Expert systems for multivariate calibration, trendsetters for the wide-spread use of chemometrics

M. Gerritsen, J.A. van Leeuwen, B.G.M. Vandeginste [1], L. Buydens and G. Kateman
Laboratory for Analytical Chemistry, Catholic University Nijmegen, Nijmegen (Netherlands)

(Received 1 August 1991; accepted 18 December 1991)

Abstract

Gerritsen, M., Van Leeuwen, J.A., Vandeginste, B.G.M., Buydens, L. and Kateman, G., 1992. Expert systems for multivariate calibration, trendsetters for the wide-spread use of chemometrics, *Chemometrics and Intelligent Laboratory Systems*, 15:171-184.

Chemometrics is less applied in practice than desirable. One of the reasons is the lack of software tools that allow an easy application of chemometrics by non-experts. A possible solution to this problem is the integration of expert systems with software for data analysis. This is illustrated with an example on high-performance liquid chromatography-ultraviolet data analysis. The integration of expert knowledge with chemometrical software will probably become an important future trend in chemometrical research as it is not restricted to multivariate data analysis.

INTRODUCTION

In chemometrics a large number of techniques has been introduced for multivariate data analysis. All techniques have their own specific features and characteristics. Application of these techniques enables the analytical chemist to analyze more complex samples or to analyze samples within a shorter time. The interest to introduce chemometrics in large scale routine analysis and to transfer this technology from the research stage into the laboratory is growing. Especially multivariate calibration receives much attention because important analytical techniques can be used in a multivariate mode. These technical improvements can only be used to their full benefit if the appropriate data analysis techniques are provided as well.

A number of factors obstruct the quick introduction of multivariate techniques in routine analysis. The most important factor is that the use of multivariate techniques requires a thorough understanding. Multivariate analysis is seen as difficult to learn and to apply and to a large extent this is true. Normally it requires specialists (chemometricians) to introduce and operate multivariate techniques in a laboratory. The current

Correspondence to: Dr. M. Gerritsen, Laboratory for Analytical Chemistry, Catholic University Nijmegen, Toernooiveld 1, 6525 ED Nijmegen, Netherlands.

[1] Present address: Unilever Research Laboratorium Vlaardingen, Postbus 114, 3130 AC Vlaardingen, Netherlands.

situation, where analysts are trained in specific analytical techniques, is likely to persist in the future. The trend will therefore be the incorporation of these techniques in the instrument.

On the other hand, application of most chemometrical techniques as a black box model is not possible. Specific chemometrical knowledge is necessary in order to determine when to use which technique and how to interpret the results correctly. This type of knowledge is mostly gathered by experience and normally lacks a theoretical background. Despite its highly theoretical nature, heuristics play an important role in multivariate analysis. They are a necessary addition to the mathematical techniques employed, to make the link to practical applications. Heuristics can be implemented in a computer program using expert systems.

Expert systems or decision support systems are relatively new to chemometrics. Although they have a long history in chemistry, it took some time before they were introduced as part of chemometrics. Expert systems contain the knowledge of an expert in a certain area and provide a user with decisions of expert quality in standard situations. Their applicability to chemometrics will primarily consist of providing expert advice in situations where consultation of a chemometrician is impossible or too expensive.

The first applications of expert systems in chemistry were 'stand-alone' expert systems in that they captured the knowledge of a single expert or of a group of experts, which was applied sequentially to solve a problem. These systems were often rule-based and did not contain the sophisticated software that experts, for instance in multivariate data analysis, use. Multivariate data analysis techniques are normally used in the form of a computer program hence a system combining these techniques with heuristic knowledge requires an approach that integrates conventional software with expert systems. The problem addressed above on the introduction of multivariate techniques in routine analysis is a typical example of a problem that could be solved through integration of the multivariate techniques with expert systems on when to use which technique. Especially if the integrated programs also contain an expert system on the interpretation of the results these programs will find their way into the laboratory confirming the practical use of chemometrics. Integration of conventional software with expert system type software allows to use the full power of both techniques.

The aim of the present paper is to show by means of an example of multivariate data analysis, the analysis of high-performance liquid chromatographic–ultraviolet (HPLC–UV) data, that some expert knowledge is required for application in practice. It will be shown that a combination of this expert knowledge with multivariate data analysis software into one expert system will enhance the possible use of multivariate data analysis techniques. First, the analytical system will be discussed arguing that an integrated approach using all available knowledge is usually necessary. Then, the area of HPLC–UV data analysis will be introduced with examples of when to apply multivariate data analysis techniques. Finally, it will be shown that this type of knowledge can be represented in an expert system.

THE ANALYTICAL SYSTEM

When a strategy has to be developed for the analysis of a specific object, it is necessary to obtain information on the three major parts of the analytical system: the object, the user and the analytical method. It is important to realize which information is available and of which type the information is. For instance, there may be information available about the analytical method in terms of detection limit and instrumental precision. This information can be considered as accurate and reliable. On the other hand, there is information available like the expected composition of the sample or the required precision defined by the user. This information may change during the analysis, for instance if expectations on the contents of the sample are proven to be wrong or if the user has to settle for a less strict required precision.

The analytical method itself can again be divided into three parts: the sampling, the measurements and the interpretation of the obtained data. It may be clear that the first two parts are directly related to the properties of the object and the wishes of the user. Depending on whether an object is homogeneous or heterogeneous, an experimental design should be proposed which concerns all relevant parts of the objects. The number of samples or measurements necessary can be directly related to the required precision. The type of measurements done of course depends on the properties of the object and the type of instruments available in the laboratory. The components to be analyzed can be more or less vaporous, have absorbances in specific regions of the wavelength domain etc.

Even the last part of the analytical method, the data analysis, is related to the object under investigation and the user of the information. The prior knowledge, the required information, the complexity of the sample etc. are important factors for knowing what information should be derived from the data and whether it is possible to extract this information. Experience in analyzing related objects can be used directly in building the optimal strategy for data analysis.

At all the stages of the analysis, decisions have to be made about the applicability of certain techniques. This may be at the stage of selecting the analytical technique but also at the stage of selecting the correct data analysis technique. Especially at the stage of data analysis much uncertainty exists. However, some general guidelines that would help the practical use of such methods can easily be constructed. As data analysis techniques are normally used on a computer, the guidelines are preferably also implemented in a computer program. Expert systems provide a good opportunity to do this. The possibilities of a combination of expert systems and multivariate data analysis will be illustrated here using quantitative analysis of HPLC–UV data as an example.

MULTIVARIATE ANALYSIS OF HPLC–UV DATA

HPLC–UV data are obtained by coupling a diode array detector to an HPLC column. In this way, spectra can be recorded within short time intervals, which gives a two-dimensional data matrix for each sample. In Fig. 1 a three-dimensional representation of such a data matrix is given.

When discussing the analysis of HPLC–UV data we assume that a number of decisions have already been made. HPLC has been chosen for separating the components and a diode array detector has become necessary because no selective HPLC method could be found to obtain complete separation and no selective wavelengths for the analytes of interest were available.

In order to extract information from this large amount of data multivariate data analysis is required. A number of multivariate techniques was developed and published in chemometrical literature [1–3]. They offer the possibility to extract information from systems that could not be analyzed previously. Yet in most commercial diode array software packages they are still absent, even though they have been known for a number of years already. The reason for this must be found in the fact that a number of decisions must be taken in order to perform the analysis. For these decisions some background knowledge about chemometrical techniques is required.

In Fig. 2 the potentialities of the techniques are shown schematically. The techniques are represented as vectors in a two-dimensional space of

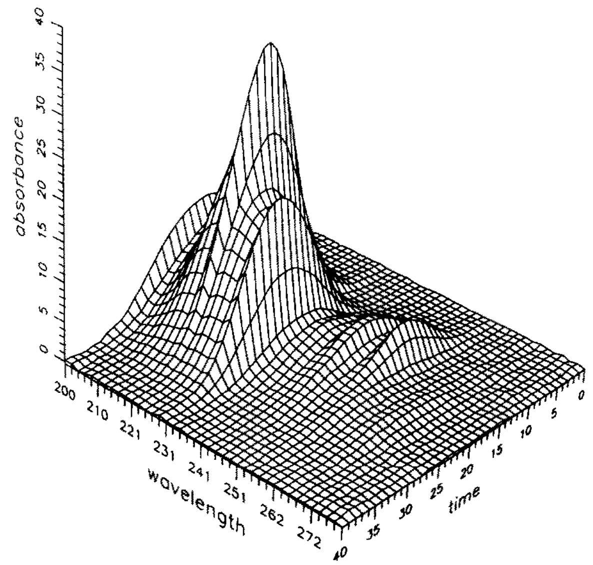


Fig. 1. HPLC-UV dataset of a three-component cluster.

knowledge. The horizontal axis corresponds to quantitative knowledge, the vertical axis to qualitative knowledge. The begin point of a vector corresponds to the knowledge that is necessary to apply that method. The end point corresponds to the knowledge obtained after multivariate data analysis has been done. When a method moves up along the vertical axis it means that qualitative knowledge is obtained. Full qualitative knowledge is defined as the situation in which the identity of all (underlying) components is known. Full quantitative knowledge means that the concentrations of all components are known. From Fig. 2 we can see that the choice of a technique depends on the knowledge present about the object (starting point), the required information (end point) and the potentialities of the multivariate techniques (connection made by method vector).

The prior knowledge usually consists of qualitative knowledge on the sample studied. The analyst e.g. knows the identity of the components present in the sample from previous studies on similar samples and he now wants to quantify them. Other possibilities are that he is interested in a specific compound or that he wants to test for the presence of a number of suspicious analytes. These starting points correspond to different points along the axis of qualitative knowledge. If the aim of the analysis is to increase the qualitative knowledge, a method of analysis should be selected that moves upward along the axis of qualitative knowledge. A typical example of such a chemometrical technique is the method of target factor analysis or target testing (TT), which was described by Malinowski [4]. This method can be used to test for the presence of a hypothetical compound. This compound is present if the corresponding spectrum lies inside the spectrum space of the measurements. Spectra of hypothetical compounds can be tested individually in the presence of interfering compounds. A requirement for applying this technique is of course the availability of the spectra of the compounds that have to be tested. This means that some prior knowledge should be present (point D in Fig. 2). Various criteria have been used to verify the hypothesis. Ho et al.'s method based on

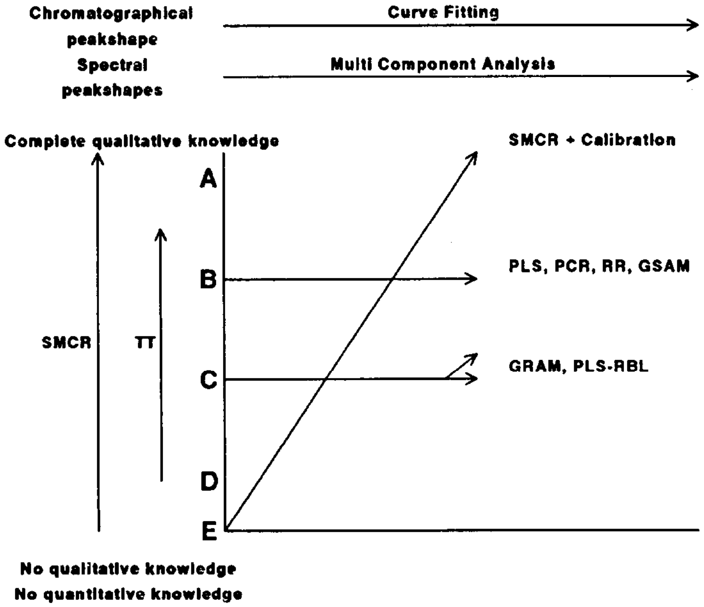


Fig. 2. Schematic representation of the possibilities of a number of multivariate techniques that can be used for the analysis of HPLC–UV data. SMCR = Self-modelling curve resolution; PLS = partial least squares; TT = target testing; PCR = principal component regression; GSAM = generalized standard addition method; RBL = residual bilinearization; RR = ridge regression; GRAM = generalized rank annihilation method.

Bessel's inequality [5] and Malinowski's SPOIL function [6] were developed first but were not based on statistics. This means that the limits of acceptance should be found empirically. Later, both Lorber [7] and Malinowski [8] developed tests based on statistics of which Malinowski's *F* test was also described for situations with missing data. A simple method like Bessel's inequality however can be very useful when large amounts of similar data have to be analyzed. Strasters et al. [9] evaluated the technique of target testing for peak tracking. He concluded that the correct components could be identified by target testing even in the presence of mobile phase effects.

A second vector that moves along the axis of qualitative knowledge is that of self-modeling curve resolution (SMCR). This name has been given to the family of algorithms that tries to calculate the underlying pure profiles by making only very general assumptions about the model behind the multivariate data. This means that the prior knowledge needed is very general and does hot depend on the specific sample of interest. The first application of self-modeling curve resolution was given by Lawton and Sylvestre [10]. Their method was able to determine estimates of pure spectra in two-component clusters and required selective regions in the spectra of the pure components. Later on, new techniques were developed which tried to avoid the disadvantages mentioned above. Examples of very general curve resolution techniques are the algorithms of iterative target transformation factor analysis (ITTFA) [11,12] and evolving factor analysis (EFA) [13,14]. These techniques are not restricted to a certain number of components and only make very general assumptions. For a practical situation, this means that these techniques can be used for the analysis of completely unknown mixtures (point E in Fig. 2). As a result of the analysis, information is obtained on the identity of all underlying compounds. Besides calculating the pure profiles, SMCR techniques can also be used for quantitative purposes if an extra calibration step is added. In this calibration step the surfaces of the spectra or concentration profiles can be related to the corresponding concentrations. In fact this should only be done for the analyte of interest which means that quantification can be done in the presence of an interferent. At the moment a large number of SMCR techniques have already been published. An overview has been given recently by Hamilton and Gemperline [3].

In some cases complete qualitative knowledge is already present. In this case we can distinguish between knowledge in the spectral domain and knowledge in the time domain. Knowledge in the wavelength domain consists of the availability of one or more of the underlying spectra. If all spectra are known they can be projected directly on the data matrix (mostly called multi-component analysis, MCA) yielding the corresponding concentration profiles. Care must be taken for deviations in the spectra due to e.g. the mobile phase. Strasters [9] showed that small deviations in the reference spectra due to mobile phase effects produced large errors in the estimated concentrations. He also showed that some improvements could be obtained by applying the non-negativity criterion as described by Lawson and Hanson [15]. Knowledge in the time domain usually consists of a certain peak model e.g. a Gaussian peakshape. By nonlinear regression these models can be fitted to the overall profile in order to obtain the pure concentration profiles [16]. In this case only one chromatogram of the complete two-dimensional data matrix is used.

A typical qualitative analytical problem that may occur in practice is peak tracking. An algorithm for peak tracking can be used to follow the retention behavior of the analytes in HPLC optimization. The techniques discussed above (TT, SMCR, MCA) can all be used for this; however, they expect different types of prior knowledge. Another type of qualitative analysis for which SMCR techniques can be used is the peak purity test. In this case the identity of the impurity is sometimes of no importance for the analyst. This means that it is enough to determine whether the number of principal components for the peak studied is higher than one and so the technique of principal component analysis can also be used. Gemperline and Hamilton [17] described the effect of relative concentrations, spectral similarity and the chromatographic resolution on the limit of detection for severely overlapped peaks using principal component analysis. They reported a method to determine the net signal due to minor components when overlapped with major components.

Two other points are marked on the axis of qualitative knowledge. Point C corresponds to the situation where there is a specific analyte of interest. This situation often occurs in practice, e.g., when the concentration of a toxic compound has to be determined in a food product. A few multivariate techniques, which have been developed recently, are able to quantify an analyte of interest in the presence of one or more interfering compounds. These techniques make use of the bilinear structure which is present in some kinds of two-dimensional data. For HPLC–UV data this structure is present if Beer's law is applicable. Ho et al. [5] presented the rank annihilation method to quantify a particular component without having to know the identity of the rest of the components. Later, this iterative algorithm was improved [18,19] and resulted in the generalized rank annihilation method (GRAM) of Sanchez and Kowalski [20], a direct calibration method that can be used for the determination of several analytes simultaneously. Recently Öhman et al. [21] developed an algorithm called residual bilinearization (RBL) which can be used in combination with a calibration technique (e.g. partial least squares) in order to remove interferents with a bilinear structure. In comparison with GRAM, RBL assumes bilinearity of the interferents present whereas GRAM assumes bilinearity of the analytes of interest. The method vectors of RBL and GRAM are split at the end. This is caused by the fact that the pure spectrum of the interferent can be calculated if there is only one interferent present. If there are more interferents, only linear combinations of their spectra can be calculated. As mentioned above, SMCR techniques can also be used for quantification if interferents are present. These techniques assume bilinearity of all components present. In comparison to GRAM and RBL, they also give qualitative information on the interferents present (higher position on the qualitative axis), which can be very useful for the validation of the results.

Above the GRAM and RBL point another starting point B is given which corresponds to the knowledge necessary for applying partial least squares (PLS), principal component regression (PCR) and ridge regression (RR). PLS and PCR are especially suited for the situation of having a lot of correlated variables, which is the case for HPLC–UV data. Wold et al. [22] showed that multiway measurements could be unfolded in order to apply ordinary multivariate calibration techniques. Because in this way the more-dimensional structure behind the data was lost, an extra restriction could be added on the rank of the loadings. Just like GRAM, RBL or SMCR, these calibration techniques can be used to quantify a number of analytes of interest. The higher position on the axis, however, suggests that more prior knowledge is required. The difference comes from the composition of the calibration data matrices. For GRAM, RBL or SMCR techniques only the analytes of interest have to be present in the calibration samples, whereas for PLS or PCR all analytes that contribute to the signal have to be present in varying amounts. This implies that in fact for PLS, more knowledge is needed to build the required calibration samples.

In practice the availability of sufficient calibration samples is not only limited by the prior knowledge but also by the available time. The advantage of calibrating in the presence of interferents is evident. In practice it occurs very often that specific compounds need to be quantified in a complex matrix. Finding out the identity of the interferents and making calibration samples of them or finding a selective method for the analyte of interest is mostly very time-consuming and expensive.

Another property of the object studied can be the presence of matrix effects. When matrix effects are present, the matrix should be incorporated in the calibration model. A well-known multivariate technique that has been developed especially for this is the generalized standard addition method (GSAM) of Saxberg and Kowalski [23]. Known amounts of all analytes present in the unknown sample should be added to the matrix of the unknown sample. This means that this technique assumes the same prior knowledge as PCR and PLS. Standard addition can also be applied in combination with SMCR techniques or GRAM. For these methods only additions of the analytes of interest are required.

After the correct calibration samples have been composed, spectra are mostly recorded in the region between 190 and 400 nm. In order to obtain better results, data pretreatment can be an important step. Mostly the lower wavelength regions should be left out due to excessive noise caused by the mobile phase absorption [9]. Mostly parts of the spectra with a very low absorbance coefficient are also eliminated. For ITTFA and EFA the removal of a constant background absorbance is an important step prior to the analysis.

After the proper multivariate technique has been selected in accordance with the problem, the chosen technique should be used in an optimal way in order to obtain as much information as possible. A part of the analysis that requires expert knowledge is the selection of the correct model parameters. A very important parameter for the methods described above is the number of factors or principal components. Although a lot of theoretical criteria have been developed in literature for the determination of the rank of a matrix or the estimation of the number of principal components, it has been shown that some empirically developed criteria give better results. In general these criteria are based on an interpretation of the results. Especially for relatively new complex techniques like ITTFA, EFA, GRAM and RBL, the interpretation of the results is not straightforward. In order to evaluate the results, some parameters have been used in several papers. Strasters et al. [24] showed that the optimal number of principal components to use for ITTFA calculations was equal to the number of analytes in the mixture, which was sometimes much lower than the mathematical rank of the matrix. They developed a criterion based on the physical interpretation of the ITTFA results in order to evaluate the ITTFA calculations and to choose the optimal dimension of the factor space. Öhman et al. [25] used the correlation between the calculated spectra and the library spectra in order to select the optimal factor space for the GRAM. They also looked at the negative parts of the residuals in order to evaluate the RBL results. These rules have been found to be useful and should be used when applying these techniques in practice.

After the results of the analysis have been obtained they should be validated. A way to get an impression of the quality of the results is to evaluate the techniques for a lot of data with varying complexity. In this way it can be investigated which factors have a large influence on the quality of the solutions and to which levels of these factors the techniques can be used. Theoretical simulations are very popular for doing this because data with varying compositions can be composed very easily and quickly. However it must be realized that the results obtained for these simulations give too optimistic a picture for usage in practice because ideal data are used which do not contain any model errors and hence can only be used as an underestimation of the error obtained in practice. In order to obtain simulations with a higher resemblance to real data, Gerritsen et al. [26] used combinations of one-component data to simulate more-component clusters (mixtures). From evaluation studies it has been shown that for SMCR techniques factors like chromatographic resolution, peak height ratios and spectral similarities have an influence on the quality of the resolution. Vandeginste et al. [27] showed that ITTFA could be used up to resolutions of 0.6 for two-component systems with about equal concentrations. These results were later confirmed by Seaton and Fell [28]. Gemperline [29] introduced an adjusted algorithm for ITTFA by introducing some automatically generated linear inequality constraints. He showed that better results could be obtained for poorly resolved clusters. Moreover the algorithm was tested for tailing peaks and varying noise ratios. Strasters et al. [9] evaluated the algorithms of multi-component analysis, iterative target transformation factor analysis and target testing for peak tracking of systems with varying levels of resolution and peak height ratios. Multi-component analysis could be used for systems with the lowest resolution but small errors in the spectra used produced large errors in the concentration estimates. Target testing showed to be less sensitive to solvent effects but more sensitive to the factor resolution. ITTFA does not require any prior knowledge on the spectra but was the most sensitive to the factor resolution. Calculations for systems with a resolution lower than 0.25 were very inaccurate. From simulation studies Strasters et al. [30] developed a quantitative model to judge the reliability of the derived UV spectra from the observed resolution, the observed spectral similarity and the observed concentration. EFA was also applied to simulated chromatographical data by Maeder [14] and Maeder and Zuberbühler [31]. Maeder and Zilian [32] applied EFA to real HPLC data and stressed the importance of baseline correction prior to analysis. A combination of EFA and rank annihilation factor analysis and EFA was reported by Gampp [13]. Recently Keller and Massart [33] showed that the EFA results could be improved for detecting small amounts of a spectrally similar impurity in the presence of a major peak, when a moving window with a fixed number of spectra was taken for the factor analysis. Curve fitting was evaluated by Vandeginste and De Galan [16]. An important requirement for applying curve fitting showed to be prior knowledge on the number of bands present. Other important factors were the correctness of the mathematical model used, the relative peak heights and the resolution of the profiles. A scheme for the limits and the applicability of curve fitting was reported. Öhman et al. [25] compared the GRAM with the combination of RBL and PLS. RBL showed to give slightly better results but the interpretation of the results was very difficult and a combined usage of techniques was proposed. A disadvantage of techniques like RBL and GRAM is that only information on the analytes of interest is obtained.

The results of these simulation studies can first be used to obtain a better theoretical understanding of the multivariate techniques. The relative influence of a number of factors on the quality of the solutions can easily be determined and the effects of small changes in the algorithm can also be evaluated. Secondly these results can be used in practice to get an impression of the error obtained for the data analyzed. However in order to use these results for evaluating the obtained results some prior information on the composition of the cluster is necessary. In practice this information can be the result of a previous analysis or is related to the kind of sample analyzed. If this knowledge is absent a SMCR technique could be used first in order to get an impression of the number of analytes, resolution etc. When the results of different evaluation studies and applications are combined rules can be developed that will help the user to decide which technique gives the best results for his specific problem. Evaluation studies in which more techniques are applied to the same data are especially useful in developing this kind of rules.

EXPERT SYSTEMS

Expert systems incorporate the knowledge of experts in a certain area into a computer program. Expert systems have a long history in chemistry. One of the first expert systems, the Dendral system, had the interpretation of mass spectrometric data as a subject. After Dendral many other systems followed, many of them pertaining to the interpretation of spectra of various kinds. Most of these systems never got beyond the stage of a working prototype. The practical acceptance of expert systems in chemistry is relatively low. Although this situation holds for other fields as well, working expert systems can be found in many application areas. In medicine for instance, the number of systems used in practice is growing.

A reason for the low acceptance of expert systems in chemistry is that their subject area may be less suitable for an expert system approach. In spectrum interpretation the amount of expertise is very large and it is difficult to separate well-defined sub-areas that are of practical importance. Therefore, the systems developed on the interpretation of spectra did not reach a level of expertise comparable to the need in practice.

Recently, expert systems with different knowledge domains were commercialized, thus proving that expert systems do have a future in routine analysis [34]. Expert systems can be especially useful if they are integrated with other 'conventional' software.

Chemometrics is one of the areas where expert systems may find successful application. Although chemometrics may be seen as a theoretical area at first, it appears that to be able to use chemometrical techniques, one needs to be experienced in the field. As can be seen in the paragraph on

TABLE 1

Examples of frames that could be used in an expert system for HPLC–UV data analysis

Required information	
Purpose:	Peak purity test
	Qualification
	Quantification
Analytes of interest:	All
	Part
Maximum error in concentrations	0–50%
Data analysis results	
Reference spectra:	Not available
	Available for all analytes
	Available for analytes of interest
Quality spectra	Recorded in same mobile phase
	Recorded in different mobile phase
Peak model:	Non available
	Approximately known
	Exactly known
Number of components:	Known
	Unknown
Concentrations:	Not available
	Analytes of interest
	All analytes
Calibration samples	
Analytes with varying concentration:	All
	Analytes of interest
Standard addition in sample matrix:	Yes
	No
Complexity unknown sample	
Expected resolution:	Unknown
	0.0–1.0
Expected peak ratios:	Unknown
	0.01–1.0
Expected spectral correlation:	Unknown
	(−1.0)–1.0
Expected matrix effects:	Yes
	No
Data analysis method	
Analysis method:	SMCR
	Target testing
	ITTFA
	EFA
	MCA
	PLS
	PCR
	RR
	GRAM
	SMCR + calibration
	Curve fitting
	PLS–RBL

the analysis of HPLC–UV data, many different mathematical and statistical techniques exist with slightly different possibilities and application areas. Quite often, a combination of these techniques is necessary to extract the desired information from the data matrix. Selecting the right data analysis method proves difficult enough but combining several data analysis methods into one data analysis strategy is even more difficult. It requires specialist knowledge of the available techniques and much practical experience to be able to design such a strategy. However, if such knowledge would be widely available, together with the mathematical data analysis techniques, many HPLC–UV analyses would yield better results and much more difficult separation issues could be solved.

In many ways, HPLC–UV data analysis presents an ideal case for an expert system. Normally, the design of a data analysis strategy is done by hand and requires highly trained specialists. If the knowledge to design data analysis strategies were more widely available, analytical chemistry would be able to solve more difficult problems quicker. The design of an expert system on quantitative analysis of HPLC–UV data will be discussed in this paper. It serves to illustrate how two chemometrical techniques, multivariate data analysis and expert systems, can be combined to produce systems that are suitable for use in routine analysis.

OUTLINE OF AN EXPERT SYSTEM ON HPLC–UV DATA ANALYSIS

The paragraph on multivariate data analysis contains a description of a number of well-known data analysis techniques and the criteria on when to use which technique. In many cases, it is not straightforward which technique must be used and a combination of techniques may be appropriate. It may even be so that the decision to use a second or third technique can only be taken after the previous analysis has been finished. The analysis of HPLC–UV data therefore requires a data analysis strategy. With the paragraph on data analysis as a source of expert knowledge, the basics of an expert system on this subject will be outlined.

TABLE 2

Examples of rules that could be used in an expert system for HPLC–UV data analysis

If	Purpose is quantitative knowledge
And	Analytes of interest is part
And	Expected matrix effects is yes
Then	Analysis method is GRAM
And	Analysis method is SMCR + calibration
And	Standard addition in sample matrix is yes
If	Purpose is quantitative knowledge
And	Analytes of interest is part
And	Analytes with varying concentrations is all
And	Expected matrix effects is yes
Then	Analysis method is GSAM
And	Standard addition in sample matrix is yes
If	Purpose is qualitative knowledge
And	Analytes of interest is part
And	Reference spectra is available for all analytes
And	Quality spectra is recorded in same mobile phase
Then	Analysis method is MCA
If	Purpose is qualitative knowledge
And	Analytes of interest is all
And	Reference spectra is available for all analytes
And	Quality spectra is recorded in different mobile phase
And	Expected resolution is > 0.25
Then	Analysis method is ITTFA
If	Purpose is qualitative knowledge
And	Analytes of interest is all
And	Reference spectra is available for all analytes
And	Quality spectra is recorded in different mobile phase
And	Expected resolution is < 0.25
Then	Analysis method is target testing
If	Purpose is quantitative knowledge
And	Analytes of interest is part
And	Analytes with varying concentrations is all
And	Standard addition in sample matrix is no
Then	Analysis method is PLS, PCR, RR
If	Purpose is qualitative knowledge
And	Analytes of interest is all
And	Reference spectra is not available
Then	Analysis method is SMCR

An expert system normally consists of three parts: a knowledge base, an inference engine and a user interface. An important difference between expert systems and conventional software is that in expert systems the knowledge how to solve a problem is separated from the technique that manipulates the knowledge to produce new information. In conventional software the two are integrated. For a complete discussion on the structure of expert systems see for instance ref. 35.

In expert systems, the expert knowledge resides in the knowledge base. The inference engine is normally chosen from a limited set of standard inferencing techniques, that are available in most expert system building environments. Hence, the distinguishing part of the expert system is the knowledge base that contains the expert knowledge. There are various types of representation techniques that can be used to represent the expert knowledge in the knowledge base. The most commonly used knowledge representation technique is the rule-frame representation scheme. In this scheme, all concepts used in the knowledge domain are represented in frames. A frame is a three-level data storage structure in which the top level is the name of the concept. At the second level the attributes that characterize the concept are listed. At the third level, the features can get actual values, thus describing an example of the concept. Examples of frames can be found in Table 1 where some frames from a data analysis knowledge base are given.

In the rule-frame representation scheme, the expertise is represented in rules. Rules with an IF THEN format describe the relations that exist between the various frames and attributes. In Table 2 some rules of the data analysis example are given. The rules in the knowledge base are not, as in conventional programs, linked to each other in decision trees. On the contrary, each rule is a separate entity. During a consultation of the expert system, the inference engine will chain the appropriate rules together to form a line of reasoning. In every consultation, a new line of reasoning can therefore be followed, adapted to the specific problem at hand.

In Table 1, the basic concepts that must be used in the expert system are given in the form of frames. The 'required information' frame represents the requirements the user puts to the data analysis process. It is important to define these requirements at the start of the design of the data analysis strategy. The aim of the data analysis strategy should be to produce the desired information with as little experimentation and calculation as possible. In the HPLC-UV example this means for instance that if the user is only interested in one specific analyte it is not necessary to extract information on all the analytes present. If the correct multivariate technique is chosen this can mean a reduction of the number of required calibration samples.

The data analysis results frame represents all the knowledge about the data that is available at the start of the design of the data analysis strategy. If, for instance, spectra of the target compounds are available, this may be a good reason to use target testing. Also, intermediate results like the results produced by a qualitative method as SMCR can be stored here. The data analysis results frame also contains an attribute that can be regarded as the ultimate goal of the expert system. If the concentration of the compound of interest is known (or all concentrations are known) the desired information has been produced.

Other information the expert system needs is information on the calibration solutions that are available or can be produced. Standard additions are important if matrix effects are expected in the unknown sample. In this case techniques like GRAM or GSAM should be used. The number of analytes with varying concentrations in the calibration samples is an important factor to consider. Especially for the analysis of complex samples or samples with a number of unknown interferents this factor should be considered in choosing a technique for data analysis.

If available, some information on the expected complexity of the unknown sample may also be very useful. From evaluation studies it is known that factors like chromatographic resolution, spectral correlations and peak height ratios have an influence on the quality of the solutions. When it has been shown that some technique performs better for e.g. data with a low resolution than another technique these results should be used in practice.

Finally the methods of analysis that can be selected by the expert system can be stored in the frame data analysis method. For some analytical problems more than one method is suited (e.g. a number of SMCR techniques). In this case a number of options should be given by the expert system. In other situations it can be advisable to use more than one technique (e.g. one for qualitative knowledge and another for quantitative knowledge, or e.g. a second technique in order to validate the results of the first technique). In this case the expert system should select more than one technique.

Now all the necessary frames have been defined from the paragraph on data analysis. To complete it, the actual expert knowledge must be added. Normally, this knowledge is represented in rules. Some example rules are given in Table 2. The rules presented are distilled from the text in the data analysis paragraph. In practice rules should be extracted from theory, evaluation studies and specific applications published in chemometrical literature. The rules can use the data stored in the frames and can produce new information from them. The rules are given in pseudo-code in Table 2 and can be easily read.

For instance, the first two rules deal with the problem of quantitative analysis in the presence of interfering compounds. If matrix effects are expected to be present the calibration samples should be made by standard additions in the sample matrix. The distinction between GSAM and SMCR or GRAM is made by the required calibration samples. Rule 3 states that MCA is a very suited technique if the spectra of all components are known very accurately. Rules 4 and 5 state that if the reference spectra are known approximately, ITTFA and target testing are better alternatives. The choice between TT and ITTFA depends on the expected resolution. Rule 6 states that PLS, PCR or RR can be used if the calibration samples contain all factors in varying amounts. Rule 7 finally gives SMCR techniques as the best alternative to start the analysis if no prior knowledge is present.

The rule-set in Table 2 is far from complete. Much more rules can be extracted from literature or in discussion with chemometrical experts. However, it is beyond the scope of this article to produce pseudo-codes for an entire expert system on HPLC–UV data analysis. This is only an example of how chemometrical techniques can be combined to form powerful systems. Examples of other areas of interest are easily found. Aspects like data pretreatment, method selection or data interpretation require a lot of expert knowledge and can be encountered also in other chemometrical areas e.g. calibration, pattern recognition, optimization and experimental design. The combination with expert systems specially developed for specific chemometrical areas could be used to make this knowledge available in routine environments.

CONCLUSIONS

In the present paper, it is stated that chemometrical techniques could be introduced more smoothly in practice if a specific kind of chemometrical knowledge is supplied with it. This knowledge could be used to decide in which situation which technique should be used, how it should be used optimally and how the results should be interpreted. As an example the analysis of HPLC–UV data was discussed. It was shown that the choice of a technique strongly depends on the prior knowledge of the user and the knowledge required. Additionally the complexity of the sample studied and the available calibration samples were important factors to be considered. The interpretation of the obtained results could be improved by making the results of simulation and evaluation studies in literature accessible to the real user of the techniques. Expert systems have proven to be successful in representing this kind of knowledge and making them accessible to a non-expert. A few examples were given of representations of chemometrical expert knowledge.

The described situation for HPLC–UV occurs often in chemometrics. For a lot of applications, complementary chemometrical techniques are available which all were used for a diversity of problems in literature. The reason, however, for the development of more than one technique is mostly that none of the techniques is generally applicable or is working equally well in every situation. Typical examples of such techniques are algorithms for pattern recognition, optimization and experimental designs. In practice the analytical problems have a large variety which makes the use of one single chemometrical technique mostly insufficient. The step to a practical situation can only be made when the knowledge presented in chemometrical literature can be translated to an accompanying expert system. Expert systems could be built for different fields in chemometrics. Only by combining the possibilities of various techniques the full power of chemometrics comes to expression.

REFERENCES

1 H. Martens and T. Næs, *Multivariate Calibration*, Wiley, Chichester, 1989.

2 P.J. Gemperline, Mixture analysis using factor analysis I. Calibration and quantitation, *Journal of Chemometrics*, 3 (1989) 549–568.

3 J.C. Hamilton and P.J. Gemperline, Mixture analysis using factor analysis. II. Self modeling curve resolution, *Journal of Chemometrics*, 4 (1990) 1–13.

4 E.R. Malinowski and D.G. Howery, *Factor Analysis in Chemistry*, Wiley, New York, 1980.

5 C.-N. Ho, G.D. Christian and E.R. Davidson, Application of the method of rank annihilation to quantitative analyses of multicomponent fluorescence data from the video fluorometer, *Analytical Chemistry*, 50 (1978) 1108–1113.

6 E.R. Malinowski, Theory of error for target factor analysis with applications to mass spectrometry and nuclear magnetic resonance spectrometry, *Analytica Chimica Acta*, 103 (1978) 339–354.

7 A. Lorber, Validation of hypothesis on a data matrix by target factor analysis, *Analytical Chemistry*, 56 (1984) 1004–1010.

8 E.R. Malinowski, Theory of the distribution of error eigenvalues resulting from principal component analysis with applications to spectroscopic data, *Journal of Chemometrics*, 3 (1988) 49–60.

9 J.K. Strasters, H.A.H. Billiet, L. De Galan, B.G.M. Vandeginste and G. Kateman, Evaluation of peak recognition techniques in liquid chromatography with photodiode array detection, *Journal of Chromatography*, 385 (1987) 181–200.

10 W.H. Lawton and E.A. Sylvestre, Self modeling curve resolution, *Technometrics*, 13 (1971) 617–633.

11 B.G.M. Vandeginste, W. Derks and G. Kateman, Multicomponent self modeling curve resolution in high performance liquid chromatography by iterative target transformation analysis, *Analytica Chimica Acta*, 173 (1986) 253–264.

12 P.J. Gemperline, A priori estimates of the elution profiles of the pure components in overlapped liquid chromatography peaks using target factor analysis, *Journal of Chemical Information and Computer Science*, 24 (1986) 206–212.

13 H. Gammp, M. Maeder, C.J. Meyer and A.D. Zuberbühler, Quantification of a known component in an unknown mixture, *Analytica Chimica Acta*, 193 (1987) 287–293.

14 M. Maeder, Evolving factor analysis for the resolution of overlapping chromatographic peaks, *Analytical Chemistry*, 59 (1987) 527–530.

15 L. Lawson and R. Hanson, *Solving Least Squares Problems*, Prentice Hall, Englewood Cliffs, NJ, 1974.

16 B.G.M. Vandeginste and L. De Galan, Critical evaluation of curve fitting in infrared spectrometry, *Analytical Chemistry*, 47 (1975) 2124–2132.

17 P.J. Gemperline and J.C. Hamilton, Conditions for detecting overlapped peaks with principal component analysis in hyphenated chromatographic methods, *Analytical Chemistry*, 61 (1989) 2240–2243.

18 C.-N. Ho, G.D. Christian and E.R. Davidson, Simultaneous multicomponent rank annihilation and applications to multicomponent fluorescent data acquired by the video fluorometer, *Analytical Chemistry*, 53 (1981) 92–98.

19 A. Lorber, Quantifying chemical composition from two-dimensional data-arrays, *Analytica Chimica Acta*, 164 (1984) 293–297.

20 E. Sanchez and B.R. Kowalski, Generalized rank annihilation method, *Analytical Chemistry*, 58 (1986) 496–499.

21 J. Öhman, P. Geladi and S. Wold, Residual bilinearization. part I: Theory and algorithms, *Journal of Chemometrics*, 4 (1990) 79–90.

22 S. Wold, P. Geladi, K. Esbensen and J. Öhman, Multi-way principal components- and PLS-analysis, *Journal of Chemometrics*, 1 (1987) 41–56.

23 B.W.H. Saxberg and B.R. Kowalski, Generalized standard addition method, *Analytical Chemistry*, 51 (1979) 1031–1038.

24 J.K. Strasters, H.A.H. Billiet, L. De Galan and B.G.M. Vandeginste, Strategy for peak tracking in liquid chromatography on the basis of a multivariate analysis of spectral data, *Journal of Chromatography*, 499 (1989) 523–540.

25 J. Öhman, P. Geladi and S. Wold, Residual bilinearization. Part II. Application to HPLC–diode array data and comparison with rank annihilation factor analysis, *Journal of Chemometrics*, 4 (1990) 135–146.

26 M.J.P. Gerritsen, N.M. Faber, M. van Rijn, B.G.M. Vandeginste and G. Kateman, Realistic simulations of HPLC–UV data for the evaluation of multivariate techniques, *Chemometrics and Intelligent Laboratory Systems*, 12 (1992) 257–268.

27 B.G.M. Vandeginste, F. Leyten, M. Gerritsen, J.W. Noor and G. Kateman, Evaluation of curve resolution and iterative target transformation factor analysis in quantitative analysis by liquid chromatography, *Journal of Chemometrics*, 1 (1987) 57–71.

28 G.G.R. Seaton and A.F. Fell, Multivariate analysis of non-homogeneous peaks in liquid chromatography, *Chromatographia*, 24 (1987) 208–216.

29 P.J. Gemperline, Target transformation factor analysis with linear inequality constraints applied to spectroscopic–chromatographic data, *Analytical Chemistry*, 58 (1986) 2656–2663.

30 J.K. Strasters, H.A.H. Billiet, L. De Galan, B.G.M. Vandeginste and G. Kateman, Reliability of iterative target transformation factor analysis when using multiwavelength detection for peak tracking in liquid chromatographic separations, *Analytical Chemistry*, 60 (1988) 2745–2751.

31 M. Maeder and A.D. Zuberbühler, The resolution of overlapping chromatographic peaks by evolving factor analysis, *Analytica Chimica Acta*, 181 (1986) 287–291.

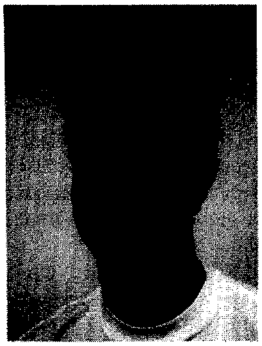
32 M. Maeder, and A. Zilian, Evolving factor analysis, a new multivariate technique in chromatography, *Chemometrics and Intelligent Laboratory Systems*, 3 (1988) 205–213.

33 H.R. Keller and D.L. Massart, Peak purity control in liquid chromatography with photodiode-array detection by a fixed size moving window evolving factor analysis, *Analytica Chimica Acta*, 246 (1991) 379–390.

34 J.A. Van Leeuwen, B.G.M. Vandeginste, L. Buydens and G. Kateman, Expert systems in chemical analysis, *Trends in Analytical Chemistry*, 9 (1990) 49–54.

35 G.F. Luger and W.A. Stubblefield, Artificial Intelligence and the Design of Expert Systems, Benjamin/Cummings, Redwood City, CA, 1989.

BIOGRAPHICAL SKETCH



Mathieu Gerritsen received his M.S. in chemistry in 1987 from the Catholic University of Nijmegen. From 1987 to 1991 he worked towards a Ph.D. under the direction of Professor G. Kateman and Dr. B.G.M. Vandeginste at the Analytical Department of the Catholic University of Nijmegen. His research interests in chemometrics include the quantitative analysis of bilinear data. At the moment he is employed as an analytical chemist by Hoogovens IJmuiden.

J.A. van Leeuwen received his Ph.D. in chemistry from the Catholic University of Nijmegen, Netherlands, in 1990. He is currently employed by AKZO Research Laboratories in Arnhem, Netherlands. His current interests are in multivariate analysis, neural networks and knowledge-based systems.

Lutgarde Buydens received her Ph.D. at the University of Brussels in 1986 at the Department of Pharmaceutical and Biomedical Analysis (Professor Massart). After a post-doctoral period at the same department she is since 1989 working as Assistant Professor at the University of Nijmegen at the Analytical Chemistry Department (Professor Kateman).