3. In the interpretation of results one also finds two substeps. The first consists of the translation of electrical signals to a list of low level data, such as chemical identities and concentrations. These must then be translated to some higher level concept; for instance, the results may indicate that the patient whose blood was analysed has a certain illness.

* Presented at the "International Symposium on Pharmaceutical and Biomedical Analysis", September 1987, Barcelona, Spain.

† To whom correspondence should be addressed.

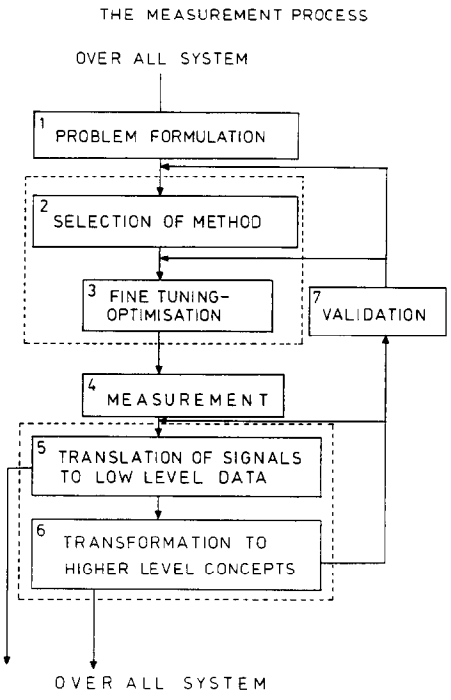

Figure 1
Schematic diagram of the measurement process.

Most processes contain feedback loops and interact with their environment. For instance, a method must be validated and quality control must be carried out. If the results of validation indicate that something is wrong, then it may be necessary to carry out the method development all over again. The results should be comprehensible to other workers, and this environmental interaction may have a consequence; for instance the results may indicate that one also needs data for another substance, therefore the method must be adapted; or another consequence could be that more rapid results are requested and this could require adaptation of the method or re-organisation of the laboratory.

Chemometric methods

Clearly the specification of the problem, step 1, is important but it has nothing to do with chemometrics. The initial selection (step 2) has only very recently been tackled by chemometric methods. The initial selection is performed by chromatographic experts using their expertise. Recent breakthroughs in the accessibility of artificial intelligence (AI) techniques have made it possible to apply expert system technology to incorporate expertise-related knowledge in computers. This is going to be important in the next few years [1–3].

Expert systems are going to be important in other areas too. Considering once again the optimization step and more specifically the optimization of mobile phases in HPLC, it is clear that for an optimization method one requires an optimization criterion (for instance, one might optimize the resolution) and an optimization design (for instance a Simplex). Many such optimization methods have been described by Schoenmakers and

Berridge [4, 5] and strategies such as the solvent triangle by Snyder and Glajch [6, 7] using the so-called overlapping resolution maps are well known in chemometrics under the name "mixture designs". Several HPLC manufacturers have one for their instrument. Unfortunately, this is usually claimed to be the one and only good optimization method. It is not well enough understood that there is no such thing and many disappointments by users are due to this misunderstanding. Different criteria and designs are needed when one optimizes a separation of nearly 100 components in a plant extract, or when one tries to separate a drug from a biological matrix. Different designs and criteria are therefore required for different applications. The problem for the non-expert user of optimization methodology is to choose the right criterion and design for a specific problem and it is to be expected that this will be done using an expert system. In fact, research about expert systems for the selection of experimental designs has already been carried out by Deming [8]. In the experimental optimization of HPLC the next significant improvement will be development of systems which contain many different criteria and designs and are driven by an expert system, that will choose the best combination for a specific situation.

The chemometrician is also involved in step 5. The output of the detector is a set of signals that must be brought into a form that is useful for the chromatographer. This first involves cleaning up the signal through noise reduction, using mathematical techniques such as smoothing, filtering, etc. In this "cleaned up" chromatogram one needs to resolve any overlapping peaks through the use of deconvolution techniques. Finally, peak areas must be related to concentration by regression or calibration methods.

There is a lot of chemometric activity devoted to regression techniques; for instance, techniques designed to detect and avoid the disastrous effect of outliers [9], and guidelines on how to carry out multivariate calibration (see below).

The data obtained in step 5 form the raw material on which conclusions can be based. Suppose that a fatty acid profile was obtained. Depending on the application, this could yield the conclusion that the bacterial isolate analysed belongs to microorganism X, or that the blood analysed indicates genetic disorder Y, or that the fat investigated could be attributed to animal species Z. In AI terminology this process is referred to as the transformation of low level data to higher level concepts, the low level data being the analytical results, while the microorganism, genetic disorder or species represent the higher level concept. Many different chemometric techniques can be used here, for example, time-series methodology to study time-dependent results, or pattern recognition or multivariate statistics to extract meaningful information in situations where many data are available on the same object.

Important topics in chemometrics are method validation and quality assurance. Much of the validation is to do with accuracy and precision and uses well established statistical procedures. Chemometricians are interested in developing methods to do this efficiently. For instance, so-called Kalman filtering techniques are used to predict how often it is necessary to check the calibration standard [10, 11]. Another trend is to include other performance criteria, apart from accuracy and precision. An example is to be found in a recent article by Mulholland [12] concerning the use of Plackett Burman designs to determine the ruggedness of an HPLC method. Finally, there is the issue of communication with the external world. There are of course very many different subjects that can be considered under this heading. One typical chemometric approach is exemplified by an operational research study by Vandeginste [13] into the organisational aspects of the laboratory, to identify bottlenecks and identify solutions to remedy them.

The initial method selection

At this stage it is interesting to consider the automatic selection of initial conditions for an HPLC method. This requires the selection of several interacting elements such as the mode of HPLC (reversed-phase, normal phase or ion-pair chromatography), the stationary phase, the composition of the mobile phase, the type of detector and the question as to whether one is going to add a buffer to the solvent or not, and the type of buffer, etc. These elements interact because they cannot be chosen independently; for instance, the use of an electrochemical detector more or less prohibits the use of normal phase HPLC.

The process leading to the selection of a method can be represented as a decision tree, as illustrated in Fig. 2. Of course, this is only an unrealistically small branch of the tree.

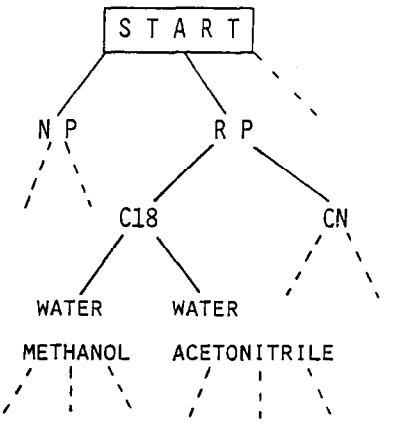

Figure 2
Example of a decision tree.

The expert's reasoning is much more complex. As already stated, the number of possible combinations of stationary phases, mobile phases, detectors, etc. is very high. This is referred to as the "combinatorial explosion". Quite clearly it is not possible to draw decision trees that permit such a large amount of potential solutions to be taken into account. One needs to reduce the number of possible decisions. Human experts use certain tactics or strategies to arrive at a solution, thereby restricting the number of possible solutions to be considered. One of the problems in constructing the decision tree is, however, that the strategies used by an expert usually exist only implicitly. The expert may have a strategy in his head but not on paper. If one is to use a decision tree then clearly the strategy must be formal and explicit.

There are two other problems with the use of a decision tree in this context. One is that many elements of the decision process are based on experience, such as the fact already cited that an electrochemical detector cannot be combined with normal-phase solvents. Finally pharmaceutical analysis does not stand still and it is therefore necessary that one should be able to add new possibilities and delete obsolete approaches.

To solve these problems the authors have investigated the possibility of using expert systems. An expert system is characterised by its structure, which consists of at least two separate parts, namely the so-called "knowledge base" and the "inference engine". The knowledge base contains knowledge about objects, such as electrochemical detection or carboxylic acids. This knowledge consists of descriptive data and relationships between

objects. The best known way of representing this kind of knowledge consists of the use of rules which are called production rules. Examples of such rules are:
"*If* X contains —COOH *then* X = acid
If X = acid *then* add acetic acid to mobile phase
If X contains —COOH *then* add acetic acid to mobile phase"
By combining these two rules one can infer that:
If X contains a carboxylic acid function then acetic acid should be added to the mobile phase.
This set of specific rules is merely an instantiation of the general logical mechanism:
"*If* A *then* B" and "*If* B *then* C" permits the inference "*If* A *then* C".

An inference engine contains such general logical mechanisms. An important characteristic of the expert system is that the inference mechanism is separate from the knowledge base containing the actual rules and facts. The inference mechanism chains the rules together to form a decision tree. These rules can consist of general or expert knowledge. Moreover, one can quite easily add rules or delete them, i.e. the knowledge base can be updated. What the expert, in this instance the chromatographer, needs to do is to formulate the rules in the first instance and enter them into the expert system.

Several expert systems for method selection in HPLC [1, 2] or for method development [3] have been proposed. The authors' expert system for method selection in HPLC of drugs contains about 120 rules at this stage. Its overall strategy [4, 5] is based on certain observations, such as the fact that all drug determinations can really be carried out on a single stationary phase, the CN bonded phase, which is suitable both for reversed-phase and normal phase operation. It is clear to the authors that, at least for drug analysis, the large number of stationary phases available is really redundant.

Although the whole system has not yet been validated, the main parts of the system have been found to give acceptable solutions in about 90% of the cases that have been chosen at random and tested.

There are two ways of developing an expert system. One can write the complete expert system including the inference engine, or one can buy the inference engine and only add the rules, i.e. the knowledge base. In the latter case one must purchase a so-called "shell" and this is of course the easiest way. This does not mean that any shell can be used. Research on the question of defining which expert shells are most suitable for chromatographic expert systems, is currently in progress in the authors' laboratory. It seems probable at this stage that pure production systems are not the best way of representing knowledge in liquid chromatography, but that a hybrid tool using both frames and rules is likely to yield the best results.

Pattern recognition

A different kind of problem is created by the surfeit of information generated by modern chromatographic or spectroscopic methods. Pattern recognition can play a role in digesting this large amount of information.

A set of chromatographic patterns yields a data matrix, consisting of the concentrations of n variables by m objects. Suppose there were only two variables. Then the objects could be visualised and their relationships studied in a two-dimensional graph. This would still be possible with three variables by making a three-dimensional graph. However, the data matrix is usually n-dimensional, n being much larger than 3, and of course it is not possible to make an n-dimensional graph. One of the aims of pattern recognition methodology is to enable the visualisation of n-dimensional data by reducing

the n-dimensional space to two-dimensional coordinates. Principal components analysis is the best known method of achieving this. Before considering this method, it should be noted that there are other methods of pattern recognition.

Supervised pattern recognition is used when one knows that the objects belong to two or more classes, for instance isolates of two different bacteria. Supervised pattern recognition is then used to derive a classification rule that permits the two classes to be differentiated.

Another application arises when it is not known whether the objects belong to more than one class, but one wants to investigate whether such classes occur. This is called unsupervised learning and the techniques used are called clustering techniques.

To understand better how the method of principal components analysis operates, one can consider a simple situation, namely the reduction of the number of dimensions from 2 to 1. This means that one reduces two-dimensional space (i.e. a plane) to a single dimension (i.e. a line) by projecting the points originally present in the plane onto a line. The question then arises as to which line one should project the points onto. In Fig. 3 two possibilities are shown. It is clear that line A (PC1) is to be preferred since the image perceived along the line is closer to the bidimensional reality. For instance, one can see along A that the points belong to two groups. This information cannot be obtained from the projection on B (PC2). One must therefore select the direction of the line so that it takes into account as much of the variance in the data as possible or, to put it in another way, so that one loses as little information as possible. In geometrical terms, the best direction is that which coincides with the axis of maximal elongation.

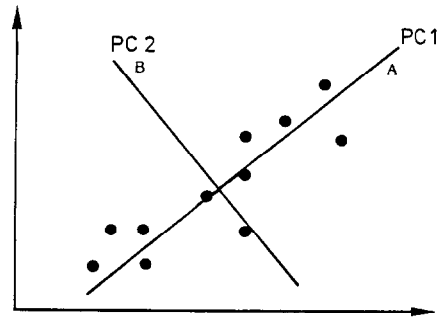

Figure 3
Reducing the dimensionality by means of principal components analysis.

The dispersion of the points along this line is then minimal. The remaining variation around the first line, which is called the first principal component (PC1), can be represented by constructing a second principal component (PC2) orthogonal to the first. In this case there is no reduction in the number of dimensions; but this is a simple transformation, such that the first dimension is more significant than the second. One can generalize this for n dimensions. One then obtains n principal components ordered in such a way that the information contained in the first is more important than that in the second, and that the second component is more significant than the third, and so on. One then reasons that the last components do not contain important information, or that they only correspond to noise, so that they can be omitted, thus reducing the number of dimensions to one, two or three and for graphical purposes, preferably to two. By plotting the sample on those two first PC axes one then obtains a two-dimensional representation of the n-dimensional data matrix. The principle of this method is simple and so much software is available, even for very small personal computers, that every

chromatographer who has to collect data sets where the chromatogram is to be used as a pattern should really be familiar with the method. An interesting application of this kind of technique is the discrimination analysis of subspecies of honeybees by means of their cuticular hydrocarbon spectra [16]. Figure 4 shows the non-linear map for these data. Non-linear mapping is a technique which is used for the same purpose as principal component analysis.

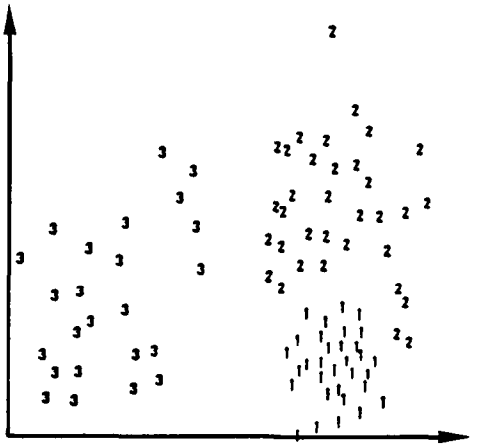

Figure 4
Non-linear map of cuticular hydrocarbon data of three honeybee subspecies: (1) Hybrid bees; (2) European bees; (3) African bees (from ref. 16).

Pattern recognition was described in the chemical literature for the first time *ca.* 1972 by Kowalski. The technique seems now to be more generally recognized. Although the applications are still limited it would seem that more and more manufacturers are starting to incorporate these methods in their instruments. One example concerns fatty acid patterns from bacteria. There is a gas chromatograph commercially available which collects gas chromatographic patterns and analyses them after pyrolysis using a pattern recognition program. The instrument is said to be able to recognize bacteria from these patterns.

Another recent example offers a solution for the detection of tampering with capsules. Pharmaceutical firms and the food industry have both been repeatedly subjected to threats that some poison would be added to their product. Unfortunately, these threats have been carried out on several occasions, the best known probably being the addition of cyanide to Tylenol in the USA in the early 1980s. Many millions of samples had to be inspected rapidly and in a non-destructive way. An unsupervised pattern recognition technique has been proposed by Lodder [17] to study the near-infrared analysis (NIRA) pattern of a capsule and thereby detect abnormalities in a simple, continuous way suitable for solving the tampering problem. In general, NIRA seems to be one of the methods where pattern recognition is destined to play an important role.

Multivariate calibration

From a chemometric viewpoint the most significant development as regards calibration in chromatography is in the use of whole spectra obtained with the photodiode array detector. This kind of calibration is multivariate in nature because the multichannel detector allows the measurement of optical absorption at many wavelengths simultaneously. The absorption at each wavelength is a variable related to the concentration of the substances present and therefore this is a multivariate situation. In

fact, one could make a data matrix of absorption at specified wavelengths against retention time, as illustrated in Fig. 5. Multivariate calibration is relevant to other topics such as peak finding or peak identification. The first operation is usually the extraction of principal components.

Again, let us first consider a simple situation in which a single substance is measured simultaneously at two UV wavelengths. The measurements fall along a line which is PC1 (see Fig. 6). One PC can therefore represent a two wavelength spectrum of a single substance. More wavelengths can be used in the same way and one can condense the whole spectrum in one PC. This by itself is not important, but now consider the situation where two substances elute under the same chromatographic peak. The pure substances would give the measurements represented by points along the broken lines and the mixtures would yield the crosses in Fig. 7.

A principal component plot can be made of these data (Fig. 8). By virtue of the fact that a significant second principal component exists, the method detects that there are two substances present. This is an initial and welcomed result and it is as far as one can go with the straightforward application of principal components. To recover the spectra of the two substances would of course be even better and to achieve this one requires factor

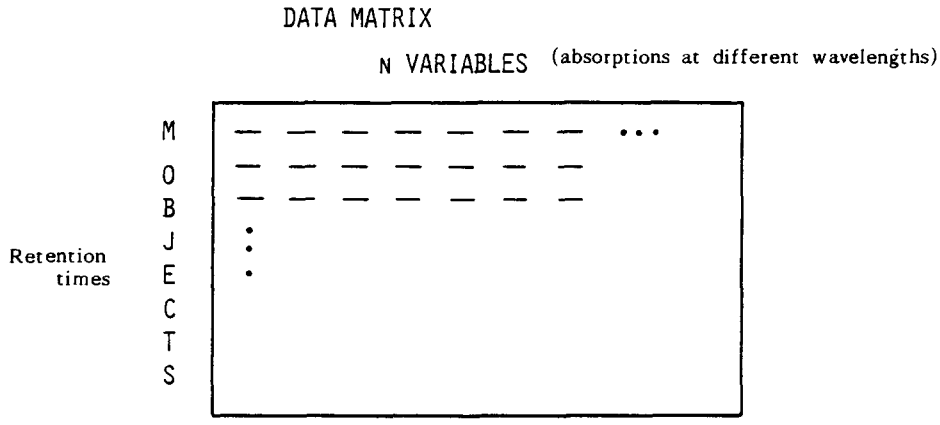

Figure 5
Data matrix of absorption against retention time.

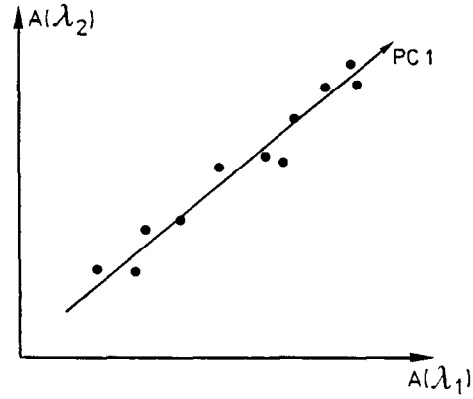

Figure 6
PC analysis of a one-component sample measured at two wavelengths simultaneously.

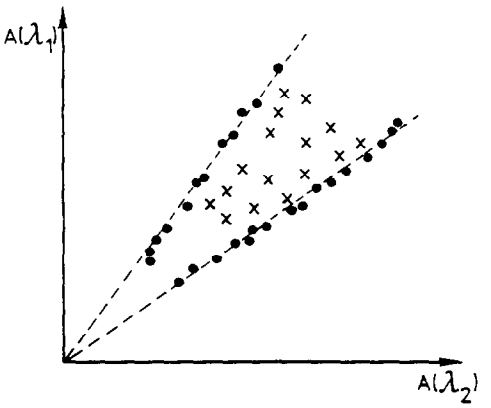

Figure 7
Analysis of a 2-component sample, measured at two wavelengths simultaneously.

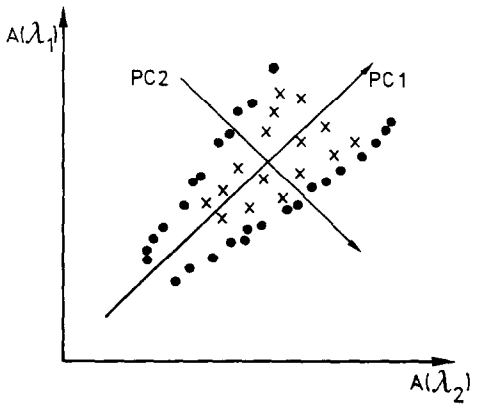

Figure 8
PC analysis of a 2-component sample, measured at two wavelengths simultaneously.

analysis. This is a method in which principal components are rotated in a meaningful fashion. One way of describing principal components is to call it a transformation of coordinates. One observes that the extraction of PC1 and PC2 requires that the centrepoint of the coordinate system be shifted from point 0 to point X, followed by rotation of the axes (Fig. 9). The angle between the axes remains the same and therefore this transformation is called orthogonal. The principal components representative of the pure substances PCA and PCB form a system of non-orthogonal axes. By transforming the principal component plot of the mixture through non-orthogonal rotation into the PCA–PCB plot, one is then able to obtain the principal component axes of the pure substances and from these one can then recover the unknown spectra. The mathematics require sophisticated matrix algebra.

Automatic learning

There are two ways of thinking, namely deduction and induction. Deductive inference is the derivation of a logical consequence from a given set of premises [18]. Chaining rules together as expert systems do, is a form of deductive reasoning. In contrast, inductive inference is a mode of reasoning that starts with specific facts and derives more general rules. Expert systems using this form of logic start with examples of situations in which certain conclusions were reached or decisions taken and try to derive from these

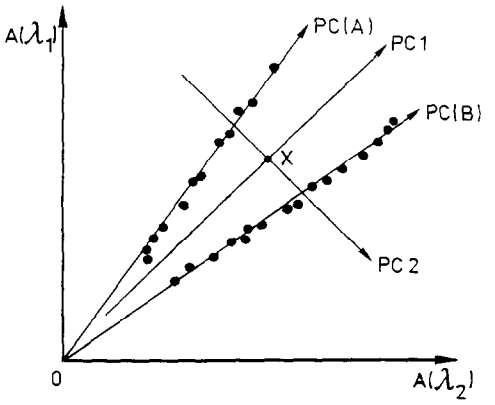

Figure 9
Orthogonal transformation of the extracted principal component-axes.

examples the rules underlying the decision process itself. The kind of expert system that is used to select methods in chromatography is a deductive system. The developer of the system has given it rules and it uses these rules to arrive at decisions. Therefore it is also called rule driven. The inductive system is not given rules: its role is to obtain them from examples. It is therefore called example drive.

An example will make clear what can be done with such expert systems. Two such systems, TIMM and EX-TRAN, were used [19] to automatically derive classification rules in a situation where a number of olive oil samples were analysed for their fatty acid composition. The samples originated from different Italian provinces and the expert system derived rules that would permit the classification of subsequent samples of unknown origin. In the particular case where it was required to discriminate between West and East Ligurian oil, the system concluded for instance, that if the linolenic acid content were higher than 15 and the oleic acid content lower than 7870, the sample would be West Ligurian, and so on. The rule set is shown in Fig. 10. Although this example stems from food analysis, it is not difficult to think of pharmaceutical or biological examples.

It is interesting to note that there is strong analogy with supervised pattern recognition, where a data matrix about two or more classes is used to derive a decision model. Indeed pattern recognition is an inductive way of thinking and the problem of

```
(100)  (LINLENIC)   :
( 46)  >   15.00    : (OLEIC)      :
             ( 45)    <   7870.00  : WE.LIG.
             (  1)    > = 7870.00  : EA.LIG.
( 54)  > = 15.00    : (LINLENIC)   :
             ( 50)    <    825.00  : (PLMTIC)   :
                           ( 2)    <   990.00 : (PLMTIC)   :
                                         ( 1)  <   785.00 : EA.LIG.
                                         ( 1)  > = 785.00 : WE.LIG.
                           (48)    > = 990.00 : EA.LIG.
                  ( 4)     > = 825.00  : WE.LIG.
```

Figure 10
The rule set, as obtained by EX-TRAN from the fatty acid composition to discriminate between East-Ligurian oil (Ea.Lig) and West-Ligurian oil (We.Lig).

classifying oils according to their origin could have been solved just as well by pattern recognition as by the use of expert systems. In fact it has been shown that certain pattern recognition methods are somewhat better than the expert systems. However, the expert systems have two big advantages. One is that they are more user-friendly; one needs less experience to be able to use them. The other and more important advantage is that pattern recognition can only be used really well with numerical data. There are, however, many situations where one needs to mix data of different types. For instance, when one uses the results of sex hormone determinations for biomedical reasons, one needs to take into account the variable sex. Such a nominal variable cannot be easily mixed with the numerical hormone data for use by pattern recognition methods. Expert systems such as EX-TRAN are better suited for this task.

Conclusion

To conclude one could state that chemometrics is or could be important at all stages of the measurement process. Chemometrics is the science that helps to make good use of information technology. It is one of the tools that will help to develop intelligent analysers, i.e. analysers that automatically select the correct method for a given problem, carry it out, validate it and interpret the results.

The implementation of some amount of logic reasoning and learning capacity is a challenging domain in which chemometricians will be most active the next few years.

Acknowledgement — This work was partly funded by the E.E.C. ESPRIT program (project nr P1570-ESCA).

References

[1] R. Bach, J. Karnicky and S. Abott, in *Artificial Intelligence Applications in Chemistry* (T. H. Pierce and B. A. Hohne, Eds), Chapter 22. ACS, Washington (1986).
[2] M. Desmet, L. Buydens and D. L. Massart, *J. Pharm. Biomed. Anal.* (1987).
[3] A. F. Fell, T. P. Bridge and M. H. Williams, *J. Pharm. Biomed. Anal.* (1987).
[4] P. Schoenmakers, in *Optimization of Chromatographic Selectivity*. Elsevier, Amsterdam (1986).
[5] J. C. Berridge, in *Techniques for the Automated Optimization of H.P.L.C. Separations*. Wiley-Interscience, New York (1985).
[6] L. R. Snyder, *J. Chromatogr.* **16**, 223–234 (1978).
[7] J. L. Glajch, J. Kirkland, K. M. Squire and J. M. Minor, *J. Chromatogr.* **199**, 57–59 (1980).
[8] S. N. Deming and S. L. Morgan, in *Experimental Optimization Methods*. Elsevier, Amsterdam (1987).
[9] P. Rousseeuw and A. Leroy, in *Robust Regression and Outlier Detection*. Wiley-Interscience, New York (1987).
[10] P. C. Thyssen, S. H. Wolfrum, H. C. Smit and G. Kateman, *Anal. Chim. Acta* **156**, 87 (1984).
[11] P. C. Thyssen, L. T. M. Prop, G. Kateman and H. C. Smit, *Anal. Chim. Acta* **174**, 27 (1985).
[12] M. Mulholland and J. Waterhouse, *J. Chromatogr.* **395**, 539–551 (1987).
[13] B. G. M. Vandeginste, J. W. A. Klaessens and G. Kateman, *Trends Anal. Chem.*, (in press) (1987).
[14] M. Desmet, G. Hoogewijs, M. Puttemans and D. L. Massart, *Anal. Chem.* **56**, 2662–2670 (1984).
[15] G. Musch and D. L. Massart, *J. Chromatogr.* **370**, 1–17 (1986).
[16] T. Mar, J. Brill, W. Bertsch, D. J. C. Fletcher and R. Crewe, *J. Chromatogr.* **399**, 277 (1987).
[17] R. A. Lodder, M. Selby and G. Hieftje, *Anal. Chem.* **59**, 1921 (1987).
[18] T. M. Michalski, J. G. Carbonell and T. M. Mitchell, in *Machine Learning: an Artificial Intelligence Approach* (R. S. Michalski, J. G. Carbonell and T. M. Mitchell, Eds), pp. 552. Tioga Publishing Company, Palo Alto, CA, U.S.A. (1983).
[19] M. P. Derde, L. Buydens, C. Guns, D. L. Massart and P. K. Hopke, *Anal. Chem.* **59**, 1868–1877 (1987).