An overview of multivariate spectral data analysis

Paul Geladi
Institute of Chemistry, Umeå University, S 901 87 Umeå, Sweden.

Introduction

The near infrared (NIR) measurement of samples has three aspects: spectroscopy, calibration and data analysis. The first near infrared experiments created a lot of interest in improving data analysis and the possibilities of efficient data analysis created new demands for spectrometers. This interaction is very powerful. In all cases, a well-designed set of calibration samples is necessary. Regression and calibration are closely connected. There have been enormous developments in the field of regression. Selection of samples and variables and the use of non-linear models are among those new developments. New data structures and larger databases pose a new challenge and demand a more holistic approach to data analysis. This paper presents some recent trends and developments in chemometrics. The literature overview given is only a small and very subjective selection. Figure 1 gives an overview of the topics that may be considered. Because of space limitations only the major topics are treated here.

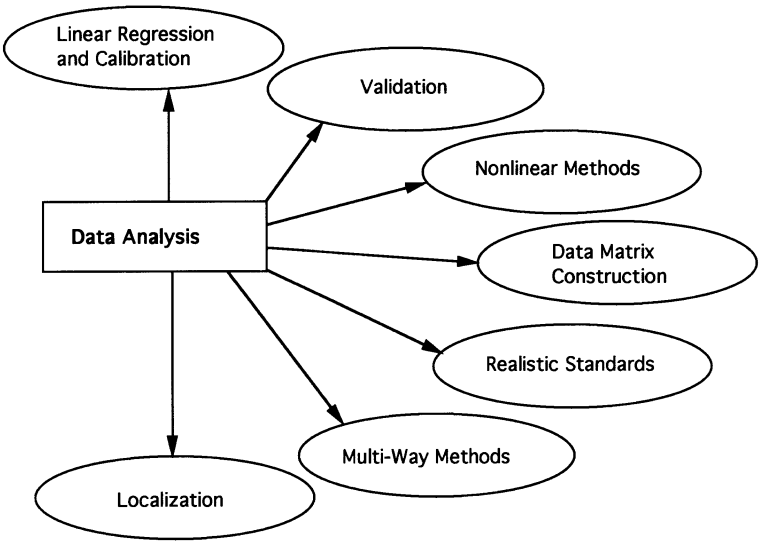

Figure 1. Important topics in data analysis. Some of these are presented in more detail in this article.

From Near Infrared Spectroscopy: The Future Waves

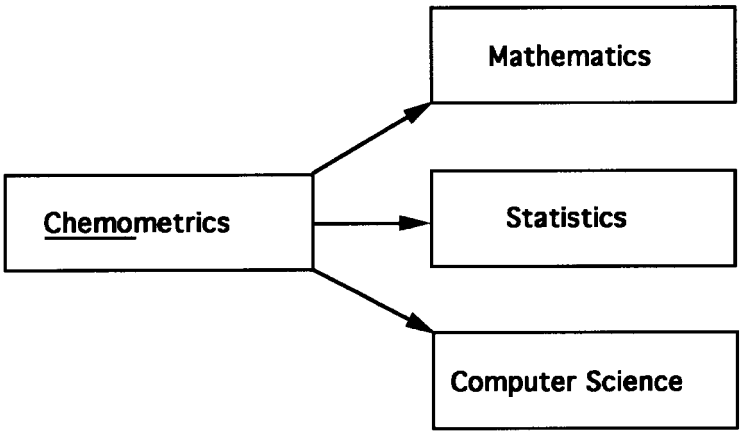

Figure 2. Chemometrics is the combination of mathematics, statistics and computer science to extract more information from chemical data, with the emphasis on a final chemical interpretation.

Chemometrics

Chemometrics is the use of mathematical, statistical and computer science methods for improving the extraction of useful information from chemical measurement data. See also Figure 2. The aspect "computer science" is gaining more and more importance. Chemometrics would not be possible without data file management. Also, many techniques have no known distributions from statistics or analytical solutions from mathematics to rely on and the use of iterative or resampling procedures prevails. The term chemo in chemometrics is very important. No matter

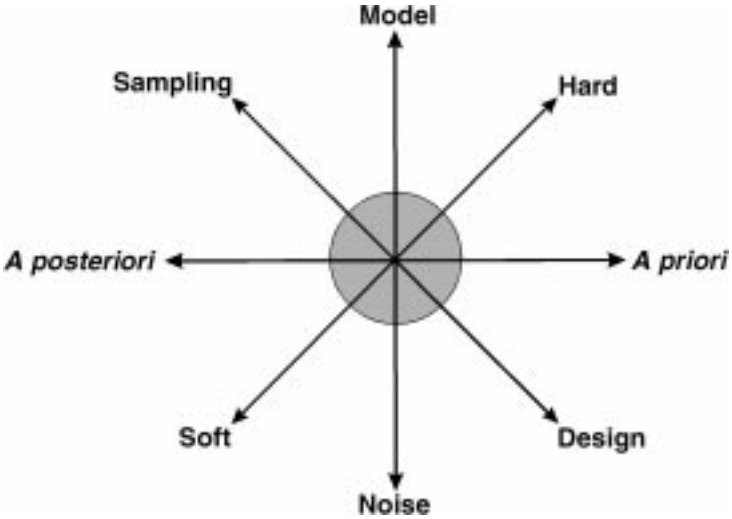

Figure 3. The rose of opposites in chemometrics. The most recent developments focus on the middle of the rose.

what an algorithm comes up with, the final interpretation of the results is the chemical one. Computer science also provides the advanced visual presentation techniques that are important for communicating results. Some of the aspects of the history of chemometrics may be found in References 1–3. In Reference 4, the rose of opposites in chemometrics was presented. It is shown as Figure 3. In the early days of chemometrics, many developments were a reaction against established statistical techniques. Chemometrics proposed soft modeling against prevailing hard models, putting emphasis on the model instead of the exact knowledge of the distribution of the noise and using *a posteriori* information instead of expecting everything to be known *a priori*. Chemometrics also tried to organize the space between random sampling and strict design. Recent work emphasizes a convergence in the middle of the rose. This may be done by including hard information in soft models, relying on partly known *a priori* information and giving equal importance to the models and the unmodeled noise. Also a better grip on using the principles of design to improve random sampling is practised.

The final goal of any chemometrical analysis should be a holistic flow chart going from problem definition over a number of steps to conclusions and a redefinition of the objectives. This is presented in Figure 4. The flowchart in Figure 4 is only one of many possible flowcharts. All the techniques of chemometrics fit in somewhere and it is no longer possible to treat one method in itself without considering the whole. A milestone in chemometrics for NIR spectrometry was the meeting "Food Research and Data Analysis" in Oslo, Norway, September 20–23, 1982. Reference 5 details the proceedings of this meeting. In a chapter (pp. 473–492) of these

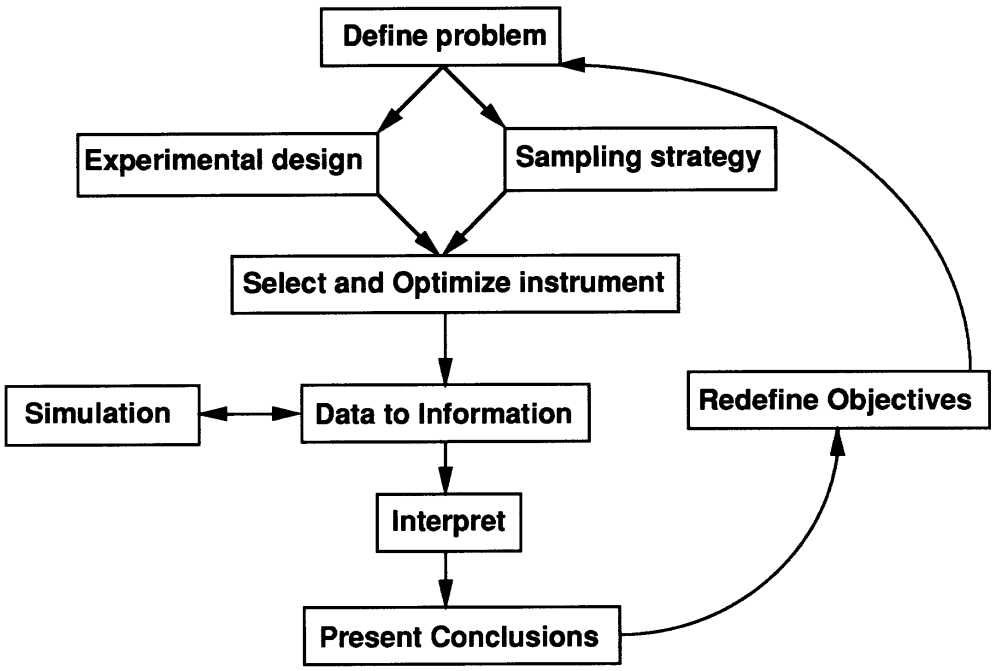

Figure 4. A possible flowchart for a chemometrical analysis. Every rectangle in the flowchart is also connected in two directions to the literature and databases but this is not shown here.

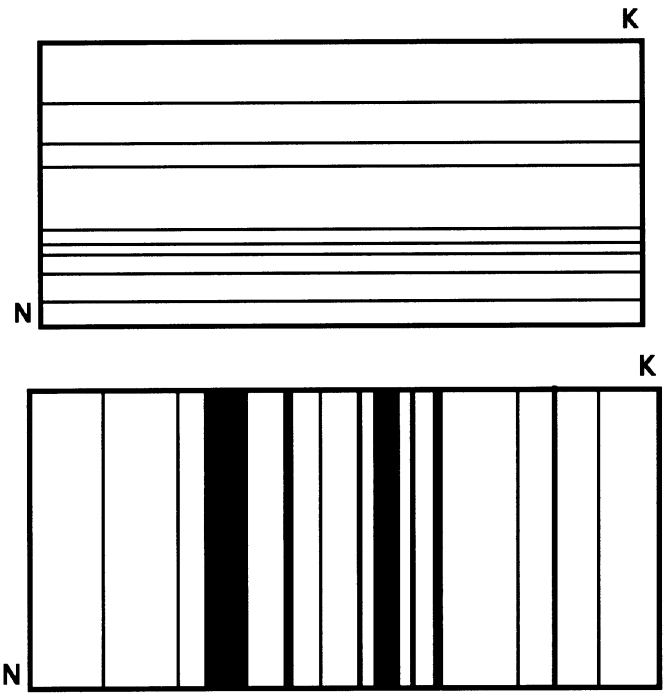

Figure 5. Above: it is possible to make subsets of a large database of samples to get a better calibration model serving a certain purpose. Below: It is possible to improve calibration models by removing certain wavelength regions. Wavelengths are usually correlated with their neighbors allowing the removal of contiguous blocks. Samples are usually organized in a non-correlated order and may be selected more randomly.

proceedings, Harald Martens, Svante Wold and Magni Martens presented "A layman's guide to multivariate data analysis". Much of what is presented in this guide is still valid and in common use, e.g. latent variable regression methods such as partial least squares (PLS) and principal component regression (PCR) but new and unforeseen directions have developed since.

Constructing data matrices by variable or sample selection

In the pioneering days of chemometrics, one had to work hard to collect objects (samples) and to increase the number of variables. Nothing that was acquired could be left out of the data analysis. Since then things have changed. Huge databases of calibration samples allow making subsets having enough samples to make a model and quick scanning allows the collection of hundreds of wavelengths in a few seconds, even with repeated scanning to reduce measurement noise. Figure 5 shows the principle of removing samples and wavelengths to obtain a better problem-oriented calibration model. Because of this, the notion that subsets of samples may be more useful for calibration emerged. A good recent example is the article by Naes and Isaksson.[6] The method is called locally weighted regression (LWR) and in Reference 6 references to some older articles are

given. The principle is that neighbours in multivariate space of the sample to be predicted are used to build the calibration model. The advantage of doing this properly is less prediction error with less components in the PCR or PLS model. A number of geometries of multivariate space have been tested, e.g. Euclidian and Mahalanobis. Also different distance-based weighting functions for the samples have been tested. The idea of locally weighted regression is a very good one. However, it requires quite a large number of calibration samples that are spread out reasonably evenly in the space of possible test samples. It is important to visualize the space of calibration samples and the spread and position of the ones selected in each calibration model.

For the removing of variables or wavelengths, the articles by Jouan-Rimbaud *et al.*[7] and by Lindgren *et al.*[8] may serve as examples. Both articles also give references to some older techniques of variable selection. An important one is the GOLPE method. A few critical remarks may be made about variable reduction. Methods like IVS-PLS and GOLPE rely heavily in cross-validation and this is not always problem-free. See the next section for more criticism of cross-validation. Another problem is that different authors use different ways of expressing prediction error. What is right: absolute errror, relative error, mean-squared error? One may also be tempted to ask what the signifcance of an improvement means? Lower prediction errors are an improvement but when does this impovement become significant and how does this relate to the number of calibration and test objects and the range of concentrations used? Also, data pretreatments such as weighting and derivation seem to have an effect. One would like to see the possibility of a complete ANOVA decomposition of the error terms as it is done in experimental design.[9]

Modifying regression models, validation

The latent variable methods PCR and PLS have been used extensively for calibration but they have also been subject to criticism and modification. An example is the article by Sun.[10] In this article it is shown for two data sets that PCR with selection of components is better than PLS or PCR with all components included. The selection is based on a correlation criterion between principal components and the concentration vector. The results are shown as the plot of RMSEP against number of components. Plots of this type are often used for showing that one method is better than another one by using less components or given a smaller prediction error. There is some danger in using these plots, especially with cross-validation results. Cross-validation does not give the absolute truth in rank (= number of components to be used). It gives a range of values for the rank where one should start looking. Again, the chemo of chemometrics comes in. One should always make sure that the chosen rank makes sense chemically, no matter what an algorithm proposes. There are many ways of doing cross-validation. One may do it one component at a time or for a whole block of contiguous or selected components. One may do leave-one-out or one may leave blocks of a certain size out. This is where the experts don't agree and there may be quite a difference in the number of components suggested. A last observation about cross-validation is that it is not distributive:

$$\mathrm{Xval}[\mathrm{A} \rightarrow \mathrm{B} \rightarrow \mathrm{C} \rightarrow \mathrm{D}] \neq \mathrm{Xval}[\mathrm{A}] \rightarrow \mathrm{Xval}[\mathrm{B}] \rightarrow \mathrm{Xval}[\mathrm{C}] \rightarrow \mathrm{Xval}[\mathrm{D}]$$

This means that cross-validating each step in a sequence of operations from A (the raw data) to D (the final results) is not the same as cross-validating the whole process. A method combining the properties of PCR and PLS was proposed by Stone and Brooks.[11] Regarding constructing linear regression vectors and comparing them, the reader is invited to look at References 12 and 13. Another improvement is the use of kernel methods for PLS. Kernel methods exist for situations with many more variables than objects or many more objects than variables. They are not a radical

From Near Infrared Spectroscopy: The Future Waves

change since the PLS models remain the same, but they allow faster calculations and a more efficient use of computer memory. They also speed up cross-validation calculations.[14–18]

Non-linear models

The classical calibration models are linear ones in which the response variable (concentration) is a weigthed sum (linear combination) of the predictor variables (absorbances). This is given as follows (mean-centering of the data allows leaving out the constant term):

$$y = Xb + f \tag{1}$$

y: the concentration for *N* calibration samples (mean-centered); *X*: the spectra (*K* wavelengths) for *N* calibration samples (mean-centered); b: the vector of *K* regression coefficients; *f*: the vector of *N* residuals. The equation is amended in the case of latent variable regression:

$$y = Td + f \tag{2}$$

T: the matrix of *A* latent variables for *N* calibration samples; *d*: a regression vector with *A* elements. Already in 1982 in Oslo (Reference 5, pp. 16–18) it was mentioned that non-linear methods would be needed for treating non-linear structures in multivariate space. Often it is found that more latent variables take care of the non-linearities in a satisfactory way. Two important ways of treating non-linearities are: non-linear transformation of the variables and non-linear transformation of the latent variables. Some possibilites are given:

$$G(y) = Xb + f \tag{3}$$

The concentrations are transformed to make the equation fit better. Exponents and logarithms may be used. This is comparable to the Box–Cox transformation in experimental design and response modeling.[9]

$$y = H(X)b + f \tag{4}$$

In biological and geological situations, the logarithmic transformation is often used because of skewed distributions. The transformation may also be something like multiplicative scatter correction.

$$G(y) = H(X)b + f \tag{5}$$

This is just a combination of what was done in Equation 3 and Equation 4. Often non-linear models may be built by adding non-linear transformations of the predictor variables:

$$y = [X\ X^2 \ldots]c + f \tag{6}$$

c is a vector of regression coefficients. It is longer than *b*, because the number of variables has increased. Other transformations of *X* (such as variable cross products, cubes, square roots etc.) may be added too. In general one may use:

$$G(y) = [H(X)\ H(X)^2 \ldots]c + f \tag{7}$$

The problem of adding many variables to a data matrix X with already many variables is that the number of variables may become difficult to handle. This makes the methods for variable selection in the previous section very useful. For latent variable methods, transformations of the latent variables and polynomials in the latent variables have been proposed to improve the handling of nonlinear data.

$$y = H(T)d + f \tag{8}$$

$$y = [H(T)\ H(T)^2 \ldots]c + f \tag{9}$$

Many non-linear methods suffer from the problems of too many parameters. It is difficult enough to select the right number of latent variables. When it also becomes necessary to select coefficients for polynomials, the amount of choices becomes confusing and models can be made to fit any situation, often without the least physical meaning. Therefore, it is considered much better to use linearization methods based on physical "hard" knowledge as in Equation 5 or Equation 7. Overviews of non-linear calibration modeling are given in References 19 and 20. Recently, some good results have been obtained with neural network calibration.[21,22]

Recalibration, calibration transfer

Recalibration and calibration transfer are very important topics. Industries and research laboratories spend a large effort in calibrating their instrument to get the best possible predictions. Very large calibration sets are sometimes built up for NIR spectrometry and there is no point to remeasuring them every day. The problem is that instruments change with time and sometimes break down completely and have to be repaired, often with spare parts that are not identical to the old ones. Getting a good calibration restored as quickly as possible is recalibration. In some industries, there are many simple spectrometers that have to be recalibrated against one more complex laboratory instrument. This is called calibration transfer. A good tutorial about the general principles of calibration transfer was written by de Noord.[23] Going to another instrument may mean a number of different situations. Baseline and sensitivity may change. Wavelength scale and wavelength resolution may change. Signal to noise ratio may change. Many of the changes may be wavelength dependent. de Noord discusses different situations from instrument matching, a hardware solution to subset recalibration a true software solution. Other recent papers on this topic are References 24 and 25. Some good work has been done and some good results were obtained.

Conclusions

All the new developments described in the preceding sections are exciting and promising in themselves and may be even more useful when combined, e.g. combining locally weighted regression with variable selection and non-linear modeling. The problem is that the situation may become very confusing even for experts, without thinking of what it would do to newcomers. A sad development seen in many publications is that the raw data are not made available to the reader, so that there is no way of confirming what the authors are claiming. To make things even worse, authors leave out details of their data analysis such as pretreatments so that even if the raw data were available, nobody would be able to repeat the calculations. A questions to be asked is: how

many authors are able to repeat their calculations and to get exactly the same results three years after the appearance in print of their article? Instrumentation will develop and improve, making the measurements more noise free and signals more specific and reducing the need for chemometrics. One may compare this with the situation where a researcher spends two years in developing software for deconvolving two overlapping peaks in a chromatogram when a new column is introduced that separates the peaks physically. On the other hand, refinement of the instrumentation always leads to the discovery of new noise and error sources, providing new challenges for chemometrics. One bottleneck in the whole process may be sampling and homogeneity of the samples. There is no point in improving instrumentation when the true source of errors in measurement is sampling, sample inhomogeneity or sample instability. Also this is a challenge, both for the chemo- and the -metrics part of chemometrics.

Acknowledgements

Fredrik Lindgren and Eigil Dåbakk are thanked for discussion and useful suggestions.

References

1. P. Geladi and K. Esbensen, *Journal of Chemometrics* **4,** 337 (1990).
2. K. Esbensen and P. Geladi, *Journal of Chemometrics* **4,** 389 (1990).
3. B. Vandeginste, *Chemometrics and Intelligent Laboratory Systems* **25,** 147 (1994).
4. P. Geladi, *Analysis Europa* **April,** 34 (1995).
5. H. Martens and H. Russwurm (Eds), *Food Research and Data Analysis*. Applied Science Publ., London (1983).
6. T. Næs and T. Isaksson, *Applied Spectroscopy* **46,** 34 (1992).
7. D. Jouan-Rimbaud, B. Walczak, D. Massart, I. Last and K. Prebble, *Analytica Chimica Acta* **304,** 285 (1995).
8. F. Lindgren, P. Geladi, S. Rännar and S. Wold, *Journal of Chemometrics* **8,** 349 (1994).
9. G. Box and N. Draper, *Empirical Model Building and Response Surfaces*. John Wiley & Sons, New York (1987).
10. J. Sun, *Journal of Chemometrics* **9,** 21 (1995).
11. M. Stone and R. Brooks, *Journal of the Royal Statistical Society B* **52,** 237 (1990).
12. P. Geladi, J. Swerts and F. Lindgren, *Chemometrics and Intelligent Laboratory Systems* **24,**145 (1994).
13. P. Geladi, in *Frontiers in Analytical Spectroscopy*, Ed by D.L. Andrews and A.M.C. Davies. The Royal Society of Chemistry, Cambridge, pp. 225–236 (1995).
14. F. Lindgren, P. Geladi and S. Wold, *Journal of Chemometrics* **7,** 45 (1993).
15. S. de Jong, *Chemometrics and Intelligent Laboratory Systems* **18,** 251 (1993).
16. F. Lindgren, P. Geladi and S. Wold, *Journal of Chemometrics* **8,** 377 (1994).
17. B. Busch and B. Nachbar, *Journal of Computer-Aided Molecular Design* **7,** 587 (1993).
18. S. Rännar, F. Lindgren, P. Geladi and S. Wold, *Journal of Chemometrics* **8,** 111 (1994).
19. S. Sekulic, M.-B. Seasholtz, Z. Wang, B. Kowalski, S. Lee and B. Holt, *Analytical Chemistry* **65,** 835A (1993).
20. J. Einax (Ed.), *Chemometrics in Environmental Chemistry* Vol 2G. Springer, Berlin (1995).
21. Z. Wang, J.-N. Hwang and B. Kowalski, *Analytical Chemistry* **67,** 1497 (1995).
22. C. Borggaard, in *Frontiers in Analytical Spectroscopy*, Ed by D. Andrews and A.M.C. Davies. The Royal Society of Chemistry, Cambridge, pp. 209-217 (1995).
23. O. de Noord, *Chemometrics and Intelligent Laboratory Systems* **25**, 85 (1994).

24. M. Forina, G. Drava, C. Armanino, R. Boggia, S. Lanteri, R. Leardi, P. Corti, P. Conti, R. Giangiacomo, C. Galliena, R. Bigoni, I. Quartari, C. Serra, D. Ferri, O. Leoni and L. Lazzeri, *Chemometrics and Intelligent Laboratory Systems* **27,** 189 (1995).
25. E. Bouveresse and D. Massart, *Analytical Chemistry* **67,** 1381 (1995).