*Pure & Appl. Chem.*, Vol. 55, No. 12, pp. 1861-1868, 1983.
Printed in Great Britain

0033-4545/83 $3.00+0.00
Pergamon Press Ltd.


# THE IMPACT OF CHEMOMETRICS ON MICROCHEMICAL ANALYSIS

D. Luc Massart and Guido Hoogewijs

Vrije Universiteit Brussel, Farmaceutisch Instituut,
Laarbeeklaan 103, B-1090 Brussel (Belgium)

Abstract - This article describes the impact of chemometrics on analytical chemistry with the example of chromatography. It focuses on applications of information theory, experimental optimization and pattern recognition.

## CHEMOMETRICS

Chemometrics has been defined as follows (1) : "Chemometrics is the chemical discipline that uses mathematical and statistical methods
a) to design or select optimal measurement procedures and experiments and
b) to provide maximum chemical information by analyzing chemical data.

In the field of analytical chemistry, chemometrics is the chemical discipline which uses mathematical and statistical methods to achieve the aim of analytical chemistry, namely the obtention in an optimal way of relevant information about material systems".

In fact, if you read this your conclusion probably is : "This is exactly what analytical chemists always have been doing". This conclusion is correct. Chemometrics is not really a new discipline. However, between 1972 and 1976, there has been a sudden increase in interest because of the simultaneous introduction of at that time rather exotic methods such as information theory, operations research, experimental design, pattern recognition and many others. This led to something of a movement and a movement needs a name : "chemometrics" was born. It is now a healthy young adult admitted in the family of analytical disciplines. Nowhere has it prospered to such a degree as in the Netherlands where it has achieved wide recognition. Figure 1 gives the official so-called "attention domains" of university research in analytical chemistry. Chemometrics is not only one of the five analytical research disciplines, it also is the one which is practised to the highest extent after research about separation methods.

| University | Atom spectrometry | Molecule spectrometry | Electrochemical analysis | Separation methods | Chemometrics |
|---|---|---|---|---|---|
| VUA | | x | x | x | |
| UvA | x | | | x | x |
| RUU | x | x | | x | x |
| KUN | | | | | x |
| THD | x | | | x | x |
| THE | | x | | x | |
| THT | | | x | | x |
| RUG | | | x | x | x |
| RUL | x | x | x | x | |

Figure 1. Attention domains in analytical chemistry in the Netherlands

This, by itself, proves that chemometrics has fulfilled a need. The first one is an intellectual one (2) : analytical chemistry has long been considered as a second hand science because it lacks a character, a philosophy of its own. Chemometrics permit to give a fundamental backbone to analytical chemistry as the science of chemical measurement.

The largest practical impact of chemometrics has been that it has relaxed, transformed or given a new sophistication to some firmly embedded but unwritten laws of analytical chemistry. You must all remember classical remarks such as :

1. Do not change more than one parameter at a time !
2. Look at your data !
3. Are you sure that that calibration line is a straight one ?
4. Does that separation procedure yield sufficiently pure fractions ?
5. The selection of a separation method is a matter of experience.

What has chemometrics made of this ?
1. You should change several parameters simultaneously (experimental design for optimization methods).
2. Keep looking at your data, even if they are more than three dimensional (principal components, clustering and other methods related to pattern recognition).
3. You don't need a straight calibration line and, anyway, straight lines are not so simple as they look (model selection, weighting schemes, non-linear calibration).
4. Never mind if the fraction is not pure : chemometrics will do the rest ! (deconvolution techniques, multicomponent resolution, the generalized standard addition method).
5. The selection of a separation method is a matter of strategy (information theory, operations research).

We will not be able to discuss all these points. In particular we will not discuss points 3 and 4. The necessity of correct calibration is widely accepted as is the need for statistical treatment to achieve it and deconvolution is performed routinely by each modern chromatographic apparatus. Therefore, we will discuss mainly points 1, 2 and 5 and apply them to chromatography. This discussion will be a highly personal one : we will discuss mainly the ideas and results of our own laboratory. One of the main themes of our research is to develop what we call an intelligent chromatograph. By this we mean an instrument that is able (i) to select its own operating conditions such as the solvents composing the mobile phase, (ii) to optimize these conditions, (iii) to treat the resulting signal to as to yield correct concentration profiles and (iv) to interpret them as a means to generate characteristic patterns. Because the time available is limited we will discuss only the following points :
1. Selection of operating conditions
2. Optimization of operating conditions
3. Generation of characteristic patterns.

## SELECTION OF OPERATING CONDITIONS IN CHROMATOGRAPHY

Unlike in GC, method development in HPLC and TLC is primarily concerned with the selection of an appropriate mobile phase. This is usually performed in two main steps. The first one consists of selecting the solvents which will make up the eluent and it also consists of choosing an initial ratio of the eluent components. The second step consists of selecting an optimal composition.

We have undertaken the selection of a first acceptable composition in HPLC by studying a very large and important family of organic substances, namely the basic drugs. There are several thousands of such compounds of pharmaceutical or medical interest and determining them in pharmaceuticals, in biological fluids, or even in cosmetics is one of the most important practical problems of analytical chemistry. Our approach was the following : 100 basic substances representing all the more important chemical families were selected and at the same time the literature was investigated and all the more promising chromatographic systems (i.e. a combination of a particular stationary phase with a particular mobile phase) were noted. After a screening step, 16 of these systems were retained and the 100 basic drugs were chromatographed in each of the systems. It remained now to select the best combination of systems i.e. the combination which allows the best separation of 100 substances. This was carried out in two steps.
Step 1 consisted of the selection of the individually "best" systems. The word "best" means that one has to select an optimization criterion and this is where information theory comes in. In step 2 one selected the best combination of systems.

For the sake of simplicity, let us consider thin-layer chromatography to explain our approach (2 - 4). In TLC the migration of a substance is characterized by its Rf-value. Assuming that substances with an Rf-value differing by at least 0.05 units can be distinguished, one can divide the Rf range into groups (0 - 0.05, 0.06 - 0.10, ...). This leads to a simplified model in which one considers two substances to be separated if they are in different blocks. This is a rough model because it considers two substances with Rf-values of 0.05 and 0.06 as separated. Nevertheless, the model allows easy calculation of approximate values of the information content. Moreover, it is not really important at this stage to see what substances are separated from each other but rather to evaluate how well the substances are spread out over the plate. If a set of 100 substances is investigated and $n_i$ substances fall into group i, the information content is calculated as

$$I = \sum_i - \frac{n_i}{n_o} \log_2 \left(\frac{n_i}{n_o}\right)$$

To understand the meaning of I it is of interest to investigate some extreme results :

1. When all the substances fall into the same group, I = 0 : no information is obtained.
2. When all the substances fall into a different group, the maximal value of I is obtained : no uncertainty is left.

These two extreme situations show that what one looks for is a system which should cause equal spreading of the Rf values over the entire range.
When one TLC system does not yield sufficient information, one must combine two systems or more. These two systems together yield an information content of

$$I = I(1) + I(2) - I(1,2)$$

where I (1,2) is the correlated information. If both system 1 and 2 separate substances A and B, then this particular bit of information is obtained with both systems and should not be considered twice : one says that correlated information is present.
The end result of this (very sketchy) theoretical treatment is that, when systems are combined, one should look for :
a) individually good systems
b) systems that yield different information.

Returning now to the HPLC of bases, this approach was applied to the 16 candidate systems and the 100 test substances. The result was the selection of a minimum number of preferred HPLC-systems which together yield a maximum of information. Concretely, two preferred HPLC-systems which should be used in priority whenever basic drugs are to be chromatographed were selected. The stationary phase is in both cases a CN-bonded phase and the two preferred eluents are hexane-dichloromethane-acetonitrile-propylamine (50:50:25:0.1) and acetonitrile -water-propylamine (90:10:0.01).

This in turn has led to the development of an entire standardized analysis scheme for basic drugs, the conception of which in part originates from our experience with practical pharmaceutical analysis. Our laboratory indeed performs a lot of analyses of pharmaceutical formulations (for the government) and of biofluids (for the industry). Since these analysis assignments are widely divergent in terms of both substances and matrices requiring analysis, we experienced that large amounts of time and effort were spent not in performing the actual determinations but during method development for each particular analysis problem. This holds true for the HPLC-conditions as well as for the work-up procedures. Hence the idea of developping a standardized analysis scheme which would be as generally applicable as possible to a large group of substances, namely basic drugs and related compounds.

The standardized analysis strategy consists of (5-7) :
1. an ion-pair extraction step always using the same solvent (chloroform) and either di (2-ethylhexyl) phosphoric acid at pH 5.5 or octylsulphate at pH 3.0 as ion-pairing reagent
2. direct injection of the extract in one of the two preferred HPLC-systems following
3. optimization of the volume ratio of the eluent components, which usually involves only a few trials with eluents of which the volume ratio of the components is slightly changed.

The ion-pair extraction using ion-pairing reagents with a relatively large C-skeleton gurantees high extraction recoveries while the large discriminating power of both preferred HPLC-systems guarantees high specificity.

The advantages of such a standardized analysis strategy, particularly for laboratories performing widely divergent pharmaceutical analyses are obvious :
1. column choice is omitted since all separations are carried out using a single stationary phase
2. optimization of the mobile phase composition is greatly facilitated since it merely consists of finetuning the volume ratio of the components
3. optimization of the extraction conditions is redundant
4. back- and re-extraction are omitted since the extracts are directly injected onto the column
5. since the extraction and chromatographic conditions are applicable to almost any basic drug, whether polar or nonpolar, fast and easy development of appropriate analysis conditions is nearly alway possible.

Some applications are given below.

TABLE 1. Applications to pharmaceutical dosage forms :
- Control of label claims and stability
- Determination of degradation products

| Analytes | Pharmaceutical dosage form | Analytes | Pharmaceutical dosage form |
|---|---|---|---|
| Ketotifen | Zaditen® Syrup | Promethazine | Phenergan® Creme |
| Oxomemazine | Doxergan® Syrup | Diphenhydramine | Caladryl® Creme |
| Diphenhydramine | Diphenhydramini Emulsio® | Lidocaine | Xylocaine® Gel |
| Ephedrine | Ephedronguent® Ointment | Fenfluramine | Ponderal® Unicaps |
| Chlorpheniramine | Polaramine® Syrup | Caffeine | Vizocaf® Coated tablets |
| Metoclopramide | Primperan® Suppositories and syrup | Tetracaine | Tablets |
| Paracetamol<br>p-aminophenol<br>Degradation products | Ben-u-ron® Suppositories | Carbinoxamine<br>Phenylephrine | Rhinopront® Capsules |
| Caffeine<br>Ergotamine<br>Ergotaminine (stereoisomer)<br>Degradation products | Cafergot-PB® Suppositories | Flupentixol<br>Melitracen | Deanxit® Coated Tablets |
| | | Mebeverine<br>Degradation products | Tablets |

TABLE 2. Application to biofluids

| Matrix | Analytes | Matrix | Analytes |
|---|---|---|---|
| Saliva | Carbamazepine and carbamazepine-10,11-epoxide | Plasma | Aprindine and desethylaprindine |
| Saliva | Aminopyrine | Plasma | Melperone |
| Urine | Mebeverine and metabolites | Plasma | Sulpiride |
| Plasma | Papaverine | Plasma | Imipramine |
| Plasma | Mepyramine | | Desipramine |
| Plasma | Thonzylamine | | 2-OH-imipramine |
| Plasma | Methapyrilene | | 10-OH-imipramine |
| Plasma | Acebutolol and diacetolol | | 2-OH-desipramine |
| Plasma | Thioridazine and mesoridazine | | 10-OH-desipramine |
| | | Plasma | Mebeverine and metabolites |

TABLE 3. Application to cosmetics

| Separation and determination of aminophenols, phenylenediamine derivatives and related compounds in hair dye products of the oxidative type |
|---|
| p-aminophenol |
| m-aminophenol |
| o-aminophenol |
| p-phenylenediamine |
| 2-nitro-p-phenylenediamine |
| 4-nitro-o-phenylenediamine |
| 4-methoxy-m-phenylenediamine |
| 2-hydroxy-4-aminotoluene |
| α-naphtol |
| β-naphtol |
| 1,2,3-trihydroxybenzene |

Presently our research in this field is directed towards expanding the applicability of the HPLC-strategy also to acidic and neutral compounds. We felt that we still could use the single stationary CN-bonded phase but that the use of quaternary mobile phases instead of the binary reversed phase and ternary normal phase eluent should be necessary. The number of mobile phase solvents was enlarged from 4 to 6 and now comprises besides hexane, dichloromethane, acetonitrile and water also methanol and tetrahydrofurane.

This is still less than the 8 solvents (and two stationary phases) proposed by Glajch, Snyder, Kirkland and colleagues (8-9) in their optimization routine for HPLC-separations but it is obvious that unlike in our strategy for basic drugs where the initial mobile phase composition is near to the final composition, the use of a quaternary mobile phase necessitates a formal optimization strategy. At this point chemometrics comes in again.

## OPTIMIZATION STRATEGIES

Let us look at the formal optimization strategies that are available. An optimization problem is one in which the value of a given criterion is optimized as a function of the experimental conditions. Suppose for the sake of simplicity that the resolution $R_S$ between two pairs is chosen as a criterion and that the composition of a three solvent mixture must be optimized. When the concentrations of two solvents are given, the third is also known so that there are two variables. If $R_S$ were to be measured for all the possible values of the two variables $x_1$ and $x_2$, the two following graphical representations might be obtained. The optimization problem is then to find the top of the response surface of figure 2.

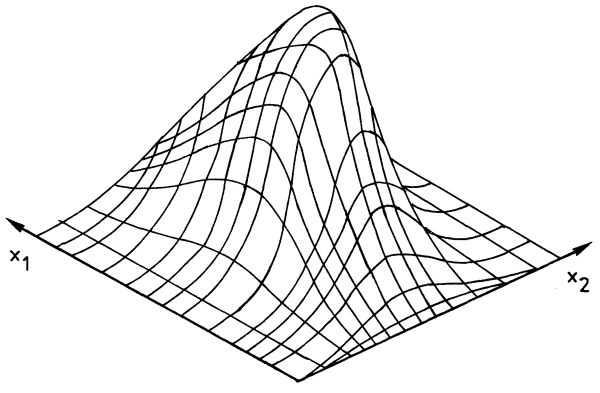


Figure 2. Response surface for two factors (from ref. 4).

One can choose among many different strategies or experimental designs (4). There are two broad categories namely the so-called simultaneous designs and the sequential designs. The first consist in pre-planning several experiments in an orderly way, to fit the surface through the experimental results and to compute the optimum. The simplest possible design for two and three variables is given in figure 3.

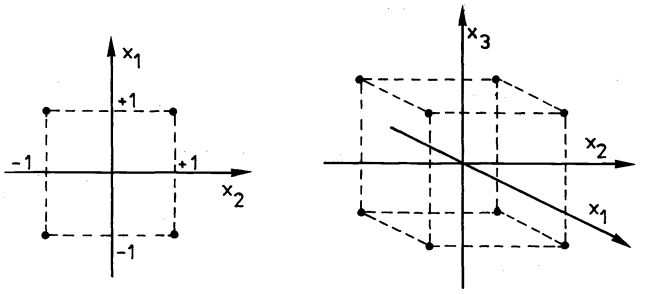


Figure 3. Factorial designs for two and three variables (from ref. 4).

More complex and more efficient designs are also possible as is shown for three factors in figure 4. These methods are generally termed factorial designs. The most simple sequential design, the Simplex, is illustrated in figure 5. In the present instance one would start with experiments 1, 2 and 3. Since 2 gives the worst result, one concludes that the optimum must be looked for in the opposite direction and one tries out 4. The worst response of triangle (simplex) 1, 3, 4 is now 3. This leads to experiment 5 and triangle 1, 4, 5. One observes that in successive triangles one moves towards the optimum O. The method as given here has been perfected : so-called modified and supermodified simplexes have been developed (10). These are still more efficient and permit to find the optimum in fewer moves.

Fig. 4

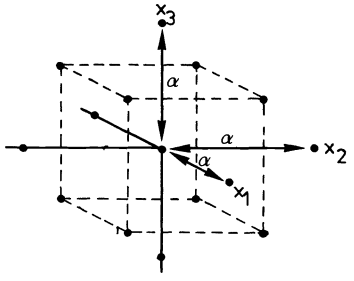


Fig. 5

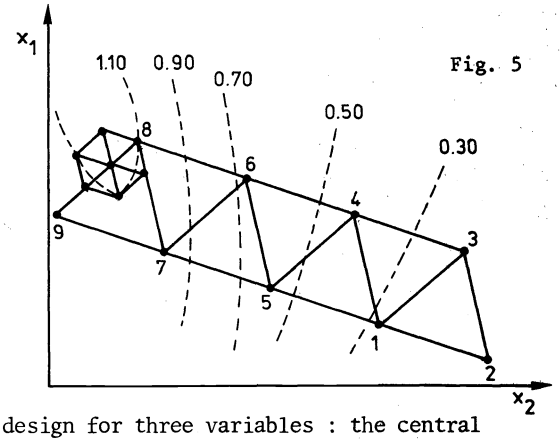


Figure 4. Example of a more complex design for three variables : the central composite design (from ref. 4).

Figure 5. Simplex optimization. The broken lines give responses, the points are experiments.

Both kinds of methods have been described here for only two variables but can be applied without problem to more variables. Whether one prefers the simultaneous or sequential methods depends on the aim of the optimization. The simultaneous experiments better describe the region around the optimum and are therefore more efficient in describing the effects of the variables. The sequential methods are more efficient in finding an optimum. In the field of chromatography both kinds of methods have recently been incorporated in instruments. Glajch (8) preferred a method which, although it is called a simplex method, is really a factorial one, while Berridge (11) preferred a sequential simplex. We are developing ourselves an instrument based on a sequential procedure. Interfacing the chromatograph to a microcomputer and the use of appropriate hard- and software for data acquisition, data processing and chemometrics together therefore make it now possible to select optimal conditions without intervention of the chromatographer. This kind of instrument is representative for a new generation of "intelligent" instruments. Similar principles have been used for instance in AAS (10) or ICP (12).

## GENERATION OF CHARACTERISTIC PATTERNS

In many instances the chromatogram is used as a pattern. An oil specialist looking at a fatty acid GLC pattern is able to tell at a glance that it belongs to, for instance, an olive oil. The specialist is using his experience as a pattern recognizer. There are mathematical techniques that also permit to recognize or classify patterns and the collection of these methods is called pattern recognition.

It is completely impossible to even attempt giving a more or less complete overview of pattern recognition as applied in analytical chemistry. We will consider here only one aspect : its application as an extension of the human eye. We know that the human eye and mind together constitute a formidable pattern recognizer, but the human eye is restricted to three-dimensional observations. This is a handicap as will be observed when we try to see how the analytical chemist interprets experimental observations. When only one variable is measured for a number of samples, the analytical chemist will make a one-dimensional graph in the hope of discerning for instance a grouping among the samples (figure 6 a). When he measures two variables a two-dimensional graph results and when he measures three variables a computer or a lot of patience may help him to make a three-dimensional representation (figures 6 b and c). However, when there are more than three variables, it becomes impossible to make such a graph. An oil chromatogram may easily contain twenty peaks and therefore a comparison of many chromatograms then becomes a twenty-dimensional problem.

Some pattern recognition methods such as principal components and non linear mapping permit to reduce the n-dimensional space to an m-dimensional one with $m << n$. The interesting application in the present context is when $m = 2$ because in that case a two-dimensional representation is obtained. Of course this must be done in such a way that the salient features of the distribution of the data in the original m-dimensional space can be recognized in the two-dimensional one, i.e., one must minimize the loss of information due to reduction of the number of variables.

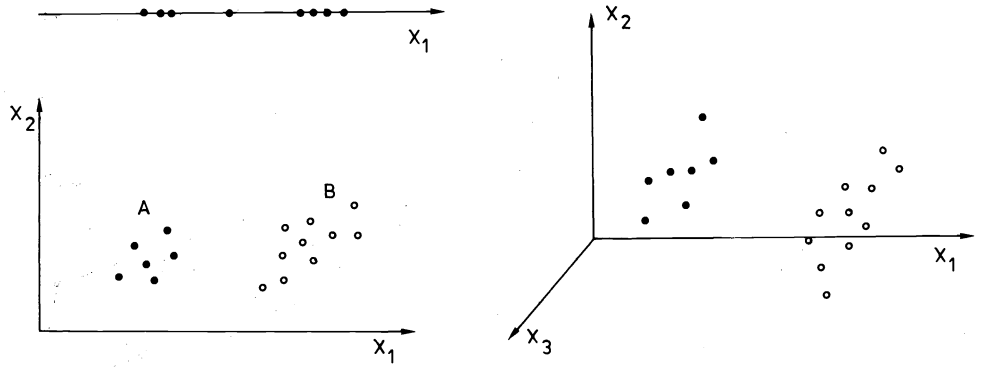


Figure 6. Plots of analytical results for resp. one variable (univariate), two variables (bivariate) and three variables.

Before having a look at a real situation, it is preferable to consider a didactical example : the reduction of a two dimensional space (a plane) to a one dimensional one (a line). The reduction is achieved by projecting the points (representing samples in the two-dimensional graph) on the line. Several directions are possible and, clearly, line a, for instance, is better than line b (figure 7) because the original structure is better preserved. Indeed the two-dimensional graph clearly shows that the samples are grouped in two groups. So does line a, but line b would lead to the (wrong) conclusion that there is only a single group of samples. This is also the philosophy of the method when going down from n dimensions to 2 (or 3) : the samples are projected on a two-dimensional plane (principal component 1 versus 2) so as to conserve as much as possible the original data structure. Several examples of results in analytical chemistry are given in a recent book (11). One example is given in figure 8. It concerns the classification of oils. The concentrations of 8 fatty acids in about 500 oil samples were determined by GLC. The oils originated from 9 italian regions. Since 8 variables were determined, this means that the oils should be looked at in an 8-dimensional space. This is impossible and the two dimensional plot of principal component 1 against 2 is shown in the figure. This figure immediately shows which classes can be separated and which classes cannot. For instance, there is no difficulty in distinguishing oils for Calabria and Umbria, but on the other hand, a discrimination between Calabria and Sicily is much more difficult. Also when a new oil, the origin of which is not known, is

Fig. 7

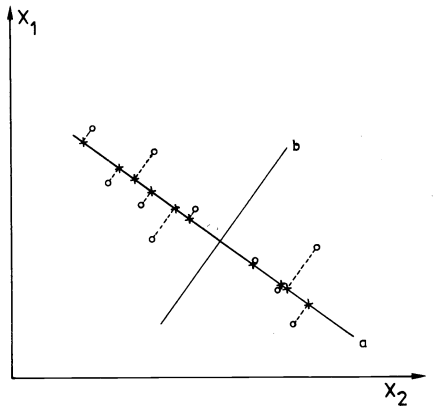


Fig. 8

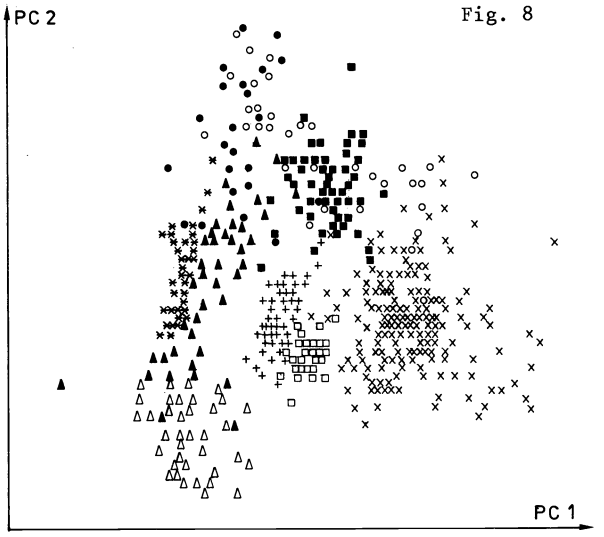


Figure 7. Objects (o) characterized by two variables $x_1$ and $x_2$ are projected (x) on line a. Line b : see text.

Figure 8. Principal components map of olive oils from nine different regions of Italy. ● = North Apulia, o = Sicily, ■ = Calabria, □ = Coast Sardine, ▲ = East-Liguria, Δ = West-Liguria, + = Inner Sardinie, x = South Apulia, ✱ = Umbria

analyzed, its situation in the PC plot can be used to give an idea of the region from which it originates. Principal components is the best known pattern recognition technique : it is, by far, not the most sophisticated one and much more detailed results can be obtained. Nevertheless, the information obtained with this rather simple method is already so large that we expect that pattern recognition techniques will one day be standard software equipment of chromatographs, so that the results delivered by a gas chromatograph will not be that x % oleic acid, y % stearic acid, etc., are present but : "This sample is a first grade olive oil from South Apulia".

## CONCLUSION

Data treatment of analytical data has changed because of the introduction of three new factors in the analytical chemists environment. These are :
- the microprocessor built in or interfaced with the instrument
- the availability of statistical or mathematical packages to run on these computers
- the advent of a new kind of analytical chemist, the chemometrician.

As they operate now they are still mixed blessings, but, as chemometricians, we feel certain that the balance is positive. Together they will develop a new generation of intelligent instruments.

## REFERENCES

1. The Chemometrics Society, Chemometrics Newsbulletin no 7 (1981).
2. D.L. Massart, Z. anal. Chem. 305, 113 (1981).
3. D.L. Massart, J. Chromatogr. 79, 157 (1973).
4. D.L. Massart, A. Dijkstra and L. Kaufman, Evaluation and Optimization of Laboratory Methods and Analytical Procedures, Elsevier (1978).
5. G. Hoogewijs and D.L. Massart, J. Pharm. Biomed. Analysis, in press (1983).
6. G. Hoogewijs and D.L. Massart, in press (1983).
7. G. Hoogewijs and D.L. Massart, J. Pharm. Belg. 38, 75 (1983).
8. R. Lehrer, Int. Lab., nov.-dec., p. 76 (1981).
9. J.L. Glajch, J.J. Kirkland, Anal. Chem. 55, 319A (1983).
10. P. Van der Wiel, L.G. Van Dongen, B. Vandeginste and G. Kateman, to appear in Laboratory Microcomputer (1983).
11. G. Kornblum, B. Vandeginste, personal communication.