JOURNAL OF CHEMOMETRICS
J. Chemometrics 2000; **14**: 187–195

Three-mode principal component analysis of monitoring data from Venice lagoon

Riccardo Leardi[1]*†, Carla Armanino[1], Silvia Lanteri[1] and Luigi Alberotanza[2]

[1]*Dipartimento di Chimica e Tecnologie Farmaceutiche e Alimentari, Via Brigata Salerno (Ponte), I-16147 Genova, Italy*
[2]*CNR, Istituto per lo Studio della Dinamica delle Grandi Masse, San Polo 1364, I-30125 Venezia, Italy*

SUMMARY

A data set obtained by 44 monthly determinations of 11 variables from 13 sampling sites in the Venice lagoon has been treated by three-mode principal component analysis. The results show that the sampling sites are grouped according to their geographical location, following an inner–outer lagoon direction. In terms of sampling periods, a very strong seasonal effect has been detected, together with an almost linear decrease in nutrients (P and NO_3^-) and increase in eutrophication. Copyright © 2000 John Wiley & Sons, Ltd.

KEY WORDS: three-mode PCA; display methods; environment

1. INTRODUCTION

When monitoring a wide range of dissolved aquatic chemical parameters, a huge number of quantitative analytical data are obtained. Usually the parameters are measured at regular intervals at a certain number of sampling sites. As a result, a three-way data array is obtained in which the three modes are variables, geographical location and time.

The aim of this paper is to study the environmental information contained in a wide data set produced by 4 years of monitoring of the waters of the Venice lagoon. More than 20 variables were measured at 13 sampling stations located inside the lagoon, once a month for the 44 months from May 1987 to December 1990.

This study is part of a much wider project on the ecosystem of the Venice lagoon; the 4 year period is long enough to obtain some preliminary information useful for the global study.

A Tucker3 model [1–8] has been used, leading to the easy identification of the effects present in each of the three modes.

* Correspondence to: R. Leardi, Dipartimento di Chimica e Tecnologie Farmaceutiche e Alimentari, Via Brigata Salerno (Ponte), I-16147 Genova, Italy.
† E-mail: riclea@dictfa.unige.it
Contract/grant sponsor: MURST

2. EXPERIMENTAL

2.1. Experimental area

The region under study includes both the city of Venice and the important industrial area of Porto Marghera, where some channels discharge their water into the lagoon. This region is crossed by many ships of various tonnages and is a hydraulically very active system.

The 13 sampling sites are representative of the different situations of the central lagoon basin (Figure 1). Sites *f, g, h* and *i* are located near the industrial harbor area of Porto Marghera; sites *d* and *e* are between the city and the industrial area; sites *b* and *c* are near the city; sites *n* and *o* are in the center of the lagoon; site *l*, near the land, is in the 'petrol channel', the route of ships; and sites *a* and *m* are the two inlets of the lagoon.

The location of the sampling sites makes it possible to detect urban wastes from Venice, numerous wastes from the industrial area, and pollutants discharged by the freshwater channels flowing into the lagoon.

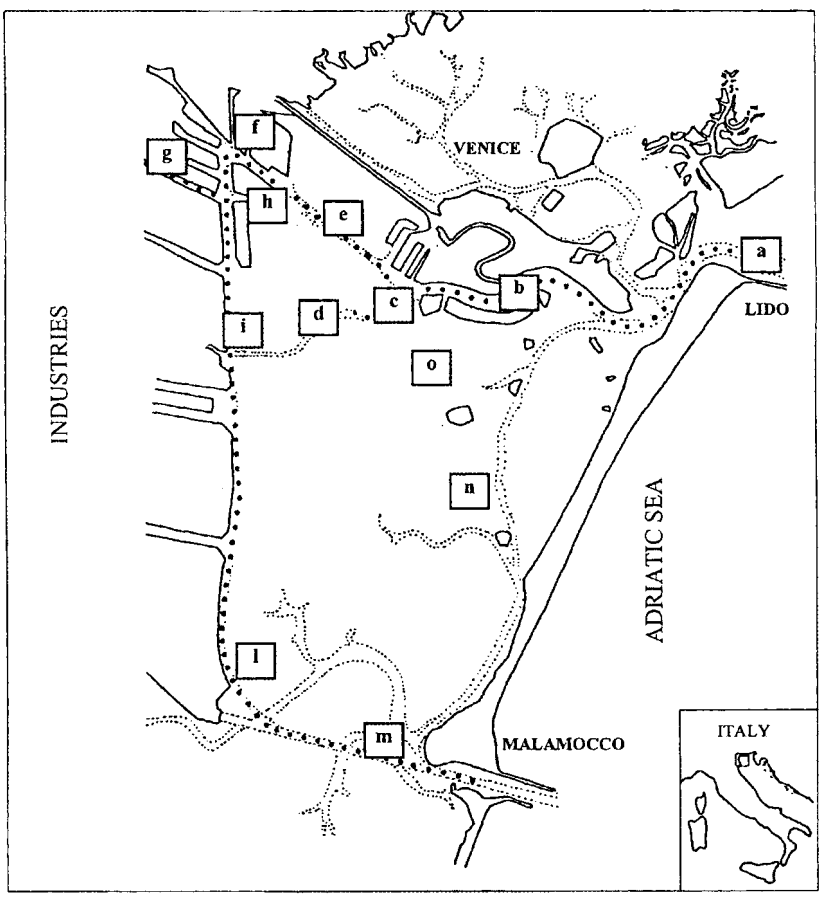

Figure 1. Sampling sites in the Venice lagoon: *a*, Lido inlet; *b*, San Marco; *c*, Giudecca channel; *d*, San Giorgio in Alga; *e*, toward the industrial area; *f, g, h, i*, around the industrial area; *l*, near the land in the 'petrol channel'; *m*, Malamocco inlet; *n*, between the Lido and San Clemente islands; *o*, between San Clemente and Giudecca islands.

2.2. *The measured parameters*

Physical and chemical parameters were measured to study the Venice lagoon environment. The data set was previously published and the analytical methods and their accuracies described by Alberotanza and Zucchetta [9–12].

Chlorophyll-a was spectrophotometrically measured after extraction with acetone; total suspended matter was weighed after drying in a 50 °C oven for 24 h; water transparency was measured by Secchi disk extinction through three colored filters (blue, green and red); fluorescence was measured with a high-flow cell fluorometer, determining (*in vivo*) the chlorophyll present in the chloroplasts of phytoplankton; turbidity has been obtained by a scatterometric measure; suspended solids were measured by a gravimetric determination of filtrate with a membrane filter; NH_4^+ was measured by a specific electrode; NO_3^- was measured spectrophotometrically after reduction with diazotization; P was measured by a gravimetric method; COD (chemical oxygen demand) was measured by a potentiometric titration; BOD_5 (biological oxygen demand) was measured by the determination of dissolved oxygen before and after a 5 day incubation.

Some other parameters had been measured, but they were not included in this study since they were always lower than the detection limit (mercury, cyanide, sulphides, phenol, aromatic hydrocarbons, chlorinated hydrocarbons, surfactants, hydrocarbons and oils); nitrites were always very near to the detection limit, and the noise was very high; pH, chloride and sulphate were not used since they had not been measured in the first 2 years of monitoring.

2.3. *Sampling period*

Sampling was performed once a month in the period May 1987–December 1990, giving a total of 44 samplings. The sampling at all 13 sampling sites was performed on the same day, corresponding to the moon quadrature, when the tidal excursion and the water exchange are the lowest.

At the beginning of 1986 some secondary and tertiary sewage treatment plants, treating both industrial and domestic sewages, became partially operational; they reached full power at the end of 1988.

3. THREE-MODE PRINCIPAL COMPONENT ANALYSIS

It can happen that the structure of a data set is such that a standard two-way table (objects versus variables) is not enough to describe it. For instance, in our case the same analyses have been performed at different sampling sites on different days; therefore a third mode needs to be added to represent the data set, which can be imagined as a parallelepiped of size $I \times J \times K$, where I is the number of sampling sites (objects), J is the number of variables and K is the number of sampling times (conditions).

To apply standard PCA [13,14], these three-way data arrays $\underline{\mathbf{X}}$ have to be matricized to obtain a two-way data table. This can be done in different ways, according to what one is interested in focusing on.

If we are interested in studying each 'sampling', a matrix $\mathbf{X}_b'$ is obtained having $I \times K$ rows and J columns. This approach is very straightforward in terms of computation, but since $I \times K$ is usually a rather large number (572 in our case), the interpretation of the resulting score plot can give some problems.

To focus on the sampling sites, the data array $\underline{\mathbf{X}}$ can be matricized to $\mathbf{X}_a'$ (I rows, $J \times K$ columns). The interpretability of the score plot is usually very high, but since $J \times K$ is usually a rather large number (484 with our data set), the interpretation of the loading plot is very difficult.

The same considerations can be made when focusing on the sampling times: in this case, $\mathbf{X}_c'$ is obtained (K rows, $I \times J$ columns).

Three-mode PCA allows a much easier interpretation of the information contained in the data set, since it directly takes into account its three-way structure. The final result is given by three sets of loadings together with a core array describing the relationship among them. If the number of components is the same for each way, the core array is a cube. Each of the three sets of loadings can be displayed and interpreted in the same way as a score plot of standard PCA.

In the case of a cubic core array a series of orthogonal rotations can be performed on the three spaces of the objects, variables and conditions, looking for the common orientation for which the core array is as much as possible body-diagonal [7].

If this condition is sufficiently achieved, then the rotated sets of loadings can also be interpreted jointly by overlapping them.

Trilinearity of the data set is also assumed, meaning that the effect of sampling location is the same at any time and that the effect of time is the same at any location.

Mathematically, it can be said that

$$x_{ijk} = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} a_{ip} b_{jq} c_{kr} g_{pqr} + e_{ijk}$$

where a_{ip}, b_{jq} and c_{kr} denote elements of the component matrices **A**, **B** and **C** of orders $I \times P$, $J \times Q$ and $K \times R$ respectively, g_{pqr} denotes the elements (p, q, r) of the $P \times Q \times R$ core array $\underline{\mathbf{G}}$, and e_{ijk} denotes the error term for element x_{ijk} and is an element of the $I \times J \times K$ array $\underline{\mathbf{E}}$

The data can also be seen as having a four-way data structure (13 sites $\times$ 11 variables $\times$ 12 months $\times$ 3 years), but a quadrilinear model was not thought to be suitable owing to the fact that seasonal variations are assumed not to be the same over different years.

The data set was analysed with a program developed by the authors in the MATLAB (The Mathworks, Natick, MA, USA) environment.

4. RESULTS AND DISCUSSION

4.1. Data pretreatment and three-mode PCA

Variables NH_4^+ and P were measured with two different detection limits: 0·1 in the years 1987 and 1988, and 0·01 in the years 1989 and 1990. To eliminate possible bias, all values $<0{\cdot}1$ were set to 0·1.

Except for COD, all the variables show a skewed distribution, and therefore a logarithmic transformation has been applied on them. This kind of distribution is very usual in environmental data, and the logarithmic transformation is commonly used since it allows one to obtain a normal distribution from highly skewed data.

In three-mode methods, scaling and centering of the data are often crucial. Several pretreatments can be applied owing to the possibility of scaling and/or centering along or across the different modes.

Our data set must undergo a pretreatment which can remove the differences among the variables (due to the different scales and measurement units) without removing the differences among the stations and among the sampling times.

Till now, this problem has been solved by performing a *j*-scaling [2,8,15,16]. To do that, the three-way array $\underline{\mathbf{X}}$ is matricized to a two-way matrix $\mathbf{X}_b$ having $I \times K$ rows and J columns. On it, autoscaling is performed; by doing that, the global variance of each variable is set to one, and the differences among the objects and the conditions are preserved. This approach has been followed also in our paper.

It has anyway to be considered that *j*-scaling calculates averages over two modes. This operation removes some offsets but at the same time may introduce some other offsets, thereby introducing

artificial variation that the model has also to model. More suitable pretreatments are therefore currently under study [17].

According to the scree test [18], two components were significant in each mode.

The model parameters have been estimated in a least squares sense and under the orthogonality constraint. The algorithm converges after four iterations, explaining 34·6% of the total variance of the *j*-scaled data. Such a rather low value is not unusual when working with environmental data, because of the very high noise related to the great variability of weather conditions and to the rather high experimental error of some of the variables.

After body diagonalization the following core array is obtained; the cubic core array is reported according to the following unfolding:

$\begin{bmatrix} c_{111} \\ c_{211} \end{bmatrix}$	$\begin{bmatrix} c_{121} \\ c_{221} \end{bmatrix}$	$\begin{bmatrix} c_{112} \\ c_{212} \end{bmatrix}$	$\begin{bmatrix} c_{122} \\ c_{222} \end{bmatrix}$
−34·94	1·99	−1·86	−1·97
1·39	2·13	−2·83	30·48

Since an almost complete body diagonalization has been obtained, it is possible to interpret jointly the three sets of loadings.

The fact that the off-diagonal terms are almost negligible indicates that a very similar result would have been obtained also with a PARAFAC model [5].

4.2. The variables

In the plot of the variables (Figure 2) the first axis shows a contrast between variables related to the presence of algae (fluorescence, chlorophyll-a and total suspended matter) and variables related to chemical pollution (NO_3^-, P and, in a lower measure, NH_4^+). This contrast is rather logical, since a higher amount of the former will decrease the concentration of the latter, NO_3^- and P being nutrients of algae. The fact that NH_4^+ is in an intermediate position can be explained by taking into account that its origin is related to the chemical degradation of NO_3^- and that it is consumed by algae.

The second axis is very clearly related to water transparency, turbidity and NH_4^+.

4.3. The sampling stations

In the plot of the objects (Figure 3) the stations are spread along the second axis. As already discussed, this axis is related to water transparency, with high values corresponding to low water transparency.

The fact that the variation of the objects and of the conditions is almost totally in a single direction does not mean that the intrinsic dimensionality of the corresponding mode is one. It has to be remembered that these plots derive from orthogonal rotations of the original solutions, in which both the objects and the conditions were lying 'diagonally' across both dimensions.

If also the map of the lagoon is taken into account, one can see that the distribution of the stations along the second axis has a very strong correspondence with the geographical location, with the direction low values–high values (high transparency–low transparency) roughly corresponding to the direction outside–inside.

In more detail, the four sites in the industrial area (*f, g, h, i*) are by far the locations where water is less transparent. Water becomes gradually more transparent when moving toward (*d, e*) and past (*c, b*) the city and in the area between Malamocco and the industries, inside the 'petrol channel' (*l*); the four stations with the clearest water are in the middle of the lagoon (*n, o*) and at the two inlets (*a, m*).

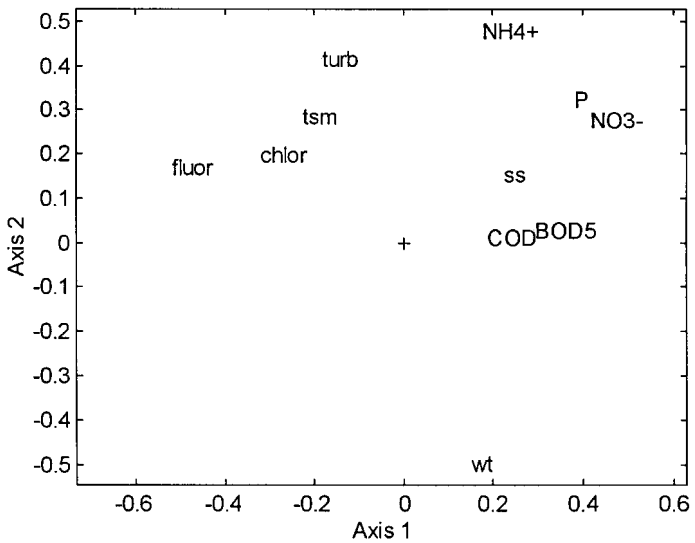

Figure 2. Plot of variables: chlor, chlorophyll-a; tsm, total suspended matter; wt, water transparency; fluor, fluorescence; turb, turbidity; ss, suspended solids; NH_4^+; NO_3^-; P; COD, chemical oxygen demand; BOD_5, biological oxygen demand.

This direction also corresponds to a decrease in salinity; it must be noticed that the variable salinity has not been taken into account so as not to mask the information about chemical and biological characteristics of water with a physical variable having the same general trend.

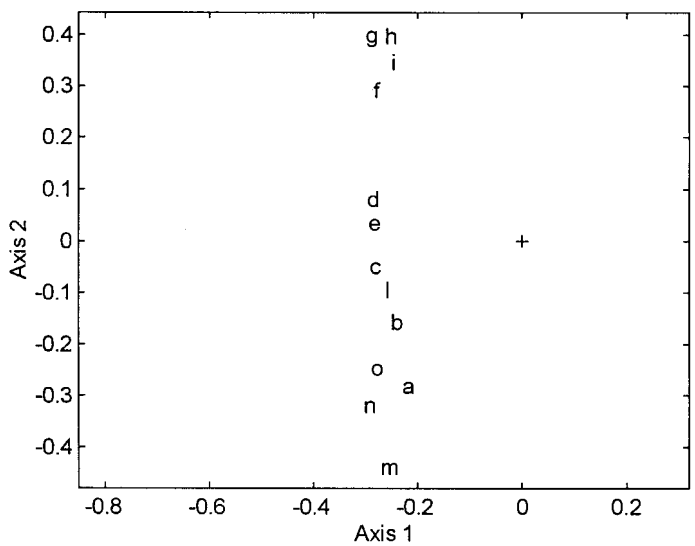

Figure 3. Plot of sampling sites (for coding, see Figure 1).

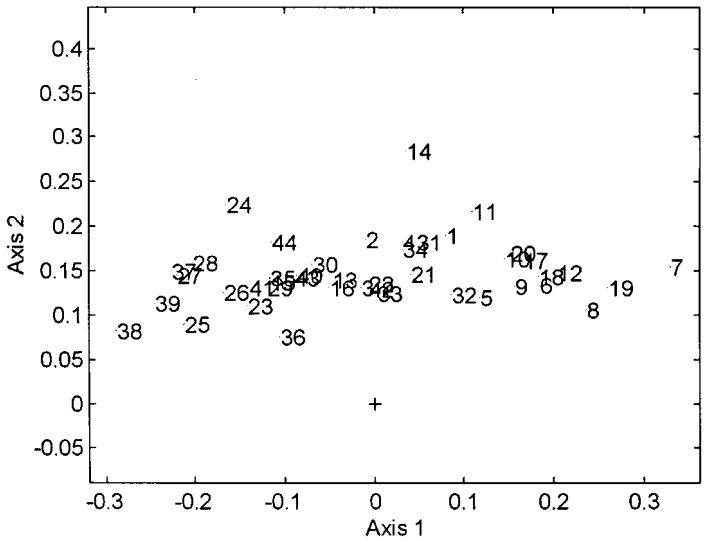

Figure 4. Plot of conditions: 1, May 1987; 2, June 1987; 3, July 1987; …; 44, December 1990.

4.4. The conditions

The interpretation of the plot of the conditions (Figure 4) is much less straightforward.

The points are elongated on axis 1, meaning that a variation in the type of pollution took place during the studied period, while the water transparency does not seem to have changed significantly. It can be noticed that indices higher than 20, corresponding to the years 1989 and 1990, have, on average, lower values on axis 1, meaning that a greater amount of eutrophication and a lower amount

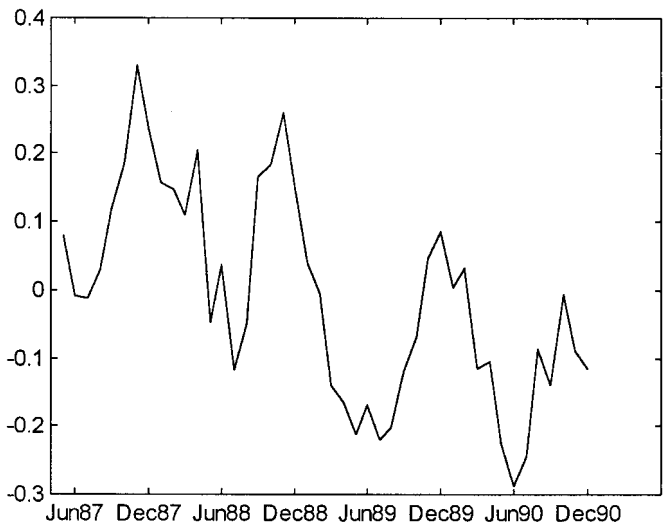

Figure 5. Plot of loadings of conditions on axis 1 versus sampling month.

of nutrients are present. This becomes more evident when plotting the loadings of the conditions on axis 1 versus the sampling time (Figure 5).

In this plot a very strong seasonal effect is shown, with maxima in winter and minima in summer, meaning that the variables that are indicators of eutrophication have a maximum in summer and a minimum in winter, whilst nutrients have a maximum in winter and a minimum in summer, when they are consumed by algae.

It must be noticed that the variable temperature has not been used so as not to mask the information about chemical and biological characteristics of water with a physical variable having by itself a very strong seasonal trend.

5. CONCLUSIONS

The application of three-mode PCA to a complex environmental data set from the Venice lagoon (13 sites $\times$ 11 variables $\times$ 44 samplings) allowed us to obtain an easy interpretation of spatial and temporal phenomena taking place in the region. As a result, it was clear that the main difference among the sites was related to water transparency. In terms of samplings, a strong seasonal effect, with eutrophication having a maximum in summer and nutrients having a maximum in winter, and a general trend of decrease in nutrients and increase in eutrophication were detected.

Of course, nothing new has been added in showing that the industrial area is the most polluted region of the lagoon. On the other hand, it has been possible to show how nutrients have decreased owing to the activity of the sewage treatment plant. The increase in eutrophication also showed that the concentration of nutrients is much higher than the limiting value. This depends on the fact that the lagoon floor, highly polluted, is a continuous source of nutrients diffusing to the surface. To improve the environmental condition of the lagoon, a cleaning of the floor has to be taken into account.

The data set is available from the authors upon request.

ACKNOWLEDGEMENTS

This research was supported by a grant of MURST ‘Progetto Sistema Lagunare Veneziano’. The authors thank P. M. Kroonenberg and R. Bro for useful scientific discussion.

REFERENCES

1. Tucker LR. Some mathematical notes on three mode factor analysis. *Psychometrika* 1966; **31**: 279–311.
2. Kroonenberg PM. *Three-mode Principal Component Analysis*. DSWO Press: Leiden, 1983.
3. Geladi P. Analysis of multi-way (multi-mode) data. *Chemometrics Intell. Lab. Syst.* 1989; **7**: 11–30.
4. Kroonenberg PM. The analysis of multiple tables in factorial ecology. III. Three-mode principal component analysis: ‘analyse triadique complète’. *Acta Oecol.* 1989; **10**: 245–256.
5. Smilde AK. Three-way analyses. Problems and prospects. *Chemometrics Intell. Lab. Syst.* 1992; **15**: 143–157.
6. Beffy JL. Application de l’analyse en composantes principales à trois modes pour l’étude physico-chimique d’un écosystème lacustre d’altitude: perspectives en écologie. *Rev. Statist. Appl.* 1992; **40**: 37–56.
7. Henrion R. Body diagonalization of core matrices in three-way principal component analysis: theoretical bounds and simulations. *J. Chemometrics* 1993; **6**: 477–494.
8. Henrion R. N-way principal component analysis. Theory, algorithms and applications. *Chemometrics Intell. Lab. Syst.* 1994; **25**: 1–23.
9. Alberotanza L, Zucchetta G. *Caratteristiche delle Acque della Laguna di Venezia—Dati Relativi al Monitoraggio Effettuato nel Corso dell’Anno 1987*. CNR-ISDGM, Tipografia Commerciale: Venezia, 1988.
10. Alberotanza L, Zucchetta G. *Caratteristiche delle Acque della Laguna di Venezia—Dati Relativi al Monitoraggio Effettuato nel Corso dell’Anno 1988*. CNR-ISDGM, Tipografia Commerciale: Venezia, 1989.

11. Alberotanza L, Zucchetta G. *Caratteristiche delle Acque della Laguna di Venezia—Dati Relativi al Monitoraggio Effettuato nel Corso dell'Anno 1989*. CNR-ISDGM, Tipografia Commerciale: Venezia, 1990.
12. Alberotanza L, Zucchetta G. *Caratteristiche delle Acque della Laguna di Venezia—Dati Relativi al Monitoraggio Effettuato nel Corso dell'Anno 1990*. CNR-ISDGM, Tipografia Commerciale: Venezia, 1992.
13. Massart DL, Vandeginste BGM, Deming SN, Michotte Y, Kaufman L. *Chemometrics: a Textbook*. Elsevier: Amsterdam, 1988.
14. Meloun M, Militky J, Forina M. *Chemometrics for Analytical Chemistry*, Vol. 1, *PC-aided Statistical Data Analysis*. Ellis Horwood: New York, 1992; Chap. 5.
15. Gemperline P, Miller KH, West TL, Weinstein JE, Hamilton JC, Bray JT. Principal component analysis, trace elements, and blue crab shell disease. *Anal. Chem.* 1992; **64**: 523A–532A.
16. Henrion R, Andersson CA. A new criterion for simple-structure transformations of core arrays in N-way principal component analysis. *Chemometrics Intell. Lab. Syst.* 1999; **47**: 189–204.
17. Bro R, Smilde A. Centering and scaling in component analysis. *J. Chemometrics*, in press.
18. Cattell RB. The scree test for the number of factors. *Multivar. Behav. Res.* 1966; **1**: 245–276.

