Available online at www.sciencedirect.com

ScienceDirect

Polar Science 8 (2014) 242–254

http://ees.elsevier.com/polar/

Investigation of Arctic and Antarctic spatial and depth patterns of sea water in CTD profiles using chemometric data analysis

Ewelina Kotwa [a,*], Silvia Lacorte [b], Carlos Duarte [c], Roma Tauler [b]

[a] *Department of Informatics and Mathematical Modelling, Technical University of Denmark, Building 305, Asmussens Alle, DK-2800 Lyngby, Denmark*

[b] *Institute of Environmental Assessment and Water Research, Spanish Council of Research, Jordi Girona 18, 08034 Barcelona, Spain*

[c] *Mediterranean Institute for Advanced Studies, Spanish Council of Research, Miquel Marqus, 21-07190 Esporles, Mallorca, Spain*

Received 7 October 2012; revised 26 February 2014; accepted 23 May 2014
Available online 27 June 2014

Abstract

In this paper we examine 2- and 3-way chemometric methods for analysis of Arctic and Antarctic water samples. Standard CTD (conductivity–temperature–depth) sensor devices were used during two oceanographic expeditions (July 2007 in the Arctic; February 2009 in the Antarctic) covering a total of 174 locations. The output from these devices can be arranged in a 3-way data structure (according to sea water depth, measured variables, and geographical location). We used and compared 2- and 3-way statistical tools including PCA, PARAFAC, PLS, and N-PLS for exploratory analysis, spatial patterns discovery and calibration. Particular importance was given to the correlation and possible prediction of fluorescence from other physical variables. MATLAB's mapping toolbox was used for geo-referencing and visualization of the results. We conclude that: 1) PCA and PARAFAC models were able to describe data in a satisfactory way, but PARAFAC results were easier to interpret; 2) applying a 2-way model to 3-way data raises the risk of flattening the covariance structure of the data and losing information; 3) the distinction between Arctic and Antarctic seas was revealed mostly by PC1, relating to the physico-chemical properties of the water samples; and 4) we confirm the ability to predict fluorescence values from physical measurements when the 3-way data structure is used in N-way PLS regression.

Keywords: Arctic Antarctic; CTD; Multi-way analysis; Fluorescence

1. Introduction

Recent changes observed in the Arctic and Antarctic Ocean regions give support to proposals that polar ecosystems are responding rapidly to processes influenced by global climate change. In these proposals the changes are driven by, for example, sea water transport, ice melt, global atmospheric circulation, and increasing concentrations of green-house gases at a global scale. Measurements from large ocean areas such as the Arctic and Antarctic are crucial for tracking and understanding the effects of ocean and environmental change in these areas and other parts of the globe.

The main aim of this paper is to examine and deliver appropriate multivariate tools for exploratory

* Corresponding author.
E-mail addresses: ek@imm.dtu.dk, ewelina.kotwa@gmail.com (E. Kotwa).

http://dx.doi.org/10.1016/j.polar.2014.05.003
1873-9652/

and regression analysis of CTD (conductivity–temperature–depth) data. Such data are obtained from CTD profilers installed in ships navigating polar ocean waters during the ice-free seasons (July–August in the Arctic and January–February in the Antarctic), and provide an important source of measurements from these areas. The resulting data, comprising a 3-way structure with variable, depth, and location modes, are often analyzed in a univariate way (one variable at a time and independently from other variables and locations). This treatment does not take into account possible underlying covariance dependencies and, therefore, does not use all of the information contained in these fairly complex data structures. Moreover, if a large number of variables are considered, the analysis could become time-consuming and inconvenient. More sophisticated 2-, 3-, or multi-way statistical methods, such as principal component analysis (PCA), the PARAFAC model, and partial least squares (PLS) regression or its multi-way version N-PLS, are known to be more relevant for extracting all information from multi-way data sets, such as those obtained by the CTD sensor (Smilde et al., 2004), and therefore will be considered throughout this paper.

The data used in this study were collected during the 2007 ATOS I (Arctic) and 2009 ATOS II (Antarctic) polar expeditions from a total of 174 locations. Depth profiles of seven variables common to both polar areas were considered and included in the analysis: temperature, conductivity, salinity, oxygen, beam transmission, fluorescence, and sea-point turbidity (Table 1).

The second objective of this work is a comparative sea water study to identify geographical differences within and between the Arctic and Antarctic seas, by considering their physical and chemical characteristics. For this purpose, PCA and PARAFAC models were applied and their results are discussed.

Additionally, within the sea water study, measured fluorescence is of interest because it is strongly related to the amount of chlorophyll (reflecting the maximum concentrations of biota and algae population), and hence, biological activity in the water. A regression study to explain and predict fluorescence values from the remaining variables is therefore the third objective of this paper, using the partial least squares regression technique with its multi-way alternative, N-PLS.

The remainder of the paper is organized as follows. Section 2 describes the two data sets and the data transformations applied before chemometric analysis. A brief methodological overview of the techniques employed in the investigation is presented in Section 3, followed by the results and discussion in (Section 4). Finally, the conclusions are given in Section 5.

Table 1
Variables measured at every station in both, the Arctic Sea and the Antarctica.

No.	Variable	Unit
1	Temperature	[ITS-68, deg C]
2	Conductivity	[mS/cm]
3	Salinity	[PSU]
4	Oxygen	SBE 43 [ml/l]
5	Beam Transmission	[%]
6	Fluorescence	arbitrary units [AU]
7	Sea-point Turbidity	[FTU]

2. Experimental data

Data samples were collected during two oceanographic expeditions, covering the area 68°N–81°N, 20°W–20°E in the Arctic and 60°S–70°S, 50°W–77°50′W in the Antarctic, as shown in Fig. 1. During both expeditions a standard, real-time CTD sensor was used, producing depth-profile data for each of the 49 locations in the Arctic (due to the position of permanent ice and the needs of the scientists, the measurement was occasionally repeated, resulting in a total number of 80 samples) and for the 93 Antarctic stations. The sensors used in the two expeditions were non-identical, measuring different ranges of variables, sometimes in different units. However, seven variables common to both locations were identified and are used in this study (Table 1). The first four variables are of a physical (temperature and conductivity) or chemical nature (salinity and dissolved oxygen) and the remaining three variables are related to radiation (beam transmission, fluorescence, and sea turbidity).

At each location the sensor was dropped down to a certain water depth (50–1000 m), enabling replicate measurements for each depth as the sensor was descending and ascending. Here, only the up-cast measurements were included as they proved to have nearly identical profiles as the down-cast data, but contained a lower number of incomplete observations and therefore provide more reliable results. The measurements were collected approximately every few seconds, but the data set was reduced by taking averages by depth, resulting in one observation per meter (arbitrary choice). Given that the greatest changes in the measured signal were from near the surface, all data collected below 100 m depth were disregarded because most of these values were effectively constant.

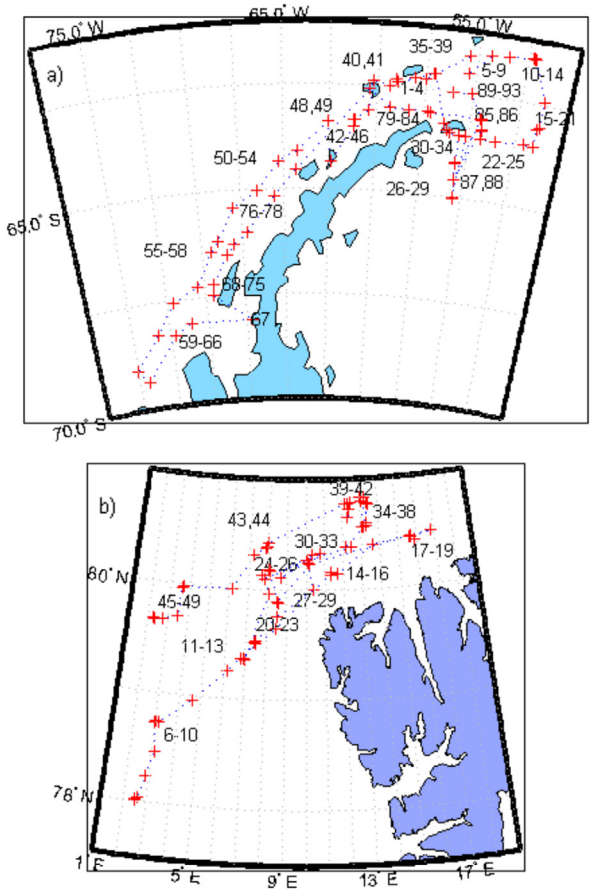

Fig. 1. Locations of the CTD measurement stations in the Antarctic (a) and in the Arctic Sea (b).

2.1. Arranging the data sets

The most 'natural' and straightforward way to arrange the whole data set was a 3-way framework, resulting in a data cube, relevant for PARAFAC and NPLS models (Fig. 2). For the PCA and PLS analyses, the data were unfolded and the two distinct unfolding directions were adapted for this study: variable-wise (to the left) and station-wise (to the right). Station 71 from the Antarctic expedition was disregarded due to the high number of missing values (>60%). In addition, the first 5 stations (10 measurements due to repetitions) were removed from the Arctic data. These measurements were taken 'on the way' to the final destination area and were not of significant interest to the overall study. In addition, their location did not fit within a 2-D 'grid' on the map (they were aligned), which could cause later plotting difficulties. Following the removal of these measurements the total number of locations sampled for the Antarctic was 92 and for the Arctic Sea was 70 (each repetition was considered as a separate measurement). This resulted in the final dimensions of the data being: [100 × 7 × 162] for 3-way methods, and [16200 × 7] and [162 × 700] for the unfolded data sets, variable- and station-direction, respectively.

2.2. Missing values and outliers

Due to technical issues during sensor data acquisition (presumably strong waves on the surface or instrumental errors during ascent or descent), the data from close to the surface are commonly corrupt or missing. Consequently, the first step of the analysis was to remove anomalous signal disturbances and to interpolate the resulting empty spaces in the data matrix. This is a necessary step before applying any of the classical least square routines, which cannot accommodate the presence of missing values. This step was performed using a variable-wise standardization method. Standardized values exceeding the 99% confidence interval were flagged as outliers and then substituted by a weighted average of neighboring values. More sophisticated methods exist for both outlier removal and missing data interpolation ((Filzmoser et al., 2008), (Rousseeuw

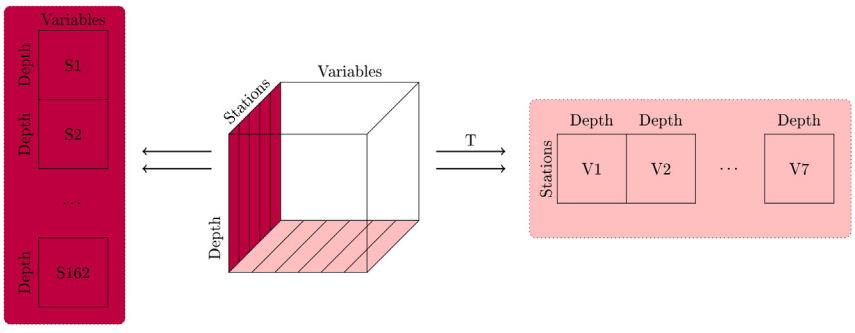

Fig. 2. Arrangement of the CTD data in a cubic structure according to three modes: depth, variables, and stations. Two unfolding directions, variable-wise (left) and station-wise (right), were adapted to fit the 2-way workframe.

et al., 2006), (Serneels and Verdonck, 2009), (Stanimirova and Walczak, 2008)) using, for example, robust statistics or an EM-algorithm; however, for the needs of this study, the adapted methodology yields satisfactory results.

In addition, variable-wise data centering and scaling were applied to the whole data set (Arctic plus Antarctic), but this did not remove the offset specific for each location. On the contrary, the aim was to maintain and emphasize the differences in variable behavior between the two sites. Fluorescence and sea turbidity variables were initially given using different units for the Antarctic and the Arctic Sea. To avoid unit differences corrupting the results, these variables were scaled prior to the pre-processing step that was applied to the whole data set.

3. Methods

As previously stated, the objective of this work is to deliver and compare multivariate statistical tools that can be used in two ways: firstly, to explore and understand interdependencies present in the CTD data, and secondly to explain the fluorescence variability by regressing it on the remaining variables. For this purpose, two approaches to data analysis are adopted. We start by considering and implementing the most commonly used 2-way chemometric techniques: PCA and PLS. The term '2-way', not to be confused with '2-dimensional', refers here to data sets that can be arranged in a matrix, having two distinct directions (modes); e.g. objects and variables. The term '3-way' denotes data accommodated in a cubic form, where a third direction (here, location) has been added. Hence, before fitting a 2-way model to 3-way data, the analyzed cube has to be unfolded (according to one of its modes) and reshaped into a matrix (see Section 2.1). It has been argued (Smilde et al., 2004) that in principal, 3-way techniques would be more suitable and beneficial for a 3-way data structure. Therefore, some of the generalizations of the 2-way methods, such as PARAFAC and multi-way PLS (N-PLS), are applied and discussed below. Throughout this paper, scalars are indicated by lowercase italics, vectors by bold lowercase characters, two-way matrices by bold capitals, and three-way arrays by underlined bold capitals.

3.1. 2-way methods

If the focus of the analysis lies in summarizing patterns, dependencies, or differences within the data set, then decomposition methods such as principal component analysis could be used for this purpose. PCA is a linear subspace-based technique, perhaps most commonly found in the chemometric literature. A PCA model is presented as follows:

$$x_{ij} = \sum_{r=1}^{R} t_{ir} p_{jr} + e_{ij} \;\; i = 1, \ldots I; \;\; j = 1, \ldots J; \tag{1}$$

where x_{ij} is an element of the matrix $\mathbf{X}(I \times J)$, $\mathbf{t}$ and $\mathbf{p}$ are the decomposed vectors, and e_{ij} contain the model residuals. In brief, PCA projects the data (objects and variables) to lower dimensional spaces where it is easier to explore and visualize them. This is completed by finding a sum of the vector products, called scores ($\mathbf{t}$) and loadings ($\mathbf{p}$), which are orthogonal and are determined by maximizing the variance explained by them. The vector products, which are linear combinations of the original variables (or objects), are called principal components, and often a small number of these components allows us to explain the data variation in a satisfactory way. More details concerning the PCA method can be found in, for example, in (Eckart and Young, 1936), (Pearson, 1901) or (Wold et al., 1987).

To cover the most important information contained in the polar data, two unfolding directions of the data cube (according to variable and location modes) are considered. Classical and robust versions of PCA (here, ROBPCA as developed by (Hubert et al., 2005)) are fitted, due to the potential influence of outliers.

In parallel, for explaining the measured fluorescence by other available variables, a PLS regression model is applied. PLS regression is a 2-way calibration method that approximates the calibration matrix $\mathbf{X}$ by r components (called factors or latent variables) and, at the same time, projects the dependent variable $\mathbf{y}$ on these components, which are constructed to obtain a compromise between fitting $\mathbf{X}$ and predicting $\mathbf{y}$. To calculate a component in the PLS regression model, a one-component model of $\mathbf{X}$ of the form (after (Bro, 1996)),

$$x_{ij} = t_i w_j \tag{2}$$

is used where the t_i are scores and the w_j weights. Weights $\mathbf{w}$ are found so that they yield a score vector $\mathbf{t}$ with maximal covariance with $\mathbf{y}$, which can be written as follows:

$$\max_{\mathbf{w}} \left[cov(\mathbf{t}, \mathbf{y}) | \min \left(\sum_{i=1}^{I} \sum_{j=1}^{J} \left(x_{ij} - t_i w_j \right)^2 \right) \wedge ||\mathbf{w}|| = 1 \right] \tag{3}$$

This algorithm is run sequentially, meaning that only one component is calculated at a time. This

component is then subtracted from the calibration matrix $\boldsymbol{X}^i = \boldsymbol{X}^{i-1} - \boldsymbol{t}^i\boldsymbol{p}^i$ and the new component is found from the residuals. A broader description of PLS and its applications can be found in (Bro, 1996), (Martens and Naes, 1992),or (Smilde et al., 2004) (Wold et al., 1984).

As stated in Section 1, the measured fluorescence carries information about the amount of chlorophyll and therefore biological activity in the water. It is well known that fluorescence is strongly correlated to other measured, 'light-related' variables, such as beam transmission or sea turbidity. It could also be expected that the biological activity is influenced by other physico-chemical conditions, such as the amount of dissolved oxygen, temperature, and salinity. It will therefore be interesting to compare the regression results of a PLS model when the predictors block ($\boldsymbol{X}$) is constructed by all of the CTD variables and, as a second scenario, by the physico-chemical variables only.

To identify the optimal number of components for PCA and PLS models, the cross-validated root mean square error (RMESCV), illustrated in Fig. 3, was calculated by means of cross-validation with 81 contiguous blocks. This corresponds to one 'split' being equal to two locations (200 observations for variable-wise and 2 for station-wise unfolded data).

3.2. 3-way data analysis

The sea water measurements analyzed in this study follow three different modes (variable, depth, and location). Therefore, by unfolding the data cube and using 2-way techniques we risk the 3-way correlation structure being 'flattened' and some information lost. A PARAFAC model, which can be perceived as one of the tri-linear extensions of PCA, is able to overcome this issue. Other 'extensions' exist, such as the Tucker3 model (Smilde et al., 1994); however, here we consider PARAFAC due to its simplicity and the easy interpretation of loadings.

The PARAFAC model, introduced by (Harshman, 1970) and popularized by (Bro, 1997) and (Smilde et al., 2004), decomposes a data cube $\underline{\mathbf{X}}(\mathrm{I} \times \mathrm{J} \times \mathrm{K})$ into a sum of triple vector products, called loadings. The most common way of writing the model is as follows:

$$x_{ijk} = \sum_{r=1}^{R} a_{ir} b_{jr} c_{kr} + e_{ijk} \tag{4}$$

where x_{ijk} is an element of $\underline{\mathbf{X}}$; $\mathbf{A}(I \times R)$, $\mathbf{B}(J \times R)$, and $\mathbf{C}(K \times R)$ are the orthogonal matrices with elements a_{ir}, b_{jr}, and c_{kr}, respectively; R is the number of components; and e_{ijk} represents the error term.Since previous studies on a robust version of PARAFAC are not without contention, only the classical version of the algorithm is presented here. A number of components are chosen according to four indices describing the performance of the model: explained variance, core consistency, number of iterations, and total elapsed time (Table 2). An overview regarding component selection criteria can be found in (Bro and Kiers, 2003).Similarly, a regression model can be generalized to fit a 3-way data structure. Multi-way PLS (or N-PLS) is a generalization of PLS regression, which predicts $\mathbf{y}$ and decomposes $\underline{\mathbf{X}}(I \times J \times K)$ into a set of triads. A triad (being a tri-linear equivalent of a bilinear factor)

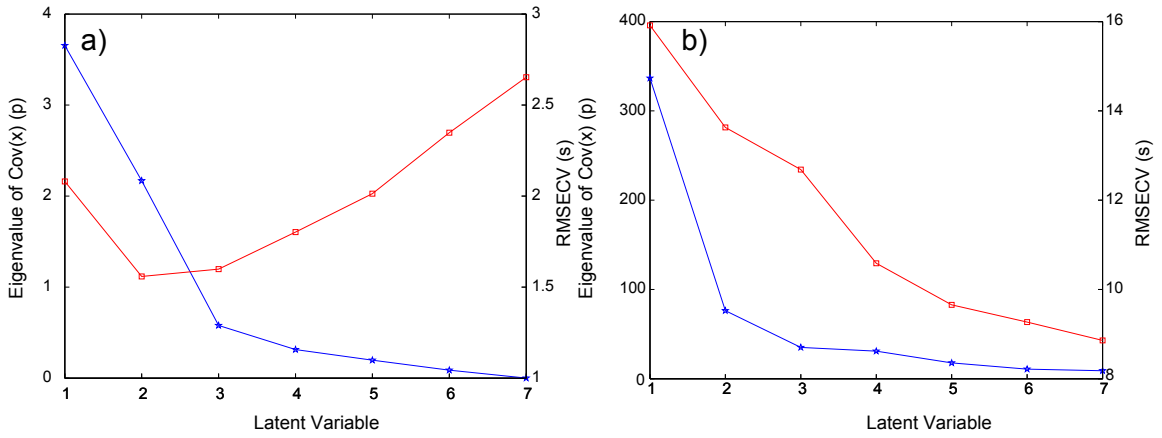

Fig. 3. Cross-validated root mean square error (red squares) and eigenvalues (blue stars) of the covariance matrix X for the PCA model X, for a) variable-wise unfolded data, and b) station-wise unfolded data. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Choosing amount of components in PARAFAC model according to four indices: explained variance, core consistancy, number of iterations and total calculation time.

No.	Variance explained	Core consistancy	Iteration	Time
1	30.32	100	5	0.56
2	63.46	100	5	0.41
3	73.70	−481	24	1.06

consists of one score vector $\mathbf{t}$ and two weight vectors $\mathbf{w}^J$ and $\mathbf{w}^K$ (Bro, 1996), (de Jong, 1998) (Smilde, 1997). The resulting model is given by

$$x_{ijk} = t_i w_j^J w_k^K \tag{5}$$

The tri-PLS model is thus defined as a problem of finding $\mathbf{w}^J$ and $\mathbf{w}^K$ in

(6)

$$\max_{\mathbf{w}^J\mathbf{w}^J}\left[cov(\mathbf{t},\mathbf{y})|\min\left(\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\left(x_{ijk}-t_i w_j^J w_k^K\right)^2\right) \wedge\left|\left|\mathbf{w}^J\right|\right| = \left|\left|\mathbf{w}^K\right|\right| = 1\right] \tag{6}$$

which is essentially a singular value decomposition (SVD) task. As N-PLS is also a sequential method, after finding the first component, both $\underline{\mathbf{X}}$ and $\mathbf{y}$ are 'deflated' (replaced by residuals of the respective models) to recommence the algorithm. The above algorithm considers only one dependent variable, $\mathbf{y}$, but its generalization, corresponding to a case of several dependent variables (as for CTD data) is straightforward (see e.g., (Bro, 1996)).

To apply N-PLS to the CTD data, we 'slice out' the $\mathbf{Y}$-block (fluorescence) from the initial data cube. The remaining variables constitute a predictor block $\underline{\mathbf{X}}$, also of a cubic shape. For selecting the number of components, a multi-way cross-validation is performed, indicating a three component model. Again, two scenarios for the predictive block are considered: including all variables, or alternatively only those relating to physico-chemical properties.

3.3. 2- or 3-way model for 3-way data

Choosing a relevant type of model is always a challenge from a practical viewpoint. Whereas the 2-way models, such as PCA and PLS, have been widely acclaimed and commonly used, their 3-way equivalents remain unpopular. The reluctance towards the use of 3-way models might have various sources such as a lack of experience and knowledge, preference for seemingly simpler solutions, or a lack of relevant software. Ideally, the 3-way data would be best described by a 3-way model; however, due to the limitations and assumptions of the model and the fact that 'real' data rarely follow these assumptions, the data analyst is forced to evaluate the pros and cons of each solution and choose accordingly.

For example, in choosing between the PCA and PARAFAC models, the following arguments might be helpful in making the final decision. If the experimental data fulfill the tri-linear assumption (required invariability of the component profiles across the different data slices with different weighting coefficients for each slice (Harshman, 1970), (Smilde et al., 2004)), the application of a PARAFAC model is usually superior to its bilinear counterpart, PCA. The reasons for this are numerous. First, PARAFAC takes into account interrelations existing in all three data directions; moreover, the problem of rotational freedom, which is common in PCA, is solved because PARAFAC provides the unique solution (up to the scaling constant, sign, and permutation ambiguities) (Smilde et al., 2004). In addition, the PARAFAC model resolves each mode separately, giving a straightforward physical interpretation for each of the profiles (there is no need to unfold the data in different directions and fit 2 or 3 different models). Finally, due to the relatively low number of degrees of freedom, the PARAFAC model is more robust and does not tend to over-fit, as is often the case with PCA (it is a rule of thumb that if several models describe the data equally well, one should choose the simplest in order to keep the model robust against overfit).

In spite of the benefits that may be gained from applying a 3-way model, some drawbacks also exist. The most important is that real data do not always conform to the tri-linear assumption (e.g., oceanographic data) and the model could return degenerated solutions. Degeneracy might also occur when a large number of factors are required that are interrelated (compare with the Tucker model). In most of these cases, the bilinear model is still appropriate and PCA or other methods such as MCR (Multivariate Curve Resolution), described by (Tauler, 1995), can be successfully applied.

3.4. Mapping method

An effective visualization tool is available when working with the PARAFAC or location-unfolded PCA model. It projects the third mode location loadings

directly onto a map and creates a loading variability image. This map uses the method known in geostatistics as 'kriging', and its conceptual background, which mathematically consists of random field interpolation techniques, can be technically complex and is beyond the scope of this paper. In brief, MATLAB's mapping toolbox, which is used in this work, allows the transfer of the loading values according to their GPS coordinates onto the specified fragment of the world map. Following this step, the 2D interpolation of these values is performed within a small convex set spanned by the location coordinates (however, the results should be used with caution because they are insensitive to water/land borders). This way of representing the loadings offers an attractive tool for a better understanding and interpretation of the spatial variability in the data.

4. Results and discussion

4.1. Data exploration

Fig. 3 shows the eigenvalues of the covariance matrix and validated RMSE for variable- and station-unfolded data scenarios. It can be seen that the two plots are dissimilar. In the first case, two or three components explain 85% and 92% of data variability, respectively, and it is clear that the model would overfit if more components were chosen. For the latter scenario, the RMSE index decreases in value gradually and it is less evident what number of components are relevant. Therefore, to obtain a similar variance explanation as in the variable-wise case (>80%), a 3-component model is chosen.

Robust and classical PCA models give similar results when the station-wise unfolding of the data is considered; therefore, only the results of the classical version are discussed. However, the situation is different for the variable-wise direction. The robust PCA attributes 62% of the explained variation to PC1, 23% to PC2, and 7% to PC3, whereas the same components account for 52%, 31%, and 8.5% of data variability, respectively, for its classical counterpart. This might be due to several reasons, such as unidentified outliers still present in the data, the fact that the data distribution does not conform to the PCA assumptions, or simply low efficiency of the robust method. Taking this into consideration, only the output from the robust version of PCA, for the variable direction, is finally reported.

In Fig. 4, scores (PC1, PC2, PC3) and loadings (on PC1 vs PC2 and PC1 vs PC3) for both fitted models are

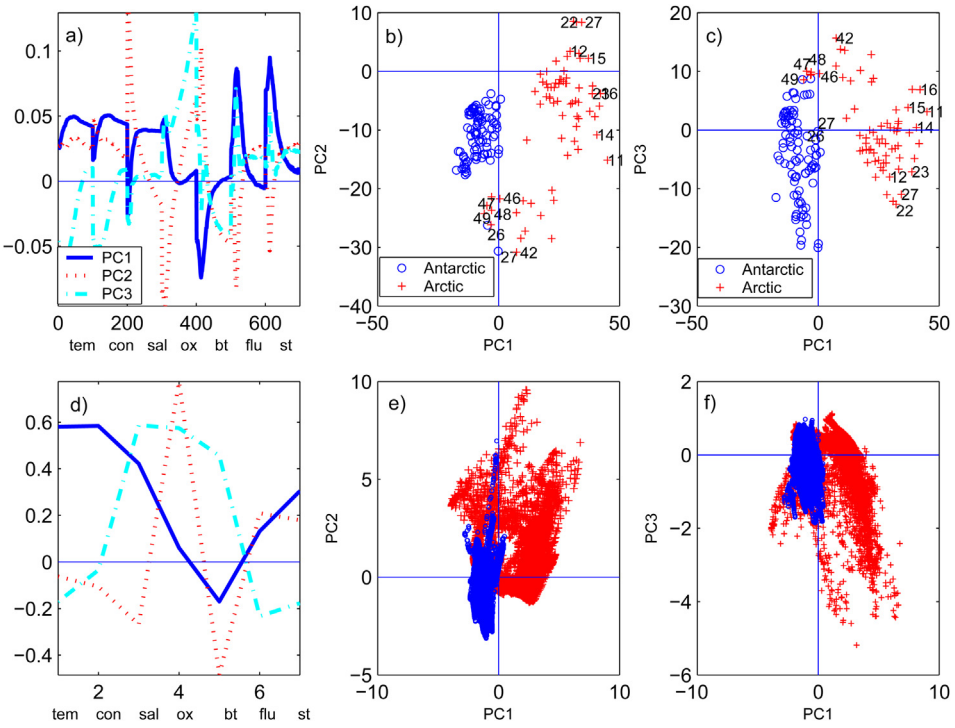

Fig. 4. Loadings and scores for the PCA models on station-wise (a–c) and variable-wise (d–f) unfolded data.

presented. The different behavior of scores corresponding to the Arctic Sea (red crosses) and the Antarctic (blue circles), mostly along the PC1 and PC2 coordinate axes, is clear to see. In Fig. 4b, almost all of the Arctic locations have positive score values, and the Antarctic locations have negative score values on PC1. In Fig. 4e, the Arctic Sea values are on average higher than the Antarctic values for both PC1 and PC2. Another interesting observation is that the Arctic scores are more widely scattered throughout the coordinate system, which indicates higher variation within that area. The third principal component does not appear to introduce any additional information and therefore it could be argued that in order to avoid over-fitting, a two-component model would be more relevant here.

From the loading plot (Fig. 4a), two major types of depth profile are apparent as indicated by PC1 (blue solid line): one for physico-chemical variables and another for light-related variables. By adding information from the score plots, (Fig. 4b and c) it is evident that temperature, conductivity, salinity, fluorescence, and sea turbidity have higher values in the Arctic Sea than in the Antarctic (as their corresponding profile loadings are positive on PC1).

Moreover, two clusters are visible within the Arctic data (Fig. 4b): the main data cloud having strictly positive values on the PC1 and being centered around zero on PC2, and a second smaller group having negative PC2 values and being closer to zero at PC1. This second cluster consists of the data from stations 4146 from the most northwest area covered by the expedition (close to the border of the ice where the ship could not move freely; Figure ??). We would therefore expect them to have different characteristics (e.g., lower temperature) than the other locations investigated in the Arctic Sea.

Finally, information obtained from the 'variable direction' confirms that the Arctic and Antarctic samples show different behaviors according to PC1, with high positive values for temperature, conductivity, salinity, fluorescence, and sea turbidity evident for the Arctic samples. Given that the Antarctic water samples are predominantly negative, it can be expected that these variables would present on average higher values in the Arctic Sea. On the other hand, PC2 is negatively correlated with salinity and positively with oxygen; however, the geographical interpretation is more difficult as the scores from both locations are spread more evenly across this component. On average, it seems that samples from the Antarctic have slightly higher values on PC2, although more statistical tests are needed to confirm this hypothesis.

The main difficulty in interpreting the two-way PCA model output is that in order to obtain the full information about each mode, the data cube should be unfolded in three different directions. However, this creates three different models (here only two are shown) and care should be taken in cross-interpreting their results as there is no certainty that, for example, PC1 in the variable direction will reflect the same information as the corresponding component in the station-wise unfolded data set. This risk can be mitigated by applying one PARAFAC model.

Table 2 presents four different indices that are normally used when determining the number of PARAFAC components. It is apparent that the whole data system can be well approximated (63% of the total data variation) using only two components, with a core consistency of 100% that converges quickly to the minimum, as the model starts degenerating once the number of components increases. This degeneration might be caused by the fact that the data do not follow the tri-linearity condition (see the discussion in Section 3.3 or simply that more components will lead to model over-fitting. Loading profiles resolved by PARAFAC in the three data modes (depth, variable, and location or station) are shown in Fig. 5. It is apparent how advantageous the properties of the PARAFAC method are when summarizing the whole data variability and its underlying interrelations using only three plots. The loading profiles of the two components in the depth mode (Fig. 5a) describe the sea water changes observed from the surface to deeper samples. The variable contributions resolved in the second mode are shown in Fig. 5b, and finally the location (geographical) profiles are presented in 5c.

The information contained in these figures can be understood as follows. The first component describes the major changes occurring in the physico-chemical characteristics of the water carried by temperature, conductivity, and salinity. The corresponding depth profile indicates an increase in the contribution of this component until c. 25 m below sea level, where it then gradually declines with depth. Moreover, from Fig. 5c we can conclude that this component has substantially higher values for the Arctic Sea samples (except for some of the last station samples, located close to the glacier), which confirms the common knowledge about the two polar locations. On the other hand, the second component is influenced positively by variables such as fluorescence, sea turbidity, and dissolved oxygen, and negatively by beam transmission. The depth profile for this component increases to reach its peak around 15 m below sea level, where the average maximum of chlorophyll (DCM) is

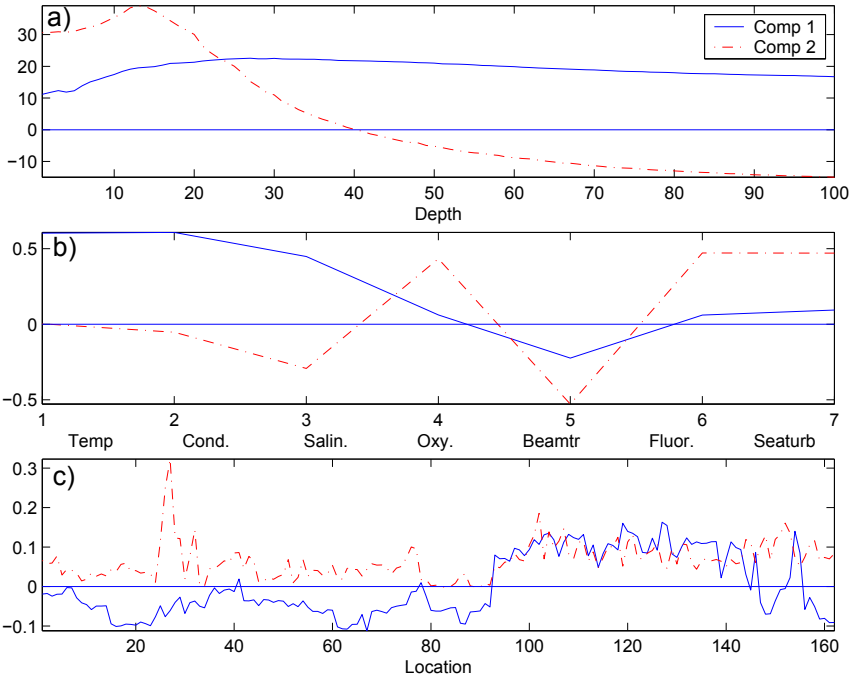

Fig. 5. Resolved profiles from a 2-component PARAFAC model according to a) depth mode, b) variable mode, and c) station mode.

expected, and then decreases exponentially with depth. The location profile emphasizes again the differences between Arctic and Antarctic samples by attributing higher values (and also higher variance) to the Arctic area, with significant exceptions for some Antarctic samples (sample 27 being the extreme southeast station). From this, we could draw an initial conclusion that biological activity, reflected by fluorescence, is richer in the Arctic Sea. In addition, the second component has low loadings for temperature (high for the first component), which indicates that it describes a completely different pattern of measured parameter changes than the first component; therefore, we will call it 'radiation related'. It is apparent that changes in dissolved oxygen and fluorescence (biological activity) in the second component are independent of changes in temperature, conductivity, and salinity, which is in contrast to the pattern depicted by the first resolved component, where these variables were positively inter-correlated. Moreover, the fact that the shapes of the depth profiles for the two components are different, with their maxima around 25 and 15 m for the first and second components, respectively, confirms the existence of two different types of phenomena and patterns, interacting differently.

It could be argued that similar information can be extracted by looking at a collection of plots, depicting one variable at a time. This might be true in this case; however, such treatment would require fairly time-consuming analysis of multiple plots, and the time demands would increase if a larger number of variables were considered. It has been demonstrated that when the more complex model structure was chosen, the easier the interpretation of its results. To sum up, the above analysis shows that: 1. similar information can be extracted by applying two (or three; not shown) PCA models or one PARAFAC model; 2. interpretation of PARAFAC results is mush easier because it delivers concise information about all three data modes in only three figures; and 3. care should be taken when choosing the number of components for both data models, as a greater number of components might lead to over-fitting the PCA model and degeneration of the PARAFAC results.

4.2. Map representation of the scores

In the previous section we showed that the first and second PARAFAC components cover the two major patterns present in the data: physico-chemical and spectral radiation-related. The map representation of these components for both the Antarctic and the Arctic areas (Fig. 6) can then be interpreted via the variability image of the two major phenomena. By looking at the first component (Fig. 6a and b) it is evident that in the Antarctic the gradient (i.e., the direction in which the values grow) is towards the north and in the Arctic

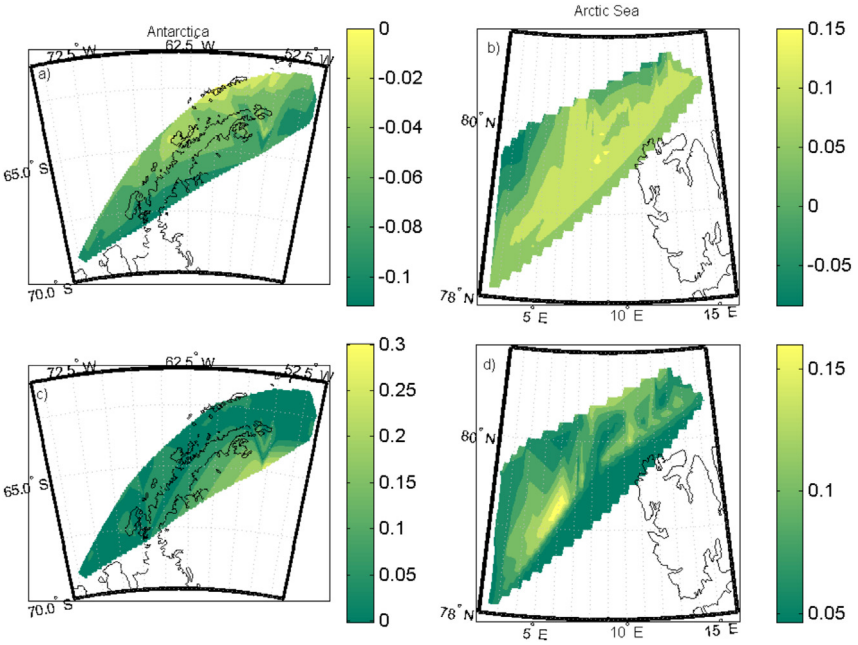

Fig. 6. Map representation of the PCA Components 1 (a and b) and 2 (c and d) for the PARAFAC model for Antarctic (a and c) and the Arctic Sea (b and d).

towards the southeast. This confirms the expectation that the temperature (and other related variables) should rise when moving away from the Poles or when approaching land (e.g., Svalbard for the Arctic). The biological activity, represented by the second component, follows considerably different patterns as illustrated in Fig. 6c. In the Antarctic region we can clearly see a comparatively high value at station 27. This is probably related to the different bio-characteristics of the region, as it is located in the most southeast part covered by the expedition, which might be more advantageous for biological activity. In the Arctic Sea the peak in this component is around stations 1011, which might be caused by a local phenomena. Additionally, the higher concentrations of bio-activity are located close to the ice border in the north, which corresponds to relatively low temperatures.

4.3. Regression

As stated previously, a high correlation between radiation-related variables and measured fluorescence is to be expected. However, the most interesting results are generated when these variables are excluded from the explanatory variable *X*-block. This will identify to what degree the fluorescence can be determined by the physico-chemical conditions of the sea water.

To identify the optimal number of components and validate the resulting models, a cross-validation routine was carried out, splitting the data set into 81 contiguous blocks. We used *K-fold cross-validation*, where each block corresponds to two geographical locations that are withheld from the training data consecutively. Calibrated (RMSE) and cross-validated (RMSECV) error measurements were then calculated, reflecting the performance of each model.

Initially a standard 2-way PLS regression model is fitted, using all the variables in the unfolded (variable-wise) data set. The validated RMSE suggests three latent variables (LVs) should be used, explaining 92% and 74% of *X*- and *Y*-block variability, respectively (see Table 3). When inspecting the model weights, which show the impact of each explanatory variable on *Y*, it is not surprising that beam transmission (negative) and sea turbidity (positive) have the highest values, both being incorporated in the first latent variable (not shown). This is expected because these variables carry the light-related information, meaning they dominate LV1. Subsequently, the influences of temperature and conductivity are taken into account by LV2, and salinity and oxygen are manifested only in LV3. As a second scenario, the radiation variables are removed from the predicting block so that the reduced, three-component model now accounts for only 25% of the observed fluorescence variation, which is a very weak result. The control plots in Fig. 7a and c show observed versus predicted *Y* values and indicate that the model is unable

Table 3
Explained variance and prediction errors (calibrated - RMSE; cross-validated - RMSECV) of PLS and N-PLS models for two variants of predictive block: 1. with all CTD variables; 2. with physico-chemical variables only.

	All variables					Physico-chemical			
	No.	X block	Y block	RMSE	RMSECV	X block	Y block	RMSE	RMSECV
PLS	1	40.54	65.01	0.6105	0.6184	52.31	20.98	0.8889	0.8939
	2	70.30	70.47	0.5291	0.5346	92.40	25.27	0.8644	0.8712
	3	92.32	74.08	0.4297	0.4359	100	25.33	0.8641	0.8717
nPLS	1	26.66	60.40	0.5325	0.5502	46.00	31.56	0.5571	0.5674
	2	65.56	76.88	0.3718	0.3883	72.04	69.73	0.4176	0.4364
	3	72.55	85.50	0.2781	0.2972	83.42	79.13	0.3252	0.3519

to identify the difference in data behavior within the Arctic Sea and the Antarctic, leading to poor predictions. At the same time. the *X*-block is fully explained, as the remaining variables in the model are highly correlated.

The results are quite different in the case of multilinear PLS. We decided on a three-component structure after considering the cross-validation results. Remarkably, the N-PLS model with only physico-chemical variables as predictors was now able to explain up to 79% of the measured fluorescence and around 83% of the *X* array. A plot of predicted versus observed values (Fig. 7b and d), together with RMSECV values, confirms the obtained improvement. The complete results for both data scenarios are presented in Table 3, from which we can conclude that 2-way PLS is largely outperformed by its 3-way alternative in predicting the fluorescence values using non-radiation-related variables.

This result can be explained by the fact that the 3-way model accounts for the interrelations existing within the data, which could have been disregarded during unfolding of the data set. Therefore, this example clearly shows the importance of choosing an adequate modeling technique.

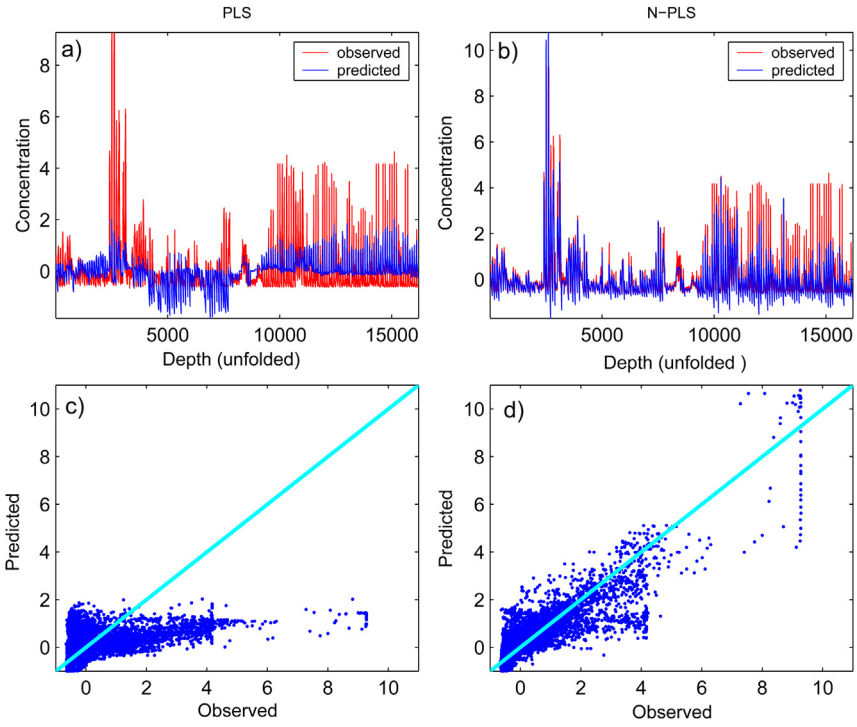

Fig. 7. Predicted versus observed values for the PLS (a and c) and N-PLS (b and d) models, with only the physico-chemical variables entering the X-block.

5. Conclusions

It has been shown that Arctic and Antarctic sea waters could be clearly differentiated according to CTD water samples collected during the 2007 ATOS I and 2009 ATOS II polar expeditions. Two principal components have been identified by PCA and PARAFAC models, which summarize well the whole data set: the 1st PC is related to physico-chemical properties and the 2nd to light-related variables. The distinction between Arctic and Antarctic seas was revealed mostly by PC1. Moreover, multi-way PLS regression confirmed the possibility of predicting fluorescence values (and therefore life presence) from measured CTD physical variables such as temperature, conductivity, salinity, and dissolved oxygen. This result was only clearly revealed when the three-way data structure was considered in the regression model, whereas it was completely hidden in the case of the classical two-way unfolded PLS method.

The pros and cons relating to 2- and 3-way chemometric methods were analyzed and discussed, and the resulting conclusions were obtained.

1. Similar sets of information can be extracted by applying two (or three; not shown) PCA models or one PARAFAC model; however, interpretation of the PARAFAC results is much more straightforward.
2. In general, 3-way models describe 3-way data better than 2-way models, if the data conform to the assumptions that underlie these models.
3. Applying a model of a structure less complex than the data structure itself, raises the risk that the underlying correlation structure will be flattened and important information lost, thereby resulting in a marked deterioration of the result quality (see the PLS correlation model).
4. Care should be taken when choosing the number of components for both data sets, as when this number increases it might lead to over-fitting the PCA model and degeneration of the PARAFAC results.

To sum up, recent instrumental developments within analytical chemistry and environmental sciences have led to an increased occurrence of high-dimensional data sets. This leads directly to an increased requirement for data analytical tools, as simple statistical methods become not only highly time-consuming, but are generally not applicable to these vast data structures. Multivariate data analysis tools, such as those presented in this paper, are therefore likely to become widely used in future studies within environmental sciences, including oceanography.

Acknowledgments

This research was part of the ATOS project, funded as part of the Spanish contribution to the International Polar Year (IPY) by the Spanish Ministry of Science and Innovation (POL200600550/CTM). We thank the ATOS participants, UTM and crew of R/V Hesprides for help with CTD sampling and logistics.

References

Bro, R., 1996. Multiway calibration. multilinear pls. J. Chemometr. 10 (1), 47–61.

Bro, R., 1997. Parafac. tutorial and applications. Chemometr. Intell. Lab. Syst. 38 (2), 149–171.

Bro, R., Kiers, H.A.L., 2003. A new efficient method for determining the number of components in parafac models. J. Chemometr. 17 (5), 274–286.

de Jong, S., 1998. Regression coefficients in multilinear pls. J. Chemometr. 12 (1), 77–81.

Eckart, C., Young, G., 1936. The approximation of one matrix by another of lower rank. Psychometrika 1 (3), 211–218.

Filzmoser, Peter, Maronna, Ricardo, Werner, Mark, 2008. Outlier identification in high dimensions. Comput. Stat. Data Anal. ISSN: 01679473 52 (3), 1694–1711.

Harshman, R.A., 1970. Foundations of the Parafac Procedure: Models and Conditions for an" Explanatory" Multimodal Factor Analysis. University of California, Los Angeles.

Hubert, M., Rousseeuw, P.J., Branden, K.V., 2005. Robpca: a new approach to robust principal component analysis. Technometrics. ISSN: 00401706 47 (1), 64–79.

Martens, H., Naes, T., 1992. Multivariate Calibration. John Wiley & Sons Inc.

Pearson, K., 1901. On lines and planes of closest fit to systems of points in space. Lond. Edinb. Dublin Philos. Mag. J. Sci. 2 (11), 559–572.

Rousseeuw, Peter J., Debruyne, Michiel, Engelen, Sanne, Hubert, Mia, 2006. Robustness and outlier detection in chemometrics. Crit. Rev. Anal. Chem. ISSN: 10408347 36 (3–4), 221–242.

Serneels, Sven, Verdonck, Tim, 2009. Principal component regression for data containing outliers and missing elements. Comput. Stat. Data Anal. ISSN: 01679473 53 (11), 3855–3863.

Smilde, A.K., 1997. Comments on multilinear pls. J. Chemometr. 11 (5), 367–377.

Smilde, A.K., Tauler, R., Henshaw, J.M., Burgess, L.W., Kowalski, B.R., 1994. Multicomponent determination of chlorinated hydrocarbons using a reaction-based chemical sensor. 3. medium-rank second-order calibration with restricted tucker models. Anal. Chem. 66 (20), 3345–3351.

Smilde, A.K., Bro, R., Geladi, P., Wiley, J., 2004. Multi-way Analysis with Applications in the Chemical Sciences, vol. 978. Wiley Online Library.

Stanimirova, I., Walczak, B., 2008. Classification of data with missing elements and outliers. Talanta. ISSN: 00399140 76 (3), 602–609.

Tauler, R., 1995. Multivariate curve resolution applied to second order data. Chemometr. Intell. Lab. Syst. 30 (1), 133−146.

Wold, S., Ruhe, A., Wold, H., Dunn III, W.J., 1984. The collinearity problem in linear regression. the partial least squares (pls) approach to generalized inverses. SIAM J. Sci. Stat. Comput. 5, 735.

Wold, S., Esbensen, K., Geladi, P., 1987. Principal component analysis. Chemometr. Intell. Lab. Syst. 2 (1), 37−52.