CHEMOMETRICS IN ANALYTICAL CHEMISTRY

S. BHATTACHARJEE*

INTRODUCTION

Over the last two decades, chemometrics has carved out a firm niche in the field of analytical chemistry. This term was supposedly coined by the Swedish Physical Organic Chemist S.Wold in 1972. Together with the Americal Analytical Chemist B.R.KowaLski, Wold formed the International Chemometrics Society. Subsequently, there have been several major reviews [1-7], ACS symposia[8,9], a National Bureau of Standards Conference [10] a NATO School [11] a series of monographs [12] and several textbooks [13,14]. Acceptace of chemometrics as a growing discipline has also been emphasized by [15,16] two international journals devoted to chemometrics.

Chemometrics in analytical chemistry is essentially application of the basic statistical methods for analysing the analytical data albeIt with a much broader scope. Similar methods, however, were used for many years by biologists, geologists and numericaL taxonomists. Ability of the modern analytical instruments to generate large amounts of data rapidly made it imperative that new approaches are needed to interpret what may be described as large multivariate data matrices.

This definition of chemometrics has generated further debate on the scope of the subject. It is argued that since the statistical packages are also linked with the expert systems, library searching, graphical and databasing routines etc., chemometrics should encompass a much broader gamut of methods rather than mere application of the statistical methods to analytical data. The central theme of chemometrics, must, however, be the laboratory instrument and how computational methods can increase the productivity of the experiments.

* *Scientist, Analytical Chemistry Division, National Metallurgical Laboratory, Jamshedpur-8 31 007*

This essay attempts to provide on overview of the chemometric methods within the ambits of analytical chemistry. There is no attempt to derive the results mathematically, though citations have been made wherever appropriate. This article will concentrate, principally, on gaining an understanding as to how chemometrics can be useful in the modern analytical laboratory.

SAMPLING METHODS

The job of an analyst in an analytical laboratory starts with sampling. These samples may be physical entities, or analytical parameters. Several statistical texts have been written over the last three decades aiming to help the experimenter optimise sampling methods [17-19] and arrive at the representative sample. Following subsections will describe in what way chemometrics can help the experimenter optimise his time and instrumentation more.

Sampling Theory : Sampling Theory in analytical chemistry has been described in detail by Kateman and Pijpers[20]. The chemist frequently encounters continuous processes often described as time series. Examples are continuous industrial processes where deviations from pre-set limits can result in poor quality of a product or sometimes industrial accidents. Naturally occuring time series found in geochemistry when measuring compounds down the core, in environmental chemistry when monitoring seasonal diurnal changes in composition, in clinical chemistry when monitoring biorhythm and finally when tracking reaction kinetics by methods such as stopped-flow.

Frequency of sampling is the most important parameter, which, in turn, depends on the type of question to be addressed. In some instances, an exact future trend is anticipated which is normally a cyclic trend. Cyclical time series are well known in spectroscopy,

economic forecasting and geological palaeoclimatic studies where the effect of the changing orientation of the earths orbit around the sun changes the sea surface temperature and influences the proportion of varius chemicals in geochemical samples of different ages. In time series analysis, the regular spacing of samples is extremely important, tthough irregular spacings can also be negotiated. Depending upon the sample size, frequency of sampling and the sample spacing, the analyst chooses his next course of action which might be a fourier transform, curve fitting or the method of interpolation, linear, spline, rectangular etc.

Simplex Method : In sampling a time series, normally a single variable is measured as a function of time. However, the analyst is often interested in more complex processes, where more than one independent variables are involved. These may be, for example, the atmospheric corrosion as a function of humidity, temperature, rainfull and sulfur-dioxide concentrations or the yield of a synthetic reaction as a function of temperature, solvent, catalysts etc. These problems require optimisation techniques. The traditional approach is to perform an experiment under certain conditions, measure the performance of one variable and optimise, change the second variable keeping the first variable at the optimum value and continue till all the variables are optimised. However, it is possible to model this process by considering the experiments as samples in multidimensional space, the dimensions being the number of independent parameters. The resultant model is also known as the objective function or the response surface. However, there is a potential risk of missing this response surface if the variables are optimised one by one.

Simplex optimisation is a method for searching these surfaces[21-23] and is based on the observation that most experimental surfaces are likely to be relatively smooth and not contain sudden

discontinuities. The first step in the analysis is to define an allowed region within which to search for the response. The second stage is to choose a step size, which a reasonable amount by which to change each variable, when searching for an optimum. The third step is to define the response surface and establish initial condition. The simplex terminates when all possibilities of making extra measurements yield a worst response. Provided that the response surface is well behaved and the initial conditions and step size have been well chosen, simplex methods should yield the best conditions using a small number of experiments.

Factorial Methods : Factorial approaches to experimental design contrast with simplex approaches in that several experiments can be performed simultaneously. There are many types of design, but the overall philosophy is to sample a number of points on the response surface : regression analysis can thus be used to fit this surface, and find the maximum if there is one.

There are many possible designs, but they are all dependent on an intuitive estimate of the shape of the response surface in advance. The region of most variability is that in which most experiments should occur. A four factorial experiment is one in which four factors which may be temperature, pressure, concentration and pH, for example, are to be varied.

Factorial methods have the advantage over simplex approaches in that they enable the entire response surface to be constructed. They are not only used to find optimum, but can also be used to examine how variables interact, which conditions are the most crucial and so on, and tend to give a much better idea of the overall effect of experimental conditions on an analytical process.

CHOICE AND OPTIMISATION OF ANALYTICAL CONDITIONS

Once the sampling is done, the next job of the analyst is to analyse the sample by the best possible means that he has. This necessitates the choice of the right instrument from an array of instruments, minimise the number of experiments and decide upon how many replicate measurements are necesary. In all such instances, there are established chemometric techniques.

Choice of Masuring Techniques : Same sample may be analysed by more than one techniques. The analyst must make a pragmatic and judicious choice. For example, if the objective of a series of measurements is to detect outliers, then the question to be addressed is "which technique detects outliers most efficiently ?" If several techniques are used, can the number of measurements be reduced as more than one provide similar informations ? Procrustes [24-25] is one approach to comparing the information in two or more sets of measurements. Principal Component Analysis (PCA), is also used for reducing the dimensionality of the data. This enables the analyst to look at the data more conceptually. An alternative approach is Partial Least Squares (PLS) [26]. Canonical correlations [27] could also be employed. A radically different approach to comparison of techniques comes from information theory [28]. However, the information theoretical approach has come under considerable critiism [29].

Replicate Measurements : Replicate measurements fall under routine practice. The normal reason for replicate analysis is to provide confidence in the analytical method and sampling strategy.

The replicate analysis, however, can be used in a far more useful manner. The replicate analysis may be used for investigating inter-laboratory, inter-analyst and inter-machine variability. The ANOVA,

analysis of variance, method [30] can be employed for the analysis of internal variability, and can be used to anwser question such as which instrument is more reliable ? A more sophisticated approach includes multi dimensional ANOVA (MANOVA) [32]. After ANOVA, the analyst can use replicate informations to establish the accuracy of his methods and determine which techiques are the most efficient. And once the analyst finds his answers with the ANOVA, he dispenses with the replicate measurements, preferring instead to use mean readings in the subsequent analysis.

DATA PROCESSING

Conventionally there is a tendency to confuse chemometrics with chemical pattern recognitions, whereas, chemical pattern recognition is only a small part of chemometrics. At the end of sampling and experimentation, the analyst is left with a vast array of numbers. Unless a problem is fairly simple, chemometrics methods can be used to interpret these large multivariate data sets. Four excellent books on pattern recognition are available [33-36] and together they provide comprehensive summaries of the literature over the last 15 years.

Classification and clustering : Classification of samples is one of the principal goals of pattern recognition. Methods of classification can be divided into unsupervised and supervised approaches. The difference between these methods is that for supervised approaches a test (training) set is required to set up a model. In unsupervised methods no prior test set is required. Cluster analysis (an unsupervised approach) was described in detail by Massart and Kaufman [37].

There are a very large number of computer packages available to perform the cluster analysis. The result is normally displayed graphically as a dendrogram. The majority of commercial softwares,

CLUSTAN, SAS, SPSS give the user the opportunity of a wide variety of options. More sophisticated softwares, however, have been developed. MASLOC is an agglomerative method [38] based on location theory. CLUPLOT [39] is a method for detection of the number of significant clusters rather than thinking of individual objects.

In the supervised approach, SIMCA [40-42] (Soft Independent Modelling of Class Analogy) is, perhaps, the most popular. In SIMCA a principal component model is established for each class. Principal component analysis aims to reduce the data from a large number of original measurements to a small number of principal trends. SIMCA method has been implemented in packages such as SAS [43]. NIPALS algorithm is used to calculate principal components rapidly [44]. Other approaches include SPHERE [45,46], UNEQ[47] AND PRIMA [48].

Correlation : Another problem frequently encountered by the analyst is to relate different sets of data. This problem can be solved by multivariate calibration. The relationship is unlikely to be simple. There are a large number of techniques in this area, but PLS (Partiod Least Squares) is the best known method [49-52]. Recently the method has been extended to version PLS2 [53]. A further sophistication is n-dimensional PLS[44].

There is considerable debate as to whether PLS really is the best approach to multivariate calibration. Historically, this method was one of the first chemometric techniques to become available as a user friendly microprocessor based package. However, there is no guarantee that the PLS approach theoritically extracts the most reliable information. Other methods such as principal component regression and canonical correlation can also be employed but are less readily available to the analyst. PLS is only really useful if it predicts new trends in analytical data. However, if PLS is misapplied, their apparent trends could only be artifacts of the data processing technique.

USE AND ABUSE OF CHEMOMETRICS

In previous sections, a formidable array of methods have been considered, nothwithstanding that many have been omitted and the descriptions are only subjective. Nevertheless, this has been clearly demonstrated that chemometrics can aid the analyst in a large variety of different ways. As instruments become more and more sophisticated and data become easier to obtain, the analyst is confronted with an even more awesome task. Chemometrics is bound to become essential to the analytical laboratory of 1990's. However, the reader must be aware that if improperly applied, chemometric yields meaningless results.

One of the common problem to confront the chemometrician is poor data. Chemometrics can only reveal the trends that are actually burried within the data. If the experiment is not well designed, measurements are not replicate, instrument is not turned properly, chemometrics is bound to arrive at a wrong conclusion. Another problem is the misinterpretation of output. An obvious example includes the use of correlation coefficient to assess the goodness of the fit to a given model. If an incorrect number of experiments are performed, it is possible to obtain almost unit correlation coefficient from any data.

Several user-friendly softwares are available for pattern recognition in commercial instrumentation. This does not, however, mean that any particular package is an automatic choice for solving any particular problem.

Another common misapprehension is that the chemometrician prefers computation approaches to traditional "eyeballing". If equivalent information can be obtained by visual inspection of data or simple methods of analysis, the sophisticated statisticl methods are likely to be redumdant.

Finally, chemometrics is not magic. Chemometrics aims to increase the efficiency of the analytical process. If it has failed to save time, money manpower, machine time or whatever, then it has been misapplied.

FUTURE

The Automated Library : As instruments become more and more sophisticated, the generation of data becomes easier, it is visualised that chemometrics will be at the heart of the automatic library. It is not possible to interpret huge voluminous data manually, expecially where time is a constraint. Chemometric methods are surely going to take over the charge.

Mathematical Chemistry : Chemometrics is a part of mathematical chemistry but only a small part of it. Mathematics and statistics have been heavily used for solving problems in quantum mechanics, statistical mechanics, spectroscopy, kinetics the structure of matter and so on. Chemometrics is only a recent development Analytical chemists have only recently had large data sets available to them, and so comparitively little time to develop their techniques. Thus chemometrics is slowly moving along parallel line to those already developed in physical chemistry. Matrix algebra and optimisation are the key to quantum mechanics and are now similarly being used in chemometrics.

CONCLUSION

This essay had attempted to present an overview of the application different statistical methods in an analytical laboratory. Starting from sampling to design of experiments and data processing, there are techniques available to negotiate them. This essay is by no means an exhaustive one and many methods have been omitted. Only those aspects have been discussed which are of direct interest to the analyst. The future is excitingly bright and the coming century should see the growth and development of chemometrics into a mature subject.

References :

1. B.R. Kowalski, *Anat. Chem.*, **52**,112R, (1980).

2. T.E. Frank and B.R. Kowalski, *Anal. Chem* **54**, 232R, (1982).

3. M.F. Delaney, *Anal Chem.*, 56, 261R, (1984).

4. L.S. Ramos, K.R.Beebe, W.P. Carey, E.Sanchez, B.C.Erickson, B.R.Wilson, L.E. Wangen and B.R.Kowalski, *Anal. Chem.*, **58**, 294R, (1986).

5. S.D.Brown, T.Q.Barker, R.J. Larivee. S.L.Monfre and H.R.Wilk, *Anal. Chem.*, **60**, 252R, (1988).

6. S.D. Brown, *Anal.Chem.*, **62**, 84R, (1990).

7. S.D.Brown, R.S.Bear, Jr., and T.B. Blank, *Anal.Chem.*, **64**, 22R, (1992).

8. B.R. Kowalski, Editor, "Chemometrics Theory and Applications" *ACS Symposium Series No.52, American Chemical Society, Washington DC* (1977).

9. D.A.Kutz, Editor, "Chemometrics Estimators of Sampling, Amount and Error", *ACS Symposium Series No.284, American Chemical Society, Washington DC,* (1985).

10. C.G.Spiegelman, R.Walters and J.Sacks, Editors, *J.Res. Natl. Bur.Stand., Special Issue,* **90**, 391, (1985)

11. *B.R. Kowalski, Editor, "Chemometric, Maths and Statistics in Chemistry." Raidel, Dordrecht,* (1984).

12. D.Bawden, Editor, "Chemometric Series," Research Studies Press, Letchworth

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

14. M.Meloun, Jiri Militky and Michele Forina, *"Chemometrics for Analytical Chemistry, 122: PC-aided statistical data analysis". Ellis Horwood, N.Y.,* (1992).

15. D.L.Massart, Editor-in-Chief, *Chemometrics and Intelligent Laboratory Systems, Elsevier, Amsterdam.*

16. B.R. Kowalski, Editor-in-chief, *Journal of Chemimetrics, Wiley, Chichester.*

17. W.G.Cocharan and G.M. Cox, 'Experimental Design" Second Edition, Wiley, N.Y. (1957).

18. O.L.Davies *'The Design and Analysis of Industrial Experiments' Oliver and Boyd, London,* (1954).

19. G.E.P. Box, W.G.Hunter and J.S.Hunter, *'Statistics for Experimenters", Wiley, N.Y.*(1978).

20. G.Kateman and F.W.Piipers,"*Quality Control in Chemical Analysis' Wiley, N.Y.,* (1981).

21. G.S.G.Beveridge and R.S.Schechter,"*Optimised Theory and Practice," McGraw-Hill, N.Y.,*(1970).

22. S.N.Deming and S.L.Morgan, *Anal.Chem.* **45,** 279A, (1973).

23. K.W.C.Burton and G.Nickless, *Chemometrics Intell.Lab.Syst.,* **1**,135,(1987).

24. W.Kristof and B.Wingersky, "P*roceedings of the 79th Annual Convention of the APA*,1 p.81. (1971).

25. J.C.Gower, Psychometrika, **40,** 33, 1975

26. M.Martens, H.Martens and S.Wold, *J.Sci.Food Agric.,* **34**,715, (1983).

27. D.N.Lawley, *Biometrika,* **46**, 59,1959,

28. D.A.Reeve and A Crozier in "*Hormonal Regulation of Development I-Encyclopedia of Plant Physiology" Ed., J.MacMillan,* **9.,** *Springer-Verlag, Berlin,* (1980).

29. I.M.Scott, *Plant Cell Environ,* **5,** 339, 1982.

30. R.Caulcutt and R.Boddy,"*Statistics for Analytical Chemistry," Chapman and Hall, London.*(1983).

31. H.Sheffe, "*The Analysis of Variance, "Wilely, New York,* (1959).

32. I.W.L. Cole and I.E. Grizzle, *Biometrics* **22,** 810, (1966).

33. P.C. Jurs and T.L.Isenhour, *"Chemical Applications of Pattern Recognition,"Wiley, New York,* (1975).

34. K.Varmuzav, *"Pattern Recognition in Chemistry," Springer-verlag, Berlin,* (1980).

35. O. Strauf, *"Chemical Pattern Recognition" Research Studies Press, Letchworth,* (1986).

36. D.D.Wolf and M.I.L. Parsons, *"Pattern Recognition Approach to Data Interpretation," Plenum, New York,* (1983).

37. D.L.Massart and L.Kaufman, I*nterpretaation of Anaytical Data by the Use of Cluster Analysis," Wiley, New York,* (1983).

38. D.L.Massart, L.Kaufman and D.Coomans, *Anal.Chim. Acta,* **122**, 347, (1980).

39. D.Coomans and D.L.Massnt, *Anal.Chim.Acta,* **133**, 225, (1981).

40. S.Wold, *Pattern Recognition,* **8,** 12, (1976).

41. S.Wold, *Technical Report No.387,University of Wisconsin,* (1974).

42. S.Wold and M.Sjostriom in "Chemometrics Theory and Practice," Ed.B.R.Kowalski, *Am. Chem.Soc. Symp. Ser.* **52**, p.243, (1977).

43. P.W.Yendle, R.G.Brerecton and S.J.Badham, *SAS European Users Group Insternation Proceedings* (1986), *SAS, Institute, Corey, NC,* 21,(1986).

44. S.Wold, P.Geladi, K.Esbensen and J.Ohman, J.*Chemometrics,* **1,** 41, (1987).

45. O.Strouf and J.Fusek, *Collect. Czech. Chem. Commun.,* **44,** 1370, (1979).

46. J.Fusek and O.Strouf, *Collect. Czech. Chem. Commun.* **44**, 1362, (1974).

47. M.P.Derde and D.L. Massart, *Anal.Chim, Acta,* **184,** 33, (1986).

48. V.Juricskay and E.G.Veress, *Anal. Chim. Acta,* **171**, 61, (1985).

49. H.Wold in, *"Mulliariate Analysis". Academic Press, New York,* (1966)..

50. S.Wold, A.Ruhe,H.Wold and W.Dunmn SIAM *J.Stat. Comput,* **5,** 735, (1984).

51. A.Naes and H.Martens, *Commun.Stat. Simul Comput.*, **14,** 735, (1984).

52. A.Lorber, L.E.Wangen and B.R. Kowalski, *J.Chemometrics*, **1**, 19, (1987).

53. R.Manne, *Chemom. Intell. Lab. Syst.*, **2,** 187, (1987).