Available online at www.sciencedirect.com
SCIENCE DIRECT®
Analytica Chimica Acta 500 (2003) 185–194
ANALYTICA CHIMICA ACTA
www.elsevier.com/locate/aca

Multivariate calibration
What is in chemometrics for the analytical chemist?

Rasmus Bro*

Department of Dairy and Food Science, The Royal Veterinary and Agricultural University, Rolighedsvej 30, DK-1958 Frederiksberg, Denmark

Received 20 March 2003; received in revised form 27 May 2003; accepted 27 May 2003

Abstract

Chemometrics has been used for some 30 years but there is still need for disseminating the potential benefits to a wider audience. In this paper, we claim that proper analytical chemistry (1) *must* in fact incorporate a chemometric approach and (2) that there are several significant advantages of doing so. In order to explain this, an indirect route will be taken, where the most important benefits of chemometric methods are discussed using small illustrative examples. Emphasis will be on multivariate data analysis (for example calibration), whereas other parts of chemometrics such as experimental design will not be treated here. Four distinct aspects are treated in detail: noise reduction; handling of interferents; the exploratory aspect and the possible outlier control. Additionally, some new developments in chemometrics are described.

Keywords: Multivariate data analysis; Exploratory; Interferents; Multi-way analysis

1. Introduction—univariate calibration

Assume that we want to build a calibration model for a specific analyte. We have a measuring device which yields the signal in Fig. 1 (left) when measuring three different pure samples of the analyte. Each profile consists of 100 distinct measurements, in this example different times. Disregarding the actual shape of these profiles, they could represent a spectrum, a sensor measurement, a chromatogram, FIA-gram, etc. Even more generally, each profile could be thought of as a set of 100 different univariate measurements (pH, concentration, flow, etc.).

In order to build a univariate calibration model, the most feasible one of these 100 variables is chosen. In a spectroscopic setting, a typical choice could be to pick a wavelength corresponding to a peak maximum of the analyte spectrum. In this example, the signal at time 50 is chosen. In order to build a calibration model through univariate linear regression, some basic assumptions must be fulfilled, two of the most important being:

- *Selectivity*: A univariate calibration model can only provide accurate results, if the measured signal does not have contributions from other sources. Hence, only the analyte of interest must contribute to the measured signal. The importance of this is significant. If other analytes contribute to the signal, the results will be biased. What is worse is that there is no way to detect from the univariate signal that incorrect results are obtained. Therefore, the adequateness of the result becomes a matter of belief (as well as experience and external validation).

* Tel.: +45-3528-3296; fax: +45-3528-3245.
E-mail address: rb@kvl.dk (R. Bro).

0003-2670/$ – see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/S0003-2670(03)00681-0

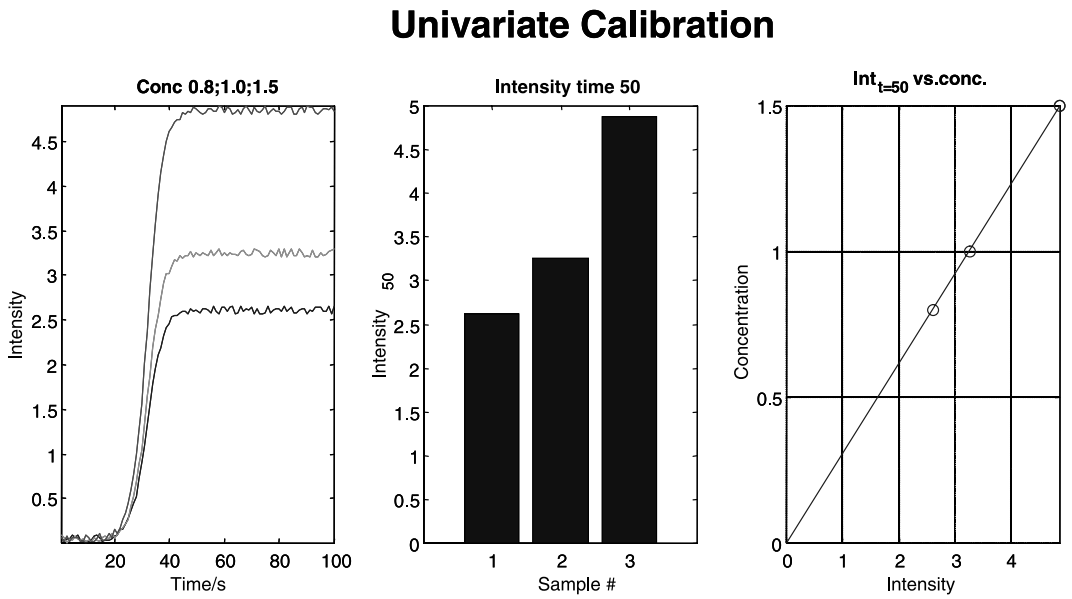

Fig. 1. Signal from one sensor/instrument over time. (Left) signals from three samples with arbitrary concentrations 0.8; 1.0 and 1.5. (Middle) signal from the three samples at time 50. (Right) correlation between signal at time 50 and concentration.

- *Linearity*: There must be a linear relation between the analyte concentration and the signal. If this is not the case, though, non-linear remedies are readily available.

From Fig. 1, it is seen that in this case the signal at time 50 provides excellent means for building a linear regression model predicting the analyte concentration.

2. Multivariate advantage 1—noise reduction

Instead of using just one out the 100 variables it makes sense to use all the measured information. This can lead to a number of distinct advantages, some of which will be elaborated on in the following. The most straightforward advantage is a noise reduction obtained by using more (redundant) measurements of the same phenomenon. This can be illustrated using *principal component analysis* on data similar to the signal in Fig. 1. The details of principal component analysis are omitted here but can be found elsewhere [18,29].

In this example with three samples and 100 variables principal component analysis will result in one significant *principal component* (also more generally called a latent variable) that describes the main variation in the data. In Fig. 2, the meaning of this latent variable is illustrated graphically. A substantial amount of noise has been added to highlight the noise reduction properties. In Fig. 2 (upper left), the original measurements are shown. As can be appreciated from the figure, all sample profiles are of the same shape (up to the noise) and differ only in the magnitude. This is exactly what the *one-component* principal component model states.

The model is given by three different parts: the loading vector (Fig. 2, lower left) which is the common shape that describes the shape of all measured profiles and (Fig. 2, upper right) the scores which is the sample-specific information; namely the amount of this loading vector in each sample. The loadings and scores are found from the measured profiles alone and in a least squares sense. The last part of the principal component model is then the residuals; the part of the profiles that deviates from the common shape. Ideally, the residuals are measurement noise. The model is called a one-component model because there is one loading vector with associated score vector. In case there is more than one type of phenomena in the measured profiles, more components can be determined

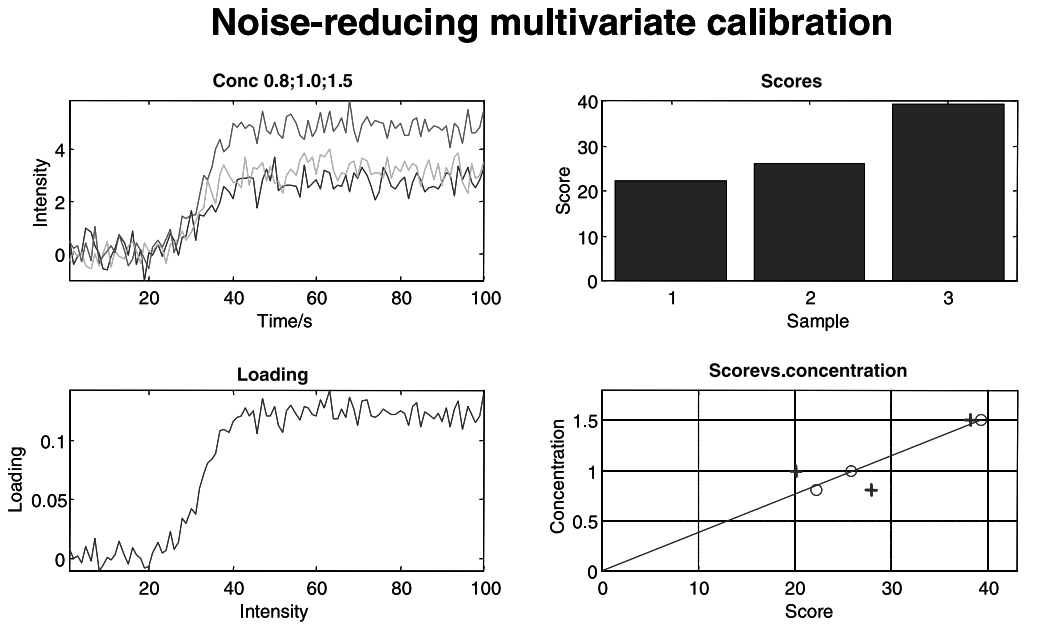

Fig. 2. Principal component analysis of a dataset with three samples and 100 variables. The three profiles (upper left) can be described by one principal component instead of the 100 original variables. The quantitative difference between the three profiles is described by the different amounts (scores—upper right) of the loading vector (lower left).

in an equivalent way. The principal component (the loadings and scores) is defined as a weighted average of *all* the original variables. The weights are given in the *loading*-vector (Fig. 2) which is the qualitative information about the samples. It describes what type of information characterizes the samples. The associated weighted averages are the scores and these then provide the quantitative information.

Another important interpretation of this model stems from realizing that instead of having 100 original variables which are difficult to assess simultaneously, these 100 variables have been replaced with *one* new variable. This new variable is the principal component and it can be treated in much the same way as any other variable. The definition of an original variable, e.g. intensity at time 50, is given by its definition (measure at time 50) and the corresponding values (the readouts of the instrument for the three samples). In the same fashion the principal component is given by its definition (the weighted average of 100 variables—the loading) and the corresponding readouts (the scores). The important difference between the original variables and the principal component is that the principal component carries information from *all* original variables simultaneously.

In the lower right part of Fig. 2, the score values (0) are plotted against the concentrations, showing a good correlation. The signal measured at time 50 (+), on the other hand, is shown to correlate quite poorly. This is due to the noise and in fact none of the original variables correlate well with the concentration. It is only by using a (weighted) average of all the measurements that a minimization of the influence of the noise is obtained yielding a robustness towards random artifacts. This is one of the clear advantages of multivariate analysis.

3. Implication of interferents

An important advantage in using multivariate data and models is that non-selective signals can be made selective by use of mathematics. Thus, interferents can be handled provided that the signal (shape) of the interferents is not completely identical to the signal from the analyte.

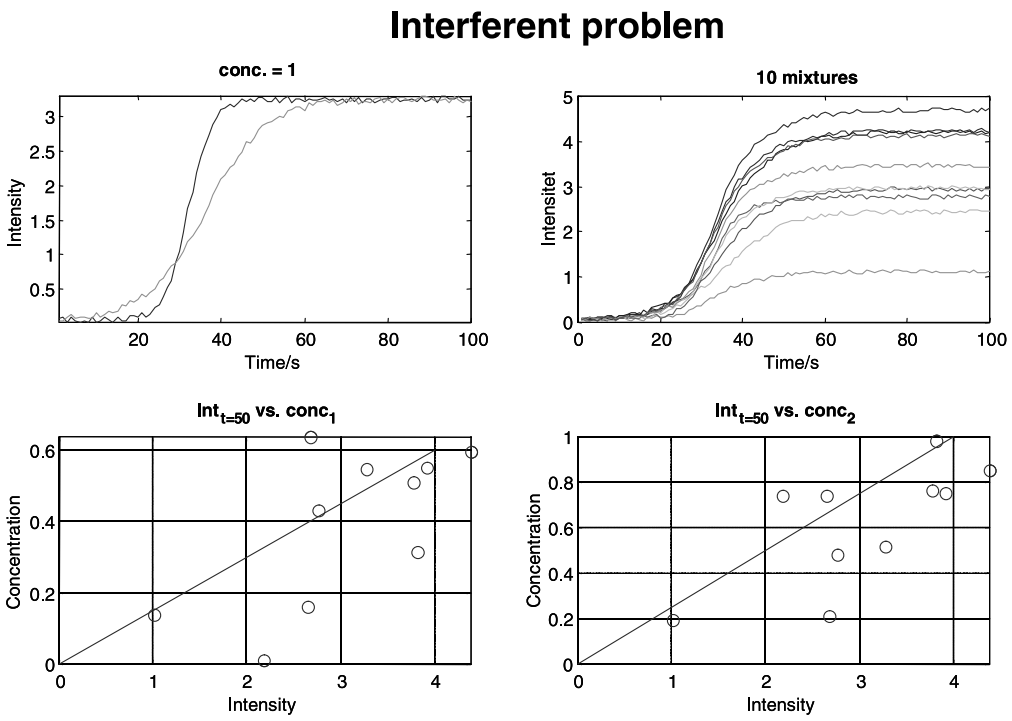

Fig. 3. Illustration of interferents. Top left, the signal is shown from the analyte of interest and from the interferent. Top right, the profiles from ten mixtures of these two substances are shown. The lower figures show scatter plots of the intensity at time 50 against the concentration of the two substances. No prediction is possible due to the lack of selective channels.

In Fig. 3, a situation is shown where the signals have contributions from both the analyte of interest as well as from an (unknown) interferent. To make it worse, the shape of the interferent profile is very similar to that of the analyte (Fig. 3, top left). There could even be more than one type of interferent. But regardless, the occurrence of interferents makes univariate calibration models impossible due to the lack of analyte-selective information. This problem does not arise from having multivariate data. On the contrary, the use of multivariate data makes it possible to handle a situation with non-selective information. The fact that it is possible to incorporate interferents in the calibration model implies

1. *Instrumental selectivity not needed*. It is possible to build calibration models even when chemical or physical selectivity is impossible to obtain.
2. *Sensors can be optimized*. In developing/choosing sensors it is not necessary to focus primarily on chemical selectivity. Other aspects such as signal-to-noise ratio or robustness could be equally if not more important when chemical selectivity can be supplemented by mathematical selectivity.
3. *Outliers can be detected*. Calibration models can be made semi-intelligent and e.g. check whether the assumptions behind the model are fulfilled. This can be done because the *shape* of the measured profile reflects the underlying profiles. If the sample is of a different constitution, the user is automatically warned that either the sample is wrong or the sensors malfunctioning. Information is even provided to the cause of the problem.

This outlier control makes it possible to see when for example a tap-water sample is being predicted by a model that is only valid for waste water. This outlier control on model-level is *not* possible when using univariate calibration models [19,22,26].

4. Multivariate advantage 2—handling interferents

Assume that the signals from the two substances in the prior example are additive. This means that the contribution from one does not affect the contribution from the other. If this holds approximately then a multivariate regression model can be made [18]. To build the model, a set of mixture measurements are needed where the analyte concentration is known. No information is needed about any other signals in the measurements (the interferent); as long as all substances vary independently in concentration. Typically, 10–50 samples are needed depending on the complexity of the signals.

A two-component principal component model of the profiles in Fig. 4 can describe the original 100 variables. As there are two phenomena varying in the samples, it makes good sense that the 100 variables can be condensed into two new variables. The mixture signals are not, however, condensed into two principal components such that the first reflects one analyte and the second the other. For mathematical reasons this is not possible from the measurements alone.

The first component is given by the smooth curve in Fig. 4 (top left). This loading describes the general trend in the data and therefore the scores of this component reflect the general level of the measured signal. This score is the upper one in Fig. 4 (upper right). Sample two, for example, has a low score on this component, which reflects the low measurement signal for this sample. This sample corresponds to the lower-most profile in Fig. 3 (upper right). The second loading vector is noisier and less straightforward interpretable. It describes the deviation from the common profile (loading one). The principal component model of any particular sample is given as loading vector one times the sample score one plus loading vector two times the sample score two (plus more in case more components are needed).

From the loading vectors in Fig. 4 and the pure profiles in Fig. 3, it is seen that by modifying the amount

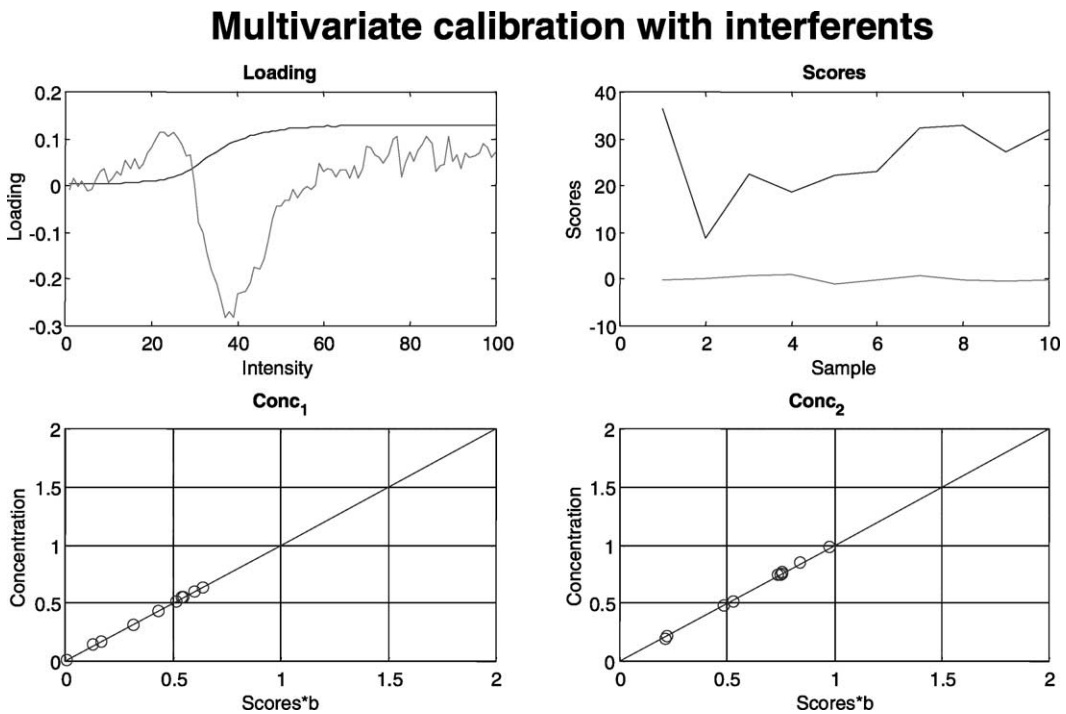

Fig. 4. Handling multivariate data. Principal component analysis of data from Fig. 3 leads to a two-component model. Loadings are shown upper left and scores upper right. The lower figures illustrate that with these new variables it is possible to determine the concentrations of both components simultaneously because their concentrations in the calibration samples are known.

of the second loading vector the first loading vector is 'corrected' towards either the profile of the analyte of interest or the interferent. Thereby it is possible to describe any mixture of the two using only these two principal components. This forms the basis for *multivariable calibration*. The scores are well conditioned (numerically) and orthogonal (statistically linear independent) and for these reasons the scores are well suited for use as independent variables in a multiple linear regression model. The original 100 variables cannot be used for the same regression purpose due to the close similarity of the different variables (Fig. 4, bottom).

A calibration model based on the scores is called a *principal component regression* model. To predict the concentration of the analyte of interest in a new sample, the mixture profile of the sample is measured. From the already known loading vectors found during the calibration stage, the two score values of this new sample can be calculated and these scores are inserted in the regression model. This yields the prediction of the analyte concentration.

There are a number of alternative techniques for finding multivariate regression models. Of the more common are: partial least squares regression [18], principal component regression, feed-forward neural networks [6], regression using radial basis functions [27], ridge regression [25] and principal covariates regression [10]. Though there are differences between these methods and though they are often presented using very different terminology, their main property is that they handle multivariate non-selective measurements and enable the utilization of all measured information rather than having to resort to pre-selecting a few discrete measurement-channels.

5. Multivariate advantage 3—the exploratory aspect

It is interesting and important that the above example of a calibration model does not only provide numerical predictions of the sought property. Besides the actual prediction, a number of informative parameters plus the residuals are obtained. These can be used in an exploratory fashion to investigate the validity of the model, to improve the model, to understand why a model does not work, to see where a sample differs from others samples, etc.

The loadings, for example, can show if some measured variables do not behave as expected. A spike in the loadings in Fig. 4 would indicate a mal-functioning sensor; a low loading would indicate an irrelevant measurement, etc. The scores can provide expected as well as unexpected information about the samples. An extreme score value indicates an extreme sample, possibly an outlier. As the scores are variable measurements just like ordinary ones (only defined as weighted averages), they can be plotted as ordinary variables. They can be plotted against time, versus residuals or against each other in scatter plots. In scatter plots, groupings of samples are easily seen and provide a simple visual aid to assess if e.g. replicates are correctly placed close to each other.

All in all, the possible visualization of principal components, hence of all the original variables, provides a significant way of *understanding* the model. This is one of the most pronounced reasons for the widespread use of multivariate chemometric methods. If all data were perfect, there would be many ways to build reasonable models, but in science and research, the most important aspect is often to understand and learn from experiments. Through understanding, means are provided for improving status quo. Even in actual run-time usage of a model, errors will occur. Detecting this and further understanding why these errors occur is important. Examples on the exploratory aspect of data analysis are numerous [9,13,21,28].

6. Multivariate advantage 4—outlier control

As mentioned above, outliers are important to detect in all data analysis. Errors are the rule rather than the exception due to for instance trivial errors, instrument errors and sampling errors. If these errors are sufficiently large either in quantity or quality, they can destroy any meaningful result or interpretation. It may seem difficult to detect outliers when complicated multivariate data are used, but in fact, the detection of outliers is greatly enhanced from having multivariate data. One example of this is provided below.

Assume that a multivariate calibration model has been made as discussed in Section 4. When this model is to be used for predicting the concentration in new samples it is necessary to ensure that these samples

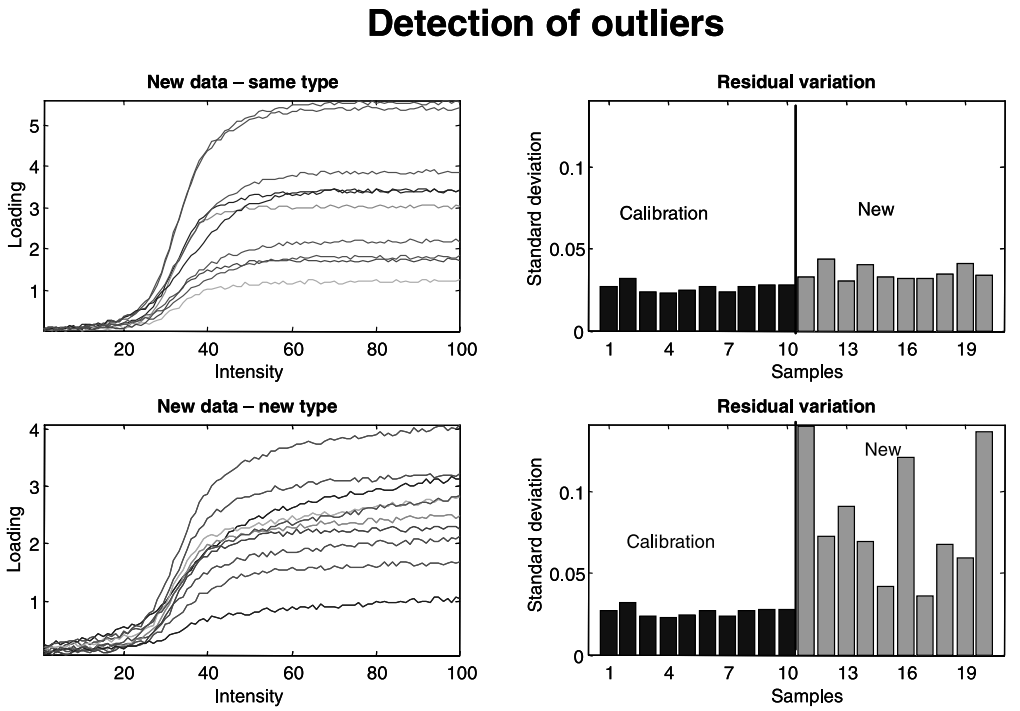

Fig. 5. Outlier control. A calibration model based on the data in Fig. 3 with one analyte of interest and one interferent. The model is used on data of the same type (top) and on data with a slightly different new interferent (bottom). The residual variation is much higher for inadequate data, providing a simple flag for the validity of the prediction.

are of the same constitution as the samples for which the calibration model was built. That is, the new profiles of the new samples should be linear combinations of the same phenomena as the samples used in calibration. In Fig. 5, an example is shown of a new set of samples that are consistent with the calibration set (top left) as well as a dataset 'contaminated' with an additional substance in the samples not calibrated for (bottom left).

The residuals of any sample are the part of the profile that cannot be described by the loadings. Ideally these residuals reflect measurement noise and are hence of the same (small) magnitude for any valid sample. Indeed, this is so for the valid samples. In Fig. 5 (top right), the standard deviation of the residuals are shown for each sample separately. The left part shows the calibration samples and the right part the residuals of the new samples. The residuals are of the same magnitude, indicating they are of the same basic constitution.

Looking instead at the samples where new substances were introduced, the residuals are huge compared to the calibration samples. This directly implies that these samples contain un-modeled variation and cannot be predicted with the current model. Predicting would correspond to predicting with a univariate regression model in the presence of interferents—only in the multivariate case, it is possible to detect that un-calibrated interferents are present! Note that the outliers cannot be visually detected univariately e.g. from the signal at time 50. This and similar kinds of outlier detection are used routinely in chemometrics and also e.g. in process monitoring to detect processes that are out of statistical control.

7. Multi-way data—new options in chemometrics

In recent years new techniques called multi-way methods have evolved in chemometrics. The methods

are interesting in e.g. analytical chemistry because they enable so-called mathematical chromatography. That means that both quantitative and qualitative analysis can be performed on measurements on mixtures. In analogy to ordinary chromatography, the multi-way methods make it possible to separate measured signals directly into the underlying contributions from individual substances.

Multivariate data are normally held in a data table with every row holding the measurements of one sample and every column representing a different measured variable. Such a table can be arranged as a matrix which is the form used in multivariate data analysis. However, there are situations where such a two-way structure is not the most natural way to arrange data.

If a sample is measured using several sensors over time, then for each sensor a profile is obtained. For one sample alone, the data can be held in a matrix. The same holds in fluorescence spectroscopy if the emission spectra are measured at several excitation wavelengths (see Fig. 6). Many other types of measurements lead to multi-way data (hyphenated methods, process data, sensory analysis, etc.). When the data from one sample yields a matrix, then data from several samples can be held in a box—a three-way array.

To analyze such three-way data, dedicated three-way models are available. Two accepted methods are the PARAFAC model (parallel factor analysis) [8,14] and Tucker model [1,3,15] which are two different extensions of principal component analysis to three-way data. Multi-way partial least squares regression [7] is the extension of the so-called partial least squares regression method.

Mathematical chromatography can be illustrated by a simple dataset of five samples containing different amounts of tryptophan, tyrosin and phenylalanine which are measured using fluorescence spectroscopy (http://www.models.kvl.dk, May 2003). The landscape in Fig. 6 represents one of these samples. Fitting a three-component PARAFAC structure to the data, a model is obtained that provides estimates of the three amino acid emission spectra, the three excitation spectra and the relative concentrations. These estimates are shown in Fig. 7. The magnitudes of the estimates are relative as is also the case for e.g. conventional chromatographic measurements. In order to perform a quantitative analysis the absolute concentration of the analyte must be known for at least of one of the samples.

From the estimated pure spectra it is seen that the leftmost (emission maximum 270 nm) component is

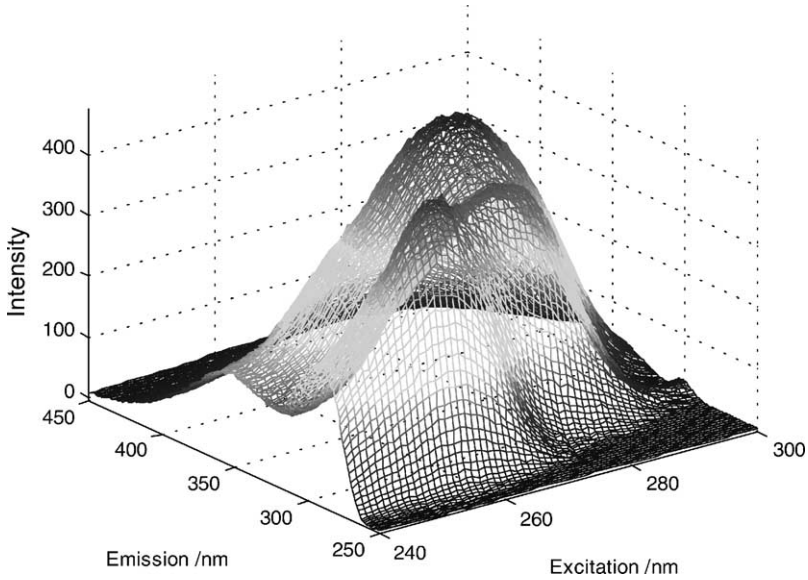

Fig. 6. Fluorescence measurements can lead to three-way data. For every sample an emission spectrum is measured at several excitation wavelengths and these data provide a fluorescence excitation-emission landscape. Several such landscapes are held in a three-way array.

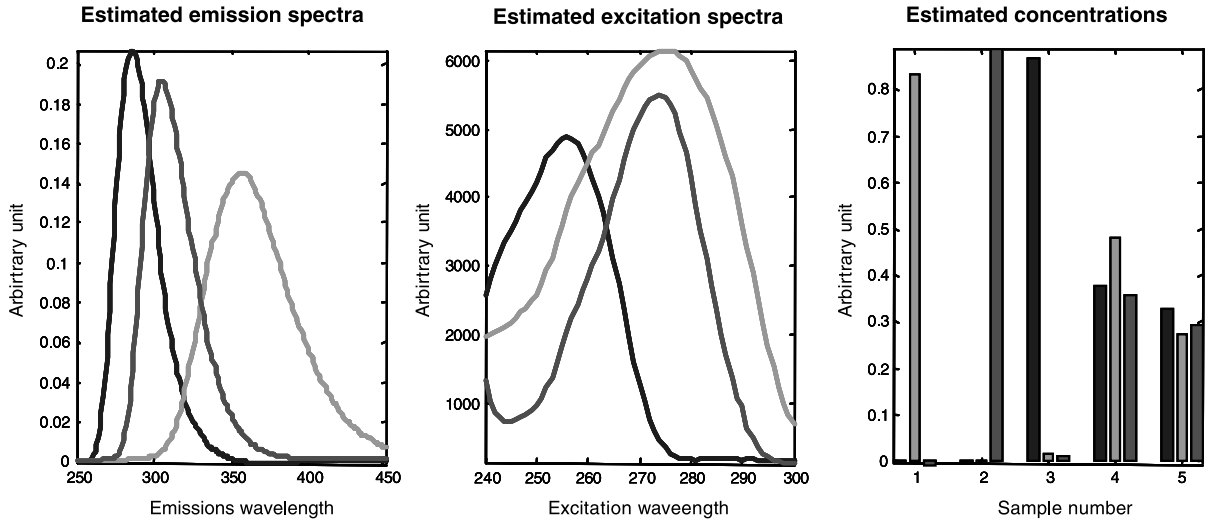

Fig. 7. Result from a PARAFAC model. The model of the five samples directly separates the mixtures into the contributions from three the individual amino acids.

phenylalanine, the middle band is tyrosin and the one to the right is tryptophan. Therefore, e.g. the five left-most bars in Fig. 7 (right) are the relative concentrations of tryptophan in the five samples. It is remarkable that the information in Fig. 7 can be found from the measured mixtures alone. If the concentration of tryptophan is known in e.g. sample one, the relative concentrations can be scaled to absolute form as shown in Table 1.

These results show that both quantitative and qualitative analysis is possible directly from mixture measurements. Using multi-way methods the best of univariate and multivariate calibration is joined: interferents can be handled as in multivariate calibration. But as in univariate calibration only few samples are needed. In the most extreme situation—as above—only one calibration sample is needed. Because the interferents are modeled specifically, there is no need to know the reference concentrations in the mixtures as long as a pure sample measurement is representative for the analyte contribution to the mixture samples.

Table 1
Quantitative results from the PARAFAC model: estimated concentrations of tryptophan in four samples based only on the measured fluorescence and the concentration of tryptophan in sample one

Sample	Estimated concentration (μM)	Reference concentration (μM)
2	0.01	0.00
3	0.05	0.00
4	1.55	1.58
5	0.88	0.88

The literature contains many examples on the practical use of PARAFAC for solving complex problems in analytical chemistry [2,4,5,9,11,12,16,17,20,23,24].

8. Conclusion

A short survey of some of the properties of multivariate data analysis tools has been given. Some relevant conclusions are

- sensor-selectivity is fine, but it is not optimal to develop or pick sensors based primarily on selectivity. The mathematical selectivity obtained from suitable modeling methods should be incorporated when assessing the selectivity. Thereby, more focus can be directed towards signal-to-noise ratio, physical robustness, cost, etc.;
- multivariate models are more adequate than univariate models. This strong statement holds in general because it is always possible to discard variables

such that a univariate approach is re-obtained. Thus, a multivariate approach adds opportunities but does not remove any;

- multivariate models can handle situations that cannot be handled univariately. In particular, it is possible to incorporate interferents and to have automatic outlier detection when building or using a model;
- multivariate models and data make it possible to supplement the traditional deductive approach with an exploratory one. Rather than using experiments to simply verify hypotheses, then new ideas, knowledge and hypotheses may come from measured data directly by properly visualizing measurements descriptive for a particular problem. As shown repeatedly in the literature, such an approach can be both time- and money-saving;
- in the future it is expected that demands will be made for effective use of information, fast analysis results, low consumptions of chemicals, and more robust methods that should even work on unforeseen sample matrices. Multi-way methods point to an interesting development in this respect, but wide-spread use will need a close collaboration with instrument makers to ensure that more instruments are 'multi-way enabled'.

References

[1] C.A. Andersson, R. Bro, Chemom. Intell. Lab. Syst. 52 (2000) 1.
[2] C.A. Andersson, Applied chemometrics, in: Høskuldsson, Agnar, Nørgaard, Lars (Eds.), Proceedings of the Conference on Applied Statistics and Chemometrics, vol. 95, Thor Publishing, 1997.
[3] C.A. Andersson, R. Bro, J. Chemom. 14 (2000) 103.
[4] J.L. Beltran, J. Guiteras, R. Ferrer, J. Chromatogr. A 802 (1998) 263.
[5] S. Bijlsma, D.J. Louwerse, W. Windig, A.K. Smilde, Anal. Chim. Acta 376 (1998) 339.
[6] T.B. Blank, S.D. Brown, J. Chemom. 8 (1994) 391.
[7] R. Bro, J. Chemom. 10 (1996) 47.
[8] Bro R, Multi-way Analysis in the Food Industry. Models, Algorithms, and Applications, Ph.D. thesis, University of Amsterdam (NL), http://www.mli.kvl.dk/staff/foodtech/brothesis.pdf, 1998.
[9] R. Bro, Chemom. Intell. Lab. Syst. 46 (1999) 133.
[10] S. de Jong, H.A.L. Kiers, Chemom. Intell. Lab. Syst. 14 (1992) 155.
[11] T. Do, N.S. McIntyre, Surface Sci. 435 (1999) 136.
[12] W.P. Gardner, R.E. Shaffer, J.E. Girard, J.H. Callahan, Anal. Chem. 73 (2001) 596.
[13] P. Geladi, Chemom. Intell. Lab. Syst. 14 (1992) 375.
[14] R.A. Harshman, M.E. Lundy, Comput. Statistics Data Anal. 18 (1994) 39.
[15] P.M. Kroonenberg, Comput. Statistics Data Anal. 18 (1994) 73.
[16] S.E. Leurgans, R.T. Ross, Statistical Sci. 7 (1992) 289.
[17] A. Marcos, M. Foulkes, S.J. Hill, J. Anal. Atom. Spectromet. 16 (2001) 105.
[18] H. Martens, T. Næs, Multivariate Calibration, Wiley, Chichester, 1989.
[19] D.L. Massart, L. Kaufman, P.J. Rousseeuw, A. Leroy, Anal. Chim. Acta 187 (1986) 171.
[20] L. Moberg, G. Robertsson, B. Karlberg, Talanta 54 (2001) 161.
[21] L. Munck, L. Nørgaard, S.B. Engelsen, R. Bro, C.A. Andersson, Chemom. Intell. Lab. Syst. 44 (1998) 31.
[22] T. Næs, T. Isaksson, NIR News 3 (1992) 8.
[23] R.P.H. Nikolajsen, K.S. Booksh, A.M. Hansen, R. Bro, Anal. Chim. Acta 475 (2003) 137.
[24] R.T. Ross, C. Lee, C.M. Davis, B.M. Ezzeddine, E.A. Fayyad, S.E. Leurgans, Biochimica et Biophysica Acta 1056 (1991) 317.
[25] R. Sundberg, Scand. J. Statistics 26 (1999) 161.
[26] B. Walczak, Chemom. Intell. Lab. Syst. 28 (1995) 259.
[27] B. Walczak, D.L. Massart, Chemom. Intell. Lab. Syst. 50 (2000) 179.
[28] C. Weihs, J. Chemom. 7 (1993) 305.
[29] S. Wold, K.H. Esbensen, P. Geladi, Chemom. Intell. Lab. Syst. 2 (1987) 37.