Fresenius J Anal Chem (1997) 357 : 522–526

ORIGINAL PAPER

R. Henrion · G. Henrion · M. Böhme · H. Behrendt

Three-way Principal Components Analysis for fluorescence spectroscopic classification of algae species

Received: 31 May 1996 / Revised: 1 July 1996 / Accepted: 4 Juli 1996

Abstract Excitation-emission matrices (EEM) from fluorescence spectroscopy may contain characteristic information about different algae species. As a result of measurements, one gets a whole stack of EEMs each of them corresponding to one species. Such a stack of matrices has to be understood as a cubic data array spanned by the dimensions ‘excitation’, ‘emission’ and ‘species’. The interpretation of higher dimensional data arrays requires efficient tools from multivariate data analysis. In this paper, it is illustrated how Three-way Principal Components Analysis as the appropriate generalization of conventional Principal Components Analysis may serve as a powerful method for classification of algae species.

1 Introduction

The usual method of characterizing the biological composition of phytoplankton samples is based on microscopic counting and on the determination of volume according to Utermöhl [1]. The limitations of this method are defined by its expense of time. Therefore, its application fails, for instance, when short-term changes of phytoplankton compositions are studied with a high temporal resolution and numerous samples have to be processed in a short period of time. On the other hand, due to the distinction of taxonomic groups according to their excitation and emission spectra, fluorescence spectroscopic measurements are appropriate in order to reduce or replace microscopic counting. Up to now, only singular attempts have been made to characterize populations of phytoplankton on a fluorescence spectroscopic basis [2], [3], [4], [5], [6]. The fluorescence behaviour reflects the pigment composition of algae [7]. The contents of chlorophyll-a and accessory pigments varies between different species [8], and, within a species, it depends on the growth conditions. Apart from pigmentation, the amount of absorbed energy being emitted as fluorescence is also a function of the physiological condition of algae [9]. According to the studies of Hilton et al. [2] and Oldham et al. [11], this influences the intensity of fluorescence only, but not the spectral patterns of single species. Therefore, appropriately scaled spectra could be useful for differentiation of algae groups, and the attempts of Oldham et al. [11] to record excitation-emission matrices (EEM) suggest not only to determine the total biomass of algae but even quantitatively to estimate the composition of algae in natural mixture samples.

In order to arrive at such characterizations of algae species and having in mind that EEMs usually produce high amounts of data, one has to employ appropriate methods from multivariate data analysis. These are well described both in statistical and chemometric literature (e.g. [12], [13], [14]). Due to their bilinear nature, EEM data can be decomposed into typical emission and excitation profiles by using Principal Components Analysis (which basically corresponds to a singular value decomposition). These profiles may serve as a characteristic fingerprint for the algae species under consideration. For comparison of different species, the advantage of using the profiles rather than the original EEMs relies on the extreme data (and noise) reduction which is achieved in this decomposition step. Nevertheless, visual inspection and comparison of even a moderate number of different profiles soon becomes a tedious work with growing risk of subjective errors. At this point, one could proceed by applying other methods of data analysis (such as cluster analysis, for instance) to the profiles obtained. To our experience, however, it is more efficient to replace such a sequential, data table oriented approach by so-called Three-

Dedicated to Prof. Dr. Rolf Borsdorf on the occasion of his 65th birthday

R. Henrion
Institute of Mathematics, Humboldt University,
Unter den Linden 6, D-10 099 Berlin, Germany

G. Henrion (✉)
Institute of Chemistry, Humboldt University,
Hessische Strasse 1–2, D-10 115 Berlin, Germany

M. Böhme · H. Behrendt
Institute of Freshwater Ecology and Inland Fisheries,
Berlin, Germany

way Principal Components Analysis, a method which directly takes account of the very three-dimensional data structure (emission × excitation × algae species). The investigations to be discussed here are restricted to a limited data base consisting of the EEMs of five algae species, in order to intimate the methodology and potentials of the described approach rather than to give a thorough interpretation of all possible details.

2 Experimental

Five species of algae were used for fluorescence measurement from different monospecies cultures: *Aphanizomenon flos-aque, Asterionella formosa, Cryptomonas sp., Monoraphidium minutum* and *Synura petersenii.* These species were selected to represent main groups of phytoplankton: Cyanobacteria, Diatoms, Cryptophytes, Chlorophytes and Crysophytes.

Fluorescence was measured on a SPEX Fluorolog 2 with a 150 Watt Xenon arc lamp and single-grating Czerny-Turner-Monochromators with focal length of 0.227 m (excitation) and 0.34 m (emission), respectively. Depending on the fluorescence intensity of the sample, all slits at both the excitation and the emission monochromators were fixed at 1 mm or 2.5 mm, respectively, to reach a spectral dispersion of 2.5 or 6.2 nm. The detector was a PMT Hamamatsu R928. During measurement, samples were magnetically stirred to keep homogeneity.

Each EEM consists of 10 excitation spectra from 350 to 630, step 1 nm. The integration time was 0.1 s/nm. The emission varied from 660 to 705 with a stepwidth of 5 nm. Eleven EEMs of each sample were averaged to reduce noise.

3 Results and discussion

3.1 Database and preliminary steps of analysis

The data base to be considered in the following consists of the EEMs of the 5 algae species mentioned in the experimental part. The EEMs cover a range of 281 excitation wavelengths and 10 emission wavelengths. The first 14 excitation wavelengths are removed from the original data base since they contain perturbations of extreme intensity which would be likely to dominate the interesting part of data information. Figure 1 shows as an example the EEM of *Cryptomonas sp.*. Some perturbation caused by the Raman emission of water is visible around the excitation wavelength of 400 nm and the emission wavelength of 650–670 nm. Since this EEM resulted from a single species, it may be assumed to be generated (up to noise) by a typical emission and excitation spectrum. If, instead, the EEM is caused by a mixture of different species, then it would be based on a whole set of typical emission and excitation spectra. The identification of the underlying spectra is made possible by methods like Principal Components Analysis or Factor Analysis. Principal Components Analysis decomposes a data table into independent pairs of column and row profiles (principal components) thereby successively exhausting maximal parts of data variation. Without noise there would be as many principal components as species in a mixture. Figure 2 shows the first two principal components of the EEM in Fig. 1. The first principal component corresponds to the typical emission (Fig. 2 a) and excitation (Fig. 2 b) spectrum. For the purpose of definition, the emission spectrum is normalized to unit length (the squares of components sum up to one). These two curves may serve as a fingerprint for the considered algae species. In contrast to the use of any of the original spectra (EEM slices at a fixed emission or excitation wavelength), they are strongly noise-reduced and also freed from minor device perturbations. This can be seen from the second principal component (Figs. 2 c, d) which at an excitation wavelength of 400 nm exhibits a sharp peak surrounded by noisy data. Actually, this peak corresponds to the perturbation mentioned in Fig. 1 (but note the different scales of intensity in Fig. 2 d compared to Fig. 2 b).

Similar fingerprints as given by the curves in Fig. 2 a, b result for the remaining 4 species. Visual comparison of these curves sometimes reveals evident differences. How-

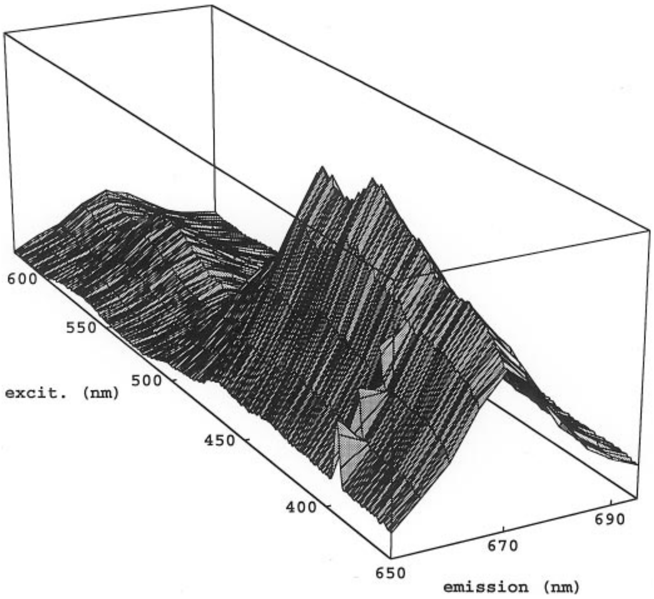


Fig. 1 Excitation-emission spectrum of a selected algae species

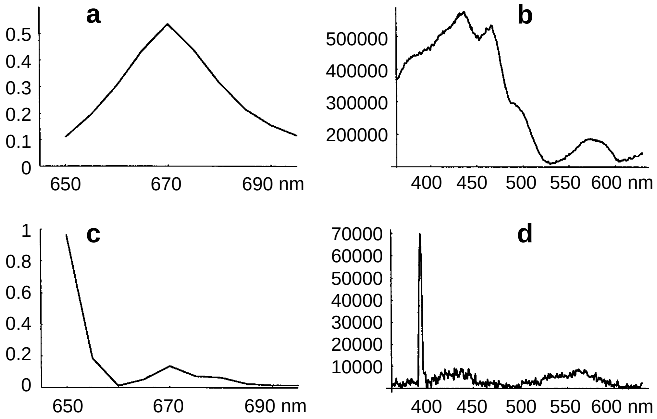


Fig. 2 a–d Principal components decomposition of the EEM in Fig. 1. The first principal component contains the emission and excitation spectra, respectively, shown in **a** and **b**. Similarly **c** and **d** refer to the second principal component ever, the more species are involved in the analysis, the more complicated it becomes to get a systematic overview of similarities or dissimilarities among species without additional tools from data analysis. One possibility is to transform the curve similarities into a taxonomic dendrogram of species by means of hierarchic clustering. Doing so, the intended classification of species is achieved by a two step procedure: the first to extract typical spectra and the second to detect similarities. In the next section, a method is described that allows classification and spectral identification in a single step by taking into account the three-dimensional structure of data.

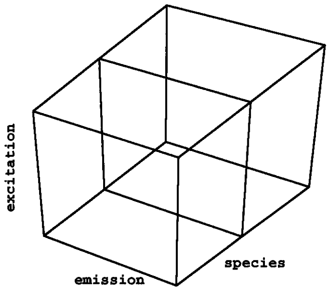
3.2 Three-way Principal Components Analysis

As a result of fluorescence measurements, one gets a collection of EEMs each of which may be viewed as a data table with excitation wavelengths defining the rows and emission wavelengths defining the columns (although of course, physically the EEMs are recorded as files on a computer). However, regarding the whole data base as a loose collection of isolated tables and separately analyzing them as intimated in the preceding section, leads to a loss of information about interrelations. It is advisable instead to understand a collection of tables with equal dimensionality as a three-dimensional array. This is illustrated in Fig. 3 where the EEMs of different species are stacked on top of each other. Accordingly, an element of a three-dimensional array is identified by x_{ijk} where k refers to the number of the corresponding EEM (species) while i, j denote the position in the data table (row, column). An appropriate generalization of Principal Components Analysis from data tables to three-dimensional arrays was first proposed in psychometric literature by Tucker [15] in 1963. The model aims at representing the measured data as a linear combination of few idealized, orthogonal factors. These idealized or latent factors need not be measurable themselves (like certain psychological categories) but rather serve as a basis to explain the measured data. In our context of algae classification, one might ask for idealized emission spectra, idealized excitation spectra and idealized algae species, which may or may not coincide with single or a group of spectra or species, but few of which are sufficient to describe the major part of information contained in the data base. More precisely, the Tucker model writes as

$$x_{ijk} = \sum_{u=1}^{r}\sum_{v=1}^{s}\sum_{w=1}^{t} g_{iu}h_{jv}e_{kw}c_{uvw} + \varepsilon_{ijk} \qquad (1)$$

Here, as noted above, the x_{ijk} refer to the measured data, the g_{iu} denote the intensity of the idealized excitation spectrum no. u at wavelength i, the h_{jv} are the intensity of the idealized emission spectrum no. v at wavelength j and the e_{kw} represent the loading of algae species no. k in the idealized species no. w. The r, s and t are preselected numbers of idealized factors of the three dimensions. Usually, a small value of r, s and t suffices to describe a major part of data variation. The coefficients c_{uvw} indicate how all these idealized factors linearly mix together to yield the measured data. Appropriately put together, the c_{uvw} built up a three-dimensional array **C** which is called the core matrix or core array. This core matrix contains valuable information for the interpretation of the factors obtained. The squared elements c^2_{uvw}, for instance, indicate how much data variance is described by the combination of the idealized factors u, v, and w, respectively. An essential feature of model (1) is that the idealized factors are required to be orthogonal to each other within a fixed dimension. More precisely, if **G**, **H**, **E** are the matrices collecting the elements g_{iu}, h_{jv}, e_{kw} in (1), then these matrices are supposed to be orthonormal. Finally, ε_{ijk} is an error term indicating the difference between measured (x_{ijk}) and explained data (triple sum). Given the measured data and the numbers r, s, t of assumed idealized factors, these factors may be identified using an Alternating Least Squares (ALS) algorithm proposed by Kroonenberg and De Leeuw [16]. In short terms, this algorithm proceeds as follows: starting with some initial orthonormal matrices **G**, **H**, **E** two of these are fixed in turn while the third one is optimized (in the sense of minimizing the error term ε). This optimization is based on a specific eigenvector problem. After convergence of these iterated matrices, containing as their columns the desired idealized factors, the optimal core matrix can be easily estimated as an independent variable. For details of this algorithm and its generalization to N dimensions, we refer to [17]. Two further aspects of Three-way Principal Components Analysis merit to be mentioned: first, the possible ways of data scaling become much more complex as compared to conventional tables. In the present analysis the cubic data array was scaled in such a way as to give all the EEMs of different species the same (unit) variance. There was no scaling carried out within the EEMs. Second, the obtained idealized factors in (1) are unique only up to rotation, i.e. postmultiplication of the factor matrices **G**, **H**, **E** will not affect the goodness of the solution. But it will change well the structure of the core matrix. Therefore, appropriate rotations of the solutions are frequently looked for as in factor analysis in order to simplify the interpretation. In particular the core matrix should contain as few significant elements as possible, in this way reducing the number of factor combinations (between different dimensions) to be interpreted. Finally, it is noted that there exist alternative ways of Three-way decomposition among which the Par-

Fig. 3 A cubic data array built up from a stack of EEMs



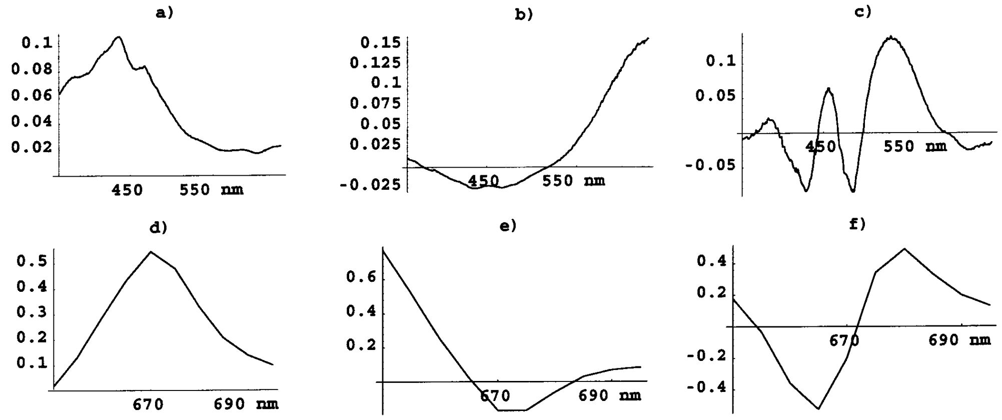


Fig. 4 a–f Plot of the first three idealized excitation (**a**, **b**, **c**) and emission (**d**, **e**, **f**) spectra calculated from a Three-way Principal Components decomposition of the considered data base

allel Factor Analysis (PARAFAC) model is likely to be the most important. For its application to fluorescence spectroscopic data see, for instance [18] or [19]. The PARAFAC model is simpler and easier to interpret than Tucker's model. It is useful, in particular, if the data array is completely determined by instrumental responses (like emission × excitation × time in a fluorescence-HPLC coupling). On the other hand, this model does not allow for interactions between factors of different dimensions. Such interactions are typical, however, if noninstrumental variables (like algae species in our context) enter the data base.

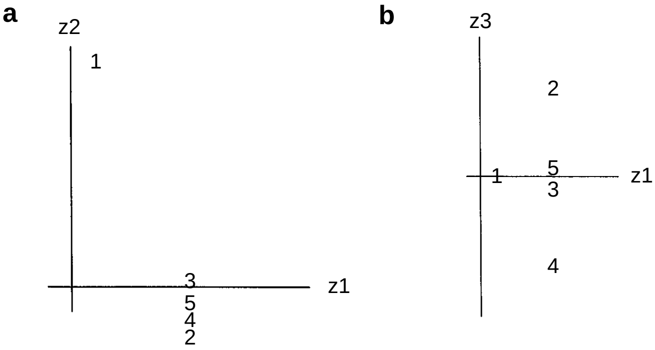


Fig. 5 a, b Plot of the loadings of the five observed algae species with respect to the first vs. the second (**a**) and the first vs. the third (**b**) idealized species

Table 1 Variance contributions of the leading 8 factor combinations as percentage related to the sum of all 27 squared core matrix elements

(u, v, w)	$c^2_{uvw}/\sum^3_{x,y,z} c^2_{xyz}$
(1, 1, 1)	79.03%
(2, 2, 2)	9.56%
(2, 1, 2)	8.46%
(1, 2, 1)	1.00%
(3, 1, 3)	0.73%
(1, 2, 2)	0.71%
(1, 3, 3)	0.18%
(2, 2, 1)	0.12%

3.3 Results

For the given data base the number of idealized factors was chosen to be three in all three dimensions (i.e. $r = s = t = 3$ in (1)). Although the general model allows for independent variation of r, s, t we restrict considerations to an equal number of components in order to keep the presentation simple. Including more than three factors was not advisable here, because then uninteresting parts of data variation (e.g. device-specific perturbations or increasing noise) enter the solutions.

Figures 4 and 5 contain visualizations of the idealized excitation and emission spectra as well as of idealized algae species. Formally, these are plots of the g, h and e- coordinates in (1). The plots are normalized to render the sum of squared coordinates equal to 1 for each factor. The first excitation and emission factor (Fig. 4 a, d) resembles very much the corresponding spectra of *Cryptomonas species* in Fig. 2, so this species is likely to be reflected in the first idealized factor. Note, however, that at least the excitation spectrum (Fig. 4 a) is again noise-reduced to the corresponding spectrum in Fig. 2 of conventional Principal Components Analysis. For the moment, we skip the remaining spectra but turn instead to Fig. 5. Here, the five algae species are plotted corresponding to their loadings with respect to the idealized species labeled, z_1, z_2, z_3. The numbers are chosen according to the ranking of species given in Section 3.1. The diagrams offer an easy way of visual classification: the first diagram suggests that there are mainly two different groups of algae species in the present data base. On the one hand, one has species (2, 3, 4, 5) seemingly coincident with z_1. On the other hand, one has the independent (recall the orthogonality of idealized factors) species 1 which might be identified with z_2. The second diagram reveals minor differences (factors are ranked according to their importance) in the bigger group mainly by opposing species 2 and 4. This difference is reflected in the idealized species z_3. But how does this classification relate to the idealized spectra in Fig. 4? To answer this question one has to study the core matrix along with the identified factors. As mentioned in the preceding section, the squared core matrix elements indicate the variance covered by a corresponding factor combination. Instead of collecting all 27 elements of the (3 × 3 × 3) core matrix (due to $r = s = t = 3$) Table 1 is reduced to the 8 major elements normalized in such a way as to render the sum of all 27 core matrix elements equal to 100%. The factor combinations (u, v, w) have to be read in the order: idealized excitation spectrum, idealized emission spectrum, idealized species. Obviously, the biggest part of data variation is covered by a combination of all first idealized factors (1, 1, 1). This means that the first idealized species z_1 – reflecting the group (2, 3, 4, 5) of observed algae species – first of all relates to the first idealized excitation spectrum (Fig. 4 a) and to the first idealized emission spectrum (Fig. 4 d). In other words, in a first approximation these spectra may be taken as typical fingerprints common to all of the four mentioned species. Concerning idealized species z_2, which may be identified with original species No. 1 (see Fig. 4 a), the major variance contribution is split into the combination (2, 2, 2) and (2, 1, 2). The first index both times being equal to 2, the idealized excitation spectrum of Fig. 4 b is identified as the one typical for this species. This excitation spectrum is clearly different from the one of Fig. 4 a which explains the classification of No. 1 as an independent species. The second index being equal to 1 or to 2 with comparable variance contributions indicates, that the measured emission spectrum of this species is an average of the idealized emission spectra in Fig. 4 d, e. The reason for this splitting is to be found in the fact that the emission spectra of species No. 1, on the one hand, and the other observed species, on the other hand, are correlated to a certain degree (in contrast to the almost uncorrelated excitation spectra). But the idealized spectra are always uncorrelated, so the real emission spectrum of species No. 1 is a weighted sum of the curves in Fig. 4 d, e. Finally, according to Table 1, the third idealized spectra (Fig. 4 c, f) are both related to the third idealized species z_3. The shape of both spectra is typical for shifting slightly (compare the magnitudes of contributions (3, 1, 3) and (1, 3, 3) with that of (1, 1, 1)) the position of peaks in the main spectra (Fig. 4 a, d) after superposition. These difference spectra give the main distinction between species No. 2 and 4 along the z_3- axis in Fig. 5 b). In contrast, species No. 3 and 5 have almost identical fingerprints corresponding to the first idealized spectra. These two species are hardly distinguished by their fluorescence behaviour.

Acknowledgement Financial support by the Fonds der Chemischen Industrie is gratefully acknowledged.

References

1. Utermöhl H (1958) Mitt Internat Verein Limnol 9 : 1–38
2. Hilton J, Rigg E, Jaworski G (1989) J Plankton Res 11 : 65–74
3. Yentsch CS, Phinney DA (1985) J Plankton Res 7 : 617–632
4. Gold VM, Gaevski NA, Shatrov IY, Popelnitskii VA (1986) Gidrobiol Zh 22 : 80–85
5. Krause H, Dirnhofer P, Gerhardt V (1987) Ergeb Limnol, Arch Hydrobiol Beih 29 : 55–62
6. Cowles TJ, Desiderio RA, Neuer S (1993) Mar Biol 115 : 217–222
7. Maske H, Haardt H (1987) Arch Hydrobiol 29 : 123–129
8. Falkowski PG, Wyman K, Ley AC, Mauzerall DC (1986) Biochim Biophys Acta 849 : 183–192
9. Heaney SI (1978) Freshwat Biol 8 : 115–126
10. Ojala A (1993) Phycologia 32 : 22–28
11. Oldham PB, Zillioux EJ, Warner IM (1985) J Mar Res 43 : 893–906
12. Dillon WR, Goldstein M (1984) Multivariate analysis. Methods and applications. Wiley, New York
13. Sharaf MA, Illmann DL, Kowalski BR (1986) Chemometrics. Wiley, New York
14. Henrion R, Henrion G (1994) Multivariante Datenanalyse. Springer, Berlin Heidelberg New York
15. Tucker LR (1963) In: Harris CW (ed) Problems of measuring change. University of Wisconsin Press, Madison, pp 122–137
16. Kroonenberg PM, De Leeuw J (1980) Psychometrika 45 : 69–97
17. Henrion R (1994) Chemom Intell Lab Syst 25 : 1–23
18. Appellof CJ, Davidson ER (1981) Anal Chem 53 : 2053–2056
19. Leurgans SE, Ross RT (1992) Statist Sci 3 : 289–319