Available online at www.sciencedirect.com
SCIENCE DIRECT®

Chemometrics and Intelligent Laboratory Systems 72 (2004) 123–132

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemolab

Procrustes rotation in analytical chemistry, a tutorial

Jose Manuel Andrade[a,*], María P. Gómez-Carracedo[a], Wojtek Krzanowski[b], Mikael Kubista[a,c]

[a] *Department Analytical Chemistry, University of A Coruña, Campus da Zapateira s/n, E-15071, A Coruña, Spain*
[b] *School of Mathematical Sciences, University of Exeter, Laver Building, North Park Road, EXETER, Devon, EX4 4QE, UK*
[c] *Department of Chemistry and Biosciences, Chalmers University of Technology, SE-41390, and MultiD Analysis AB, Göteborg, Sweden*

Received 17 October 2003; received in revised form 10 December 2003; accepted 14 January 2004
Available online 1 April 2004

Abstract

Most analytical chemists are well acquainted with collecting multivariate data and analyzing them by basic tools provided in standard chemometric software. But sometimes this is not enough to extract the desired information. Two typical cases are when data sets need to be compared or when a subset of the measured variables shall be ranked. Both these cases can be addressed by so-called Procrustes rotation and its generalized forms. Procrustes rotation is conceptually a rather simple procedure, and it is available in some chemometric software. In this tutorial, we present the basics of Procrustes rotation, we exemplify its application on some selected examples and, finally, we review the literature. The goal is to make this very powerful technique more popular and accessible to the broader chemometric community.

Keywords: Procrustes rotation; Trilinear decomposition; Variable selection; Spectral analysis; Environmental; Analysis; Chemometrics

1. Introduction

A major goal in Science is to unravel information in measured data that answers the questions that have been raised. Often the problem studied and the questions asked are less well defined. Data related to quality control, environmental analysis, studies of new materials, clinical analyses and the like are particularly difficult to analyze because they are multivariate and it is hard to decide on the relative importance of the measured variables. Even when the overall problem is defined, for example, evaluating the consequences of an oil spill such as that in Galicia, Northwest of Spain, last year, it is not clear what samples should be taken and what parameters should be measured to obtain the best reflection of the catastrophe and its consequences. Even when the problem is clearly defined, it may be unclear what objectives should be reached. Therefore, a very large number of samples is often taken and as many analytical parameters as possible are measured, in the hope that the very large amount of data collected should hold the information requested. This strategy leads to the problem of later selecting which of the measured variables reflect the features of interest and which should be excluded. The mathematical and statistical tools to extract useful information in multivariate data are termed ‘‘chemometrics’’. In this tutorial, we focus on two problems that are of general interest in chemometrics, and that are efficiently addressed by Procrustes rotation methods.

The first problem is the selection of variables. In general when characterizing samples the first time, we measure a large number of variables, many of which may not be very informative. In fact, some may even be unrelated to the problem of interest, and blur the picture instead of making it clearer. Such variables should be removed in analysis and from further measurements. Also less informative variables may be removed for cost and time saving reasons. In cases where measurement time is important, such as on-line monitoring, one may need to remove most of the variables, keeping only the few most informative ones. For these purposes, we need methods to decide how many variables to keep and which ones. Procrustes rotation is a very efficient method to select variables.

The second problem is to find correlations among data sets. It can be wines assessed against a range of descriptors by experts, environmental samples collected at different places and time, spectroscopic measurements on related samples, etc. In these cases, the goal is to find correlations in the measured data. This problem is also efficiently addressed by Procrustes rotation.

* Corresponding author. Fax: +34-981-167065.
E-mail address: andrade@udc.es (J.M. Andrade).

0169-7439/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2004.01.007

It was Hurley and Cattell [1] who first coined the name “Procrustes Analysis” for a set of techniques to compare different sets of principal components. They named the technique after the Damastes innkeeper (Polyphemon, in some legends) in Greek mythology who insisted that travellers staying the night should fit his bed exactly. During their sleep, he either stretched them or chopped off their limbs to ensure they fit the bed. “Procrustes” has became an adjective meaning “hammering in order to elongate (stretch)”. From early applications in Factor Analysis, Procrustes rotation was opened up to more general multivariate usage by Schönemann and Carroll [2] and by Gower [3].

In this tutorial, we describe the basic principles of Procrustes rotation and exemplify it with three case studies. One is an environmental study where 12 different characteristics were measured on 95 sampling points during four sampling seasons. The goals of the analysis are to select a minimum number of variables to describe the main observations and to compare the results of the four sampling seasons. The second example is characterization of the binding of a fluorescent dye to DNA. Here, goals are to determine the number of different complexes formed, their relative amounts, and also the spectral characteristics of the bound dye in the different complexes. The third example is a set of six test samples, each one characterized by two-dimensional fluorescence excitation/emission measurements. Here goals are to determine the spectra and concentrations of the sample components without having access to standards.

Procrustes rotation is usually performed on the principal components that describe the data instead of on the data itself. We therefore start by describing the principal component analysis (PCA) that precedes Procrustes rotation.

2. Principal component analysis

Suppose that p variables have been observed on n individuals to yield *multivariate* data sets that each being arranged in a data matrix $\mathbf{X}$ with n rows (the samples) and p columns (the variables). Insight into many multivariate techniques is helped by geometric visualisation of the data as n points in a space of p dimensions (see note below). This geometrical model of the data matrix is a basis for inspecting the data, in order to uncover any patterns or anomalies that might be present. However, it first needs some simplification in order for it to be of practical use. In particular, we generally need a *reduction in dimensionality,* so that the points may be more readily plotted and inspected. Such a reduction can usually be achieved without excessive loss of information by principal component analysis (PCA).

A note of caution is in order here as there is some confusion in literature concerning the use of the word “dimension” in chemometrics. In many classical works, data arrays of $(n \times p)$ are considered as n points in a p dimensional space, hence, the data are said to be p-dimensional. The term dimension, however, is also used for the dimensionality of the array. An $(n \times p)$ array is two-dimensional (also termed two-way data), while an $(n \times p \times m)$ array is three-dimensional (three-way data). In this tutorial, we adhere to the traditional nomenclature and refer to an $(n \times p)$ space as p-dimensional, describing $(n \times p)$ arrays as bilinear (two-way) and $(n \times p \times m)$ arrays as trilinear (three-way).

Correlations among the different measured variables usually cause most points to lie within a subspace that occupies fewer than p dimensions, with rather few points, outliers, lying outside this subspace. Therefore, in most multivariate data sets, the bulk of the information will lie in a subspace of relatively low dimensionality, while residual noise will be scattered through the rest of the space. The first task is to identify the subspace that contains the significant features [4–6].

Of conceptual importance in PCA classification is the fact that only the configuration of points is fixed while the coordinate axes are of rather arbitrary positions. They can be translated or rotated with the points being kept at their positions. To identify the signal subspace, we first move the axis to the centroid of the points, and then rotate them into a position such that the first few contain the greatest spread of points while the remainder have the points close to the centre. Any such axis defines a new variable, or *component,* that is a linear combination of the original variables. So if the original variables are denoted $x_1, x_2,\ldots, x_p$, and the new components are denoted $y_1, y_2,\ldots y_p$, then each y can be obtained from the x’s by means of a simple formula $y_I = a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{ip}x_p$, where $i = 1, \ldots, p$ [6].

Moving the axes to the centroid of the points simply corresponds to mean-centering the data. The condition that the first few components should represent the greatest spread of the points is fulfilled by calculating the principal components of the data. Its coefficients, also called loadings, a_{ij}, can be easily computed using standard software. For example, there is a mathematical decomposition of any matrix known as the *singular value decomposition* (svd). This will decompose the data matrix $\mathbf{X}$ into a product of three matrices: $\mathbf{X} = \mathbf{UDV}^{\mathrm{T}}$. Here $\mathbf{U}$ has n rows and p columns, $\mathbf{D}$ is a $(p \times p)$ diagonal matrix and squares, d_i^2, of its elements represent the amount of information corresponding to each principal component, and $\mathbf{V}$ is a $(p \times p)$ orthogonal matrix with columns containing the loadings, a_{ij}, of the ith principal component. The superscript T denotes matrix transpose. When $\mathbf{X}$ is subject to svd, the best r-dimensional approximation of the p dimensional space in a least-squares sense is given by a plot of points on the first r principal components as axes. The coordinates of the points on these axes are called principal component scores. These are also obtained from the singular value decomposition of $\mathbf{X}$ (they are the first r columns of $\mathbf{UD}$) [5,6].

2.1. Selecting the number of components

The first important question in multivariate data analysis is to decide how many PCs should be used to model the data. We want r to be large enough to contain the essential features in the data, but we also want it to provide a reasonably simple structure for plotting and inspecting the data. Many different approaches have been suggested to answer this question, but no definite single best approach has yet emerged. Among these approaches, the method of Eastment and Krzanowski [4] takes a predictive and quite intuitive standpoint and utilises the above singular value decomposition in a direct way. Let x_{ij}, u_{ij}, d_I, v_{ij}, denote elements of the matrices **X**, **U**, **D** and **V**. If the first r components correspond to "signal" and the remaining (p–r) components represent "noise", we can write: $x_{ij} = \sum_{k=1}^{r} u_{ik} d_k v_{jk} + \varepsilon_{ij}$ where ε_{ij} is the residual noise. A predictor of x_{ij} is calculated as

$$\tilde{x}_{ij} = \sum_{k=1}^{r} u_{ik} d_k v_{jk}. \qquad (1)$$

The goodness of fit for a particular r can then be obtained from the predictors of all elements of **X** by calculating the overall discrepancy between observed and predicted elements:

$$\text{PRESS}(r) = \frac{1}{np} \sum_{I=1}^{n} \sum_{j=1}^{p} (\tilde{x}_{ij} - x_{ij})^2.$$

To avoid bias, i.e., not to use each data point in both prediction and assessment, the data point x_{ij} should not be used in the prediction of $\tilde{x}_{ij}$ [4,5] (nevertheless, as much of the original data as possible should be used in predicting x_{ij}). This is avoided as follows: Delete the ith row of **X**, mean center the columns and denote the result $\mathbf{X}^{(-\mathrm{i})}$. Likewise delete the jth column of **X**, mean center the columns and denote the matrix $\mathbf{X}_{(-\mathrm{j})}$. Perform singular value decompositions of these two new matrices [4] $\mathbf{X}^{-1} = \bar{\mathbf{U}}\bar{\mathbf{D}}\bar{\mathbf{V}}^{\mathrm{T}}$ with $\bar{\mathbf{U}}$=($\bar{u}_{\mathrm{st}}$), $\bar{\mathbf{V}} = \bar{v}_{\mathrm{st}}$, and $\bar{\mathbf{D}} = \mathrm{diag}(\bar{d}_1,\ldots,\bar{d}_p)$, and $\bar{\mathbf{X}}_{(-j)} = \tilde{\mathbf{U}}\tilde{\mathbf{D}}\tilde{\mathbf{V}}^{\mathrm{T}}$ with $\tilde{\mathbf{U}}$=($\tilde{u}_{\mathrm{st}}$), $\tilde{\mathbf{V}}$=($\tilde{v}_{\mathrm{st}}$), and $\tilde{\mathbf{D}}$=diag ($\tilde{d}_1, \ldots, \tilde{d}_{(p-1)}$).From these construct, the predictor [4–6]:

$$\hat{x}_{ij} = \sum_{k=1}^{r} \left(\tilde{u}_{ik} \sqrt{\tilde{d}_k} \right) \left(\sqrt{\bar{d}_k} \bar{v}_{jk} \right). \qquad (2)$$

In difference with the predictor in Eq. (1), this new predictor does not make use of x_{ij}. To find optimum value of r, PRESS(r) is calculated for different values of r from 1 to $p - 1$. Eastment and Krzanowski [4] suggested to calculate $W_r = ((\text{PRESS}(r-1) - \text{PRESS}(r))/D_r) \div (\text{PRESS}(r)/D_{\mathrm{res}(r)})$, where D_r is the number of degrees of freedom required to fit the rth component and $D_{\mathrm{res}(r)}$ is the number of degrees of freedom that remains after fitting the rth component. $D_r = n + p - 2r$ and $D_{\mathrm{res}(r)}$ is obtained by successive subtractions starting with $(n-1)p$ degrees of freedom for the mean-centered matrix **X**, i.e., $D_{\mathrm{res}(1)} = (n-1)p$ and $D_{\mathrm{res}(r)} = D_{\mathrm{res}(r-1)} - (n + p - 2(r-1))$ [4–6]. Formed this way, W_r represents the increase in predictive power when adding the rth component, divided by the mean predictive information in the remaining components. Hence, the number of significant components r is given by the highest value for which W_r is greater than 1.

3. Procrustes analysis

Having determined the number of significant PCs, we are ready to compare data sets by Procrustes rotation. This is better achieved by using the principal component subspace where we have signal rather than noise. So, let **T** and **Z** the $n \times r$ scores matrices for two original data sets (**X** and **Y**, respectively) and let one of them, **T**, be fixed and transform the other data set, **Z**, to match **T**. Geometrically, this is done

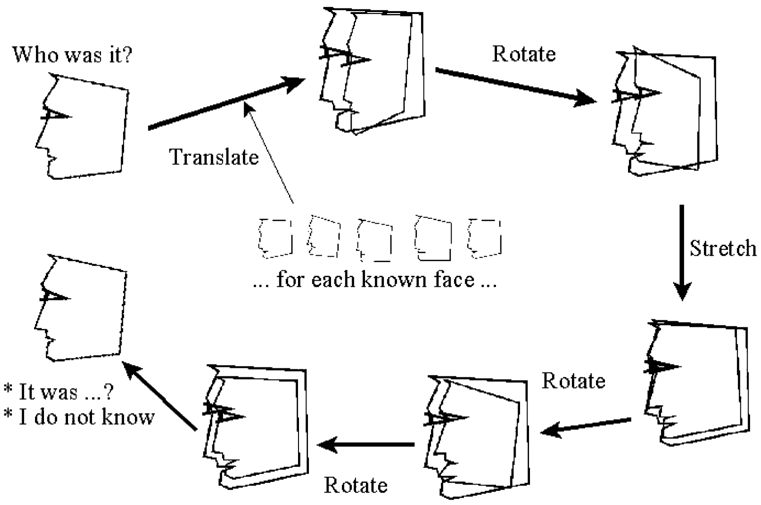

Fig. 1. Illustration how human brain may use Procrustes rotation to identify faces.

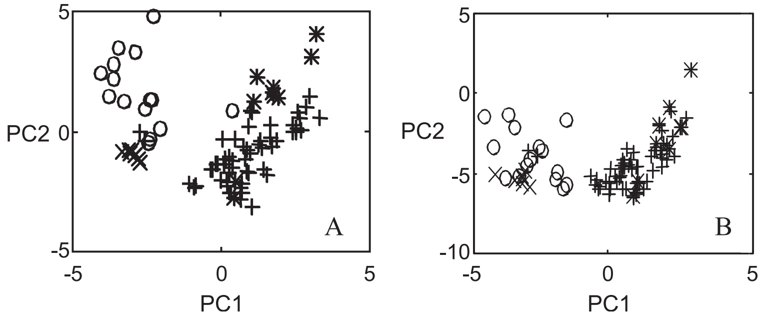

Fig. 2. Principal component scores of roadsoils considering 12 variables. Only results for fall (A) and spring (B) are shown. Highway (○), city gardens (+), transects (×), main avenue (*).

by translating, rotating, and then stretching/shrinking it such that the sum of squared distances, M^2, between the elements of $\mathbf{Z}$ and the corresponding elements of $\mathbf{T}$ is minimized. The smaller value of M^2, the more similar are the two configurations. A perfect match gives $M^2=0$. Below we briefly describe Procrustes rotation. For thorough derivation, see Krzanowski [6].

The first step, translation, is effected by mean-centering the original $\mathbf{X}$ and $\mathbf{Y}$ (if no data pretreatment was carried out in advance, the scores should be mean-centered). For rotation and stretching, we first perform singular value decomposition of the product $\mathbf{T}^{\mathrm{T}}\mathbf{Z}$ to obtain $\mathbf{UDV}^{\mathrm{T}}$. Rotation and reflection is then performed by multiplication with the matrix product $\mathbf{VU}^{\mathrm{T}}$. The discrepancy between $\mathbf{T}$ and $\mathbf{Z}$ (so between $\mathbf{X}$ and $\mathbf{Y}$) is finally calculated as $M^2=\mathrm{trace}(\mathbf{TT}^{\mathrm{T}}+\mathbf{ZZ}^{\mathrm{T}}-2\mathbf{D})$. Note that $\mathrm{trace}(\mathbf{TT}^{\mathrm{T}}+\mathbf{ZZ}^{\mathrm{T}}-2\mathbf{D})=0$ when $\mathbf{T}=\mathbf{Z}$.

Apart from its mathematical principles, Procrustes rotation can be seen as a natural way to compare objects and might mimic what our brains employ in daily life. Think, for instance, what happens when we obtained a blurred vision of a face of a person we pass when walking fast. Probably, our brain will make some kind of "Procrustes comparison" by considering in turn each of the faces stored in our memory to identify the "fuzzy" face (Fig. 1). A Procrustes approach is being used in the analysis of pictures taken by CCTV cameras of the suspect in the murder case of Swedish former foreing Minister Anna Lindh (http://chalmersnyheter.chalmers.se/Article.jsp?article = 2352).

To select variables, let us first suppose that the number of variables r has been decided, either from substantive prior knowledge or, for example, by the W_r statistic. This implicitly suggests that we anticipate the "true" configuration dimensionality to be (no greater than) r. The first r principal components are extracted from the data matrix $\mathbf{X}$, and the coordinates of the data points on these components are taken as the "basis configuration". Let us denote this configuration by $\mathbf{Z}$. Each variable is then removed in turn from the data set and the first r principal components are extracted from the reduced $n\times(p-1)$ data matrix. Let the coordinates of the data points on these components be denoted by $\mathbf{Z}_{(j)}$ when the jth variable has been removed. Comparing $\mathbf{Z}_{(j)}$ with $\mathbf{Z}$ by means of Procrustes Analysis gives the discrepancy value $\mathbf{M}^2_{(j)}$. The variable which causes the least disturbance to the data configuration when it is omitted, is the one that has the smallest $\mathbf{M}^2_{(j)}$. value. This variable is removed from the set to leave $p-1$ variables. The procedure is then repeated on the reduced set of variables to find the one out of $p-1$ variables that can be eliminated with least disturbance. The procedure goes on until only r variables remain. These will be the "best" r variables, in the sense that they are the r variables that best capture the structure of all original p variables.

Experimental research often has to cope with qualitative variables which makes the comparisons a bit more complex. Krzanowski [7] has therefore extended the Procrustes rotation technique to situations where all variables are categorical, so that correspondence analysis must be employed

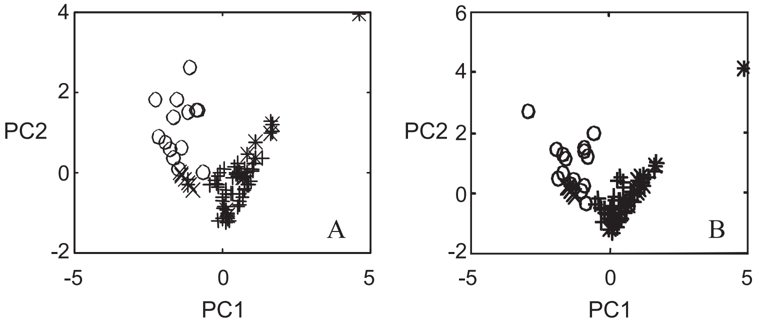

Fig. 3. Principal component scores of roadsoils considering only the two Procrustes-selected variables. Fall (A) and spring (B). Highway (○), city gardens (+), transects (×), main avenue (*).

rather than principal component analysis to form the basis configuration.

4. Example 1: variable selection with Procrustes rotation

Samples were collected in the medium-size city (ca. 300,000 inhabitants) of A Coruña, NW Spain and its surroundings. Soils were sampled at 95 points representing green areas, along a main avenue with high traffic density (ca. 100,000 vehicles per day), along a highway and along some perpendicular transects of the highway (both cultivated and uncultivated fields). Samples were collected in fall, winter, spring and summer, and analyzed for the heavy metals Cd, Co, Cu, Cr, Fe, Mn, Ni, Pb and Zn, humidity, loss on ignition (LOI), and pH.

For all sampling season, the 12 measured variables produce distributions in which the 95 sampling points divide into two main blocks that can be identified as gardens within the city and samples pertaining to the highway. This main division is found in the principal component score plots PC1–PC2 for each season. The first principal component accounts for about 38% of the total information and somewhat depends on sampling season (Fig. 2). Detailed inspection of the loadings [8] suggests that metals related to the lythological characteristics of the soils are a main discriminator. Hence, the main difference between the two groups of samples can be attributed to natural soil variation.

PC2 instead discriminates between samples that have different pollution characteristics. Negative PC2 scores and scores close to zero reflect samples collected at the border of the highway or at the main avenue. Positive PC2 scores are from the transect samples along a highway and most of the city gardens (Fig. 2A). The differences are larger during fall and winter possibly due to accumulation of pollutants during summer and late spring (that bring the groups closer) (Fig. 2B).

Table 1
Loadings for the two most significant consensus vectors and the angles they form with the corresponding principal components for the four seasons

Variable	⟨PC1⟩	⟨PC2⟩
Cd	0.46	– 0.06
Co	– 0.16	– 0.49
Cu	0.43	– 0.10
Cr	0.02	– 0.54
Fe	– 0.24	– 0.40
Mn	– 0.17	– 0.16
Ni	0.07	– 0.48
Pb	0.46	– 0.15
Zn	0.45	– 0.03
Angles	$\alpha 1$	$\alpha 2$
Autumn	6.6	24.0
Winter	6.4	19.0
Spring	3.7	17.7
Summer	6.6	10.4

In the study, a total of 12 variables were measured. If some could be left out without compromising measured information, too much workload and money could be saved. This can be tested by Procrustes rotation. Using the W_r statistic, we find that two PCs reflect most of the features of the four seasons, hence, two variables should be sufficient to describe the system. By Procrustes rotation, these were identified as Co and Pb (Cd) in autumn, Co and Pb (Cd) in winter, Co and Pb (Cd) in spring, and Pb and Cd (Co) in summer. The elements in brackets are the third most important variables in each season, respectively.

It is noteworthy that a lithological and an anthropogenic variable are selected. This most likely reflect the main difference between the two sample groups ("city" vs. "highway"). From these variables, we can predict increased pollution in summer time, reflected by high values of Pb and Cd.

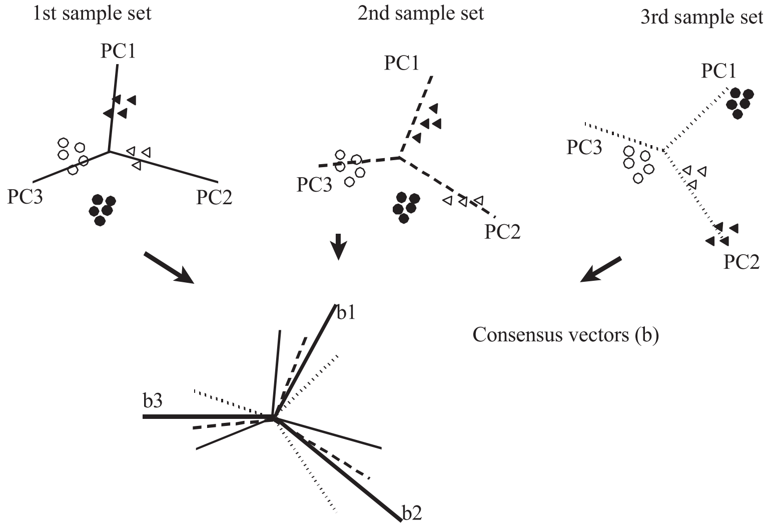

Fig. 4. Concept behind generalized Procrustes rotation. Angles b-PC denote closeness amongst the consensus and original spaces. "Loadings" give chemical interpretation of the consensus vectors.

Consider the principal component scores plots (PC1 vs. PC2) constructed using only the two selected variables. The agreement with the original PC scores plots is satisfactory as the main sample patterns can still be observed (Fig. 3) [8]. By visual inspection, we see similarities between the four sampling seasons. By Procrustes rotation, we can quantify them. Although regular Procrustes rotation compares two data sets, four data sets can be compared by applying generalized Procrustes rotation [6].

The idea behind generalized Procrustes rotation is to find a consensus subspace to which the individual subspaces are compared simultaneously. The methodology depends on whether points or axes are used to compare the subspaces; here we use axes. If r principal components have been extracted for each data set, then we can derive r consensus vectors, $\mathbf{b}$, that resemble all data sets as closely as possible. The consensus subspace is given by [6]:

$$\boldsymbol{H} = \sum_{g=1}^{G} \boldsymbol{L}_g{}' \boldsymbol{L}_g$$

where G is the number of subspaces compared and columns of $\mathbf{L}_g$ are the principal component loadings in the gth subspace. We can even quantify the similarity between the data sets and the consensus subspace by calculating the angle between its principal components and the corresponding consensus vectors. For the first consensus vector, the angle is $\cos^{-1}\{(\mathbf{b}_1'\mathbf{L}_g'\mathbf{L}_g\mathbf{b}_1)^{1/2}\}$. Fig. 4 illustrates these conceptual ideas.

Table 1 shows the two most significant consensus vectors for the four sampling seasons. The first consensus vector is dominated by Pb, Cd, Zn and Cu. These are the variables for which the sample scatters (groups) are similar during the year. Note that the angles between the 1st consensus vector and the original PCs are generally very low. This means that these variables behave similarly in the four seasons. The second consensus vector is mainly associated with Cr, Co and Ni. The angles between this consensus vector and the corresponding PCs are larger, reflecting lower correlations.

5. Example 2: Procrustes rotation in spectral analysis

Spectroscopic measurements are popular in the study of test samples. Usually one is interested in identifying the substances present, which is typically done by comparing

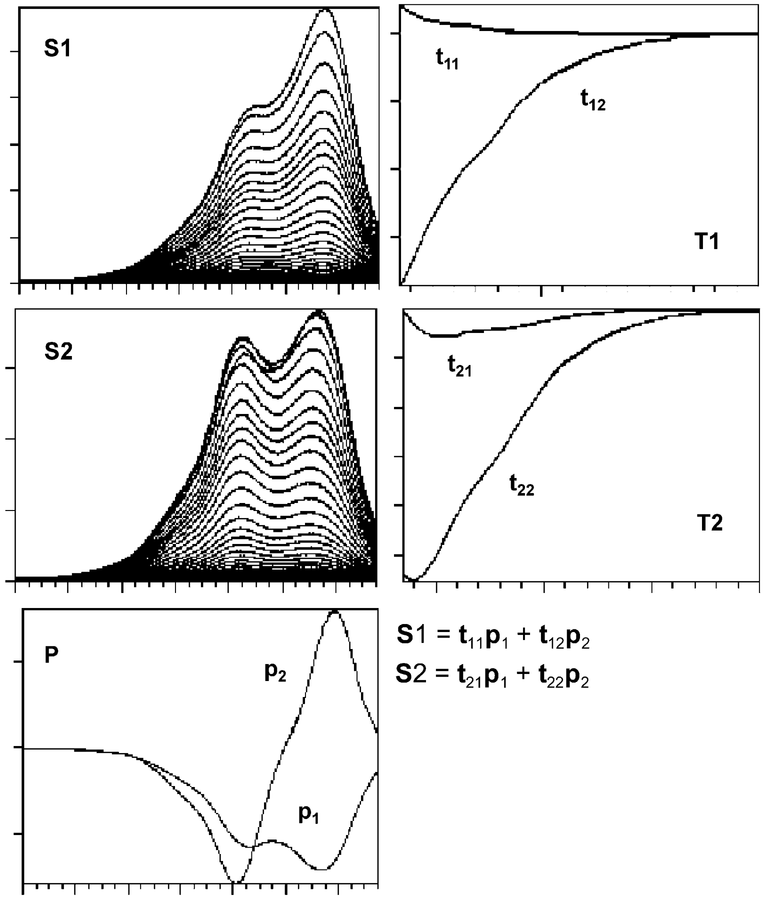

Fig. 5. Application of Procrustes rotation to spectral data to identify pure profiles and relative concentrations. **S**1 and **S**2 are excitation spectra measured at different emission wavelenghts for samples containing 1 dye per 40 and 20 base pairs, respectively; **T**1 and **T**2 are their scores matrices, **P** are common loadings.

measured spectra with those of pure substances. However, reference spectra may not always be available. In such cases, the substances can be identified by spectroscopic measurements that are related to Procrustes rotation. The approach requires a minimum of two data sets that are suitably correlated [9]. It can be two spectra measured on each of a set of samples, or it could be a two-dimensional spectroscopic measurement on each of two samples [10]. It could also be a single sample studied by a suitable three-dimensional technique [11]. The measured data should be correlated such that the contribution of each substance to the two spectra has the same wavelength dependence but different spectral magnitude. Typical combinations are excitation spectra measured at different emission wavelengths and emission spectra measured at different excitation wavelengths.

When designed this way, the experiments generate two correlated sets of spectra that are described by $\mathbf{S}1=\Sigma(\mathbf{c}_k\mathbf{v}_k)$ and $\mathbf{S}2=\Sigma(\mathbf{c}_k d_k\mathbf{v}_k)$, where $\mathbf{c}_k$ is the concentration dependence of chemical k, $\mathbf{v}_k$ is its normalized spectrum, and d_k is the relative amplitude of the responses of chemical k in the second measurement compared to the first. In matrix notation, $\mathbf{S}1=\mathbf{CV}$ and $\mathbf{S}2=\mathbf{CDV}$.

By first analyzing the two data sets separately by singular value decomposition, we can then relate them by Procrustes rotation. This determines **C**, **V** and **D** [9]. Briefly, PCA gives $\mathbf{S}1=\mathbf{T}1\mathbf{P}^{\mathrm{T}}$ and $\mathbf{S}2=\mathbf{T}2\mathbf{P}^{\mathrm{T}}$. Then $\mathbf{T}1=\mathbf{S}1\mathbf{P}$ and $\mathbf{T}2=\mathbf{S}2\mathbf{P}$. Procrustes rotation transforms **T**1, **T**2 and $\mathbf{P}^{\mathrm{T}}$ into **C**, **CD** and **V** by calculating a rotation matrix, **R**, such that $\mathbf{C}=\mathbf{T}1\mathbf{R}^{-1}$ and $\mathbf{V}=\mathbf{RP}^{\mathrm{T}}$. This is done as follows: **T**1 is Procrustes rotated to yield **T**2 as $\mathbf{Q}=(\mathbf{T}1^{\mathrm{T}}*\mathbf{T}1)^{-1}*\mathbf{T}1^{\mathrm{T}}*\mathbf{T}2$, which defines **D** and **R** by the similarity transform $\mathbf{D}=\mathbf{RQR}^{-1}$. Procrustes rotation analysis is readily performed using (e.g.) DATAN, available at www.muldid.se.

As an example, we analyze the binding of the fluorescent asymmetric cyanine dye thiazole orange to DNA. Asymmetric cyanine dyes have fluorescence only in bound state, which makes them popular labels of biomolecules. They are used, for example, in the LightUp probes for detection of nucleic acids in real-time PCR (see, www.lightup.se for more details) [12].

Two mixtures of thiazole orange and DNA were prepared with one dye bound per 20 and 40 base pairs, respectively, and fluorescence excitation spectra were measured at thirty-four different emission wavelenghts of each sample (Fig. 5).

First, the number of principal components, which should correspond to the number of independent fluorescent species that are present, was determined. Two significant components were found, which nicely account for the

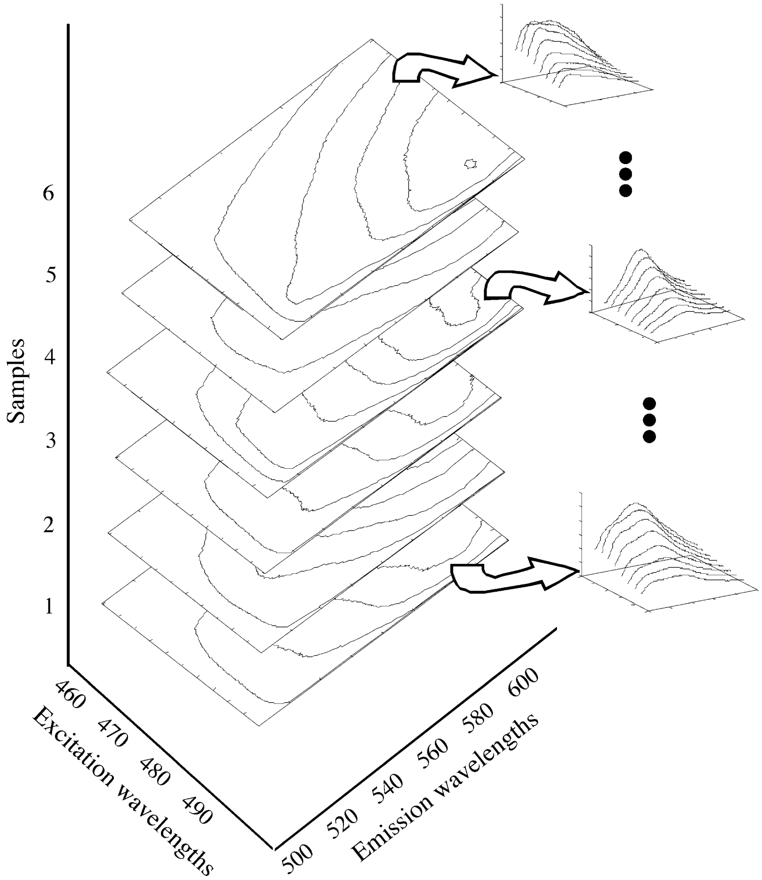

Fig. 6. Trilinear decomposition of a three-array data set. Excitation/emission scans of six samples shown as contour plots (each one represents a bilinear data set). For three of the samples intensity plots are shown in two-dimensional surface plots.

measured spectra (Fig. 5). The scores vectors however do not represent real spectra. To calculate components' spectra, the score vectors are rotated by Procrustes rotation. This produces the spectra shown in Fig. 5. The calculated spectra have distinct features and can be identified as bound monomer and dimer, respectively [13,14]. Their concentrations, calculated from the *d*-values, reveal that at low binding ratios (1 dye per 40 base pairs) 77% of the dye binds as monomer, while at high binding ratios (1 dye per 20 base pairs) only 41% is bound as monomer.

6. Example 3: trilinear decomposition

Trilinear decomposition is an special form of generalized Procrustes rotation. It compares several data sets [13] but also requires the components' contributions to be factorizable.

$$\boldsymbol{I}(\alpha,\beta,\gamma)=\sum_{i=1}^{r}\boldsymbol{I}_i(\alpha,\beta,\gamma)=\sum_{i=1}^{r}\boldsymbol{I}_i(\alpha)\boldsymbol{I}_i(\beta)\boldsymbol{I}_i(\gamma)$$

Trilinear decomposition is also called Parallel Factor Analysis (PARAFAC), which was developed by Harshman in 1970 [15]).

Typical cases where trilinear decomposition is powerful are studies of chemical equilibria and chemical reactions by spectroscopic methods and for analysis of test samples by two-dimensional measurement techniques [13]. The latter is illustrated here by an example.

Six samples containing Fluorescein, Eosin Y, Rhodamine 6G and Rhodamine B at different concentrations were prepared and their emission spectra were recorded at eight excitation wavelenghts. The measured data are described by:

$$\boldsymbol{I}(\lambda_{\text{ex}},\lambda_{\text{em}},\boldsymbol{c}(\text{pH}))\propto\sum_{i=1}^{n}\boldsymbol{I}_i(\lambda_{\text{ex}},\lambda_{\text{em}},\boldsymbol{c}(\text{pH}))$$
$$=\sum_{i=1}^{n}\boldsymbol{c}_i(\text{pH})\boldsymbol{I}_i(\lambda_{\text{ex}},\lambda_{\text{em}}) \qquad (3)$$

Since the emission spectrum of a pure fluorescent species is, in general, independent of the excitation wavelength used and, vice versa, the excitation spectrum is independent of the wavelength of emission, the excitation/emission matrix can be factorized:

$$\boldsymbol{I}(\lambda_{\text{ex}},\lambda_{\text{em}},\boldsymbol{c}(\text{pH}))\propto\sum_{i=1}^{n}\boldsymbol{I}_i(\lambda_{\text{ex}},\lambda_{\text{em}},\boldsymbol{c}(\text{pH}))$$
$$=\sum_{i=1}^{n}\boldsymbol{I}_i(\lambda_{\text{ex}})\boldsymbol{I}_i(\lambda_{\text{em}})\boldsymbol{c}_i(\text{pH}) \qquad (4)$$

Fig. 6 shows the excitation/emission scans of the six samples as contour plots. For three of the samples, the data are also shown as two-dimensional surface plots. The collec-

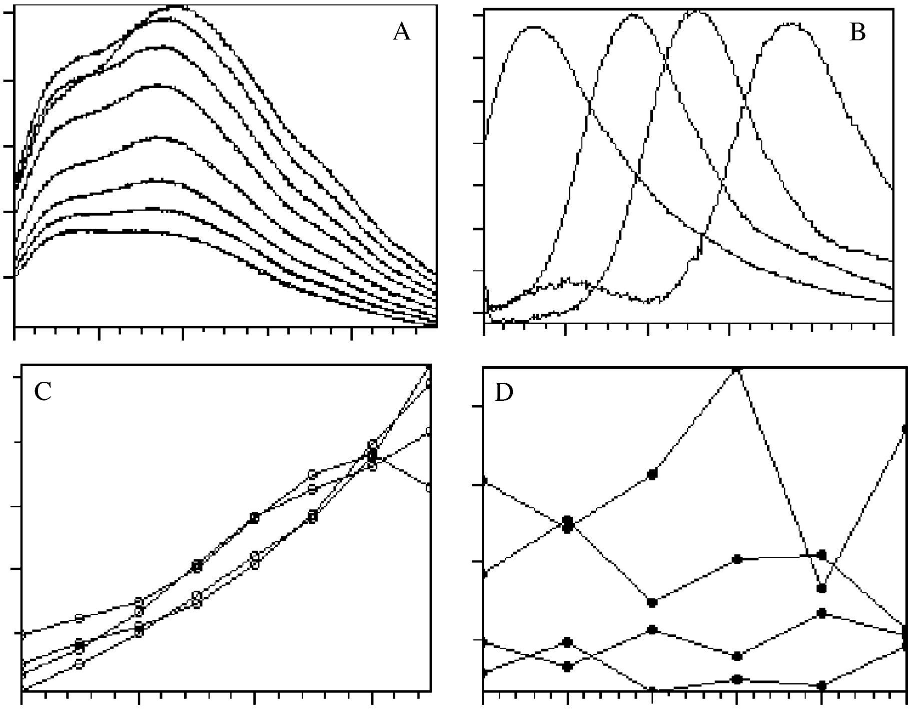

Fig. 7. Decomposition of the data in Fig. 6. Subplot (A) shows excellent agreement between measured and reproduced spectra assuming four principal components (measured and reconstructed spectra coincide to such degree that they are hard to distinguish). Data shown are emission spectra measured at eight different excitation wavelengths of one of the four-compounds mixtures. Equally good agreement was found for the other five mixtures. Subplots (B) and (C) show calculated emission and excitation spectra, respectively, for the four compounds [13]. Subplot (D) shows the calculated concentrations of the chemicals in each of the six mixtures.

tion of these six contour plots, each one being a bilinear data set, makes up the trilinear data set (three-array data set).

Four principal components reproduce the measured spectra (only one set is shown). By trilinear decomposition the emission spectra, the excitation spectra and the concentrations for all substances are calculated (Fig. 7). Comparing with spectra of standards, which were not used in the analysis, showed that the calculated spectra are in excellent agreement with the true ones.

7. More applications of Procrustes rotation in analytical chemistry

We have presented three examples of Procrustes rotation in some detail above. Short description of some more selected applications of Procrustes analyses follows below with references to original work.

In environmental studies, it is commonly important to identify the essential measurement variables to develop simpler and more economic monitoring schemes [16]. In airborne-related studies, King and Jackson [17] used Procrustes rotation to select variables that best describe the sample groups and Richman and Vermette [18] applied it to compare data space (measured variables) with target space (i.e., a list of pollution sources). In this way, Procrustes rotation is an attractive way to source apportionment.

Industrial quality control is another important field where Procrustes rotation methods are becoming important. In industry, homogeneous products are generally fabricated and it is seldom meaningful to group samples. This makes it hard to identify problems and many PCs are usually required to adequately describe the system. For quality control, it is therefore advisable to combine Procrustes rotation with industrial know-how and expertise. Deane and McFie [19] applied Procrustes rotation to quality control of kerosene, and Andrade et al. [20] verified that almost the same variables were useful to describe kerosene quality from a different refinery. The good agreement between the two studies supports the notion that Procrustes rotation is a valuable tool for quality control in production.

Scarponi et al. [21] determined total phenols, polymerizable phenols, antocyanins and leucoantocyanins in Italian red wines that were subjected to different treatments. By Procrustes rotation, the main features of the data were identified and used to define a subset of variables that described the main features.

Procrustes analysis was used to determine the number of significant masses in GC-MS by Demir et al. [22] and Bessant et al. to compare data collected by two detectors in a triply coupled diode array liquid chromatography electrospray mass spectrometry device [23]. The latter study determined consensus chromatogram and consensus spectra that accounted for the data measured by both detectors. In a following study [24], the authors reduced the mass spectra of 2- and 3-hydroxypyridines to 20 significant masses by comparing the data collected by the two detectors. Schulze and Stilbs [25] employed Procrustes rotation to analyze Fourier-transform pulsed-gradient spin-echo NMR data to identify specimens with distinct diffusion coefficients. A refined approach was later employed for magnetic resonance imaging by Antalek and Windig [26].

Robert and Carbo [27] used Procrustes rotation to describe organic molecules. They compared molecular quantum similarity measures to find redundancies in the information. In another molecular classification Tomas et al. [28] classified newly-synthesised fulvene molecules into groups defined by the Cambridge Structural Database using a reduced set of descriptors identified by Procrustes rotation.

Guo et al. [29] presented a method based on Procrustes rotation and genetic algorithms to select a subset of variables in PCA. A similar approach was used by Guo et al. [30] to select a subset of variables in sequential projection pursuit to preserve as much sample clustering information as possible.

Application of Procrustes rotation to resolve spectra was pioneered by Scarminio and Kubista [31], and has since then been applied to many systems [11]. Applications include the characterisation of pairs of samples by two-dimensional spectroscopy [32] and the characterisation of a single sample by a three-dimensional measurement [11].

Antonov et al. [33] compared the Procrustes-based approach with a technique of simultaneous resolution of overlapping bands and found them to be similar for studies of monomer-dimer equilibria. Tang and Wang [34] resolved overlapping spectra of nitrites and nitrates using Procrustes rotation.

Vigneau et al. [35] related NIR spectra of oils to mid-IR spectra by Procrustes rotation. Anderson and Kalivas [36] presented some fundamentals for calibration transfer using piece-wise Procrustes rotation and constrained Procrustes analysis. They compared these methods with piece-wise direct standardisation and direct standardisation.

González-Arjona et al. [37] developed a Procrustes discriminant analysis method and compared it to Discriminant Partial Least Squares, Class Modeling Linear Discriminant and Artificial Neural Networks algorithms. They found that Procrustes rotation performs at least as well as discriminant PLS and ANN when the classes were linearly distributed.

Heberger and Andrade [38] compared variable selection by Procrustes rotation with the new nonparametric Generalized Pair-wise Correlation Method (GPCM), and found them to produce similar results.

8. Concluding remarks

Being a rather unknown and “exotic” technique some 10 years ago, Procrustes rotation and its generalized form, trilinear decomposition, is rapidly gaining in popularity. Main reason is computerization of instruments, which makes it simple to collect multidimensional data. Another reason is that scientists are becoming more used to think ‘‘multidimensional’’ and are learning to appreciate the great advantages multidimensional techniques have to offer compared to traditional measurements in fewer dimensions. Finally, the rapid development of user friendly and easily workable software for muldimensional analysis is making life easier. The popular Matlab© software today provides several powerful commands and toolboxes for multidimensional analysis, and DATAN© from MultiD Analyses AB (www.multid.se) has a very intuitive user interface that gives easy access to most functions needed for multidimensional analysis. We therefore expect multidimensional measurements to become very important in the future and methods, such as Procrustes rotation, to become routinely associated to analytical methods that are taught at under graduate university level.

Acknowledgements

Prof. Mikael Kubista acknowledges a grant from the Xunta de Galicia (Visiting Professor at the University of A Coruña, DOG-15/07/2003).

References

[1] J.R. Hurley, R.B. Cattell, Behav. Sci. 7 (1962) 258–262.
[2] P.H. Schönemann, R.M. Carroll, Psychometrika 35 (1970) 245–255.
[3] J.C. Gower, Statistical methods of comparing different multivariate analyses of the same data, in: F.R. Hodson, D.G. Kendall, P. Tautu (Eds.), Mathematics in the Archaeological and Historical Sciences, Edinburgh Univ. Press, Edinburgh, Scotland, 1971, pp. 138–149.
[4] H.T. Eastment, W.J. Krzanowski, Technometrics 24 (1982) 73–78.
[5] W.J. Krzanowski, Biometrics 43 (1987) 575–584.
[6] W.J. Krzanowski, Principles of Multivariate Analysis; A User's Perspective, Revised ed., Clarendon Press, Oxford, England, 2000.
[7] W.J. Krzanowski, Appl. Stat. 42 (1993) 529–541.
[8] A. Carlosena, J.M. Andrade, M. Kubista, D. Prada, Anal. Chem. 67 (1995) 2373–2378.
[9] M. Kubista, Chemometr. Intell. Lab. Syst. 7 (1990) 273–279.
[10] R. Sjöback, J. Nygren, M. Kubista, Biopolymers 46 (1998) 445–453.
[11] M. Kubista, J. Nygren, A. Elbergali, R. Sjöback, Crit. Rev. Anal. Chem. 29 (1999) 1–28.
[12] N. Svanvik, G. Westman, W. Dongyuan, M. Kubista, Anal. Biochem. 281 (2000) 26–35.
[13] Datan for Chemistry. Users manual, version 3.0, September, 2003; available on www.multid.se.
[14] J. Nygren, J.M. Andrade, M. Kubista, Anal. Chem. 68 (1996) 1706–1710.
[15] R.A. Harshman. Foundations of the PARAFAC procedure: Models and conditions for an ‘‘explanatory’’ multi-modal factor analysis. UCLA Working Papers in Phonetics, 16 (1970) 84.
[16] J.M. Andrade, D. Prada, S. Muniategui, E. Alonso, P. López, P. De la Fuente, M.A. Quijano, Anal. Chim. Acta 292 (1994) 253–261.
[17] J.R. King, D.A. Jackson, Environmetrics 10 (1999) 67–77.
[18] M.B. Richman, S.J. Vermette, Atmos. Environ. 27 (1993) 475–481.
[19] J.M. Deane, H.J.H. Macfie, J. Chemom. 3 (1989) 477–491.
[20] J.M. Andrade, S. Muniategui, P. Lopez-Mahia, D. Prada, Fuel 76 (1997) 51–59.
[21] G. Scarponi, I. Moret, G. Capadaglio, J. Chemom. 4 (1990) 217–240.
[22] C. Demir, P. Hindmarch, R.G. Brereton, Analyst 121 (1996) 1443–1449.
[23] C. Bessant, R.G. Brereton, S. Dunkerley, Analyst 124 (1999) 1733–1744.
[24] S. Dunkerley, J. Crosby, R.G. Brereton, K.D. Zissis, R.E. Escott, Analyst 123 (1998) 2021–2033.
[25] D. Schulze, P. Stilbs, J. Magn. Reson. 105 (1993) 54–58.
[26] B. Antalek, W. Windig, J. Am. Chem. Soc. 118 (1996) 10331–10332.
[27] D. Robert, R. Carbo-Dorca, J. Chem. Inf. Comput. Sci. 38 (1998) 469–475.
[28] X. Tomas, J.M. Andrade, A. Alvarez-Larena, Talanta 48 (1999) 781–794.
[29] Q. Guo, W. Wu, D.L. Massart, C. Boucon, S. de Jong, Chemometr. Intell. Lab. Syst. 61 (2002) 123–132.
[30] Q. Guo, W. Wu, D.L. Massart, C. Boucon, S. de Jong, Anal. Chim. Acta 446 (2001) 85–96.
[31] I. Scarminio, M. Kubista, Anal. Chem. 65 (1993) 409–416.
[32] J. Nygren, A. Elbergali, M. Kubista, Anal. Chem. 70 (1998) 4841–4846.
[33] L. Antonov, G. Gergov, V. Petrov, M. Kubista, J. Nygren, Talanta 49 (1999) 99–106.
[34] G. Tang, B. Wang, Fenxi Huaxue 24 (1996) 1437–1440.
[35] E. Vigneau, M.F. Devaux, M. Safar, J. Chemom. 9 (1995) 125–135.
[36] C.E. Anderson, J.H. Kalivas, Appl. Spectrosc. 53 (1999) 1268–1276.
[37] D. González-Arjona, G. López-Pérez, A.G. González, Talanta 49 (1999) 189–197.
[38] K. Heberger, J.M. Andrade, Croat. Chem. Acta (in press).